A University of Sussex PhD thesis

Available online via Sussex Research Online:

http://sro.sussex.ac.uk/

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details
An investigation of genome stability in human female meiosis by genome-wide chromosome fingerprinting and copy number analysis.

Robert Christopher Blanshard

Submitted for the degree of Doctor of Philosophy
Genome Damage and Stability Centre, School of Life Sciences
University of Sussex, Brighton, UK.

September 2018
Declaration

I hereby declare that this thesis has been composed by myself and that the work, whether in the same or different form, has not been previously submitted to this or any other university for a degree.

I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Chapter 5, *Direct detection of gene conversions*, includes work completed in collaboration with Ivan Vogel (University of Copenhagen). My contribution to the work is clearly stated within the text, and the conclusions drawn are my own. This forms part of the article submission, currently available on BioRxiv (Vogel *et al.*, 2018), and submitted for peer review:

- Vogel I, Blanshard RC, Hoffmann ER. SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data. *bioRxiv* [Internet] 2018; Cold Spring Harbor Laboratory. Available from: https://www.biorxiv.org/content/early/2018/08/20/393256

Robert Blanshard
Dedication

To my parents, Rosie and Chris for their love and support, and for giving me the best opportunities to learn, to grow and to follow my passions.

To my elder siblings, Emily and Jonathan, and my cousins, aunties and uncles for their loving guidance in all aspects of life. I wouldn’t have been able to get to this stage without them.
Acknowledgements

Firstly, I would like to thank all the women who have kindly donated their eggs for this research. This resource of such precious material is pivotal to developing our basic understanding of the human germline.

Thank you to my collaborators at the Hoffmann Lab, GENERA (Valle Giulia Clinic, Italy), The Bridge Centre (London, UK), the Laboratory of Reproductive Biology (Rigshospitalet, Denmark) and each laboratory that provided data for Chapter 3; for providing samples, guidance, expert support and of course, friendship.

I would like to acknowledge the support and advice of my thesis committee, and their understanding as I faced obstacles during this project.

In particular I thank Eva Hoffmann for her dedicated supervision, and her mentorship both in science and life in general. Also, a thank you to Ben and the boys for hosting my many visits to Brighton and Copenhagen, and for allowing me to steal Eva’s precious weekend time to complete our work.

Thank you to Alan Handyside for supporting me at Illumina, and for giving me a fantastic insight into clinical embryology and cytogenetics.

I would like to thank Illumina for their support in my career development, in particular my colleagues and supervisors, Anthony Brown, Tristan Orpin and Philippa Burns, without whom this would not have been possible. I extend my thanks to the wider Assay Development and Bioinformatics teams in Cambridge for their technical guidance, patience and moral support whilst I managed my studies in between our other projects; I am truly grateful.

I also thank my dear friends for the frequent and necessary emotional support, and for sharing their own doctoral experiences which helped me greatly manage my research.

Last but not least, I thank my siblings, Emily and Jonathan, of whom I am so proud of their own achievements. You both inspire me every day, which always helps me push through when times are tough!
List of academic publications

  **Appendix A1.1.**

- Vogel I, **Blanshard RC**, Hoffmann ER. SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data. Submitted. **Appendix A1.2.**


*Contributed equally to the work
Abstract

In both natural and assisted human conceptions, the incidence of extra or missing chromosomes (aneuploidy) increases dramatically in women of advanced maternal age (35 years and above). An estimated 5-30% of all natural conceptions are thought to be aneuploid, of which 70-90% of supernumerary chromosomes are expected to be maternally inherited. In addition, structural abnormalities lead to segmental aneuploidies, which are poorly understood. In both cases, the resulting genomic imbalance is a major cause of infertility, pregnancy loss, and congenital disorders in human. Within the last five years, technological advances have improved the resolution for genome content analysis in trophectoderm biopsies (3-10 cells) taken from preimplantation embryos. Using these technologies and improving them to single cell resolution will allow unprecedented studies of the rates and origins of aneuploidies in the human germline.

In this thesis, I have assessed and discussed the clinical incidence and impact of mosaic and segmental chromosome abnormalities in preimplantation embryos (Chapter 3); the requirement for single cell genomics, and develop of a pipeline for concurrent NGS and SNP microarray analysis on individual products of meiosis (Chapter 4); and development of a high-precision, single cell genotyping algorithm through collaboration, to ask questions about genome diversification and human genome evolution (Chapter 5). Collectively, my findings show that the chromosomes are highly unstable in the human eggs and preimplantation embryos and discuss the implication for embryo testing.
Table of contents

Declaration ........................................................................................................................................... i
Dedication ............................................................................................................................................... ii
Acknowledgements .............................................................................................................................. iii
List of academic publications ............................................................................................................... iv
Abstract ............................................................................................................................................... v
List of tables .......................................................................................................................................... x
List of figures .......................................................................................................................................... xi
Abbreviations ........................................................................................................................................ xiii

Chapter 1: General Introduction ....................................................................................................... 1
  1.1 Aneuploidy is a major cause of pregnancy loss and congenital birth defects ........... 1
  1.1.1 The origins of cytogenetics ............................................................................................... 1
  1.1.2 Five percent of clinically recognised pregnancies from natural conceptions are aneuploid ........................................................................................................................................ 3
  1.1.3 Human female meiosis requires ordered segregation of chromosomes to generate the haploid gamete ....................................................................................................................................... 6
  1.1.4 The J-curve for maternal age-related aneuploidy is observed in human embryos .... 10
  1.1.5 Comprehensive chromosome screening of trophectoderm biopsies reveals mosaic and sub-chromosomal abnormalities with high dynamic range .................................................................... 16
  1.2 Requirement for single cell genomics to understand the origin of aneuploidy in meiosis ........................................................................................................................................... 21
  1.2.1 SNP markers are highly informative for inheritance patterns .................................... 22
  1.2.2 Methods of WGA limit inferences from single cell genomics .................................. 28
  1.2.3 Sources of human oocytes ............................................................................................... 33
  1.3 Single cell studies of genome diversification are confounded by technical limitations of high-resolution genotyping ......................................................................................................... 35
  1.3.1 SNP genotyping allows highly informative, cost-effective genome analysis ......... 35
  1.3.2 Genotype calling of single cell DNA is confounded by noise introduced by WGA .... 38
  1.3.3 Gene conversions facilitate genome diversification by breaking up the linkage disequilibrium between variants ................................................................................................................ 38
  1.4 Research proposal ..................................................................................................................... 40

Chapter 2: General materials and methods ..................................................................................... 41
  2.1 General laboratory equipment and materials ................................................................. 42
  2.2 DNA controls for assay development ............................................................................. 44
  2.2.1 Coriell permissions ....................................................................................................... 44
  2.2.2 Extraction of genomic DNA from cultured cell lines .............................................. 44
  2.2.3 Description of cell lines used for assay controls ..................................................... 44
Chapter 2: Methods

2.3 Clinical aneuploidy data derived from human embryos ........................................ 45
2.4 Collection of human oocytes and matching polar bodies ..................................... 46
2.4.1 Assisted reproductive technology pipeline ...................................................... 46
2.4.2 Fertility preservation pipeline ......................................................................... 46
2.5 Culture of cell lines for single cell isolation and genomic DNA extraction .......... 47
2.5.1 Cell line culture procedures at Illumina ......................................................... 48
2.5.2 Extraction of genomic DNA from cell cultures .............................................. 50
2.6 Single cell isolation from cultured cell lines ...................................................... 50
2.6.1 Single cell isolation buffer preparation ......................................................... 51
2.6.2 Dead cell removal from live cell cultures ...................................................... 51
2.6.3 Cell isolation and tubing .................................................................................. 52
2.7 Whole-genome amplification: multiple displacement amplification ................. 53
2.8 Quality assessment of DNA samples ................................................................. 54
2.8.1 Quality assessment by DNA fragment analysis .............................................. 55
2.8.2 Quality assessment by DNA quantification ..................................................... 56
2.9 Genotyping by SNP microarray ........................................................................... 57
2.9.1 Infinium™ HumanKaryomap-12 BeadChip Array .......................................... 58
2.9.2 Primary analysis of SNP microarray data ...................................................... 59
2.10 Next generation sequencing .............................................................................. 60
2.10.1 Copy number variation detection in SureMDA products .............................. 60
2.10.2 Whole-genome deep sequencing ................................................................... 63

Chapter 3: Chromosome constitution of human embryos ....................................... 66

3.1 Overview ............................................................................................................ 66
3.2 Materials and Methods ...................................................................................... 67
3.2.1 Study design ................................................................................................... 67
3.2.2 Embryo Culture and Biopsy ......................................................................... 68
3.2.3 Next-Generation Sequencing ....................................................................... 68
3.2.4 Primary analysis ............................................................................................ 69
3.2.5 Exclusion criteria .......................................................................................... 70
3.2.6 Classification of chromosomal abnormalities .............................................. 71
3.2.7 Statistical analysis ....................................................................................... 72
3.3 Results ............................................................................................................... 72
3.3.1 A dataset of 2,083 embryo biopsies after stringent quality control ............. 72
3.3.2 Whole chromosome aneuploidies increase with maternal age, reducing the incidence of embryos affected by only mosaic or segmental abnormalities .................. 73
3.3.3 Mosaic and segmental abnormalities do not share the chromosome bias profile of whole chromosome aneuploidy

3.3.4 Segmental aneuploidies frequently affect an entire chromosome arm.

3.3.5 Independent validation of chromosomal abnormalities reveals the reported incidence of mosaic abnormalities is affected by technical limitations.

3.3.6 Embryos affected by only mosaic or segmental abnormalities are present in cycles with a maternal age of 30-44 years.

3.4 Study conclusions

Chapter 4: Gross chromosomal rearrangements in human oocytes

4.1 Overview

4.2 Materials and Methods

4.2.1 Generating an MDA-specific reference database for NGS

4.2.2 NGS pipeline validation

4.2.3 Oocyte duos and trios

4.2.4 Exclusion criteria

4.3 Results

4.3.1 Reference database curation and pipeline validation using cell lines

4.3.2 Structural CNV detection by MDA and low-pass NGS identifies reciprocal segmental abnormalities in meiotic duos and trios and allows extraction of their breakpoints.

4.4 Study conclusions

Chapter 5: Direct detection of gene conversions

5.1 Overview

5.2 Materials and Methods

5.2.1 Cell lines

5.2.2 Whole genome amplification

5.2.3 BeadChip microarrays and SNP genotyping using GenCall

5.2.4 Generation of reference genotypes on the HumanKaryomap-12 v1.0 BeadChip Arrays

5.2.5 Genotyping in the single cell environment

5.2.6 Oocytes and polar bodies for mapping of gene conversions

5.2.7 PCR-free deep sequencing

5.3 Results

5.3.1 Generation of two single cell datasets from using MDA and SNP microarrays with stringent quality control

5.3.2 Improved genotype precision in single cell array data using two-stage machine learning
5.3.3 PCR-free NGS yields high genome coverage for single nucleotide variant calling in single cell MDA products .................................................................144
5.3.4 SureTypeSC accurately predicts putative gene conversions in individual human female meioses ......................................................................................148
5.4 Study conclusions .........................................................................................150
Chapter 6: Discussion .........................................................................................152
5.5 Addressing the research proposal .................................................................152
5.6 Solving the noise floor in single cell genomics ..............................................152
5.7 Limitations of the work ..................................................................................155
5.8 Conclusions, wider implications and future work .......................................156
References ............................................................................................................158
Appendix A1.1 .......................................................................................................176
Appendix A1.2 .......................................................................................................177
Appendix A1.3 .......................................................................................................178
Appendix A1.4 .......................................................................................................179
Appendix A1.5 .......................................................................................................180
Appendix A2 .........................................................................................................181
List of tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Historical overview of human aneuploidies resulting in live birth, their clinical features and reference to the first description and cytogenetic discovery.</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Sample stratification for the HapMap Project.</td>
<td>35</td>
</tr>
<tr>
<td>2.1</td>
<td>General laboratory equipment.</td>
<td>42</td>
</tr>
<tr>
<td>2.2</td>
<td>General laboratory materials.</td>
<td>43</td>
</tr>
<tr>
<td>2.3</td>
<td>Description of EBV-lymphoblastoid and fibroblast cell lines.</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>Cell culture-specific equipment and materials.</td>
<td>47</td>
</tr>
<tr>
<td>2.5</td>
<td>Materials for single cell isolation.</td>
<td>51</td>
</tr>
<tr>
<td>2.6</td>
<td>Materials for quality assessment of DNA.</td>
<td>54</td>
</tr>
<tr>
<td>2.7</td>
<td>Materials for genotyping by SNP microarray.</td>
<td>58</td>
</tr>
<tr>
<td>2.8</td>
<td>Materials for next generation sequencing.</td>
<td>60</td>
</tr>
<tr>
<td>3.1</td>
<td>embryo stratification categories and proposed ‘traffic-light’ system for clinical risk classification of embryos for transfer.</td>
<td>72</td>
</tr>
<tr>
<td>3.2</td>
<td>Stratification of PGT-A samples by collaborating laboratory.</td>
<td>73</td>
</tr>
<tr>
<td>3.3</td>
<td>Independent analysis of chromosomal abnormalities reported by Lab D.</td>
<td>88</td>
</tr>
<tr>
<td>3.4</td>
<td>Independent analysis of chromosomal abnormalities reported by Lab D, whilst excluding samples that were removed during independent review due to uneven NGS traces.</td>
<td>88</td>
</tr>
<tr>
<td>4.1</td>
<td>Sample stratification for MDA-NGS pipeline validation using cell lines of known karyotype.</td>
<td>97</td>
</tr>
<tr>
<td>4.2</td>
<td>Oocyte and polar body sample stratification for MDA-NGS pipeline.</td>
<td>99</td>
</tr>
<tr>
<td>4.3</td>
<td>Validation of CNVs in genomic DNA and single cells of known karyotype.</td>
<td>106</td>
</tr>
<tr>
<td>4.4</td>
<td>Segmental breakpoints detected in oocytes and polar bodies.</td>
<td>124</td>
</tr>
<tr>
<td>5.1</td>
<td>Stratification of cell line samples used for algorithm training and validation.</td>
<td>128</td>
</tr>
<tr>
<td>5.2</td>
<td>Metrics used to assess validation of the SureTypeSC genotyping tool.</td>
<td>131</td>
</tr>
<tr>
<td>5.3</td>
<td>Summary of genotype calls from single cells.</td>
<td>139</td>
</tr>
<tr>
<td>5.4</td>
<td>10-fold cross-validation using dataset GM12878.</td>
<td>141</td>
</tr>
<tr>
<td>5.5</td>
<td>Performance of tested classifiers using cell line GM12878.</td>
<td>142</td>
</tr>
<tr>
<td>5.6</td>
<td>Proportion of correct genotype calls when using GenCall and SureTypeSC on single cell array data from cell line GM12878.</td>
<td>143</td>
</tr>
<tr>
<td>5.7</td>
<td>Allele drop-in, allele drop-out and call rate with GenCall and SureTypeSC</td>
<td>144</td>
</tr>
<tr>
<td>5.8</td>
<td>Validation of SNP conversions detected using SureTypeSC by NGS.</td>
<td>150</td>
</tr>
</tbody>
</table>
List of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Chromosomal abnormalities detected by chorionic villus sampling after first trimester pregnancy loss.</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>Human female meiosis.</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Resumption and completion of meiosis II in natural conceptions.</td>
<td>9</td>
</tr>
<tr>
<td>1.4</td>
<td>Development and biopsy of human oocytes and preimplantation embryos.</td>
<td>11</td>
</tr>
<tr>
<td>1.5</td>
<td>Chromosomal abnormality types detected by PGT-A in trophectoderm biopsies of day-5 preimplantation embryos.</td>
<td>18</td>
</tr>
<tr>
<td>1.6</td>
<td>Meiotic origin of non-mosaic human trisomies.</td>
<td>21</td>
</tr>
<tr>
<td>1.7</td>
<td>Collecting the three single cell products of female meiosis.</td>
<td>25</td>
</tr>
<tr>
<td>1.8</td>
<td>Inferring the meiotic origin of chromosome segregation errors.</td>
<td>27</td>
</tr>
<tr>
<td>1.9</td>
<td>Whole-genome amplification noise in single cell DNA.</td>
<td>31</td>
</tr>
<tr>
<td>1.10</td>
<td>Single cell genomics workflow.</td>
<td>32</td>
</tr>
<tr>
<td>1.11</td>
<td>Homologous recombination gives rise to crossover and non-crossover events in meiosis.</td>
<td>39</td>
</tr>
<tr>
<td>2.1</td>
<td>Nunc™ IVF ICSI dish containing microdroplets for cell dilution and single cell isolation.</td>
<td>53</td>
</tr>
<tr>
<td>2.2</td>
<td>PCR program for Nextera*-based NGS library preparation from MDA products.</td>
<td>61</td>
</tr>
<tr>
<td>3.1</td>
<td>Proportion of biopsies categorised by PGT-A result.</td>
<td>75</td>
</tr>
<tr>
<td>3.2</td>
<td>Frequency of chromosomal abnormalities reported in 2,071 embryo biopsies following PGT-A.</td>
<td>77</td>
</tr>
<tr>
<td>3.3</td>
<td>Characterisation of uniform and mosaic segmental abnormalities in 2,083 embryo biopsies, stratified by size, chromosome arm affected (p or q) and copy number variation (gain or loss).</td>
<td>79</td>
</tr>
<tr>
<td>3.4</td>
<td>NGS trace of trophectoderm biopsy 44128_2, showing copy number variations with low technical noise following PicoPLEX WGA.</td>
<td>81</td>
</tr>
<tr>
<td>3.5</td>
<td>NGS traces of trophectoderm biopsies with discordant uniform segmental copy number variations, following PicoPLEX WGA, between PGT-A Lab D and an independent review.</td>
<td>82</td>
</tr>
<tr>
<td>3.6</td>
<td>NGS traces of trophectoderm biopsies with discordant mosaic segmental copy number variations, following PicoPLEX WGA, between PGT-A Lab D and an independent review.</td>
<td>83</td>
</tr>
<tr>
<td>3.7</td>
<td>NGS traces of trophectoderm biopsies with discordant whole chromosome aneuploidies, following PicoPLEX WGA, between PGT-A Lab D and an independent review.</td>
<td>85</td>
</tr>
<tr>
<td>3.8</td>
<td>NGS traces of trophectoderm biopsies with discordant mosaic, whole chromosome aneuploidies, following PicoPLEX WGA, between PGT-A Lab D and an independent review.</td>
<td>87</td>
</tr>
<tr>
<td>3.9</td>
<td>Incidence of reported embryo abnormalities per cycle, as a function of maternal age.</td>
<td>90</td>
</tr>
<tr>
<td>4.1</td>
<td>Quality control of WGA and library preparation for genomic DNA and cell lines of known karyotype.</td>
<td>103</td>
</tr>
<tr>
<td>4.2</td>
<td>Box plot for primary low-pass, next-generation sequencing metrics for bulk genomic DNA and single cell isolates from four cell lines.</td>
<td>105</td>
</tr>
<tr>
<td>4.3</td>
<td>CNV validation in MDA products from gDNA samples of known karyotype.</td>
<td>107</td>
</tr>
<tr>
<td>4.4</td>
<td>CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM04927 (47, XY, +21).</td>
<td>108</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM04927 (47, XY, +21), with additional CNVs detected in chromosome 1 (+1q) and chromosome 16 (-16q).</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>CNV validation following MDA whole genome amplification and NGS, in single cells from cell line GM00526 (47, XY, +13).</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>CNV validation following MDA whole genome amplification and NGS, in single cells from cell line GM10985 (46, XX, -3p).</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM50121 (46, XY, -18p +18q).</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>Quality control of MDA products derived from oocytes and polar bodies.</td>
<td></td>
</tr>
<tr>
<td>4.10</td>
<td>Copy number variation NGS traces from oocytes collected from two different clinics</td>
<td></td>
</tr>
<tr>
<td>4.11</td>
<td>Copy number variation NGS traces for oocyte-Polar body trio 942.44.</td>
<td></td>
</tr>
<tr>
<td>4.12</td>
<td>Copy number variation NGS traces for oocyte-polar body duo PT1053-1.</td>
<td></td>
</tr>
<tr>
<td>4.13</td>
<td>Copy number variation NGS traces for oocyte-polar body duo 933.21, aligned with the assumptions of diploid and triploid ground states.</td>
<td></td>
</tr>
<tr>
<td>4.14</td>
<td>Copy number variation NGS traces for oocyte-polar body duo PT1052-9.</td>
<td></td>
</tr>
<tr>
<td>4.15</td>
<td>Copy number variation NGS traces for oocyte-polar body duo GEN128-50.</td>
<td></td>
</tr>
<tr>
<td>4.16</td>
<td>Copy number variation NGS traces for oocyte-polar body duo PT1050-16.</td>
<td></td>
</tr>
<tr>
<td>4.17</td>
<td>Copy number variation NGS traces for oocyte-polar body duo 860.2.</td>
<td></td>
</tr>
<tr>
<td>4.18</td>
<td>Fragmentation of the first polar body following in vitro maturation of GV-stage oocytes.</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Signal-noise detection in whole-genome amplified DNA from single cells.</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Quality control of single cell MDA.</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Genotype call rates for single cells and bulk genomic DNA hybridised to HumanKaryomap-12 BeadChip arrays and analysed in GenomeStudio.</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>SNP microarray assay controls exported from GenomeStudio for arrays scanned using an iScan System.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>SNP microarray assay controls exported from GenomeStudio for arrays scanned using a NextSeq550 System.</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>SureTypeSC improves the performance for single cell genotyping.</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Post-fragmentation DNA library quality control using bioanalyzer.</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>Post-library preparation quality control using bioanalyzer.</td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td>Distribution pattern and frequency of genome-wide markers on the HumanKaryomap-12 SNP array.</td>
<td></td>
</tr>
<tr>
<td>5.10</td>
<td>Direct detection of gene conversions in an oocyte-PB1-PB2 trio.</td>
<td></td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCGH</td>
<td>Array comparative genomic hybridisation</td>
</tr>
<tr>
<td>AOA</td>
<td>Artificial oocyte activation</td>
</tr>
<tr>
<td>ART</td>
<td>Assisted reproductive technology</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary sequence alignment map (*.bam)</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variation</td>
</tr>
<tr>
<td>DLR</td>
<td>Derivative log ratio</td>
</tr>
<tr>
<td>DSB</td>
<td>Double strand break</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double-stranded or duplex DNA</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescence in situ hybridisation</td>
</tr>
<tr>
<td>FN</td>
<td>False negative</td>
</tr>
<tr>
<td>FP</td>
<td>False positive</td>
</tr>
<tr>
<td>GCR</td>
<td>Gross chromosomal rearrangement</td>
</tr>
<tr>
<td>GDA</td>
<td>Gaussian discriminant analysis</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic DNA</td>
</tr>
<tr>
<td>GV</td>
<td>Germinal vesicle</td>
</tr>
<tr>
<td>ICSI</td>
<td>Intracytoplasmic sperm injection</td>
</tr>
<tr>
<td>IVM</td>
<td>In vitro maturation</td>
</tr>
<tr>
<td>LCL</td>
<td>Lymphoblastoid cell line</td>
</tr>
<tr>
<td>MDA</td>
<td>Multiple displacement amplification</td>
</tr>
<tr>
<td>MI</td>
<td>Meiosis I</td>
</tr>
<tr>
<td>MII</td>
<td>Meiosis II</td>
</tr>
<tr>
<td>NDJ</td>
<td>Non-disjunction</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>NTC</td>
<td>'No template' control</td>
</tr>
<tr>
<td>PB1</td>
<td>First polar body</td>
</tr>
<tr>
<td>PB2</td>
<td>Second polar body 2</td>
</tr>
<tr>
<td>PGT-A</td>
<td>Preimplantation genetic testing for aneuploidy</td>
</tr>
<tr>
<td>PGT-M</td>
<td>Preimplantation genetic testing for monogenic disease</td>
</tr>
<tr>
<td>PGT-SR</td>
<td>Preimplantation genetic testing for structural rearrangements</td>
</tr>
<tr>
<td>PSSC</td>
<td>Precocious separation of sister chromatids</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative PCR</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RSB</td>
<td>Resuspension buffer</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SPRI</td>
<td>Solid phase reversible immobilisation</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole genome amplification</td>
</tr>
</tbody>
</table>
Chapter 1: General Introduction

1.1 Aneuploidy is a major cause of pregnancy loss and congenital birth defects

1.1.1 The origins of cytogenetics

Scientific research is an iterative process, in which our knowledge of biological systems is often built on the work of others, underpinned by technological developments that drive future discovery. In order to push the envelope regarding the origins and characterisation of human genome changes, we must first reflect on the well-reviewed origin of human cytogenetics, and the milestone iterations that paved the way for high resolution, molecular studies (Jacobs, 2014; Ferguson-Smith, 2015). Cytogenetics is the study of chromosomes; in relation to their inheritance, structure and function, particularly in cases of disease or disorder. Therefore, the term aneuploidy (meaning the presence of an abnormal number of chromosomes in a cell), is redundant without first establishing the euploid state. In 1956, Tjio & Levan and then Ford & Hamerton, challenged Painter’s 1924 conclusion and revealed the true chromosome number in human to be 46 rather than 48. Surprisingly, despite having the technology to observe and count individual chromosomes, the requirement for high quality cell material, and refined culture and chromosome preparation techniques still led to debate over the number of chromosomes observed in a given cell. Advanced techniques to arrest cells at metaphase allowed easier observation and counting of individual chromosomes. It was not until late 1958 that Jacobs & Strong reported a 47, XXY chromosome constitution during a blinded study, in a patient with suspected Klinefelter syndrome; the first aneuploidy observed in human (Table 1.1). In 1959, the link between congenital disorders and chromosome imbalance was in the spotlight, with aneuploid karyotypes confirmed in quick succession for patients with Down syndrome (Jacobs et al., 1959b; Lejeune et al., 1959), Turner syndrome (Ford et al., 1959a), Triple X syndrome (Jacobs et al., 1959a), trisomy 13 (Patau syndrome) (Patau et al., 1960), trisomy 18 (Edward’s syndrome) (Edwards et al., 1960) and XYY (Jacob’s) syndrome (Hauschka et al., 1961). These observations were all made from analysis of chromosome preparations from bone marrow or fibroblast cultures and were prompted by the severe morphological and developmental difficulties of the affected individuals.
Table 1.1: Historical overview of human aneuploidies resulting in live birth, their clinical features and reference to the first description and cytogenetic discovery.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Clinical features</th>
<th>Cytogenetic discovery</th>
<th>Karyotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klinefelter syndrome</td>
<td>Male with tall stature</td>
<td>Jacobs and Strong, 1959</td>
<td>47, XXY</td>
</tr>
<tr>
<td>Klinefelter et al., 1942</td>
<td>Slow motor development of new-borns</td>
<td>Ford et al., 1959b</td>
<td>47, XXY / 46, XX (mosaic)</td>
</tr>
<tr>
<td>Down's syndrome</td>
<td>Mental impairment</td>
<td>Lejeune et al., 1959</td>
<td>47, +21</td>
</tr>
<tr>
<td>Down, (1866)</td>
<td>Stunted growth</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dysmorphism of the head and neck</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turner syndrome</td>
<td>Female with short stature</td>
<td>Ford et al., 1959a</td>
<td>45, X</td>
</tr>
<tr>
<td>Turner, 1938</td>
<td>Short neck with webbed appearance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Absence of puberty, ovarian dysfunction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Problems with mathematical learning and visual-spatial coordination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple X syndrome</td>
<td>Female with tall stature</td>
<td>Jacobs et al., 1959a</td>
<td>47, XXX</td>
</tr>
<tr>
<td></td>
<td>Learning and behavioural disabilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delayed development of speech and language skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delayed motor development, weak muscle tone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patau syndrome</td>
<td>Mental impairment</td>
<td>Patau et al., 1960</td>
<td>47, +13</td>
</tr>
<tr>
<td>Bartholin, 1656</td>
<td>Holoprosencephaly, microcephaly.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dysmorphism of the face, hands and feet.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Abnormal development of nervous, musculoskeletal, and urogenital systems.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edward's syndrome</td>
<td>Severe mental impairment</td>
<td>Edwards et al., 1960</td>
<td>47, +18</td>
</tr>
<tr>
<td></td>
<td>Structural heart defects, cardiac anomalies</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delayed growth and development</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Microcephaly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jacob's syndrome</td>
<td>Male with tall stature</td>
<td>Hauschka et al., 1961</td>
<td>47, XYY</td>
</tr>
<tr>
<td></td>
<td>Delayed motor development</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delayed or difficult speech</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consistent with Autism spectrum disorder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Learning and behavioural disabilities</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The technology of the time relied on the collection and growth of cell cultures, and incubation with colchicine to cause mitotic arrest. By mixing the metaphase chromosomes with a hypertonic solution, they could be spread under a coverslip and viewed in a single optical plane with a light microscope. The resolution was sufficient for Nowell and Hungerford (1960) to identify a consistent chromosomal abnormality in chronic granulocytic leukaemia. The loss of half the q-arm of this small chromosome was attributed to the neoplastic character of the leukaemia cells. With the development of protocols for air-dried preparations from lymphoblast cultures, the resolution of chromosome analyses increased. However, further developments were required to improve reliability, since chromosome analysis had begun to be used for prenatal diagnosis in the late 1960s (Steele and Breg, 1966). A breakthrough that allowed chromosomes to be differentiated by features other than size came in 1969, with the discovery of chromosome banding in the bean plant, *Viva faba* (Caspersson et al., 1969). The phenomena resulted from differential intercalation of quinacrine into DNA, seen as light and dark ‘Q-bands’ under UV light. This important observation was later found to correspond directly with the banding pattern produced by selective Giemsa staining of denatured chromosomes, where each chromosome exhibits a unique pattern of light and dark bands (Seabright, 1971; Sumner et al., 1971). Trypsin G-banding was quickly adopted for diagnostic cytogenetics, and with it the original discovery of Nowell and Hungerford, (the Philadelphia chromosome) was resolved to be a reciprocal translocation between chromosomes 9 and 22 \( t(9;22)(q34; q11) \) (Rowley, 1973).

This is just one demonstration of how technological improvements develop basic knowledge and understanding in science.

1.1.2 Five percent of clinically recognised pregnancies from natural conceptions are aneuploid.

Following the first reports of aneuploidy, scientists looked at pregnancy loss since many of the clinical manifestations were severe. The notion that aneuploidy causes pregnancy loss is consistent with one third of pregnancies in Western and Asian populations ending in loss (Wilcox *et al.*, 1988; Wang *et al.*, 2003). This rate is very high, and we now know that at least 50-60% of clinical pregnancy losses are caused by chromosome abnormalities. By 1965, David Carr had twice demonstrated that approximately one quarter of spontaneous abortions contained supernumerary or missing chromosomes, and that these imbalances were a probable cause of pregnancy loss (Carr, 1963, 1965). Whilst we now know the rate is much higher than this, this was the first estimation of the effects of aneuploidy on pregnancy loss. The analysis of cultured foetal tissue revealed monosomy X, trisomy of a single autosome, and triploidy to be common abnormalities. In a seminal paper by Boué *et al.* (1975), it was concluded that out of the 15% of
clinically recognised pregnancies which spontaneously abort, half (7%) were lost due to a chromosome abnormality. The study correlated prospective information about the pregnancy (<12 weeks gestation) with retrospective analysis of spontaneously aborted foetus karyotypes for 1100 women, and supports recent data derived from chorionic villus sampling of 1,011 pregnancy losses occurring within the first trimester (Soler et al., 2017). To this end, given 54% of the observed abnormalities were trisomy or double trisomy events, Boué et al. hypothesised that an equal number or monosomy events (excluding XO observed in affected Turner’s syndrome) may also occur; however, these zygotes are lost prior to establishing as clinically recognised pregnancies. Considering the incidence of triploidy (20%), tetraploidy (6%), structural abnormalities (4%) and those seen in livebirths, the group estimated that 50% of natural human conceptions could have a chromosomal abnormality. More specifically, autosomal trisomy’s are more frequently observed for chromosomes 13, 15, 16, 18 and 22, with chromosomes 16 and 22 accounting for 37% of trisomy’s detected in pregnancy losses (Figure 1.1A) (Soler et al., 2017).

Boué et al. also observed a maternal age effect for trisomy, that was statistically supported by studying the effect of both maternal or paternal ages in 4000 trisomy 21 cases (Erickson, 1978). The large dataset enabled the age effect of each sex to be investigated, whilst controlling for age in the other. The study concluded that birth order and maternal, but not paternal age, had a significant effect on the incidence of trisomy 21. These observations were further supported by Hassold & Chiu (1985) with a larger study of 2,264 pregnancy losses. In summary, 50.4% of losses were chromosomally abnormal with trisomy accounting for 30% of aneuploidies. When stratified by maternal age, the percent of spontaneous abortions (assuming an overall rate of 15%), was constant between the ages of 20-30 years (10-15%) but rose to 40% by the age of 40 years. Interestingly, a slightly higher percentage (15%) of chromosomally abnormal pregnancy losses were identified in young women <18 years which reduced to 11% at the age of 20 years. This was consistent with Erickson (1978) and gave a distinct ‘J’ shape to the age-effect curve. Trisomy affected 20-33% of pregnancy losses from women <35 years, and >67% of losses from women ≥40 years. Chromosome 16 was the only chromosome to follow a uniform maternal-age effect, increasing steadily with age as opposed to a subdued level at 20-35 years followed by an exponential increase. This data is supported by Soler et al. (2017), whereby a significant increase in the incidence of autosomal trisomy is observed between maternal patients <35 or ≥35 years, with the exception of chromosomes 13, 18 and 16 (Figure 1.1B). In addition, the incidence of triploidy in pregnancy losses significantly decreases with maternal age, whilst the incidence of monosomy X is constant, accounting for ~10% of losses in both groups. Taken together, no less than 5% of all natural, clinically recognised pregnancies are
aneuploid. Of the live births, an estimated 0.3% are aneuploid, of which 0.13% account for trisomy 21 (Summarised by Hassold & Jacobs, (1984)). Therefore, aneuploidy is a leading cause of natural pregnancy loss and a major cause of congenital birth defects.

**Figure 1.1:** Chromosomal abnormalities detected by chorionic villus sampling after first trimester pregnancy loss. (A) Frequency of single trisomies by autosome, detected after analysis of 1,011 pregnancy losses. (B) Proportion of chromosome abnormalities by maternal age group. *Indicates significant difference between age groups (p<0.05). Error bars show Standard Error of Proportion. Data adapted from Soler et al., (2017).
1.1.3 Human female meiosis requires ordered segregation of chromosomes to generate the haploid gamete

Molecular analyses have been used to identify the parental origin of aneuploidies. Restriction fragment length polymorphisms were used to infer haplotypes based on the variation in fragment lengths generated by DNA restriction digestion between individuals. Each fragment length is considered an allele and can be resolved by gel electrophoresis. Single nucleotide polymorphism (SNP) genotyping by microarray is the modern high-resolution equivalent and has determined that the majority of aneuploidies are of maternal origin, with the exception of the sex chromosomes (Hassold et al., 1990). The detection of the unique SNP fingerprints from both maternal homologues in the peri-centromeric region (where crossovers are unlikely to form), provides evidence of chromosome missegregation in the first meiotic division (Ottolini et al., 2015). This, together with the evidence supporting the maternal age effect of aneuploidy, led researchers to study human female meiosis for the aetiology of chromosome imbalances (Hassold et al., 2007).

Meiosis is used by sexually reproducing organisms to halve the chromosome sets in their gametes. This is achieved by using a single round of DNA replication followed by two nuclear divisions. The products of human gametogenesis, spermatozoa and oocytes, constitute the respective male and female counterparts. Upon fertilisation the zygote inherits one compliment of chromosomes from the egg and the other from the sperm, restoring the diploid state (Hassold and Hunt, 2001). Whilst spermatogenesis takes ~60 days in the adult testis, in human females, meiosis takes more than a decade to complete, which is believed to contribute to the age-related increase in aneuploid eggs in human (Hassold and Hunt, 2001). Chromosome missegregation is more frequent in mammals than in other model organisms. However, the error rate in human oocytes (20-30%) is far greater than those of the mouse (1-2%) and yeast (0.01%) (Hassold and Hunt, 2001). To study chromosome missegregation, it is important to first establish the meiotic program that gives rise to euploid gametes.

Human female meiosis is initiated in oocyte precursor cells at 11-12 weeks gestation in the foetal ovary (Gondos et al., 1986). This is in contrast to male meiosis, which initiates in puberty and continues without interruption. Each homologous chromosome (homolog) is replicated and the two sister chromatids are held together by cohesion, mediated by protein complexes (Hunt and Hassold, 2008). Early in prophase I, homologs pair and cross over, the reciprocal exchange of genetic material that gives rise to recombinant chromosomes (Figure 1.2; left).
Figure 1.2: Human female meiosis. A single round of DNA replication is followed by two nuclear divisions. Sister chromatids are held together lengthwise by cohesin linkages. Homologous chromosomes pair and crossover during prophase I before arresting at the diplotene stage (dictyate arrest). Crossovers manifest as chiasmata and together with sister chromatid cohesion, maintain the bivalent structure until the oocyte is recruited for maturation (10-50 years later). At meiosis I, cohesion is lost along chromosome arms but maintained at the centromere. Homologs segregate and half are extruded in the first polar body (PB1). Meiosis II initiates but arrests during metaphase II, until activation by sperm or exogenous stimuli in vitro. Once activated, all cohesion is lost, sister chromatids segregate and half are extruded in the second polar body (PB2).

Crossovers can be detected cytologically as chiasmata, which together with sister chromatid cohesion form the cruciform or bivalent structures that promote correct homolog segregation at meiosis I (MI) (Hassold and Hunt, 2001). Several lines of evidence support the notion that chiasmata are important for accurate segregation of homologs at MI in females. First, in mice that fail to form or maintain chiasmata, there is an increased risk of precocious separation of sister chromatids (PSSC) in MI, or non-disjunction (NDJ) in either MI or meiosis II (MII) (Hodges et al., 2005; Hunt, 2006). Second, in human, Down Syndrome individuals show an increased preponderance of two maternal chromosomes 21 that are inferred to have failed to crossover (Lamb et al., 1996). Finally, the position of the crossover may also influence segregation, where those which are too far from centromeres (‘too cold’) or too close (‘too hot’) predispose failure to maintain chiasmata (Hassold and Hunt, 2001; Nagaoka et al., 2012). Therefore, changes in both the rate and location of recombination events between homologous chromosome pairs, predispose segregation errors in either MI or MII.

Unequal nuclear division at either MI or MII is widely attributed to the inefficient maintenance of the bivalent structure during the first meiotic arrest. The chiasmata, in combination with cohesin linkages, are important for establishing the bivalent structure; for each homologous chromosome pair. Bivalent formation is associated with the entry into the first meiotic arrest at the diplotene stage of prophase I. This protracted dictyate arrest lasts for a
staggering 10-50 years, at which stage the bivalent structure that is made up of chiasmata and sister chromatid cohesion, must be maintained. During the arrest, the germinal vesicle (GV) stage oocytes become surrounded by somatic pregranulosa cells forming primordial follicles; however only a fraction of these will survive to term with a large proportion of oocytes lost to atresia (apoptotic death of follicles) (Hassold and Hunt, 2001; Martins da Silva et al., 2004). In the mouse, the presence of meiosis-specific cohesion transcripts do not facilitate cohesion loading past the foetal ovary stage (Revenkova et al., 2010). Similarly, a cleavage assay using an ectopic transgene of the cohesion complex kleisin subunit, Rec8, indicates there is no, or very low turnover of the original cohesion proteins during dictyate arrest in the mouse (Tachibana-Konwalski et al., 2010). In contrast, new cohesive linkages following S-phase are required in Drosophila, for faithful segregation of chromatids at MI and MII (Weng et al., 2014). However, no such model for cohesion rejuvenation has been demonstrated in humans. The original prophase cohesion proteins are understood to be necessary and sufficient to maintain the bivalent structure throughout dictyate arrest. Therefore, premature loss of chromatid cohesion predisposes chromatid segregation errors in meiosis.

When the woman reaches menarche (reproductive maturity) the follicle stimulating hormone, in conjunction with pituitary gonadotropins and luteinising hormone, signal the recruitment, growth and maturation of primordial follicles. Until menopause, follicle recruitment is a continuous process, such that one follicle completes its growth each month. Meiosis I resumption (maturation) of the oocyte is regulated by a surge in luteinising hormone levels. This triggers the removal of cohesin from chromosomal arm regions such that the chiasmata can be dissolved and the two homologs released from each other (Nagaoka et al., 2012). Cohesion is maintained between the centromeres of sister chromatids, such that they are physically associated until completion of the MII. One set of chromosomes remain in the egg, while the other is extruded in the very small polar body (PB1) (Figure 1.2; centre) (Nagaoka et al., 2012). After completion of MI, the secondary oocyte initiates MII without delay, but arrests shortly after the chromosomes align on the metaphase plate. In natural cycles, the MII arrest lasts from ovulation until fertilisation; typically, 24-72 hours.

At fertilisation, a capacitated sperm penetrates the zona pellucida and fuses with the oolemmal membrane by means of two highly conserved receptor orthologues (Izumo1 and Juno in mouse) (Bianchi et al., 2014) (Figure 1.3A).
Figure 1.3: Resumption and completion of meiosis II in natural conceptions. (A) Natural oocyte activation by fertilisation. A capacitated sperm fuses with the oolemma and injects phospholipase C zeta (PLCζ) into the cytoplasm. PLCζ facilitates production of PtdIns(1,4,5)P₃ (IP₃). IP₃ binds the receptor IP₃R on the endoplasmic reticulum (ER), triggering oscillating waves of calcium (Ca²⁺) release. Calcium signals the resumption and completion of meiosis II; including extrusion of the PB2 and transition to mitotic cell division. (B) Exit from meiosis. In the zygote, maternal and paternal haploid pronuclei (PN) fuse, followed by transition to mitotic cell division (cleavage stage of early embryo).

Figure A was re-drawn and adapted from Clift & Schuh (2013); green and yellow chromosome shading represent maternal haplotypes, for visualisation of recombinant chromatids; chromosomes not to scale.

Human oocyte activation is initiated by phospholipase C zeta (PLCζ), a sperm-specific protein that is injected into the cytosol of the oocyte. PLCζ has been shown to hydrolyse phosphatidylinositol 4,5-bisphosphate (PIP₂) to inositol 1,4,5-trisphosphate (IP₃), which acts as a second messenger to release Ca²⁺ from the endoplasmic reticulum in the oocyte (Saunders et al., 2002). Ca²⁺ oscillations trigger the resumption and completion of MII, whereby sister chromatids segregate. One set of chromatids remains in the oocyte and the other is extruded in the second polar body (PB2) (Figure 1.2; right) (Swann and Yu, 2008; Berridge, 2009; Clift and Schuh, 2013). The enzymatic content of the cortical granule causes the hardening of the zona pellucida (Balakier and Casper, 1993; Bianchi et al., 2014), which in conjunction with the dissociation of the Juno orthologue (in humans) from the oolemma, blocks further entry of sperm and thus prevents the formation of polyploid embryos (Bianchi et al., 2014). Following extrusion of the second polar body, the male and female pronuclei form and migrate towards each other and fuse, thus restoring the diploid state in the zygote. This marks the transition from meiotic to mitotic cell divisions; the initial 2-3 divisions are driven by maternal mRNA and proteins until embryonic genome activation facilitates transcription. This is followed by epigenetic reprogramming and the differentiation of cell lineages (Figure 1.3B) (Dobson et al., 2004; Niakan et al., 2012).

In conclusion, human female meiosis is highly protracted in comparison to spermatogenesis, in which the meiotic division occur sequentially without delay. The dictyate
arrest arguably places stress on genome stability and the premature loss of cohesion is hypothesised to predispose chromosome segregation errors in meiosis I, as a factor of maternal age.

1.1.4 The J-curve for maternal age-related aneuploidy is observed in human embryos

The incidence of aneuploidy in clinically recognised pregnancies from natural human conceptions increases dramatically in women aged ~35 years and over (Hassold and Hunt, 2001). However, this population does not represent the aneuploid embryos that fail to implant in the uterine wall or miscarry shortly after. Natural conceptions reveal only the tip of the iceberg with regards to the frequency of chromosome transmission errors in human gametogenesis or early embryogenesis. In 1975 Boué et al. predicted that for every trisomy seen in clinically recognised pregnancies, there was an equal monosomy. With the advent of assisted reproductive technology (ART), it is possible to have broad access to human oocytes from immature stages of meiosis (GV-stage), up to the 6-7th day of embryo development prior to transfer to the uterine wall. This allows the molecular analysis of a larger proportion of eggs prior to in vivo selection, however these are still limited to the clinical population and ovarian stimulation regimes for follicle collection. What is seen, is a similar curve for age-related increase in aneuploidy as observed for the incidence of trisomy in natural conceptions, however in the clinic the curve is shifted both in amplitude and towards higher age (Hassold and Hunt, 2001; Franasiak et al., 2014; McCoy et al., 2015).

1.1.4.1 The ART setting allows observations from oocyte to blastocyst, providing unique opportunities for molecular genetic content analysis

The early embryonic cell divisions in humans were first characterised by Hertig and Rock (Hertig et al., 1954, 1956), and much of our understanding is gained in vitro in the ART setting; which follows the embryo from fertilisation (day 0) up until transfer (Day 6 or 7). Following Development and biopsy of human oocytes and preimplantation embryos fusion of the maternal and paternal pronuclei, the embryo cleaves in a series of cell divisions in which the net size of the embryo remains the same (Figure 1.4A). The cells replicate equilaterally, forming 2-, 4- and 8-cell intermediates. Cleavage to the 8-cell stage typically competes on day 3, after which cells compact, forming a symmetrical ball of 16 blastomere cells called the morula. A major wave of embryonic transcriptional activity is reported between the 4- and 8-cell stages; indicating activation of the embryonic genome (Dobson et al., 2004). Symmetry is broken following compaction, where further cell divisions and cavitation yields the blastocyst on day 5, comprising
of the inner cell mass with a trophectoderm (TE) surround. The inner cell mass further differentiates into the epiblast and primitive endoderm prior to implantation; the trophectoderm will give rise to the placental tissues. On day 6, the blastocyst 'hatches' from the zona pellucida, to facilitate implantation in the uterine wall on day 7 (Niakan et al., 2012). The process is morphologically similar in mice, however less protracted, with blastocyst formation occurring on day 3-4. In addition, the trophectoderm of the mouse oocyte gives rise to a pool of extra-embryonic ectoderm cells through budding, in comparison to an invasive villous structure of cytotrophoblast cells in human oocytes (Niakan et al., 2012). The ART setting provides unique opportunities to study human oocytes and embryos directly.

Figure 1.4: Development and biopsy of human oocytes and preimplantation embryos. (A) Human preimplantation embryo development. (B) Biopsy technique for day-5 embryo. Briefly, a hole is made in the ZP by targeted laser pulses (red dashed line). The TE herniates through the opening in the ZP, allowing removal of 3-10 cells by aspiration with the biopsy pipette and targeted laser pulses at cell junctions. The biopsy is washed and tubed prior to whole-genome amplification. (C) Sequential polar body biopsy and activation of a human oocyte (Sample PT1063-68) with calcium ionophore. Biopsy is performed using the approach of Figure C, to yield an oocyte-PB1 duo (without activation) or an oocyte-PB1-PB2 trio (with activation). Figure A was re-drawn and adapted from Niakan et al., (2012); Not to scale. TE = Trophectoderm; ICM = Inner cell mass; ZP = Zona pellucida; PN = Pronucleus.
The culture of preimplantation embryos in vitro provides a unique opportunity to perform molecular studies of genome content and aneuploidy assessment. The technique of human embryo biopsy for preimplantation genetic diagnosis was pioneered by Handyside et al. (1989) and involved erosion of the zona pellucida by acid Tyrode’s solution on day 3, and extraction of a one or two blastomeres by micropipette. More recently, the biopsy of 3-10 trophectoderm cells from the day-5 blastocyst has been favoured in the ART setting due to improved cytogenetic accuracy and pregnancy outcomes, when compared to biopsy and analysis of multiple single cell day-3 blastomeres. Modern biopsy procedures typically use targeted laser pulses to open the zona pellucida and separate the biopsy from the embryo (Figure 1.4B). In a similar fashion, polar bodies can also be biopsied. To ensure accurate sample tracking, the polar bodies are biopsied sequentially, such that the PB1 is removed from the MII-arrested oocyte prior to oocyte activation, and the PB2 is biopsied upon its extrusion (Figure 1.4C). Biopsy of polar bodies, day-3 blastomeres and day-5 trophectoderm samples is an invasive procedure, and can result in degradation of the oocyte, embryo or biopsy itself. Taken logically, the day-5 biopsy is being increasingly used to assess genomic integrity of embryos, since it covers the latest stage of embryo development prior to transfer. Furthermore, genetic analysis of a multi-cell sample offers a lower noise profile when compared to single cell day-3 blastomeres. This assumption is supported by improved implantation and pregnancy outcomes in the ART setting when compared to polar body or blastomere analysis, and efficiency of the trophectoderm biopsy procedure is high amongst highly trained embryologists (Cimadomo et al., 2016).

1.1.4.2 Historical aspects of embryo screening

The non-destructive molecular analysis of human embryos has valuable clinical applications in improving pregnancy rates, and detection of monogenic disease prior to embryo transfer in the ART clinic. In 1989, it was demonstrated that PCR analysis of single blastomeres could accurately sex embryos based on detection of Y-chromosome-specific amplicons (Handyside et al., 1989). The significance of this was that couples who were known carriers of an inheritable X-linked disorder, could elect not to transfer male embryos following IVF. This approach, termed preimplantation genetic testing for monogenic disease (PGT-M), mitigates the dilemma to abort an established male pregnancy following sexing by chorionic villus sampling (11-14 weeks) or amniocentesis (15-20 weeks). The technique was first used clinically in 1990 and resulted in the birth of five healthy girls in three cases (Handyside et al., 1990). Moreover, data at the time indicated that biopsied day-3 embryos had a similar developmental potential to the blastocyst
stage as unbiopsied embryos (~70%), albeit known that embryo morphology was not predictive of successful implantation and pregnancy (Hardy et al., 1990).

With the concern that contamination or mis-sampling of anuclear embryo fragments could lead to clinical misdiagnosis using PCR, researchers looked to hybridisation of molecular probes for visual quantification of sex chromosomes. Of the available technologies, radiolabelled oligonucleotide probes took several days to visualise, necessitating improved outcomes following embryo freezing; and the rapid staining of biotinylated probes bound to streptavidin-linked alkaline phosphatase lacked sensitivity in embryo samples. Instead, hybridisation of fluorescently labelled oligonucleotide probes to interphase or metaphase chromosomes was soon adopted for its ability to screen multiple chromosomes at the same time with different fluorochromes (Nederlof et al., 1989; Griffin et al., 1991, 1992). Fluorescent in situ hybridisation (FISH) was inexpensive, highly sensitive and enabled quantitation of chromosome copy number at the single cell level. In early use, FISH was used for sex determination in PGT-M cases (X and Y chromosomes), but alpha-satellite probes were soon utilised for around half of the autosomes. The clinical benefit of aneuploidy detection by FISH in preimplantation embryos was soon realised for common ‘viable’ aneuploidies in chromosomes 13, 18, 21, X and Y (Delhanty et al., 1993; Munne et al., 1993).

Studies also provided supporting evidence for the age-related incidence of maternal trisomy, including that of chromosome 16 (Munné et al., 1995; Benadiva et al., 1996). The widespread use of FISH as an adjuvant therapy to ART marks the origin of preimplantation genetic testing for aneuploidy (PGT-A) as a discipline within fertility treatment. PGT-A describes the genome-wide testing for chromosome copy number variation and is distinct from preimplantation genetic testing for monogenic disease (PGT-M), which is a targeted screen to identify embryos as risk of carrying a specific monogenic disorder. Classical PGT-A requires embryo biopsy, and therefore poses a risk to embryo development. For this reason, PGT-A is offered only to patients at high-risk of aneuploidy (i.e. recurrent miscarriage or advanced maternal age) and is not routine procedure in ART. The aim of PGT-A is to identify chromosomally abnormal embryos within a given cycle and prioritise euploid embryos for transfer to the uterus. Undetected chromosome abnormalities may prevent the implantation and development of an embryo, leading to miscarriage, spontaneous abortion or congenital birth defects. In theory, PGT-A reduces the financial, physical and psychological strains of infertility treatment for the patient, by reducing miscarriage rates.

Whilst the clinical utility and popularity of FISH has been demonstrated for tens of thousands of PGT-A cases (Verlinsky et al., 2004; De Rycke et al., 2015), the technique suffers
several limitations, namely the number of fluorochromes available and the requirement for high quality chromosome preparations. Furthermore, misdiagnosis could occur since it is possible for fluorescent signals to overlap, depending on position of chromosomes during preparation. With an efficiency of ~90% aneuploidy detection rate and a 12% misdiagnosis rate per embryo, multiple single cell samples were often required to validate results (Munné et al., 2010). Misdiagnoses lead to the transfer of aneuploid embryos (false negative), or the lack of transfer and even destruction of euploid embryos (false positive). The use of the technique declined following a seminal paper that reviewed eight randomised control trials, which demonstrated the negative effect on clinical pregnancy outcome of using FISH for aneuploidy detection when combined with day-3 biopsy of two or more blastomeres (Mastenbroek et al., 2011). Although 24-chromosome analysis by FISH was eventually developed by means of multiple rounds of hybridisation (Ioannou et al., 2011), the technique was superseded by array and plate-format assays with higher throughput and efficacy. Comparative genomic hybridisation involves the hybridisation of fluorescently labelled ‘test’ DNA to euploid metaphase spreads or to arrays of bacterial artificial chromosome probes printed onto a glass surface. One drawback of comparative genomic hybridisation was the requirement for large amounts of DNA (0.2-1.0 µg) to perform the analysis, however the development of advanced whole genome amplification (WGA) methods demonstrated this was possible (Wells and Delhanty, 2000).

Array-based comparative genomic hybridisation (aCGH) is a sensitive protocol that requires linear whole genome amplification of the biopsied material from the embryo whilst preserving copy number variations from the original sample. The amplified DNA and DNA euploid reference samples (male or female) are fragmented and fluorescently labelled with either Cy3 and Cy5 dyes, respectively. The two samples are mixed in equal concentrations and hybridized to an array of oligonucleotide probes, immobilised on a glass surface. Each unique oligonucleotide is represented thousands of times in a single array location, and together make up a genome-wide backbone that represents all 24 chromosomes at a resolution determined by the number of array features. After array scanning and image analysis, the log₂ ratio between the expected Cy3 and Cy5 channel intensities is used to infer the sample DNA representation compared to the reference. The variation in channel intensity correlates with copy number variations from the expected diploid state, and the normalised log₂ ratio values are plotted by chromosome and molecular probe location for analysis. Although aCGH does not allow parental origin of aneuploidy to be inferred, nor can it be used to differentiate between euploid, haploid or polyploid samples, it was soon regarded as the gold standard for PGT-A. This was due to the relatively modest cost for genome-wide resolution (at 1 Mb) for all 24 chromosomes and an accuracy of 98% (Gutiérrez-Mateo et al., 2011). Alternative technologies included SNP
microarrays for analysis of heterozygosity and SNP coverage, and a quantitative PCR approach that replaces the need for whole genome amplification (Treff et al., 2010, 2012). As with FISH, the clinical utility of aCGH and SNP microarrays were widely demonstrated by their use in thousands of PGT-A cases. Comprehensive screening of all 24 chromosomes lead to the 66% improved ongoing pregnancy rate following transfer of single euploid embryos in good prognosis patients, which underpins aneuploidy as a major cause of miscarriage in the ART clinic (Yang et al., 2012).

1.1.4.3 The J-curve for age-related aneuploidy is shifted in amplitude and towards increased maternal age in ART populations

Two studies have been used to estimate the incidence of chromosome missegregation, which can be detected during preimplantation embryo development in the ART setting. Franasiak et al. (2014) reviewed 15,169 day-5 trophectoderm biopsies with analysis by SNP microarray and qPCR for aneuploidy. The study showed that women aged 26-35 years had the lowest rate of aneuploid embryos (at 25-35%), and women <23 years showed elevated aneuploidy rates of 45%. The incidence of whole chromosome aneuploidy increased dramatically in women >35 years, rising steeply to almost 90% aneuploidy rate per embryo at 44 years. In addition, an equal number of monosomies to trisomy events were observed (n= 4,513 vs 4,376), although with a slight maternal age-effect towards trisomy with advancing age. Similarly, McCoy et al. (2015) assessed 18,387 trophectoderm biopsies and 28,052 single cell day-3 blastomere biopsies, analysed by SNP microarrays. They reported that the incidence of aneuploidy in biopsies taken on day-3 (single blastomere) is greater than on day-5 (multi-cell TE), indicating a selection against aneuploidy that occurs at the time of embryonic genome activation. The aneuploidy curve for day-5 biopsies that is comparable to that of Franasiak et al. (2014) for a maternal age of 25-45 years. In contrast, aneuploidy rates in day-3 biopsies display a much higher amplitude (40% for 25-35 years), climbing steeply to almost 100% at 45 years thereafter. The difference can be explained by that almost half of embryos arrest prior to blastocyst formation, or otherwise ‘correct’ for chromosome imbalance before blastocyst formation. Embryo survival to day-5 is an indicator of developmental competence, although this alone does not prove that aneuploidy alone causes developmental arrest in preimplantation stages (Fragouli et al., 2013).

One important difference is the early part of the J curve. In day-3 biopsies the incidence of aneuploidy for the youngest women included in the study (20 years) appears to rise. In contrast, day-5 biopsies show a 2-fold reduction in aneuploidy rates at 20 years compared to 21-30 years, which is not in agreement with the linear regression model. Taken together, this effect may be explained by a small sample size, as younger women are less likely to seek fertility
treatments. However, the J-curve seen in clinical conceptions is both shifted upwards in amplitude and to the right of the age scale compared to the natural conceptions. These differences may be due to selection against embryos that are aneuploid in natural conceptions and the contributions of different chromosomes to the age-related incidence of aneuploidy. For instance, chromosomes with a more delayed age-dependence would be observed in embryos but not natural conceptions, since most chromosomal aneuploidies are incompatible with viability of the foetus. Furthermore, in vitro culture of embryos and gonadotrophin stimulation of women could also influence aneuploidy rates in ART settings, causing an elevation compared to natural conceptions.

1.1.5 Comprehensive chromosome screening of trophectoderm biopsies reveals mosaic and sub-chromosomal abnormalities with high dynamic range

The development and ever-reducing cost of next generation sequencing (NGS) assays and the high throughput demand for PGT-A paved the way of NGS-based comprehensive chromosome screening. This high throughput alternative to aCGH provides a highly accurate, rapid, streamlined workflow with unparalleled sample multiplexing compared to array-based methods (Fiorentino et al., 2014a, 2014b; Zheng et al., 2015). A targeted quantitative PCR approach to comprehensive chromosome screening also provides a cost-effective alternative to aCGH with a simplified workflow and has been extensively validated (Treff et al., 2012; Scott et al., 2013; Capalbo et al., 2015b); although with a lower chromosomal resolution compared with aCGH and NGS technologies, and an inability to detect partial aneuploidies. The comprehensive genome coverage provided by NGS is limited only by the whole genome amplification method and depth of sequencing (the number of times a base is read). As a result, NGS was clinically validated by a number of independent ART laboratories and is arguably the emerging gold standard methodology for PGT-A today (Sermon et al., 2016). The transition from aCGH-based chromosome screening to NGS methods saw a change in CNV analysis from a log2 ratio scale of image intensities, to a linear scale of sequence read counts. This provided a higher dynamic range in which to observe variations in copy number, thus making it easier to detect low level mosaicism and sub-chromosomal (segmental) abnormalities in trophectoderm biopsies.

Mosaicism has been a known phenomenon in human since the first observation in 1959 (Ford et al., 1959b). The use of FISH to accurately identify chromosome mosaicism in interphase nuclei of single blastomeres raised the question of its clinical significance in embryo development (Delhanty et al., 1997). The origins of aneuploidy can largely be designated to one of three occurrences; (i) somatic chromosome imbalances in the germline give rise to aneuploid gametes; (ii) chromosome segregation errors in meiosis also give rise to aneuploid gametes; (iii)
chromosome transmission errors in the mitotic cell divisions during preimplantation development give rise to aneuploid cell lineages. In the first two cases, the PGT-A analysis of a day-5 trophectoderm biopsy would reveal a uniform CNV deviation from the expected diploid state that is consistent with chromosome loss or gain in every cell in the trophectoderm biopsy (Figure 1.5A, 1.5C, 1.5E). In the third case, however, depending on the developmental stage at which a mitotic error occurs, a higher or lower proportion of the biopsied trophectoderm cells would be expected to be aneuploid as the error is propagated in subsequent cell divisions (Vera-Rodriguez and Rubio, 2017). PGT-A analysis would reveal an incomplete or partial deviation from the diploid baseline (Figure 1.5B, 1.5D, 1.5E), and has been demonstrated using a cell line mixture model with known euploid/aneuploid karyotypes down to 16-17% aneuploidy (Goodrich et al., 2015, 2017). Embryo mosaicism is also defined by non-concordant or reciprocal uniform aneuploid results, from two separate trophectoderm biopsies of the same embryo (Vera-Rodriguez and Rubio, 2017) (Figure 1.5F). This questions whether data from single trophectoderm biopsies underestimate the incidence of embryo mosaicism. Taken together, chromosomal mosaicism affects 2% of chorionic villus samples, of which only ~10% are confirmed in foetal tissue (Malvestiti et al., 2015). The incidence of mosaicism in preimplantation embryos has been estimated as high as 90%, indicating that mosaicism may be a major cause of developmental arrest; however, the sensitivity of the technology and the reporting criteria used to estimate the incidence of mosaicism should be carefully considered (Taylor et al., 2014).
Figure 1.5: Chromosomal abnormality types detected by PGT-A in trophectoderm biopsies of day-5 preimplantation embryos following PicoPLEX whole-genome amplification (WGA). (A) Whole chromosome gain of Chr. 16. (B) 40% Whole chromosome mosaic loss of Chr. 13. (C) Uniform segmental loss of Chr. 8p. (D) 40% Mosaic segmental loss of Chr. 2q. (E) Schematic diagram of a 3-cell TE biopsy, and the expected PGT-A result depending on the genomic content of each biopsied cell when analysed together. Examples shown are for chromosome loss, but the reciprocal can be expected for chromosome gain. The level of detected mosaicism is expected to be proportional to the number of biopsied cells containing the abnormality. Red and green curved lines represent parental haplotypes of a single chromosome, with each example showing loss of the ‘red’ parental haplotype. (F) A random distribution of aneuploid cells within the TE may provide a mosaic PGT-A result when two biopsies are analysed separately, giving discordant results. Chr = Chromosome; TE = Trophectoderm; ICM = Inner cell mass.
The improved resolution and dynamic range of NGS in detecting copy number variation may help elucidate the incidence, and clinical significance of a mosaic PGT-A result. Recent guidance for classification of embryos based on PGT-A results (PGDIS, 2016; CoGEN, 2018) has led some groups to prioritise embryos for transfer based on the level of mosaicism detected, which appears to be correlated with developmental potential (Munné et al., 2017). Several live births have resulted from the transfer of mosaic embryos, following appropriate genetic council and consideration of known congenital birth defects (Greco and Fiorentino, 2015). Additionally, it has been shown in mouse embryos that some aneuploid cell lineages are eliminated by apoptosis or otherwise outcompeted if a sufficient number of euploid cells exist after the blastocyst stage, demonstrating developmental potential for mosaic preimplantation embryos (Bolton et al., 2016). However, there is concern regarding the technology and criteria used to identify embryo mosaicism, such that over-estimation may result in potentially euploid or viable embryos being discarded. A single trophectoderm biopsy will never give a complete picture of blastocyst genomic constitution (Vera-Rodriguez and Rubio, 2017), although trophectoderm has been shown to be representative of the inner cell mass that gives rise to the embryo proper (Capalbo et al., 2013a). The safety of embryo biopsy, in addition to in vitro embryo culture and ovarian stimulation should also be considered. Recently, studies of non-invasive embryo testing have shown promise but require further optimisation (Fragouli and Wells, 2018; Vera-Rodriguez et al., 2018). Taken together, the clinical significance of a mosaic PGT-A result should be carefully reviewed to maintain the highest standard of patient care. Where multiple biopsy/disaggregation of the embryo is not possible, validated procedures for abnormality calling should be used to interpret challenging results (Capalbo et al., 2016; Vera-Rodriguez and Rubio, 2017). Interpreting the true incidence of mosaicism requires standardised and validated procedures in embryo biopsy and genetic assessment.

Whilst embryo mosaicism is prevalent in human preimplantation embryos, subchromosomal aberrations represent a smaller, however significant proportion of abnormalities detected in 4% of spontaneous abortions, and ~10% of PGT-A embryos (Vera-rodríguez et al., 2016). The origin of segmental abnormalities is less understood; but their occurrence appears to be independent of whole chromosome aneuploidy and maternal age. The high dynamic range of NGS-based PGT-A reveals both uniform (Figure 1.5B) and mosaic segmental abnormalities (Figure 1.5D), ranging from small fragments (5 Mb) to whole chromosome arms. NGS could provide a higher resolution analysis of segmental abnormalities when compared to a recent study using aCGH PGT-A data, which limited detection to events larger than 15 Mb, and only considered mosaic abnormalities in a small subset of day-3 (n=8) and day-5 (n=7) embryos by FISH analysis (Vera-rodríguez et al., 2016; Babariya et al., 2017). However, these technologies
pose a risk to overdiagnosis, due to the occurrence of small CNVs such as indels and variants of unknown significance which may not affect clinical outcome during embryo development (Capalbo et al., 2017).

Segmental aneuploidies are largely assumed to occur following chromosome breakage during mitosis, with frequencies higher in day-5 trophectoderm biopsies than in day-3 blastomeres (Voullaire et al., 2000; Wells and Delhanty, 2000; Babariya et al., 2017). However, segmental aneuploidies could also be balanced and unbalanced translocations. In all cases there must be a chromosome break, which may involve a single strand or both strands and involve different repair mechanisms. A small proportion of events detected in human oocytes are inferred from polar bodies to have a meiotic origin, with an increased proportion of events occurring after the second meiotic division as opposed to meiosis I (Babariya et al., 2017). This is in contrast to whole chromosome age-related aneuploidies that predominantly arise in meiosis I. In the ART setting, meiosis II is completed after ovarian stimulation with exogenous hormones and occasional oocyte culture in vitro. This raises the questions to whether ART procedures cause meiosis II-derived segmental aneuploidies. It has previously been shown that incubation of lymphocytes in folate-deficient media gives rise to chromosome breaks at the Xq28 fragile site in Fragile X syndrome patients (Branda et al., 1984). Therefore, the conditions under which oocytes are handled prior to and during fertilisation may give rise to unknown chromosome instability. The meiotic origin of segmental aneuploidies is interesting because the meiotic divisions lack a standard non-homologous end joining pathway. Therefore, it is unclear whether there are unbalanced translocations arising in meiosis, in the absence of canonical mechanisms to ligate two chromosome fragments together. The process may occur though homologous recombination if there is sufficient homology that chromosomes pair incorrectly. To this end, the incidence and meiotic origin of segmental aneuploidies could be further probed using NGS and SNP technology. The ability to detect segmental (and mosaic segmental) abnormalities with higher dynamic range raises similar questions to whole chromosome mosaics, regarding the clinical significance of a positive PGT-A result. This is perhaps most critical for the mosaic segmental abnormalities, which can be challenging to distinguish from technical noise such as artefacts from whole genome amplification.
1.2 Requirement for single cell genomics to understand the origin of aneuploidy in meiosis

In the 1980s, restriction fragment length polymorphisms were a useful tool to demonstrate that most sex chromosome aneuploidies originated from the maternal germline, with the exception of monosomy X that was due to a missing paternal X chromosome 80% of the time (reviewed by Jacobs (2014). Similar studies have shown both the parental origin, and importantly, meiotic division in which trisomies arise. These are summarised in Figure 1.6, redrawn from Hassold et al. (2007), where the majority of events occur in maternal MI, with a higher proportion occurring in maternal MII for chromosome 18. Events of paternal origin occur in <15% of cases. These data support the maternal age-effect for aneuploidy in clinical pregnancies, however the development of single cell genomics allows studies to be conducted directly on oocytes and preimplantation embryos. Single cell genomics is an umbrella term for genotyping of individual cells from a heterogeneous population. When designing a single cell genomics experiment, one must consider the methodologies to use based on their qualities and limitations. The sections written here include my contribution towards two recent publications and have been expanded where appropriate. Sections I have authored will match the published text (Blanshard et al., 2018; Vogel et al., 2018).

Figure 1.6: Meiotic origin of non-mosaic human trisomies. Data from a summary of studies, redrawn from Hassold et al., (2007). MI = Meiosis I; MII = Meiosis II.
1.2.1 SNP markers are highly informative for inheritance patterns

Comprehensive chromosome screening of both polar bodies by aCGH is particularly effective for inferring meiotic stage of chromosome missegregation because the reciprocal chromosome gains and losses could be observed in each cell. From this it was shown that premature separation of sister chromatids (PSSC) (or premature predivision) in meiosis I was a common mode of chromosome imbalance in oocytes of advance maternal age (Handyside et al., 2012). Unfortunately, the use of polar body analysis for PGT-A is limited because no direct information is gained about the oocyte, and simple copy number analysis does not distinguish between parental homologues.

Alternatively, SNP microarray technology provides the genotype information that is required to infer parental and meiotic origin of aneuploidies. In the clinic, SNP arrays are known for PGT-A (Treff et al., 2010), but also for preimplantation genetic testing for monogenic disease (PGT-M) of ~7000 monogenic disorders (Handyside et al., 2010). Briefly, haplotype information is generated from genomic DNA of both parents and a reference individual (i.e. a related sibling or grandparent of known disease/carrier status). This information is then used to phase the SNP markers derived from embryo biopsies during ART. Each allele is represented as AA, BB or AB types, and the parental data are filtered to generate a subset of ‘informative’ alleles for each parent. For example, if the mother is AA and the father is AB at a single locus, then we say the father is informative for the ‘B’ allele because if the embryo inherits a ‘B’ allele, then we know it came from the father. If the reference also contains the ‘B’ allele, then the embryo is ‘in phase’ with the reference; conversely if the reference does not contain the ‘B’ allele then the embryo is ‘out of phase’ with the reference. This process is repeated for all informative SNPs for both parents. Phasing assigns a fingerprint of alternating haplotypes to each chromosome, based on the parental meiotic recombination patterns. As the position of the disease gene is known, and the disease/carrier status of the parents and reference are known, we then ask whether the embryo has inherited the same SNP constitution as the reference or an alternate constitution within the disease region. Embryos are then selected for transfer based on the desired phase with the reference haplotypes. Considerations are made for inefficiencies in WGA, namely the random non-amplification of one allele, which is discussed later. To this end the clinical utility of SNP linkage analysis (Karyomapping) has been demonstrated (Handyside et al., 2010; Natesan et al., 2014) but it’s application for analysis of female meiosis is limited by the requirement for activation of the oocyte to trigger resumption of the MII arrest, and extrusion of the PB2 for analysis.
Assisted activation of the meiosis II-arrested oocyte has applications in both clinical and basic science

Atypical expression, localisation or structure of the sperm-specific phospholipase C zeta (PLCζ) protein has been linked to male infertility and oocyte activation deficiency, in cases where the MII-arrested oocyte fails to activate with the clinical use of assisted reproductive technology (ART); namely, intra-cytoplasmic sperm injection (ICSI) (Kashir et al., 2010, 2013; Swann and Lai, 2013). Fertilization failure causes infertility and is due to both male and female factors, although many cases remain idiopathic. Male factors include globozoospermia (sperm lacking an acrosome), which contributes to the ~3% of total fertilisation failures after ICSI (Liu et al., 1995; Flaherty et al., 1996; Shinar et al., 2014). Total fertilization failure is attributed to oocyte activation deficiencies, which are not associated with sperm-oolemma membrane fusion and therefore not addressed by standard ART therapies. A mouse oocyte activation test with sperm from the male patient detects male-factor oocyte activation deficiency (Heindryckx et al., 2005). Failure of the patient sperm to fertilize the oocyte is likely to be caused by defective PLCζ signalling (Yoon et al., 2008; Heytens et al., 2009). Female-factor oocyte activation deficiency has also been reported, but is not well characterised (Tesarik et al., 2002).

To overcome total fertilization failure, adjuvant therapies which mimic the calcium signalling response of the oocyte at fertilisation have been employed. Collectively referred to as artificial oocyte activation (AOA), therapies include the use of calcium ionophores such as A23187 (calcimycin) (Rybouchkin et al., 1997) and ionomycin (Heindryckx et al., 2008), which allow the diffusion of Ca$^{2+}$ ions across membranes down their electrochemical gradient into the cytoplasm of the oocyte. Whilst the treatment does not cause sustained calcium oscillations, it provides the first transient rise in Ca$^{2+}$ concentrations to promote oocyte activation. The application of AOA using ionophores typically includes incubation of the oocyte in media supplemented with ionophore directly after ICSI (Rybouchkin et al., 1997; Tesarik et al., 2000; Ebner et al., 2012), although the concentration and exposure time vary slightly between clinics and research groups. The direct microinjection of CaCl$_2$ (Heindryckx et al., 2008; Nikiforaki et al., 2014), or ionophore (Tejera et al., 2008) into the oocyte during ICSI, followed by incubation with the ionophore have also been reported. Several case reports suggest the use of ionophore results in live birth (Eldar-Geva et al., 2003; Chi et al., 2004; Kyono et al., 2009; Terada et al., 2009; Kim et al., 2015), however the efficiency is not known. Limited data from children conceived using AOA suggest that they complete neurodevelopment normally (Vanden Meerschaut et al., 2014). In addition, AOA has been used to overcome oocyte activation failure after ICSI, when no morphological sign of activation (PB2 extrusion or pronuclei formation) is presented after 24 hours. This type of ‘rescue AOA’ has been demonstrated in the clinic,
resulting in live birth (Lu et al., 2012a), and in the research laboratory with an activation efficiency of 72% (Eftekhar et al., 2012).

The ability to activate the MII-arrested oocyte without the use of sperm has applications in basic science. This is important for two reasons. Firstly, sample availability is often limited to unwanted oocytes from ART cycles, which are donated for research by fully consenting patients. This limits both the number and origin of oocytes available for research. ART cycles typically involve administration of recombinant follicle stimulating hormone for ovarian hyperstimulation, followed by a dose of exogenous human chorionic gonadotropin which promotes ovulation. Stimulation regimes allow recruitment and maturation of many mature antral follicles per ART cycle. However, it is currently unclear whether the use of these hormones perturb chromosome segregation in MI, and therefore account for why 50% of transferred embryos fail to produce a healthy baby in the clinic. Furthermore, collection of oocytes from ART settings largely limits the study population to women of advanced maternal age (>35 years), who are already presenting with infertility. Secondly, using sperm to trigger the resumption and completion of meiosis II for the purpose of research requires extensive ethical approval. Although ethical approval was recently granted for the use of CRISPR/Cas9 to evaluate the safety and efficacy of gene correction for heritable diseases, the use of human embryos for research remains extremely limited (Ma et al., 2017). This is because the process involves creating embryos that are subsequently and intentionally destroyed. Taken together, the use of calcium ionophore mitigates some of the practical and ethical limitations of oocyte availability. The activation of MII-arrested oocytes with calcium ionophore in the absence of sperm, triggers faithful segregation of chromosomes in the second meiotic division, and extrusion of the PB2 (Capalbo et al., 2015a).

1.2.1.2 Collection of single cell products of meiosis allows genome-wide analysis of chromosome recombination and segregation patterns

Karyomapping is a powerful analysis tool with clinical applications for PGT-M. For the study of meiosis, SNP markers can also be used build genome-wide maps of recombination and chromosome segregation. The methodology of Ottolino et al., (2015) enables analysis of all four chromatids involved in both nuclear divisions, by capturing the three single cell products meiosis. Briefly, the PB1 is biopsied from the MII-arrested oocyte for integrity of sample tracking. The oocyte is then activated with 100 µM calcium ionophore for 40 min, which triggers the completion of MII and extrusion of the second polar body (Figure 1.7A). Once a single female pronucleus forms in the oocyte, the PB2 is removed and the oocyte is isolated. The three single
cells are tubed separately for analysis, and collectively constitute an oocyte-PB trio (Figure 1.7B) (Ottolini et al., 2016). A similar approach has also been described, which instead requires isolation of the female pronucleus immediately following fertilisation by sperm (Hou et al., 2013); the limitations of this approach have been discussed previously.

**Figure 1.7:** Collecting the three single cell products of female meiosis. (A) Assisted oocyte activation. The meiosis II-arrested oocyte is incubated with a lipophilic, divalent cation, calcium ionophore which passes freely through the zona pellucida and incorporates into the oolemma. The ionophore allows passive diffusion of extracellular Ca2+ into the cytoplasm. A single transient Ca2+ rise is sufficient to trigger completion of meiosis and PB2 extrusion. (B) Following activation, a single maternal pronucleus forms in the absence of the paternal complement. Both matching polar bodies are biopsied, and the oocyte is isolated. single cells are tubed individually for analysis, and collectively constitute an oocyte-PB trio. Figure A was adapted from Clift & Schuh (2013); green and yellow shading represent maternal haplotypes, to visualise recombinant chromatids; chromosomes not to scale.
Analysis of oocyte-PB trios allows inferences to be made about the meiotic stage or origin of chromosome transmission errors (Hou et al., 2013; Ottolini et al., 2015). It should be considered that aneuploid oocytes may result from germ line mosaicism (Kovaleva, 2010), and mitotic segregation errors give rise to aneuploidy during embryo development. However, a combination of single nucleotide polymorphism (SNP) genotyping and mapping techniques on oocyte-PB trios of female meiosis, allows quantitative analysis of parental origin for all chromatids in the meiotic pathway.

On the most part, this provides little evidence in the clinic to suggest that germ line expansion is attributable to embryonic aneuploidies. The ability to perform single cell genomics analysis prior to the initiation of embryo development, adds further evidence to suggest that most chromosome transmission errors occur during female meiosis, warranting further study to understand their origin and fate. Heterozygous SNP loci derived from maternal genomic DNA are used to phase the maternal haplotypes across individual meioses. Briefly, the single cells of each oocyte-PB trio were genotyped by SNP microarray and each cell was phased using the sibling approach, similar to that used in Karyomapping. Provided at least two oocyte-PB trios are analysed per woman, any haploid cell (PB2 or oocyte) can be used as the reference to phase the remaining samples. Changes in haplotype phase along chromosome arms reveals the location of crossovers, and the genome-wide chromosome fingerprints reveals recombination maps and segregation patterns that are consistent with the origin of meiotic errors (Figure 1.8A). The meiotic divisions summarised in Figure 1.2, describe the segregation of homologous chromosomes at MI, and the segregation of sister chromatids at MII.

Due to the rare prevalence of crossovers adjacent to centromeres, we can probe the pericentromeric SNPs for heterozygosity to detect whether chromosomes missegregated in the first or second division. Figure 1.8B shows an example of each segregation pattern, stratified by meiotic origin. The MeioMap view shows the haplotype information detected adjacent to the centromere (green and yellow represent both maternal haplotypes), where blue shading represents regions of heterozygosity. Ottolini et al., (2015) reported a novel segregation pattern in which both homologues segregate their sister chromatids in MI, leading to either balanced and unbalanced divisions at MII. This is important because an reverse segregation-abnormal error appears to be MII non-disjunction by simple content analysis but is actually an MI error as seen by SNP analysis. Taken together, SNP markers are highly informative for inferring the origin of chromosome missegregation in female meiosis.
Figure 1.8: Inferring the meiotic origin of chromosome segregation errors. (A) Interpreting a SNP MeioMap for meiotic segregation patterns. An example of MII NDJ resulting in aneuploidy (chromosome loss) in the oocyte. Regions of heterozygosity (blue shading) are not expected in haploid cells (PB2 or oocyte). (B) MeioMapping reveals that haplotype fingerprinting at pericentromeric regions only (MeioMap View - shading around the centromere only) can be used to infer chromosome segregation patterns in both meiotic divisions (Single Cell View). Pericentromeric heterozygosity in the PB1 indicates the presence of alternate haplotypes, resulting from RS or PSSC with retention in the PB1; note that this configuration of PSSC would result in loss in either the PB2 or oocyte, distinguishing it from RS. Six core segregation patterns are defined, although there are multiple configurations of each. PB1 = First polar body; PB2 = Second polar body; RS = Reverse segregation; MI = Meiosis I; MII = Meiosis II; NDJ = Non-disjunction; PSSC = Precocious separation of sister chromatids. Green and yellow shading are used interchangeably to represent alternate maternal haplotypes (homologs).
1.2.2 Methods of WGA limit inferences from single cell genomics

Single cell genomics describes the deconvolution of mixed populations, allowing the detection of genetic diversity within a smaller population of cells. Detecting genome changes in single cells is a sensitive procedure, complicated by the often rare, unique and precious nature of the starting material. Applications cover many disciplines from sequencing the complete genomes of microorganisms that are challenging to culture in the laboratory to de novo mutation detection in tumour cells isolated from circulating blood. The unique genome reshuffling that occurs in single meioses is of interest when correlated with treatment procedures (e.g. ovarian hyperstimulation, in vitro maturation), external factors (e.g. inhalation of cigarette smoke or bisphenol-A exposure) or maternal age.

When studying human meiosis, sourcing material for research is challenging because all four chromatids for each chromosome must be captured. This typically involves the isolation of both the meiosis II-arrested oocyte and the PB1 (oocyte-PB1 duo), or the isolated oocyte/embryo biopsy and both the PB1 and PB2 (oocyte-PB trio). Firstly, all patients must provide fully informed consent for the genomic analysis of their donated cell material – which may be withdrawn at any time. The use of such cell material from patients receiving ART therapies is governed in the UK by the Human Fertilisation and Embryology Authority. Secondly, the activation rate for oocytes treated with calcium ionophore is ~85% (Ottolini et al., 2015). When considering that both (MII-arrested oocyte-PB1 duo), or all three (oocyte/embryo-PB trio) products of meiosis are required for complete analysis, the risk of sample deterioration due to biological or technical factors must be considered. Therefore, a complete meiotic duo/trio is a precious source for analysis of unique genomic alterations and the analysis method should be carefully selected to test a specific hypothesis.

For single cell genomic analysis, a whole genome amplification (WGA) step is typically required, to generate enough copies of the target DNA to serve as input for downstream applications. Since the amplification step is destructive of the cell and cannot be technically repeated, the efficacy of WGA relies on three key properties: genome coverage, replication fidelity and the level of technical noise (i.e. systematic or stochastic amplification bias). Due to the inherent characteristics of available DNA polymerases, a trade-off exists between high fidelity genome coverage and amplification bias. Depending on the research application, currently available technologies facilitate two pipelines for simple content analysis and SNP detection described here.
1.2.2.1 Copy number analysis requires WGA with linear amplification and tolerates systematic bias profiles

The principles for copy number variation (CNV) detection, whether performed by aCGH or next generation sequencing involve comparing target DNA hybridisation or sequence read counts to that of a reference genome consisting of many pooled individuals. The requirement for WGA is reproducible genome coverage, where systematic bias is tolerated, whether variable-representation of GC-rich regions or frequency of PCR duplicates. The genome coverage determines the maximum resolution that can be achieved, which is further controlled by the number of probes on a microarray, or the size following sequence alignment and sequencing depth. The sequencing depth refers to the frequency that sequence reads align to a specific nucleotide or genomic region. For CNV detection, a sequencing depth of 0.01× is sufficient to detect copy number variations in regions as small as 14 Mb (Fiorentino et al., 2014a). In both microarray and NGS applications, algorithms for normalisation and smoothing of copy number traces can be optimised for systematic WGA bias profiles, to correct for copy number artefacts.

The current gold-standard WGA methodology for CNV detection takes a quasi-linear approach to mitigate the underrepresentation of heterochromatic and G/C or A/T rich regions of the genome. Commercially available applications include PicoPLEX (Takara Bio, Inc.) or SurePlex (Illumina Inc.), and both have been independently clinically validated for whole-chromosome aneuploidy screening with both microarray (Gutiérrez-Mateo et al., 2011) and NGS (Fiorentino et al., 2014a, 2014b). Alternatively, the Multiple Annealing and Looping Based Amplification Cycles (MALBAC) method (Yikon Genomics) works similarly (Huang et al., 2014). In brief, SurePlex and MALBAC use a two-stage amplification. In both cases, a limited cycle PCR with degenerate primers is used for ‘linear’ pre-amplification. The PCR primers are abutted to complimentary handle sequences, designed to form hairpin structures in subsequent cycles. The hairpin structures reduce the amplification of PCR products, allowing preferential amplification of the original template DNA. Pre-amplification is then followed by an exponential amplification, whereby the handle sequences are specific targets for priming. The main difference is that while SurePlex uses an exponential PCR reaction, MALBAC uses an isothermal reaction similar to that used for multiple displacement amplification (MDA). More recently, the in vitro use of RNA polymerase, rather than PCR has been used for linear amplification of single genomes, with coverage comparable to that of MDA (Chen et al., 2017; Blanshard et al., 2018). However, the solution is not currently commercially available.
1.2.2.2 High-fidelity WGA with maximal genome coverage is required for accurate genotyping of single cells

Variant calling at single nucleotide polymorphism (SNP) loci, or de novo mutation detection, require maximal genome coverage with high-fidelity amplification, particularly in cases where the single nucleotide variant position is unknown and cannot be targeted. Whist downstream applications are designed to achieve higher sequence depth (30×) or microarray probe replicates (15-30×), a stochastic amplification bias can be tolerated. The Phi (φ) 29 polymerase first used by Lasken and colleagues for multiple displacement amplification (MDA) of human DNA, is favoured for its ability to displace secondary structures such as hairpin loops that would cause other polymerases to stall or dissociate from the template DNA (Dean et al., 2002). The result is long amplification products averaging >10 Kb that are generated from random hexamer priming. A further 3′→5′ exonuclease activity enables proofreading and improves the fidelity of amplification. MDA kits are commercially available, and some such as the REPLI-g SC Kit (QIAGEN) have been optimised for use with single cells. At the time of writing, the REPLI-g SC Kit was also available under the name of SureMDA Amplification System (Illumina Inc.); a QC assured kit marketed for ‘Research Use Only’, that has been independently clinically validated for the purpose of whole genome genotyping for Karyomapping (Natesan et al., 2014).

Phasing using SNP genotypes is limited by the efficiency of whole genome amplification and the SNP detection reactions. Ideally, one would phase the DNA directly where possible using grandparental information or long sequencing reads (e.g. Nanopore technologies). Specifically, allele drop out (ADO) describes the random non-amplification of one allele (Figure 1.9). For informative SNP loci where the embryo inherits a homozygous constitution, it is unclear whether the embryo has suffered ADO for the informative allele. The allele is therefore segregated in analysis to provide only secondary information supporting the haplotype phase. All informative SNP loci where the embryo is heterozygous, are classified as primary evidence supporting the haplotype constitution (Handyside et al., 2010).
Figure 1.9: Whole-genome amplification noise in single cell DNA. (A) Allele drop out (ADO) results from the random non-amplification of one allele leading to signal drop-out by microarray (or sequencing). The SNP loci is incorrectly genotyped as homozygous. (B) Allele drop in (ADI) typically affects homozygous loci, where poor amplification results in reduced signal above the background and an incorrect heterozygous genotype call. On the SNP microarray, A/T bases are detected in the red channel and classified as A-type genotypes; C/G bases are detected in the green channel and classified as B-type genotypes. PB1 = first polar body; WGA = whole-genome amplification.

1.2.2.3 Current pipelines for single cell genomics

The current workflow for generating CNV and SNP data from single cells is shown in Figure 1.10. Copy number variation detection is achieved by SurePlex amplification and low-pass NGS (VeriSeq® PGS; Illumina Inc.). High quality SNP genotyping cannot be performed from SurePlex amplification products for de novo mutation assembly. This is because the shorter amplification products (~1000 bp) generated by SurePlex amplification result in a lower genome coverage, increasing the incidence of allele drop-out (amplification of only one allele at a given loci). The result is a reduction in heterozygosity as well as overall call rate, and therefore a reduced resolution. Additionally, the lower fidelity of PCR-based amplification increases the rate of false discovery, and therefore an alternate pipeline is required for single nucleotide variant detection. Instead, single cells are amplified by MDA and subsequently analysed by SNP microarray (e.g. HumanKaryomap-12 DNA Analysis Kit; Illumina Inc.) or high depth NGS (e.g. TruSeq DNA PCR-Free Library Prep Kit; Illumina Inc.).
1.2.2.4 Knowledge gap and requirement for concurrent CNV and single nucleotide variant analysis

Single nucleotide variant detection is highly informative of recombination events and allows inferences to be made about segregation patterns in meiosis, by revealing a unique fingerprint of haplotypes at SNP loci. Understanding the parental origin of each haplotype block, allows us to distinguish between segregation errors occurring in both meiotic divisions. The limitation is...
that the WGA reaction is destructive of the cell and original template DNA cannot be recovered. Additionally, SNP genotyping by microarray is cost-effective, but is not sufficient for high resolution CNV detection, particularly in cases of chromosome gain or cell line mosaicism within a multi-cell sample. Therefore, concurrent CNV detection from the same WGA product would serve as an important validation of nucleic content by means of an alternate technology (e.g. by NGS) (Blanshard et al., 2018).

1.2.3 Sources of human oocytes

Originally, human oocytes were obtained from ovarian tissue from women undergoing hysterectomies (Jagiello et al., 1975; Michelmann and Mettler, 1985; Martin et al., 1986). The oocytes showed differing stages of follicle growth, of which the majority were immature and required in vitro maturation. The use of gonadotrophin stimulation and IVF practices, results in a high proportion of large or mature follicles (15-22 mm) upon collection, and typically 90% of oocytes are mature (arrested at metaphase II). Whilst this mitigates the requirement for in vitro maturation of oocytes, it is difficult to obtain mature oocytes for research since they are primarily used for fertility treatment. However, material is available from donors and women with surplus vitrified oocytes.

Immature oocytes (GV) are heterogenous in their maturation rates when obtained from large or mature follicles. However, the oocytes that mature within 24 hours (‘fast maturing’ fraction) have similar properties to mature oocytes when obtained from the large or mature follicle stages (Escrich et al., 2012). GVs can also be obtained from 8-10 mm follicles when women at risk of ovarian hyperstimulation receive very limited gonadotrophin stimulation. Moreover, immature GV oocytes can also be obtained from small antral follicles (0.5 to 3 mm) directly from ovarian tissue.

1.2.3.1 Fertility preservation

For treatment of blood disorders and a range of cancers, exposure to radiation or cytotoxic drugs can have lasting effects on the patient’s reproductive health (e.g. ovarian reserve), depending on the dose and duration of administration. Increasing cancer survival rates present an elevated requirement for fertility preservation procedures, to help patients retain their fertility post-treatment. Cryopreservation of sperm samples and oocytes collected via exogenous hormone stimulation have been used as alternatives to fertility treatments using donor gametes. More recently, techniques involving the excision of an entire ovary prior to radiotherapy or
chemotherapy, facilitate the in vivo maturation of large antral follicles following post-treatment, auto-transplantation of ovarian tissue. Briefly, following excision, the ovarian cortex is separated from the medulla and cryopreserved. The cortex contains large antral follicles, and following auto-transplantation, evidence suggests that the hypothalamus-pituitary-ovarian axes is re-established in 95% of women, indicating that their ovarian tissue functions normally (Donnez et al., 2013; Yin et al., 2016). The small antral follicles present in the medulla are not suitable for fertility preservation, due to a correlation between small follicle size (<5 mm) and reduced developmental potential (Rosen et al., 2008; Gruhn et al., 2018; Wirleitner et al., 2018). As such, this material is often available for research with informed patient consent. A high number of GV stage oocytes can be collected from the surplus ovarian tissue, from which mature metaphase II oocytes can be obtained by in vitro maturation (IVM) (Gruhn et al., 2018). This facilitates collection of large numbers of oocytes, without exogenous hormone stimulation, from a wide range of ages including girls who have yet to begin cycling. Recent data suggests that GV stage oocytes that enter and complete meiosis I within 24 hours after removal of their protective cumulus cells, are capable of entering and completing both meiotic divisions in vitro, with similar efficiency to those from large antral follicles (Gruhn et al., Appendix 1.3). Furthermore, the activation of unfertilised oocytes in vitro, to trigger the resumption and completion of the second meiotic division, does not have a widespread impact on faithful segregation of chromatids when compared to fertilised embryos (Capalbo et al., 2015a).
1.3 Single cell studies of genome diversification are confounded by technical limitations of high-resolution genotyping

1.3.1 SNP genotyping allows highly informative, cost-effective genome analysis

The completion of the human genome led to the identification of >1000 highly heritable genes involved in rare, single-gene (mendelian) disorders. However, at the time medical research into the pathologies of common disorders was limited by their multifactorial nature (involving several DNA susceptibility variants), and the lack of resources to systematically test common genetic variants for their role in disease. The low power offered by family linkage-based studies and the density of markers required to effectively detect sequence variation in population-based studies, provoked the launch of the HapMap Project in October 2002. The HapMap Project was initiated to facilitate the design and analysis of medical genetic studies, through the capture of human SNP variation. The project was largely delivered in two phases, using 270 samples derived from 4 localised populations, grouped into three analysis groups (Table 1.2).

Table 1.2: Sample stratification for the HapMap Project

<table>
<thead>
<tr>
<th>Code</th>
<th>No. of samples</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>90</td>
<td>30 parent-offspring trios from Yoruba in Ibadan, Nigeria</td>
</tr>
<tr>
<td>CEU</td>
<td>90</td>
<td>30 parent-offspring trios from Utah, USA, Centre d'Etude du Polymorphisme Humain (CEPH) collection</td>
</tr>
<tr>
<td>CHB*</td>
<td>45</td>
<td>Han Chinese in Beijing, China</td>
</tr>
<tr>
<td>JPT*</td>
<td>45</td>
<td>Japanese in Tokyo, Japan</td>
</tr>
</tbody>
</table>

*Combined for some analysis given the similarity between allele frequencies.

Identification and characterisation of SNPs was performed through a multinational consortium of contributors, each responsible for a particular chromosome or genomic region. Despite the use of six different genotyping platform technologies, quality was maintained on three levels; a calibration exercise to ‘bench-mark’ each technology, a mid-project centre evaluation, and a blind analysis of a random sample of the complete Phase I dataset. Additionally, strict QC filters were employed throughout the project to ensure completeness, consistency, mendelian inheritance, and Hardy-Weinberg equilibrium. Failing SNPs were excluded from the HapMap datasets, but still released in the database with annotation.

The Phase I dataset included 1.007 million SNPs that passed QC criteria in all three analysis groups, at a density of one SNP every 5kb. Phase II of the project increased this total to 3.6 million SNPs, of which 3.1 million were also polymorphic in at least one analysis group, at a density of one SNP every 1kb. Common SNPs were defined as those with a minor allele frequency ≥ 0.05 in at least one of the three analysis groups, of which there were 2.8 million in the Phase II dataset. The HapMap resource was retired in June 2016, and users are now directed
to the 1000 genomes database (www.1000genomes.org). However, the dbSNP database (www.ncbi.nlm.nih.gov/projects/SNP) remains a valuable resource for designing assays for genome-wide association studies and screening tests for specific disease markers (Delano et al., 2010). The requirement for specific, accurate and efficient SNP detection for clinical and research applications has driven the rapid advancement in genotyping technologies.

The sequencing depth required to accurately detect SNP variation is not yet cost-effective in comparison to the targeted approach offered by microarray technologies. A combination of efficient assay design and technological advances in the manufacture of SNP-dense arrays, creates a powerful tool for genome-wide haplotype analysis. SNP feature selection in array design is carefully conducted to combine even genome coverage whilst exploiting linkage disequilibrium (LD), to reduce the burden of the genotyping effort. LD is used to describe the non-random association of alleles at two or more sites (SNPs) along the genome in a population, such that testing one SNP is equivalent to testing another. The phenomenon arises where any given SNP is likely to be inherited together with nearby SNPs. Collectively, the combination of alleles in a region of a chromosome is called a haplotype. The inheritance of particular haplotype blocks can be predicted to a certain degree by the fact that 60% of recombination in the human genome occurs at hotspots, that in turn span only 6% of the genome (The International HapMap Consortium, 2007). In terms of whole-genome genotyping, the result is long stretches of SNPs with strong LD and a high level of genotyping redundancy. Haplotype blocks and the LD associated with them will be eroded over time, due to recombination and de novo mutation. However, the rate is slow enough in contemporary chromosomes, that the correlation between SNPs is a useful tool for efficient genotype analysis studies, whilst maintaining power of detection.

The completion of the Phase II HapMap, greatly increased the resolution of SNP coverage in the human genome, from which individual markers could be selected for inclusion in study design. The criteria for inclusion can be defined by a minor allele frequency and threshold. The HapMap consortium used a minor allele frequency $\geq 0.05$ to define common SNPs, and an LD $\geq 0.8$ to determine whether a SNP site was a suitable proxy for other SNPs within a binned region. The group report that a minimum of $\sim 500,000$ SNPs are required to capture all common Phase II SNPs in CEU and CHB+JPT analysis groups, and twice that for the YRI group. These SNPs are described to have the highest tagging power and are referred to as ‘Tag SNPs’. The Tag SNP selection methodology for genotyping in a disease study, makes use of genotyping redundancies to effectively capture the genetic variation in the genomic region(s) of interest. The whole-genome genotyping array used previously for MeioMapping (Ottolini et al., 2016) is comprised of 12 subarrays, used to process 12 samples. Each subarray features 299,140 markers
selected from Phase I HapMap SNPs, that include 200,000 highly informative tag SNPs for haplotyping with even backbone genome coverage. The backbone SNPs are supplemented to provide 220,000 markers for targeted cytogenetic analysis at high density, in additional genes, sub-telomeric and pericentromeric regions, and sex chromosomes. The array provides genome coverage of 70%, 73% and 32% in CEU, CHB+JPT and YRI HapMap analysis groups, respectively (LD ≥ 0.8) (Illumina Technical Note, 2009), thus providing efficient coverage for cost-effective cytogenetic screening. Additional to the SNP content, the unique two-colour single base extension (SBE) biochemistry used to identify A/T and C/G base incorporation at each SNP position, makes the BeadChip arrays ideal for high precision genotyping (Steemers et al., 2006).

1.3.1.1 Single nucleotide polymorphism detection by single base extension and staining

For each assay within a subarray, 50-bp oligonucleotide probes are designed with a unique address, to hybridise to the template DNA. Each probe ends one base short of the target SNP location at the 3’ end. An additional unique 25-bp sequence is also included at the 5’ end, which is used for a decoding process to identify the ‘bead type’ and location following assembly into the array substrate. The 75-bp oligonucleotide DNA probes are immobilised on 2-micron paramagnetic beads via a 5’ reactive moiety synthesised in the last step of a 3’ → 5’ oligonucleotide synthesis reaction. Each bead is coated with thousands of identical probes. The number of bead types that can be manufactured is essentially limitless, facilitating unparalleled multiplexing in array design. The selected bead types are pooled to make a library, before the bead pool library is self-assembled into pattern-etched microwells on a silicon-based substrate surface. A built-in redundancy in microwells to bead pool concentration facilitates a bead type representation of 15-30x in each sub-array. Bead type information for each BeadChip Array is stored in a bead pool manifest file (*.bpm), and each BeadChip undergoes a hybridisation-dependent decoding process to generate a unique Decode Map file, revealing the locations of each bead type within a specific array. A single-tube sample preparation and direct hybridisation is followed by an allele-specific, on-array, SBE end-point reaction. The incorporated bases are detected by a two-dye fluorescent signal amplification step and scanning. The two-channel signal intensities are reported in Intensity Data files (*.idat), for both red and green fluorescence. Taken together the BeadChip microarray is a cost-effective, targeted alternative to high depth NGS.
1.3.2 Genotype calling of single cell DNA is confounded by noise introduced by WGA

Genotype calling from BeadChip array intensity data is performed using dedicated software (e.g. GenCall algorithm (GenomeStudio; Illumina Inc.)). Alternately, independent algorithms are publicly available for interpreting the intensity thresholds, normalising and calling genotypes from bulk gDNA with high call rates and accuracy (Ritchie et al., 2011; Li et al., 2012). In all cases these algorithms are limited for single cell genomics because they are typically trained on bulk genomic DNA that does not contain ADO and ADI noise introduced from amplification of the single cell DNA (Blanshard et al., 2018). One solution would be to alter the genotype calling thresholds to exclude poorly performing SNPs, but this would result in significant data loss. Alternatively, dedicated software allows curation of independent cluster files (*.egt) that represent specific sample populations. For single cell genomics, this would require extensive work involving multiple cells lines to capture single cells, ensuring representation genotype diversity at each locus. Taken together the analysis of single cell array data presents a significant technical gap in high precision genotyping following WGA.

1.3.3 Gene conversions facilitate genome diversification by breaking up the linkage disequilibrium between variants

Gene conversions are described by the localised, non-reciprocal transfer of genetic information between homologs during meiosis. They result from non-crossover resolutions of double strand break (DSB) repair, which is required to promote homologous recombination prior to the first meiotic division (Figure 1.11). Whilst crossovers are important for chromosome segregation during meiosis, non-crossovers are predicted to occur at a 10-fold higher rate in mice and human, with specific roles in mediating accurate homologue pairing and distribution of COs (Baudat et al., 2013). In addition, gene conversions play an important role in genome diversification, by breaking down local linkage disequilibrium (LD). The frequency of NCOs in human has recently been characterised in pedigrees using SNP array and NGS data. The incidence of meiotic gene conversions, based on the number of informative heterozygous SNPs analysed was $5.9 \times 10^{-6}$ per bp, per meiosis (Williams et al., 2015). The majority of confirmed gene conversions co-localised within DSB hotspots, which is consistent with the notion that they occur in meiosis. Gene conversions can be identified by mapping the inheritance patterns of alleles, but this remains challenging to due to their short length. Furthermore, the requirement for multiple generation pedigrees to map the incidence of gene conversions is rate limiting to discovery. Mapping gene conversions directly by capturing all four chromatids per meiosis would facilitate larger studies of their incidence and is currently untested due to the technical limitations of single cell genotyping. To this end, technology-driven research would facilitate
comparisons to be made between individuals but also and between gametes of the same individual.

Figure 1.11: Homologous recombination gives rise to crossover and non-crossover events in meiosis. SPO11 induces a double strand break (DSB) and the 5' end is resected to generate a 3' overhang, which invades the homolog (single end invasion) and form a D-loop. Left: Second end capture results in the formation of a double Holliday junction (dHJ) that is preferentially resolved into a crossover (interference-dependent crossover). Non-crossovers are less likely by resolution or dissolution based on budding yeast models. Right: If the D-loop is disassembled after limited DNA synthesis, the newly synthesized DNA can be used to repair the broken chromosome. This is referred to as synthesis-dependent strand annealing (SDSA). Resolution of the single end invasion by XPF-family structure specific endonucleases such as MUS81 can lead to interference-independent crossovers (also referred to as ‘mitotic-like’ crossovers). Note that all non-crossover events display a short tract of unidirectional transfer, depending on the homologous template that is used for synthesis. Figure re-drawn from Youds & Boulton (2011).
1.4 Research proposal

The objective of this thesis work was to improve the single cell DNA sequencing capacities in order to investigate the origins of human aneuploidies in eggs and embryos. I investigated the reported incidence and clinical significance of chromosomal abnormalities in preimplantation embryos, following PGT-A with the current state-of-the-art technology (Chapter 3). I then developed a single cell DNA sequencing assay that facilitates both SNP genotyping and low-coverage sequencing on the same MDA sample, which allowed me to probe the incidence of segmental aneuploidies in human eggs (Chapter 4). Finally, I have developed new technologies to type SNPs with high precision which has allowed us to detect gene conversions directly in human oocytes and their polar bodies. To our knowledge, this is the first direct detection of gene conversion events in human gametes. Collectively, this thesis demonstrates that technology-driven research allows the exploration of basic features of human genome evolution and genetic health directly in our germline.
Chapter 2: General materials and methods

In this chapter, I have summarised materials and methods that are required to reproduce the data presented in this thesis. General laboratory equipment and materials are presented together in Section 2.1, and assay-specific or specialised equipment are presented in the relevant sections thereafter. At the time of writing, the supplier and material/catalogue part numbers are correct. Methods are written for universal application, where methods in cell culture, DNA extraction and amplification, embryo biopsy etc are common to the studies summarised in the results chapters.

I have developed an end-to-end methodology for preparing, sequencing and validating DNA libraries from single cell, whole-genome MDA products for copy number variation analysis of oocyte and polar body samples. The method utilises library preparation strategies and materials from referenced commercially available kits, but in an optimised workflow for our application. The methods have been used to generate data for three original research papers (submitted) and one published book chapter (Blanshard et al., 2018), demonstrating a significant contribution to the single cell genomics community.

- **Vogel I, Blanshard RC**, Hoffmann ER. SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data. *Submitted*. Appendix A1.2

*Contributed equally.
### 2.1 General laboratory equipment and materials

#### Table 2.1: General laboratory equipment

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10 ErgoOne® single-channel pipette</td>
<td>S7100-0220</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>P20 ErgoOne® single-channel pipette</td>
<td>S7100-2200</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>P200 ErgoOne® single-channel pipette</td>
<td>S7100-0510</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>P1000 ErgoOne® single-channel pipette</td>
<td>S7110-1000</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>P10 F1-ClipTip™ 12-channel pipette</td>
<td>4661150N</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>P50 F1-ClipTip™ 12-channel pipette</td>
<td>4661160N</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>P300 F1-ClipTip™ 12-channel pipette</td>
<td>4661180N</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Powerpette Pro</td>
<td>JENC266-185</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>96-well PCR cooler</td>
<td>022510525</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>1.5 ml microcentrifuge tube cooler</td>
<td>03-410-497</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>MicroAmp™ 96-Well Base</td>
<td>N8010531</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Veriti™ 96-Well Fast Thermal Cycler</td>
<td>4375305</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Hybex® System for Illumina</td>
<td>1057-30-2</td>
<td>SciGene, CA, USA</td>
</tr>
<tr>
<td>Dri-Block® digital block heater</td>
<td>UY-36620-05</td>
<td>Bio-Techne, MN, USA</td>
</tr>
<tr>
<td>Mars Runner 900, Class II Microbiological Safety Cabinet</td>
<td>703075</td>
<td>Labogene, Allerød, Denmark</td>
</tr>
<tr>
<td>AirClean 600 microscope cabinet</td>
<td>AC632A</td>
<td>AirClean Systems Inc., NC, USA</td>
</tr>
<tr>
<td>CKX31SF Inverted Microscope</td>
<td>-</td>
<td>Olympus, Tokyo, Japan</td>
</tr>
<tr>
<td>M125 C Stereo microscope</td>
<td>-</td>
<td>Leica Biosystems, Wetzlar, Germany</td>
</tr>
<tr>
<td>New Brunswick™ Galaxy® 170S incubator</td>
<td>C017101002</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>CP Grade Carbon Dioxide N4.5</td>
<td>111095-Y</td>
<td>BOC, Guildford, UK</td>
</tr>
<tr>
<td>5424 Microcentrifuge</td>
<td>5404000060</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>5804 Centrifuge</td>
<td>5804000060</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>Heraeus™ Megafuge™ 40</td>
<td>75004503</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>iScan System</td>
<td>SY-101-1001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>MiSeq System</td>
<td>SY-410-1003</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NextSeq 500/550 System</td>
<td>SY-415-1001 / SY-415-1002</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NovaSeq 6000 System</td>
<td>20012850</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Material</td>
<td>Part#</td>
<td>Supplier</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>10/20 µl XL graduated, filter tip</td>
<td>S1180-3710</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>20 µl bevelled, filter tip</td>
<td>S1120-1810</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>200 µl graduated, filter tip</td>
<td>S1120-8810</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>1000 µl XL graduated, filter tip</td>
<td>S1122-1830</td>
<td>STARLAB Ltd., Milton Keynes, UK</td>
</tr>
<tr>
<td>ClipTip 20, filter, low retention</td>
<td>94420213</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>ClipTip 50, filter, low retention</td>
<td>94420253</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>ClipTip 300, filter, low retention</td>
<td>94420513</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Corning™ Stripette (5 ml)</td>
<td>07-200-573</td>
<td>Corning, NY, USA</td>
</tr>
<tr>
<td>Corning™ Stripette (10 ml)</td>
<td>07-200-574</td>
<td>Corning, NY, USA</td>
</tr>
<tr>
<td>Corning™ Stripette (25 ml)</td>
<td>07-200-575</td>
<td>Corning, NY, USA</td>
</tr>
<tr>
<td>50 ml BD Plastipak™, Luer lock, syringe</td>
<td>613-3925</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>20 ml BD Plastipak™, Luer lock, syringe</td>
<td>613-3922</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>50 ml Falcon™ centrifuge tube</td>
<td>734-0448</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>15 ml Falcon™ centrifuge tube</td>
<td>734-0452</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>Thin-walled 0.2 ml PCR tubes</td>
<td>B77301</td>
<td>BIOplastics, Landgraaf, The Netherlands</td>
</tr>
<tr>
<td>Thin-walled 0.2 ml PCR tubes; 8-tube strip</td>
<td>B72911L</td>
<td>BIOplastics, Landgraaf, The Netherlands</td>
</tr>
<tr>
<td>DNA LoBind Tube 1.5 ml</td>
<td>0030108051</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>DNA LoBind Tube 5 ml</td>
<td>0030108310</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>0.2 ml Non-skirted PCR plate</td>
<td>AB-0600-L</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>0.8 ml 96-well MIDI plate</td>
<td>AB-0859</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Adhesive sealing sheets</td>
<td>AB-0558</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>96-well Sealing Cap Mats</td>
<td>AB0566</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Whatman™ Lens cleaning tissue</td>
<td>2105-841</td>
<td>GE Healthcare, IL, USA</td>
</tr>
<tr>
<td>UltraPure™ DNase/RNase-Free Distilled Water</td>
<td>10977035</td>
<td>ThermoFisher Scientific</td>
</tr>
<tr>
<td>Sodium hydroxide solution 10 M</td>
<td>72068-100ML</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>Ethanol absolute ≥99.8%, Molecular biology grade</td>
<td>437433T</td>
<td>VWR International, PA, USA</td>
</tr>
<tr>
<td>Resuspension buffer (RSB)</td>
<td>15068696</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Human euploid male genomic DNA</td>
<td>G1471</td>
<td>Promega, WI, United States</td>
</tr>
</tbody>
</table>
2.2 DNA controls for assay development

2.2.1 Coriell permissions

The cell lines described here were obtained with permission from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, NJ, USA. The most recent material transfer agreement (MTA) and statement of research intent (SORI) documents for the collaboration between Illumina Fulbourn UK and The Hoffmann Laboratory (University of Copenhagen, Denmark) are attached in Appendix 2 (A2.1 to A2.5). Cell line GM00526 was also purchased under the Illumina material transfer agreement, but personally identifiable genetic information was not shared outside the Illumina laboratory under this collaboration; therefore, cell line GM00526 does not appear in the collaborator Statement of Research Intent A2.4 or A2.5.

2.2.2 Extraction of genomic DNA from cultured cell lines

Human euploid male genomic DNA (Part: G1471; Promega, WI, United States) was used for positive control of whole genome amplification. Promega male genomic DNA was also used to generate a euploid reference dataset following whole genome amplification and NGS analysis. For assay controls of known karyotype, genomic DNA was extracted from cell lines (Coriell, NJ, USA), using the QIAamp® DNA Mini Kit (Part: 51304; QIAGEN, Hilden, Germany), according to the manufacturer’s instructions. Cells were harvested during the logarithmic phase of growth. Following a cell count and viability assessment, a cell pellet containing $3.0 \times 10^6$ cells was isolated by centrifugation at 1000 rpm for 5 min and the supernatant removed. Genomic DNA was extracted according to Appendix B of the QIAamp® kit; spin protocol for cultured cells. Samples were eluted into 1.5 ml DNA LoBind microcentrifuge tubes (Part: 0030108051; Eppendorf, Hamburg, Germany), and quality was assessed using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific Inc., MA, USA), prior to storage at -80 °C until use.

2.2.3 Description of cell lines used for assay controls

The cell lines used as controls for validation of single cell genomics pipelines are described in Table 2.3.
Table 2.3: Description of EBV-lymphoblastoid and fibroblast cell lines.

<table>
<thead>
<tr>
<th>Cell line ID</th>
<th>Karyotype</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM07224</td>
<td>46,XX</td>
<td>LCL: Cystic Fibrosis; CF Carrier</td>
</tr>
<tr>
<td>GM07225</td>
<td>46,XY</td>
<td>LCL: Cystic Fibrosis; CF Carrier</td>
</tr>
<tr>
<td>GM07228</td>
<td>46,XY</td>
<td>LCL: Cystic Fibrosis; CF Affected</td>
</tr>
<tr>
<td>GM10608</td>
<td>46,XY,del(20)(p12p11.2).arr 20p12.2p11.23(9820603-17979469)x1</td>
<td>LCL: Chromosome 20, p arm 8.16 Mb deletion.</td>
</tr>
<tr>
<td>GM10985</td>
<td>46,XX,del(3)(p25).arr 3p26.3p25.3(35333-10305377)x1</td>
<td>LCL: Chromosome 3, p arm 10.27 Mb deletion.</td>
</tr>
<tr>
<td>GM50121</td>
<td>46,XY,i(18)(q11.1).ish i(18)(q11.1)(D18Z1+,18qtel11++).arr 4q31.22(145061541-145163047)x1,18p11.32p11.21(1542-15391751)x1,18q11.1q23(16795644-76116029)x3</td>
<td>LCL: Chromosome 18, p arm trisomy</td>
</tr>
<tr>
<td>GM04927</td>
<td>47,XY,+21.arr[hg19] 6p22.2p21.32(26340784-32682724)x2 hmz.21p11.1q22.3(11039259-48096945)x3</td>
<td>LCL: Chromosome 21, trisomy</td>
</tr>
<tr>
<td>GM00526</td>
<td>47,XY,+13.arr[hg19]13q11q13.3(19045719-39052799)x3,13q13.3q21.3(39061889-72476988)x3,13q21.33(72480592-115108385)x3</td>
<td>Fibroblast: Chromosome 13, trisomy</td>
</tr>
</tbody>
</table>

2.3 Clinical aneuploidy data derived from human embryos

Four commercial laboratories provided anonymised clinical aneuploidy outcomes for 2,177 human embryos from 215 couples, that received preimplantation genetic testing for aneuploidy (PGT-A) analysis during assisted reproductive therapy cycles between 10th November 2014 and 3rd December 2016. PGT-A was performed on a single trophectoderm biopsy per embryo, using an NGS-based, copy number variation detection assay (VeriSeq PGS®; Illumina Inc., CA, USA). The retrospective analysis of aneuploidy data did not influence clinical practice, and all embryo biopsies were performed by qualified embryologists. The PGT-A assay was performed by a commercial testing facility as part of individual patient treatment. For observed chromosomal abnormalities, copy number status (gain/loss), percentage observed mosaicism, affected chromosome arm (p/q) and size of the segmental event (Mb) were recorded. The data were further stratified by maternal patient age (years), date of oocyte collection and treatment cycle number. The data were compiled in Microsoft Excel and stratified using custom macros in Visual Basic for Applications.
2.4 Collection of human oocytes and matching polar bodies

Oocytes and polar bodies were supplied through external collaborations from two clinical pipelines. In both cases, whole genome amplification of single cell material was performed at the source laboratory. The amplified DNA products were then shipped to me at Illumina Fulbourn UK for quality control and molecular genetic analysis.

2.4.1 Assisted reproductive technology pipeline

Surplus unfertilised oocytes were donated with fully informed consent, by patients receiving infertility treatment at the GENERA center for Reproductive Medicine, Rome, Italy. The protocol for embryo culture, activation and biopsy has been described previously (Ottolini et al., 2016). Briefly, the surplus oocytes were derived from ICSI cycles and the oocytes were all vitrified/thawed prior to activation. With regards to the Discussion in the relevant chapters regarding the source, we previously showed that MII vitrification has no effect on meiosis I segregation, as expected (Ottolini et al., 2015). Meiosis II-arrested oocytes were activated by exposure to 100 µM calcium ionophore for 40 minutes, to trigger resumption and completion of meiosis II. Polar bodies were sequentially biopsied upon their extrusion to ensure integrity of sample tracking. The oocyte and matching polar bodies were tubed separately in 0.2 ml PCR tubes and snap frozen in liquid nitrogen until whole genome amplification.

2.4.2 Fertility preservation pipeline

Novel techniques in fertility preservation involve the isolation of ovarian cortex tissue, prior to administration of cancer treatments that may otherwise decrease the ovarian oocyte reserve. Patients electing to donate the surplus ovarian medulla tissue, provide a source of immature oocytes that can be collected and matured in vitro, without the use of exogenous hormones that are used in assisted reproductive therapies. The protocol for in vitro maturation (IVM) of immature germinal vesicle (GV) oocytes to mature MII-arrested oocytes has been described previously (Gruhn et al., 2018). Briefly, small antral follicles containing cumulus oocyte complexes were isolated from surplus ovarian medulla tissue under saline solution. The cumulus oocyte complex consists of a GV oocyte surrounded by cumulus cells. Cumulus oocyte complexes were washed in ‘LAG’ media and incubated in ‘IVM’ media supplied with the MediCult IVM System (Part: 82214010; Origio, Måløv, Denmark) at 37 °C, 5% CO₂, and ambient O₂ for 24-48 h. Following IVM incubation, cumulus cells were removed by rinsing in 80 IU/ml hyaluronidase solution with a denuding pipette. Oocytes that had extruded the first polar body
were either biopsied to yield MII oocyte-PB1 duos, or activated with calcium ionophore using the protocol of Ottolini et al. (2016), to yield oocyte-PB trios as previously described. Whilst it is important to consider that the environmental differences of IVM could impact the fidelity of chromosome segregation and stability during meiosis, the evidence presented in this thesis suggests that aneuploidy rates in human small and large antral follicles are not dissimilar.

2.5 Culture of cell lines for single cell isolation and genomic DNA extraction

The culture of cell lines was performed according to the supplier’s recommendation, as described here. Specific equipment and materials required for cell culture are listed in Table 2.4, which complements Tables 2.1 and 2.2. All cell line culture was performed by me, for the purpose of this thesis.

Table 2.4: Cell culture-specific equipment and materials

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Brunswick™ Galaxy® 170S incubator</td>
<td>C017101002</td>
<td>Eppendorf, Hamburg, Germany</td>
</tr>
<tr>
<td>CP Grade Carbon Dioxide N4.5</td>
<td>111095-Y</td>
<td>BOC, Guildford, UK</td>
</tr>
<tr>
<td>Water bath</td>
<td>JBA5</td>
<td>Grant Instruments Ltd., Royston, UK</td>
</tr>
<tr>
<td>1×PBS pH 7.4</td>
<td>10010023</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Trypsin-EDTA (0.25%), phenol red</td>
<td>25200056</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Minimum Essential Media (MEM), Non-essential amino acids (NEAA)</td>
<td>10370047</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>(-)L-glutamine; (+)Phenol Red; (-)HEPES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GlutaMAX™ Supplement, 200 nM</td>
<td>35050038</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>RPMI Medium 1640</td>
<td>61870010</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>(+)GlutaMAX™; (+)Phenol Red; (-)HEPES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetal Bovine Serum, qualified, US origin</td>
<td>26140087</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Trypan blue solution (0.4%)</td>
<td>15250061</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>MycoSensor™ qPCR Assay Kit</td>
<td>302107</td>
<td>Agilent, CA, USA</td>
</tr>
<tr>
<td>QIAamp DNA Mini Kit</td>
<td>51304</td>
<td>QIAGEN, Hilden, Germany</td>
</tr>
<tr>
<td>Millpore® 0.22 µm polyethersulfone (PES) membrane, syringe driven filters</td>
<td>SLGP033RS</td>
<td>Merck, Darmstadt, Germany</td>
</tr>
<tr>
<td>T25 Corning® cell culture flask</td>
<td>430168</td>
<td>Corning, NY, USA</td>
</tr>
<tr>
<td>C-Chip™ disposable Haemocytometer</td>
<td>DHC-N01</td>
<td>NanoEnTek, Seoul, South Korea</td>
</tr>
</tbody>
</table>
2.5.1 Cell line culture procedures at Illumina

EBV-lymphoblastoid cell lines (LCLs) derived from whole blood, are classified by the Health and Safety Executive: Advisory Committee on Dangerous Pathogens, as Hazard Group 2, potentially infectious material. Therefore, manipulation of all cell cultures was performed in a Mars Runner 900, Class II Microbiological Safety Cabinet (Part: 703075; Labogene, Allerød, Denmark), within a Containment Level 2 laboratory. Cell lines were cultured under standard conditions (37.0°C; 5.0% CO₂) in a New Brunswick™ Galaxy® 170 S incubator (Part: C017101002; Eppendorf, Hamburg, Germany).

2.5.1.1 Receipt of live cell cultures

Cell lines were shipped in culture media at ambient temperature for three days. Upon arrival, shipment flasks incubated overnight. EBV-LCL culture flasks were stood upright to allow cell to settling for concentration by media reduction. The date of resuscitation and receipt, passage number and culture conditions were recorded.

2.5.1.2 Culture media preparation

EBV-LCLs were cultured in Roswell Park Memorial Institute (RPMI) 1640 media + 2 mM L-glutamine GlutaMAX™ (Part: 61870010; Thermo Fisher Scientific Inc., MA, USA), supplemented with 15% foetal bovine serum (Part: 26140087; Thermo Fisher Scientific Inc., MA, USA). Fibroblast cell lines were cultured in minimal essential media (Part: 10370047; Thermo Fisher Scientific Inc., MA, USA), supplemented with 15% foetal bovine serum and 2 mM L-glutamine GlutaMAX™ supplement (Part: 35050038; Thermo Fisher Scientific Inc., MA, USA). Stock bottles of RPMI 1640 and minimal essential media were stored at 4 °C.

Upon first use of stock bottles, single-use aliquots of foetal bovine serum and GlutaMAX™ supplements were prepared to minimise the risk of contamination from repeated tube-opening. Foetal bovine serum aliquots were stored at -20 °C and thawed in a water bath (Part: JBA5; Grant Instruments Ltd., Royston, UK) at 37 °C for 20 min immediately before use.

Complete media was prepared in sterile 50 ml centrifuge tubes (Part: 734-0448; VWR International, PA, USA), and sterile filtered using Millex® 0.22 μm polyethersulfone membrane, syringe driven filters (Part: SLGP033RS; Merck, Darmstadt, Germany) with 50 ml Luer-lock syringes (Part: 613-3925; VWR International, PA, USA). Complete media was labelled with a batch identifier and stored at 4 °C for no longer than 21 days from preparation.
2.5.1.3 **Cell count and viability assessment**

Cell count and viability assessments were performed by mixing 10 µl cell culture 1:1 with 0.4% Trypan blue solution (Part: 15250061; Thermo Fisher Scientific Inc., MA, USA), and manual counting using a C-Chip™ disposable haemocytometer (Part: DHC-N01; NanoEnTek, Seoul, South Korea) under an inverted microscope (Model: CKX31SF; Olympus, Japan).

2.5.1.4 **Cell line maintenance**

Prior to initial passage after cell line receipt, EBV-LCL cultures were concentrated by media reduction to a final volume of 10 ml, without disturbing cell clumps settled at the bottom of the flask. Upon initial passage of all cell lines, an aliquot containing 100,000 cells was taken following cell count. The cells were pelleted by centrifugation, washed in 1×PBS and assessed for the presence of mycoplasma using the MycoSensor™ qPCR Assay Kit (Part: 302107; Agilent, CA, USA), according to the manufacturer’s instructions. The MycoSensor™ assay was performed on a StepOnePlus™ Real-Time PCR System (Part: 4376600; Thermo Fisher Scientific Inc., MA, USA).

For all passages of EBV-LCL cultures, cell cultures were gently mixed with an electronic pipette (Part: JENC266-185; VWR International, PA, USA) with a 5 ml serological pipette tip (Part: 07-200-573; Corning, NY, USA) until no cell clumps were visible. A 500 µl aliquot was transferred to a 1.5 ml microcentrifuge tube for cell counting and viability assessment. On initial passage, EBV-LCL cultures were seeded at 0.4×10^6 cells ml⁻¹ in T25 flasks (Part: 430168; Corning, NY, USA) in a final volume of 4 ml RPMI 1640 + 15% foetal bovine serum, pre-warmed to 37 °C in a heat block (Part: UY-36620-05; Bio-Technne, MN, USA). In subsequent passages, 0.2×10^6 cells ml⁻¹ were seeded into T25 flasks in a final volume of 4 ml complete media. Culture flasks were mixed by gentle rocking prior to incubation. EBV-LCLs were typically cultured for a maximum of 4 days before passage, where the logarithmic growth phase was maintained by feeding on day two with 2-3 ml fresh, complete media. When feeding, sufficient fresh media was added to bring the cell concentration to 0.5×10^6 cells ml⁻¹. Confluency is achieved at 1.0-1.5×10^6 cells per ml.

For fibroblast cultures, culture media was removed with an electronic pipette and a 5 ml serological pipette tip, prior to washing of the cell sheet with 5 ml 1×PBS (Part: 10010023; Thermo Fisher Scientific Inc., MA, USA) to remove excess serum. The 1×PBS was discarded, and cells were dissociated from the flask by incubation with 1.5 ml Trypsin-EDTA (0.25%) (Part: 25200056; Thermo Fisher Scientific Inc., MA, USA) for 5 min at 37.0 °C, followed by gentle tapping of the flask and observation under an inverted microscope. Trypsin was inactivated by addition of 3 ml minimal essential media + 15% foetal bovine serum + 2nM GlutaMAX™ and pipette mixing. The cell culture was centrifuged in a sterile 15 ml conical tube at 1000 rpm for 5
min, and the supernatant removed. The cell pellet was resuspended in 1 ml fresh complete media by pipette mixing, and 10 µl was taken for cell count and viability assessment as described previously. Fibroblast cultures were seeded at 2.5×10^5 cells per ml in T25 flasks in a final volume of 6 ml minimal essential media + 15% foetal bovine serum + 2nM GlutaMAX™ pre-warmed to 37 °C. Cells were cultured under standard conditions prior to routine passage 80% adherent confluency on the lower flask wall; typically, every 72 hours.

2.5.2 Extraction of genomic DNA from cell cultures

Following a cell count as described previously, a cell pellet containing 3.0×10^6 cells was harvested following by centrifugation of liquid culture at 1000 rpm for 5 min using a 5424 Microcentrifuge (Part: 5404000060; Eppendorf, Hamburg, Germany), and removal of the supernatant. Genomic DNA was extracted using the QIAamp DNA Mini Kit, including the Appendix B protocol for cultured cells (Part#: 51304, QIAGEN, Hilden, Germany), according to the manufacturers’ instructions. Sample elute quality was assessed using a NanoDrop 1000 spectrophotometer (Part: ND-1000; Thermo Fisher Scientific Inc., MA, USA), prior to storage at -80°C until use.

2.6 Single cell isolation from cultured cell lines

The isolation of single cell samples from cell cultures is described here. Specific equipment and materials required for are listed in Table 2.5, which complements Tables 2.1 and 2.2. All cell isolations were performed by me, for the purpose of this thesis.
### Table 2.5: Materials for single cell isolation

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS-20x</td>
<td>9808</td>
<td>Cell Signaling Technology, MA, USA</td>
</tr>
<tr>
<td>Polyvinylpyrrolidone (PVP) Clinical Grade, 10%</td>
<td>10905000</td>
<td>Origio, Måløv, Denmark</td>
</tr>
<tr>
<td>Dead Cell Removal Kit</td>
<td>130-090-101</td>
<td>Miltenyi Biotech, Bergisch Gladbach, Germany</td>
</tr>
<tr>
<td>MACS MS columns</td>
<td>130-042-201</td>
<td>Miltenyi Biotech, Bergisch Gladbach, Germany</td>
</tr>
<tr>
<td>Trypan blue stain (0.4%)</td>
<td>15250-061</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Thin-walled 0.2 ml PCR tubes</td>
<td>B72911</td>
<td>BIOplastics, Landgraaf, The Netherlands</td>
</tr>
<tr>
<td>Nunc™ IVF ICSI dishes</td>
<td>150265</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>The STRIPPER® Micropipette</td>
<td>MXL3-STR</td>
<td>Origio, Måløv, Denmark</td>
</tr>
<tr>
<td>STRIPPER® tips, 125 µm</td>
<td>MXL3-IND-125</td>
<td>Origio, Måløv, Denmark</td>
</tr>
<tr>
<td>QIAamp® DNA Mini Kit</td>
<td>51304</td>
<td>QIAGEN, Hilden, Germany</td>
</tr>
</tbody>
</table>

#### 2.6.1 Single cell isolation buffer preparation

Sterile filtered 1×PBS and 1×PBS + 0.1% polyvinylpyrrolidone isolation buffers were prepared fresh for each isolation session from stock solutions. Two 1×PBS buffers labelled ‘Wash PBS’ and ‘Cell PBS’ were prepared by diluting 20×PBS in microbiological grade dH₂O to a final volume of 50 ml, each. Stock clinical grade polyvinylpyrrolidone was first diluted 1:1 with microbiological grade dH₂O to a final concentration 5% by pipette mixing. One 1×PBS + 0.1% polyvinylpyrrolidone buffer was prepared by diluting 20×PBS and 5% polyvinylpyrrolidone in microbiological grade dH₂O to a final volume of 10 ml. All buffers were sterile filtered using 0.22 µm polyethersulfone membrane, syringe driven filters as described previously.

#### 2.6.2 Dead cell removal from live cell cultures

Dead cell removal was performed 24-48 hours following passage, 1.0×10⁶ cells were harvested in the logarithmic phase of growth, and dead cells were removed using the MACS magnetic, microbead-based Dead Cell Removal Kit (Part: 130-090-101; Miltenyi Biotech, Bergisch Gladbach, Germany) with MACS MS column separation (Part: 130-042-201; Miltenyi Biotech, Bergisch Gladbach, Germany) according to the manufacturer’s instructions. Eluted cells were washed twice by centrifugation at 1000 rpm for 4 min using a 5424 Microcentrifuge and
resuspended in 1 ml 1×PBS + 0.1% polyvinylpyrrolidone by pipette mixing. A cell count and viability stain were performed before and after dead cell removal, as previously described. Only cell preparations with viability of ≥98% after dead cell removal were used for SCI. The final cell preparation was stored at 4°C until use.

2.6.3 Cell isolation and tubing

Immediately following dead cell removal, isolation buffers, thin-walled 0.2 ml PCR tubes (Part: B72911; BIOplastics, Landgraaf, The Netherlands) and Nunc™ IVF ICSI dishes (Part: 150265; Thermo Fisher Scientific Inc., MA, USA) were exposed to UV light for one hour in a cleaned Class II Microbiological safety cabinet. Numbered PCR tubes were then primed with 3 µl 1×PBS, labelled “Cell PBS”, with care taken to place the micro-droplet at the base of the tube, avoiding the inner sides of the tube. Nunc™ IVF ICSI dishes were prepared using two 10 µl, and up to 48× 6 µl 1×PBS, 0.1% polyvinylpyrrolidone micro-droplets as shown in Figure 2.1. An aliquot of 1 µl washed cell preparation was added to the stock dilution drop (red), and the stock washed cell tube was returned to 4°C.

Cells were isolated by manual micromanipulation using a STRIPPER® Micropipette (Part: MXL3-STR; Origio, Måløv, Denmark) with an aspiration volume of 1 µl and STRIPPER® tips with a 125 µm lumen (Part: MXL3-IND-125; Origio, Måløv, Denmark, under an stereo microscope (Model: M125C; Leica, Wetzlar, Germany) in a cleaned AirClean 600 microscope cabinet (Part: AC632A; AirClean Systems Inc., NC, USA). Between cell manipulations, the pipette tip was rinsed pipette mixing in the 1×PBS buffer labelled “Wash PBS”, followed by pipette mixing in the 1×PBS buffer labelled “Cell PBS”. All cell manipulation was performed by reverse pipetting, with the pipette tip loaded with “Cell PBS” prior to cell capture.

Cells were diluted further in the working dilution droplet (pink) before isolating a single cell into each isolation micro drop (white). All planes of focus in each isolation drop were observed to confirm the presence of a single cell. Immediately prior to cell tubing, each single cell was washed in an unused wash micro drop (yellow) by gentle dispensing and aspiration into the micropipette tip. For tubing, each cell was aspirated and held close to the micropipette tip before dispensing directly into the 3 µl priming droplet in the bottom of each isolation tube. The final volume of cell isolates did not exceed 4 µl 1×PBS + 0.025% polyvinylpyrrolidone. Capped isolation tubes were stored in a -20°C, 96-well tube rack (Part: 022510525; Eppendorf, Hamburg, Germany) for the duration of the isolation session. The tubing time and tube ID for each isolated cell were recorded. Additional ICSI plates were prepared for the remainder of the cell collections as described above. An isolation buffer negative control and 3-cell positive control were also
collected. Upon completion of isolation, cell samples were transferred to -20°C for storage prior to MDA amplification within 48 hours.

![Figure 2.1: Nunc™ IVF ICSI dish containing microdroplets for cell dilution and single cell isolation. Red: stock dilution drop, Pink: working dilution drop, White: isolation drop. Orange: wash drop.](image)

### 2.7 Whole-genome amplification: multiple displacement amplification

Sample amplification by multiple displacement amplification (MDA), was performed using the SureMDA DNA Amplification System (Part: PR-40-405102-00; Illumina Inc., CA, USA), according to the Infinium™ Karyomapping Assay Protocol Guide (Part: 15052710 Rev.B; Illumina Inc., CA, USA). The double stranded/duplex DNA (dsDNA) concentration of human euploid male gDNA (Part: G1471; Promega, WI, United States) was determined in triplicate, using a fluorometric method described in [here](#). Based on the average concentration, a fresh preparation of gDNA was made by serial dilution from stock to 12.5 pg/µl in kit-supplied 1×PBS, for each MDA experiment. For each positive amplification control, 4 µl final genomic DNA was added to a thin-walled 0.2 ml PCR tube, giving a final mass of 50 pg. A 4 µl negative ‘no template’ control (NTC) was prepared by adding kit-supplied 1×PBS to a thin-walled 0.2 ml PCR tube.

All sample pipetting steps were performed over a pre-cooled, -20°C tube rack (Part: 022510525; Eppendorf, Hamburg, Germany) in a Class II Microbiological safety cabinet within a designated pre-PCR laboratory. Reagents were dispensed onto the inner wall of each sample tube, above the droplet containing the sample, so as not to remove the sample material when withdrawing the pipette tip. Reagents were collected by centrifugation at 280 × g for 5 sec. Immediately before use, single cell isolates and isolation controls were removed from -20 °C storage and centrifuged at 300 × g for 3 min. 3 µl lysis mixture was added to each sample tube prior to incubation at 65 °C for 10 min, and cooling to 10 °C on a Veriti™ 96-Well Fast Thermal Cycler (Thermo Fisher Scientific Inc., MA, USA). Immediately after lysis, sample tubes were transferred back to a -20 °C tube rack and 3 µl neutralisation buffer was added to each tube. Due to the number of samples being processed, the sample tube rack was periodically tapped on the work surface to collect reagents together. The tubes were centrifuged briefly, before 40
µl master mix was added. Samples were amplified for 2 hours at 30 °C on a Veriti™ 96-Well Fast Thermal Cycler in a post-PCR laboratory. SureMDA products were stored at -20 °C prior to resolution by gel electrophoresis and quantitation of dsDNA.

2.8 Quality assessment of DNA samples

Analytical methods for the quality assessment of DNA products can be used as a control to exclude samples based on poor quality that may otherwise give misleading data. Several commercially available methods are available and may be used for multiple applications, either prior to molecular analysis or as in-line process controls. I have summarised the general methodologies here in two groups; namely, DNA fragment analysis, and DNA quantitation. Specific equipment and materials are listed in Table 2.6, which complements Tables 2.1 and 2.2.

Table 2.6: Materials for quality assessment of DNA

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vari-gel MAXI system</td>
<td>VG-SYS</td>
<td>Scie-Plas Ltd, Cambourne, UK</td>
</tr>
<tr>
<td>Agarose</td>
<td>A9539</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>10×Tris-borate-EDTA buffer</td>
<td>T4415</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>GelRed™ stain</td>
<td>41003</td>
<td>Biotium Inc., CA, USA</td>
</tr>
<tr>
<td>Gel loading buffer</td>
<td>G2526</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>Invitrogen™ 1kb DNA Extension Ladder</td>
<td>11568626</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>BioDoc-It™ Imaging System</td>
<td>-</td>
<td>Ultra-Violet Products Ltd., Cambridge, UK</td>
</tr>
<tr>
<td>Agilent High Sensitivity DNA Kit</td>
<td>5067-4626</td>
<td>Agilent, CA, USA</td>
</tr>
<tr>
<td>2100 Bioanalyzer Instrument</td>
<td>G2939BA</td>
<td>Agilent, CA, USA</td>
</tr>
<tr>
<td>MiSeq System</td>
<td>SY-410-1003</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>MiSeq Reagent Kit v3 (150-cycle)</td>
<td>MS-102-3001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NanoDrop® ND-1000 spectrophotometer</td>
<td>-</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Quant-iT™ dsDNA High-Sensitivity Assay Kit</td>
<td>Q33120</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>FLUOStar® Omega Plate Reader and MARS software</td>
<td>-</td>
<td>BMG Labtech, Ortenberg, Germany</td>
</tr>
<tr>
<td>96-well microplate, F-bottom, non-binding</td>
<td>655900</td>
<td>Greiner Bio-One, Kremsmünster, Austria</td>
</tr>
<tr>
<td>KAPA Library Quantification Kit for Illumina® Platforms</td>
<td>KK4828</td>
<td>Roche Holding AG, Basel, Switzerland</td>
</tr>
<tr>
<td>StepOnePlus™ Real-Time PCR System</td>
<td>4376600</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>MicroAmp™ Fast Optical 96-Well Reaction Plate with Barcode, 0.1 mL</td>
<td>4346906</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>MicroAmp™ Optical Adhesive Film</td>
<td>4311971</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
</tbody>
</table>
2.8.1 Quality assessment by DNA fragment analysis

2.8.1.1 Gel electrophoresis

MDA products, including isolation and amplification controls were resolved by gel electrophoresis for quality assessment of WGA. A 0.8% agarose gel (Part: A9539; Sigma-Aldrich, MO, USA) in 1×Tris-borate-EDTA buffer (Part: T4415, Sigma-Aldrich, MO, USA) with 1×GelRed™ stain (Part: 41003; Biotium Inc., CA, USA) was prepared on the day of use and allowed to set. Samples were diluted 1:1 with loading buffer (PART: G2526; Sigma-Aldrich, MO, USA), such that approximately 1 µg DNA was loaded per lane in 5 µl. Samples were run against a 1kb DNA extension ladder (Part: 11568626; Thermo Fisher Scientific Inc., MA, USA) at 1 µg in 2 µl, for 50 min at 120 V in a 30 cm tank supplied with powerpack and accessories (Part: VG-SYS; Scie-Plas Ltd, Cambourne, UK). The gels were imaged using a BioDoc-It™ Imaging System (Ultra-Violet Products Ltd., Cambridge, UK), then scanned for annotation.

2.8.1.2 Automated electrophoresis

Automated electrophoresis was used for high-sensitivity quality control of fragmented DNA during library preparation for NGS. Libraries were analysed following fragmentation of input DNA, either by tagmentation or mechanical sheering; and/or following addition of barcode indices, either by PCR or PCR-free methods (Illumina Inc., CA, USA). Samples were diluted to fall within the detection limit of the assay and processed using the Agilent High Sensitivity DNA Kit (Part: 5067-4626; Agilent, CA, USA) on a 2100 Bioanalyzer instrument and 2100 Expert software (Part: G2939BA; Agilent, CA, USA). Libraries that did not have a DNA fragment size distribution for NGS flow cell attachment were not included in library pools for sequencing.

2.8.1.3 Sequencing-based QC

To accurately determine the insert size distribution of NGS libraries prior to deep sequencing, a low pass NGS method was used. The insert size is required for accurate assessment of library concentrations, and subsequent quantitation-based normalisation of library pools following the qPCR method described here.

A library pool containing a nominal volume of each library was prepared, by addition of 5 µl of each library into a 1.5 ml DNA LoBind centrifuge tube. 10 µl of library pool was added to 90 µl hybridisation buffer in a 0.2 ml PCR tube and denatured at 96 °C for 3 min and cooled to 4 °C on a Veriti™ Thermal Cycler. 100 µl denatured library pool was added to 600 µl hybridisation buffer, that had been pre-cooled on wet ice in a 1.5 ml DNA LoBind centrifuge tube; vortexed
and pulse spun. 600 µl of denatured, diluted library pool was analysed by paired-end sequencing of 36 bases (2×36 bp) on a MiSeq instrument (Part: SY-410-1003; Illumina Inc., CA, USA) using the MiSeq Reagent Kit v3 (Part: MS-102-3001; Illumina Inc., CA, USA) with Amplicon chemistry. The mean insert size was extracted from *.bam files during secondary analysis, performed onboard the instrument.

2.8.2 Quality assessment by DNA quantification

2.8.2.1 Spectrophotometric quantitation
The quality and concentration of gDNA extracted from cell lines was assessed by spectrophotometry using a NanoDrop 1000 instrument (Model: ND-1000; Thermo Fisher Scientific Inc., MA, USA), for quality control of input material used as assay controls in single cell genomics pipelines. A sample volume of 1.5 µl was loaded onto the instrument, and the pedestal cleaned with a Whatman™ Lens cleaning tissue (Part: 2105-841; GE Healthcare, IL, USA) between samples. Genomic DNA samples with absorbance 260/280 values <1.5 (1.8 expected) and 260/230 values < 1.8 (2.0-2.2 expected) were not used for further experimentation. For gDNA samples extracted using the QIAamp DNA Mini Kit, the kit-supplied elution buffer was used as blank; otherwise microbiological grade dH₂O was used as a blank.

2.8.2.2 Fluorometric quantitation
Fluorometric quantitation of dsDNA in MDA products, NGS library preparations and bulk genomic DNA samples was performed using the Quant-IT™ dsDNA High-Sensitivity Assay Kit (Part: Q33120; Thermo Fisher Scientific Inc., MA, USA), according to the manufacturers’ instructions. Sample DNA was serially diluted in either microbiological grade dH₂O or RSB (Part: 15068696; Illumina Inc., CA, USA) to bring samples within the detection range of the assay. 190 µl Working Solution was added to each well of a flat bottom, 96-well microplate (Part: 655900; Greiner Bio-One, Kremsmünster, Austria). Supplied standard tubes mixed by pulse vortexing and centrifugation, and were aliquoted into 0.2 ml, 8-tube strips to allow multichannel pipetting. Standards were analysed in duplicate, where 10 µl standard was added to each well containing Working Solution. Depending on the assay requiring quantification of dsDNA, 2.5-10 µl of each sample was added to wells containing Working Solution, and the dilution factor of samples to standards used, was considered during calculation of sample concentrations. The assay plate was analysed using a FLUOStar® Omega Plate Reader and FLUOStar® Mars Data Analysis
software (BMG Labtech, Ortenberg, Germany), before exporting to Microsoft Excel for secondary analysis.

2.8.2.3 Quantitative PCR

For accurate quantitation of functional NGS libraries, the KAPA Library Quantification Kit for Illumina® Platforms was used for quantitation-based normalisation of NGS library pools, according to the TruSeq® DNA PCR-Free Library Prep Reference Guide (Part: 15036187; Illumina Inc., CA, USA). Briefly, samples were serially diluted 1/10,000 and 1/20,000 in RSB and each concentration for each sample was run in duplicate, together with 6 NTC samples. Kit-supplied standards were prepared as directed and run in duplicate. Reactions were assembled in a 0.1 ml MicroAmp™ Fast Optical 96-Well Plate (Part: 4346906; Thermo Fisher Scientific Inc., MA, USA). The four concentration (pM) readings per sample were averaged, and the size-adjusted concentration was calculated using the mean insert size obtained following sequencing-based QC of NGS libraries, using the following formula where 452 denotes the bp length of the standard:

\[
\text{Size-adjusted concentration (pM)} = \text{Average concentration (pM)} \times \left(\frac{452}{\text{Mean fragment length (bp)}}\right)
\]

Note: DNA NGS libraries cannot be accurately quantified using fluorometric methods following bead-based normalisation. In addition, qPCR facilitates targeted amplification of functional NGS libraries, to measure only DNA templates that have successfully incorporated the adapter sequences required for flow cell attachment on the sequencing instrument.

2.9 Genotyping by SNP microarray

Genotype information at sites of single nucleotide polymorphisms (SNPs) were determined by microarray analysis, using products of WGA from single cells or bulk genomic DNA extracted from cell lines as input. The methodology described below uses Infinium™ BeadChip array technology (Illumina Inc.), and specifically the HumanKaryomap-12 and Multi-Ethnic Global-8 arrays. Specific equipment and materials are listed in Table 2.7, which complements Tables 2.1 and 2.2.
Table 2.7: Materials for genotyping by SNP microarray

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinium™ option starter package</td>
<td>WG-15-305</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Infinium™ HumanKaryomap-12 DNA Analysis Kit</td>
<td>RH-103-1001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Infinium™ Multi-Ethnic Global-8 Kit</td>
<td>WG-316-1001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Hybex® System for Illumina</td>
<td>1057-30-2</td>
<td>SciGene, CA, USA</td>
</tr>
<tr>
<td>Illumina hybridization oven</td>
<td>SE-901-1002</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Sodium hydroxide solution ≈0.1 M</td>
<td>72079</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>2-Propanol</td>
<td>I9516</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>Ethylenediaminetetraacetic acid disodium salt solution</td>
<td>E7889</td>
<td>Sigma-Aldrich, MO, USA</td>
</tr>
<tr>
<td>OmniPur® Formamide, Deionized</td>
<td>LC 4610</td>
<td>Merck, Darmstadt, Germany</td>
</tr>
<tr>
<td>iScan System</td>
<td>SY-101-1001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NextSeq550 System</td>
<td>SY-415-1002</td>
<td>Illumina Inc., CA, USA</td>
</tr>
</tbody>
</table>

2.9.1 Infinium™ HumanKaryomap-12 BeadChip Array

Single cell MDA products and gDNA extracted from cell lines were genotyped using the Infinium™ HumanKaryomap-12 DNA Analysis Kit (Part: RH-103-1001; Illumina Inc., CA, USA), according to the manufacturer’s instructions (Infinium™ Karyomapping Assay Protocol Guide, Part: 15052710 Rev.B). Laboratory equipment required for processing Infinium™ BeadChip arrays is supplied as a single package (Part: WG-15-305; Illumina Inc., CA, USA).

Briefly, gDNA samples were adjusted to 50 ng/µl in RSB. Single cell MDA products were not diluted. 8 µl of each sample was denatured with ≈0.1 N NaOH (Part: 72079; Thermo Fisher Scientific Inc., MA, USA) for 10 min prior to isothermal amplification at 37 °C for 2 hours on a Hybex® System (Part: 1057-30-2; SciGene, CA, USA). Infinium™ amplification products were enzymatically fragmented 37 °C for 30 min and subjected to isopropanol precipitation (Part: I9516; Sigma-Aldrich, MO, USA) by centrifugation at 3000 ×g for 20 min at room temperature, before resuspension in hybridisation buffer. Fragmented DNA was denatured at 95 °C for 20 min, and 15 µl of each sample was loaded onto the included HumanKaryomap-12 v1.0 BeadChip Arrays. Microarray slides were placed inside sealed Illumina hybridisation chambers and incubated overnight in an Illumina hybridisation oven equipped with a rocking table (Part#: SE-901-1002, Illumina Inc., CA, USA) at 48 °C for 16-20 hours.

BeadChip arrays were washed using kit-supplied reagents before assembly into Te-Flow flow-through chambers, and inserted into a Tecan Chamber Rack (Tecan, Männedorf,
Switzerland). Single base extension of Infinium™ oligonucleotide probes was performed at 44 °C, prior to denaturation with 95% formamide + 1 mM EDTA to remove the template DNA. Staining of biotin and dinitrophenyl labelled nucleotides for signal amplification with two-colour master mix and anti-stain was performed at 32 °C. BeadChip arrays were immediately washed and coated before vacuum drying at >0.9 bar for 50 min. Arrays were stored in the dark until scanning.

(Processed arrays were scanned using either the iScan System (Part: SY-101-1001; Illumina Inc., CA, USA) or NextSeq550 System (Part: SY-415-1002; Illumina Inc., CA, USA). Prior to chip loading into the required adapter, the back of each chip was wiped with 70% ethanol v/v in microbiological grade dH₂O, to remove residual slide coating and ensure the chips lay flat in the adapter for scanning. Scanning requires array chip-specific Decode Map files (*.dmap), denoting the random assortment of oligonucleotide probe-bound beads on the array surface; downloaded using Decode File Client Utility Software (Version: 3.0.2; Illumina Inc., CA, USA). The Intensity Data files (*.idat) generated by both scanning systems were converted using AutoConvert software (Version: 2.0.1; Illumina Inc., CA, USA) or Beeline software (Version: 2.0.3; Illumina Inc., CA, USA) to generate Genotype Call files (*.gtc) for analysis. Generation of Genotype Call files requires product-specific manifest files (*.bpm) and cluster files (*.egt) for data annotation, available at www.illumina.com.

2.9.2 Primary analysis of SNP microarray data

Intensity data (*.gtc and *.idat) files were imported into the genotyping module of the GenomeStudio software (Version: 2.0.2; Illumina Inc., CA, USA) for genotype calling using the Illumina GenCall algorithm. The default GenCall threshold for genotype calling was set at 0.15, although this variable was adjusted depending on the requirements of secondary analysis, as detailed in later chapters of this thesis. Total call rates and heterozygous call rates were used to screen sample quality. The full data table was exported in *.csv format for secondary analysis.

2.9.2.1 Infinium™ assay quality controls

The Infinium™ controls dashboard, within the GenomeStudio software were assessed for sample and assay performance, according to the Infinium™ Assay Lab Setup and Procedures Guide (Part: 11322460 v02; Illumina Inc., CA, USA). Briefly, Infinium™ BeadChip Arrays include the sample independent control probes for staining, single-base extension, target removal and hybridisation. The assay also includes sample-dependent control probes for stringency, non-
specific binding and non-polymorphic sites. The controls dashboard was used to troubleshoot assay performance versus sample quality.

2.10 Next generation sequencing

Preparation of NGS libraries was performed using commercially available kits and reagents with some workflow with modifications. Libraries were sequenced using Illumina systems. NGS-specific equipment and materials are listed in Table 2.8, which complements Tables 2.1 and 2.2.

Table 2.8: Materials for next generation sequencing

<table>
<thead>
<tr>
<th>Material</th>
<th>Part#</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic stand-96</td>
<td>AM10027</td>
<td>Thermo Fisher Scientific Inc., MA, USA</td>
</tr>
<tr>
<td>Tagment DNA Buffer (TD)</td>
<td>15027866</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Tagment DNA Enzyme 1-X (TDE1-X)</td>
<td>20010676</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Stop Tagment (ST)</td>
<td>15033016</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Sample Purification Beads (SPB)</td>
<td>15041032</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>Nextera® Library Amplification Mix 3 (NLM3)</td>
<td>20000465</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>TG VeriSeq PGS v2-NextSeq Index Kit</td>
<td>20000440</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>MiSeq Reagent Kit v3 (150 cycles)</td>
<td>MS-102-3001</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>M220 sonicator</td>
<td>500295</td>
<td>Covaris, MA, USA</td>
</tr>
<tr>
<td>TruSeq® DNA PCR-Free High Throughput Library Prep Kit</td>
<td>20015963</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NovaSeq 6000 S4 Reagent Kit (300 cycles)</td>
<td>20012866</td>
<td>Illumina Inc., CA, USA</td>
</tr>
<tr>
<td>NextSeq 500/550 High Output v2 Kit</td>
<td>FC-404-2004</td>
<td>Illumina Inc., CA, USA</td>
</tr>
</tbody>
</table>

2.10.1 Copy number variation detection in SureMDA products

The following describes a custom NGS library preparation workflow for genomic content analyses of single cells by copy number variation detection.

2.10.1.1 Sample input

MDA products from single cells were quantified according to the fluorometric Quant-iT method described previously and diluted to a final concentration of 5 ng/µl in RSB.
2.10.1.2 Tagmentation

Each library was tagmented according to the Nextera® DNA Library Prep Reference Guide (Part: 15027987 v01; Illumina Inc., CA, USA) with the following modification; 100 ng MDA product in 20 µl RSB was used for sample input. Tagmentation was stopped by addition of 5 µl ST (Part: 15033016; Illumina Inc., CA, USA) to each sample well as described in the manufacturer’s protocol.

2.10.1.3 Clean-up tagmented DNA

Tagmented DNA was cleaned up by 1.8× Solid Phase Reversible Immobilisation (SPRI) on Sample Purification Beads (Part: 15041032; Illumina Inc., CA, USA), according to the Nextera® Rapid Capture Enrichment Reference Guide (part: 15037438 v01; Illumina Inc., CA, USA), with the following modifications: Clean-up was performed in a 0.8 ml 96-well midi plate (Part: AB-0859; Thermo Fisher Scientific Inc., MA, USA); 90 µl Sample Purification Beads and 50 µl of each tagmented DNA library were used as input for clean-up; following removal of 80% ethanol, samples were dried on a heat block (Part: UY-36620-05; Bio-Techne, MN, USA) set to 50 °C for 5 min; samples were resuspended in 25 µl RSB, and 20 µl cleaned-up supernatant was transferred to a 0.2 ml 96-well PCR plate (Part: AB-600-L; Thermo Fisher Scientific Inc., MA, USA) for PCR.

2.10.1.4 PCR amplify tagmented DNA

DNA libraries were amplified and barcoded by PCR according to the Nextera® Rapid Capture Enrichment Reference Guide (part: 15037438 v01; Illumina Inc., CA, USA), with the following modifications: Indexes used were of the 96-sample NextSeq Index Kit (Part: 20000440; Illumina Inc., CA USA); amplification master mix was NLM3 (Part: 20000465; Illumina Inc., CA USA); the used PCR program described in Figure 2.2.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>72°C</td>
<td>3 min</td>
</tr>
<tr>
<td>98°C</td>
<td>3 min</td>
</tr>
<tr>
<td>98°C</td>
<td>30 sec</td>
</tr>
<tr>
<td>60°C</td>
<td>30 sec</td>
</tr>
<tr>
<td>72°C</td>
<td>30 sec</td>
</tr>
<tr>
<td>72°C</td>
<td>5 min</td>
</tr>
<tr>
<td>10°C</td>
<td>Hold</td>
</tr>
</tbody>
</table>

*Figure 2.2:* PCR program for Nextera®-based NGS library preparation from MDA products
2.10.1.5  Clean-up amplified DNA

PCR products were cleaned-up 1× SPRI on Sample Purification Beads (Part: 15041032; Illumina Inc., CA, USA), according to the VeriSeq® PGS Library Prep Reference Guide (Part: 15052877 v03; Illumina Inc., CA, USA) with the following modification: Following removal of 80% ethanol, samples were dried on a heat block (Part: UY-36620-05; Bio-Techne, MN, USA) set to 50 °C for 5 min. Cleaned-up amplified libraries were quantified using the Quant-iT method described previously for QC. Samples with a concentration ≥5 ng/µl were used for normalisation.

2.10.1.6  Normalise and pool libraries

Cleaned-up amplified DNA was subjected to bead-based normalisation as described in the VeriSeq® PGS Library Prep Reference Guide (Part: 15052877 v03; Illumina Inc., CA, USA) without modification. Libraries were typically pooled by addition of 5 µl each normalised library to a 1.5 ml LoBind microcentrifuge tube. The library pool was mixed by pulse vortexing and centrifugation, before aliquoting into single-use 0.2 ml PCR tubes in 20 µl volumes. Library pool aliquots were maintained on wet ice or stored at -20 °C until sequencing.

2.10.1.7  Sequencing on MiSeq System

Pooled libraries were sequenced using a MiSeq instrument (Part: SY-410-1003; Illumina Inc., CA, USA) with MiSeq Reagent Kit V3 consumables (Part: MS-102-3001; Illumina Inc., CA, USA). 8-15 µl of library pool was added to hybridisation buffer in a 0.2 ml PCR tube, with a final volume of 100 µl, and denatured at 96 °C for 3 min and cooled to 4 °C on a Veriti™ Thermal Cycler. The input volume of library pool was adjusted for a target cluster density of 1200-1400 K clusters/mm². 100 µl denatured library pool was added to 600 µl hybridisation buffer, that had been pre-cooled on wet ice in a 1.5 ml DNA LoBind microcentrifuge tube; vortexed and pulse spun. 600 µl of denatured, diluted library pool was analysed by paired-end sequencing of 36 bases (2×36 bp) using V3 Amplicon chemistry.

2.10.1.8  Primary analysis

Primary analysis of NGS data is performed on the MiSeq instrument, following demultiplexing and alignment of sequence reads to the Homo sapiens Whole Genome FASTA, hg19 reference genome (University of California, Santa Cruz), using Burrows-Wheeler Aligner (BWA) within the MiSeq Reporter software. Cluster density, total reads, reads passing filter and Q30 quality scores
were assessed for library QC. Binary sequence alignment map files (*.bam) were used for secondary analysis.

2.10.1.9 Secondary analysis
Sequencing *.bam files were imported into BlueFuse® Multi software (Version: 4.3; Illumina Inc., CA, USA), and processed using a custom annotation and reference database specifically for use with sequencing data generated from single cells following WGA by MDA. Briefly, bash scripting, BEDtools (Quinlan and Hall, 2010), and SAMtools (Li et al., 2009) were used to remove unmapped reads, duplicate reads, poorly mapped reads (mismatch threshold). Sequence reads were counted into genome-wide fixed variable bins with a median of 1 Mb. Read counts per bin were normalised for GC content and bias was removed according to the MDA-specific reference database of euploid samples. Data was smoothed using a 13-bin sliding median. Copy number variation was determined using a gaussian probability function (copy numbers 0-4) with a standard deviation of 0.33, assuming the median autosomal read count correlated to a copy number of two (Fiorentino et al., 2014a). In our case, both normalised and raw aligned sequence counts per bin were used for assessment of copy number variations in each sample. Results were recorded using Microsoft Excel. Further details of this analysis are described fully in later chapters of this thesis.

2.10.2 Whole-genome deep sequencing
2.10.2.1 PCR-free library preparation
MDA products from single cells were prepared for sequencing according to the TruSeq® DNA PCR-Free Library Prep Reference Guide (Part: 15036187; Illumina Inc., CA, USA) and associated reagents (Part: 20015963; Illumina Inc., CA, USA), with the following conditions:

i. Samples were diluted to 20 ng/µl in 60 µl using the Quant-iT method described previously, for 1 µg DNA input in 50 µl.

ii. Input DNA was fragmented with a Covaris M220 sonicator (Part: 500295; Covaris, MA, USA), with duty factor: 20%, peak power: 50 W, duration: 55 sec and temperature: 20 °C.

iii. Cleaned-up fragmented DNA was stored at -20 °C for 5 days before proceeding with the protocol.

iv. End repair control (CTE), A-tailing control (CTA) and Ligation control (CTL) reagents were included.
2.10.2.2 Quality control of library preparation

In addition to the kit-supplied CTE, CTA, CTL process controls, sample libraries were assessed by 2100 Bioanalyzer following clean-up of fragmented DNA, and on final sequencing libraries as previously described. Final sequencing libraries were quantified by qPCR, and insert size was assessed by low pass 2×36 bp NGS on a MiSeq System as previously described.

2.10.2.3 Normalise libraries

Libraries were diluted to 4 nM in kit-supplied hybridisation buffer (HT1), based on qPCR data and average insert size determined by MiSeq QC analysis. 5 µl of each 4 nM library was pooled in a 1.5 ml DNA LoBind microcentrifuge tube. Library pools were stored at -20 °C until sequencing.

2.10.2.4 Sequencing on NovaSeq System

Pooled libraries were sequenced using a NovaSeq 6000 instrument (Part: 20012850; Illumina Inc., CA, USA) with a NovaSeq 6000 S4 Reagent Kit (300 cycles) (Part: 20012866; Illumina Inc., CA, USA). 120 µl of 1 nM library pool in HT1 was prepared for sequencing according to the NovaSeq 6000 Sequencing System Guide (Part: 1000000019358 v05; Illumina Inc., CA, USA) with direct flow cell lane loading (Xp protocol). 130 µl of sequence-ready library pool was loaded onto each of the 4 flow cell lanes, and the sequencing run was started within 30 min. Paired-end sequencing of 151 bases (2×151 bp) proceeded with recipe version 1.3.0.

2.10.2.5 Sequencing on NextSeq System

Individual libraries were diluted to 1 nM in HT1. A serial dilution of 156 pM library in 100 µl HT1 was denatured at 96 °C for 3 min and cooled to 4 °C on a Veriti™ Thermal Cycler. 15-30 µl of denatured library was added to 1.3 ml HT1 that had been pre-cooled on wet ice and mixed by vortexing and pulse centrifugation. The final library was sequenced on a NextSeq500 System (Part: SY-415-1001; Illumina Inc., CA, USA) or NextSeq550 System (Part: SY-415-1002; Illumina Inc., CA, USA) using a NextSeq 500/550 High Output v2 kit (300 cycles) (Part: FC-404-2004; Illumina Inc., CA, USA). The input volume of denatured library pool was adjusted between runs for a target cluster density of 260 K/mm². For NextSeq runs, one sample was loaded per flow cell to achieve maximum sequencing depth.
2.10.2.6 Primary analysis

Sequencing output folders were uploaded to BaseSpace Sequence Hub (www.basespace.illumina.com) for primary analysis. BCL (Base Call) file demultiplexing and FASTQ generation was performed using the FASTQ Generation application (version 1.0.0). Alignment of sequence reads to the *Homo sapiens* hg19 reference genome (University of California, Santa Cruz) was performed using the BWA Aligner application (version 1.1.4). The resulting *.bam files were downloaded for secondary analysis using the GATK pipeline v4.0 (McKenna *et al.*, 2010).

2.10.2.7 Secondary analysis

The *.bam files were sorted and indexed using SAMtools v1.3.1 (Li *et al.*, 2009). Subsequently, the HaplotypeCaller from GATK pipeline with default parameters was used to genotype the alignments. In order to obtain a full list of genotypes and not only the variant sites, the output format for the HaplotypeCaller was changed from VCF to GVCF.
Chapter 3: Chromosome constitution of human embryos.

3.1 Overview

Whole chromosome aneuploidy is a major cause of miscarriage and congenital birth defects. In the clinic, the use of assisted reproductive technology (ART) provides a unique opportunity to biopsy the day-5 blastocyst for preimplantation genetic testing for aneuploidy (PGT-A). PGT-A describes the genomic assessment of embryo quality, that is used to prioritise euploid embryos for transfer back to the uterus during an IVF cycle. Recent developments in technology have seen a transition from microarray-based PGT-A to NGS platforms. Therefore, chromosome copy number is assessed on a linear scale of sequence read counts, rather than a log scale of array channel intensity ratios. The higher dynamic range for copy number detection using NGS is more sensitive to detection of low-level mosaicism and sub-chromosomal abnormalities within a single trophectoderm (TE) biopsy, than microarray technologies. This raises two questions. Firstly, can NGS-based PGT-A be used to make accurate inferences about the incidence of mosaicism and segmental aneuploidy in human preimplantation embryos? Secondly, what is the clinical significance of a mosaic or segmental aneuploid PGT-A result?

Mosaic aneuploidy detection in the clinic is limited by the invasive procedure of embryo biopsy, and the use of trophectoderm cells to make inferences about the foetal proportion of the embryo. The distribution of mosaic cells within the TE, and or inner cell mass may lead to sampling errors, and multiple biopsies are not common in clinical practice. An estimated 10-50% of mosaic PGT-A results are discordant when retested by further biopsy of the inner cell mass, trophectoderm or analysis of the whole remaining blastocyst (Vera-Rodriguez and Rubio, 2017). Furthermore, the discordance between PGT-A results from the same embryo are challenging to distinguish from damage induced by the biopsy procedure or potential technical limitations of PGT-A assays. Recent studies have demonstrated a reduced implantation rate following the transfer of mosaic embryos, however live births have been reported (Greco and Fiorentino, 2015; Munné et al., 2017). Guidance on the transfer of mosaic embryos from two international consortia have been published (PGDIS, 2016; CoGEN, 2018). However, variation in ART procedures, embryo biopsy, PGT-A platforms and aneuploidy calling criteria between clinics are reflected in the highly heterogenous estimates of embryo mosaicism within the community (Taylor et al., 2014).

Sub-chromosomal, or ‘segmental’ aberrations represent a smaller, however significant proportion of abnormalities detected in 4% of spontaneous abortions, and ~10% of PGT-A embryos (Vera-rodríguez et al., 2016; Babariya et al., 2017). The improved dynamic range in
copy number variation (CNV) detection with NGS, reveals both mosaic and uniform segmental CNV events. Characterising the incidence of segmental aneuploidies in the ART setting can give insight to the origin of segmental aneuploidies, where uniform copy number changes are largely assumed to be of meiotic origin owing to the proportion of biopsied cells affected. Secondly, the incidence of segmental aneuploidies may have clinical implications in prioritisation of embryos for transfer particularly with CNVs <15 Mb which may be hard to distinguish from technical noise or whole-genome amplification (WGA) artefacts (Babariya et al., 2017).

In this chapter, I have analysed data from four PGT-A testing laboratories, assessing the incidence of chromosomal abnormalities in trophectoderm biopsies. CNVs were detected using the same NGS-based assay, where the percentage of mosaicism and size of the CNV were also reported. We find that mosaic and segmental abnormalities do not mirror the maternal age-effect, nor share the same chromosome bias profile seen for whole chromosome aneuploidies. In addition, embryos that contain only mosaic or segmental abnormalities are prevalent in cycles from women aged 30-44 years and are found in the absence of euploid embryos in ~5% of cycles. Segmental aneuploidies frequently affect whole chromosome arms, with preferential q-arm loss. Independent reanalysis of segmental aneuploidies from Lab D revealed that segmental aneuploidy calling is prone to discordant reports between reviewers; most likely due to the noise introduced from sample handling, whole-genome amplification or NGS.

The independent data analysis of segmental aneuploidies from a single laboratory described here, and the mapping of chromosome breakpoints to fragile sites and recombination hotspots forms my contribution to the following manuscript (submitted).


3.2 Materials and Methods

3.2.1 Study design

Aneuploidy reporting data were provided by four different laboratories that performed preimplantation genetic testing for aneuploidy (PGT-A) by next generation sequencing (VeriSeq® PGS; Illumina Inc., CA, USA) between November 2014 and December 2016. The laboratories were provided with a Microsoft Excel template to capture maternal age, number of embryos
tested per cycle and observed chromosomal abnormalities. For each abnormality, the affected chromosome, level of mosaicism, whole or segmental aneuploidy and size of the segmental aneuploidy were reported. A further free-text field was provided for additional information such as ‘no result’ and triploidy.

3.2.2 Embryo Culture and Biopsy

Oocytes were retrieved following ovarian hyperstimulation and fertilised according to standard protocols, as determined by the individual patient’s prognosis. The resulting embryos were cultured at 37°C, 6% CO₂, and 5% O₂ in either single or sequential embryo culture medium. A biopsy of trophoderm (TE) material was performed on embryos surviving to the blastocyst stage of development, typically on day-5 or day-6 post-fertilisation. A small hole was created in the zona pellucida, though which a biopsy pipette was inserted. A biopsy of 3-10 trophoderm cells was performed by gentle aspiration and laser-assisted removal of the biopsy from the embryo with a biopsy pipette. The trophoderm cells were washed in a phosphate-buffered saline solution and transferred to sterile 0.2 ml thin-walled PCR tubes. Samples were snap frozen in liquid nitrogen, prior to storage and shipment to the testing laboratory for PGT-A by NGS. This retrospective analysis study did not influence clinical practice. Embryo biopsy and sample handling were conducted by trained, clinical embryologists.

3.2.3 Next-Generation Sequencing

DNA extraction and whole-genome amplification (WGA) were carried out using the SurePlex DNA Amplification System (Part: PR-40-415101-00; Illumina Inc., CA, USA) according to the manufacturer’s instructions. Sequencing libraries were prepared from WGA products using the VeriSeq® PGS Kit (Part: RH-101-1001; Illumina Inc., CA, USA) according to the manufacturer’s instructions. Briefly, SurePlex products were tagmented (tagged and fragmented) using the VeriSeq® PGS transcriptome. Samples were then amplified using a limited-step PCR program before purification and bead-based normalisation. Pooled libraries were sequenced on a MiSeq System (Illumina Inc., CA, USA) at 0.01× depth with a 1×36 bp recipe.
3.2.4 Primary analysis

3.2.4.1 Abnormality scoring by PGT-A laboratories

Primary data analysis and chromosomal abnormality reporting was performed by the PGT-A laboratories. Data were analysed using BlueFuse® Multi (BFM) software (Illumina Inc., CA, USA). Sequencing reads were aligned to a reference human genome, divided into fixed variable bins with a median of 1 Mb. Filtered reads mapping to individual bins were counted, and copy number was determined in each sample on a per chromosome basis using proprietary BFM algorithms. Chromosomes with a relative copy number variation of ±0.5 over the majority of the chromosome length were automatically called as aneuploid by the BFM software but were manually analysed for the purpose of this study. The resolution achieved by a median of 1 Mb bins, allowed segmental aneuploidies to be manually identified when a CNV was present in only part of the chromosome.

All four laboratories additionally assessed the apparent level of mosaicism for both whole chromosome and segmental aneuploidies. Mosaicism reporting was performed using the reporting criteria proposed in the Preimplantation Genetic Diagnosis International Society (PGDIS) newsletter dated 19th July 2016 (www.pgdis.org/docs/newsletter_071816.html). Briefly, samples with CNV values of ≥1.8 to ≤2.2 were considered euploid and those with CNV values of <1.2 or >2.8 were classified as uniformly aneuploid. Samples with CNV values of ≥1.2 to <1.8 or >2.2 to ≤2.8 were labelled as having a mosaic loss or gain, respectively. The relative mosaicism of whole chromosome or segmental aneuploidies with in single trophectoderm biopsies was then reported by the PGT-A laboratory as a percentage gain or loss of one chromosome copy, from an expected copy number of 2.0. For example, a copy number of 2.4 would be scored as a 40% gain, and a copy number of 1.6 would be scored as a 40% loss. Validation of the same PicoPLEX WGA and NGS analysis pipeline has been independently validated for the detection of mosaicism using cell line mixture models, to a resolution of 17% variation (Goodrich et al., 2016). The complete PGT-A reporting data were anonymised at the source laboratory and transferred to me for secondary analysis.

3.2.4.2 Independent validation of scored abnormalities

There is currently no standardised scoring protocol for apparent mosaic or segmental CNVs. Using the relative deviation from an expected copy number alone is not sufficient to distinguish between mosaic abnormalities and technical noise, such as non-linear whole genome amplification. Regarding segmental aneuploidies, small events (low number of bins) and low-level mosaicism (low copy number separation between adjacent bins on the same
chromosome), make abnormality reporting particularly challenging at telomeric or GC-rich regions, where the bin size is typically larger than 1 Mb and technical noise is expected.

To assess concordance for segmental abnormalities, we independently assessed 522 samples from the complete dataset. This included 131 biopsy records containing 158 out of 168 (94.0%) segmental abnormalities reported by one laboratory. The remaining 391 records pertained to cycle-matched biopsies that were not reported to contain a segmental abnormality and were used as a control. Segmental aneuploidies were targeted for validation as we expected these to be most affected by the technical resolution of the NGS assay.

To this end, the binned sequence data were extracted from BlueFuse® Multi software at Lab D and reanalysed by me and one other independent scorer using a different charting software. Discordances between the original PGT-A reports, and the validation outcomes were recorded. Disagreements between validation reviewers were added to a list requiring future review, but the original report by Lab D was not overturned. In addition, a reduction in telomeric sequencing reads may be the result of inefficient WGA coverage or S-phase replication in these regions, rather than segmental chromosome breaks (Figure 3.5C). Apparent mosaic ‘telomeric drops’ were not counted as segmental aneuploidies during validation, unless they had been reported by Lab D. As a result, the incidence of segmental aneuploidies may be underestimated for fragments < 10 Mb, as they are hard to interpret at the current resolution. This is particularly problematic for mosaic segmental events, which can also be inferred as technical noise.

To assess concordance for whole chromosome and whole chromosome mosaic abnormalities, I further validated PGT-A data from the first 20 patient IDs found in the BFM records supplied by Lab D. This comprised 101 samples, of which 30 were reported to contain whole chromosome aneuploidies, and 16 were reported to contain a whole chromosome mosaic events; in total 86 whole chromosome abnormalities were expected.

3.2.5 Exclusion criteria

We received 2,177 samples from 545 couples from the four PGT-A laboratories. Several cycles included the use of donor oocytes where the maternal age of the oocyte donor was not provided. Following communication with the collaborating laboratory, since no patients ≥45 years received donor oocytes, donor cycles were excluded by means of an upper age filter of ≥45 years. In addition, embryos that gave no PGT-A result due to failure of the whole genome amplification or NGS were also excluded from secondary analysis. A further 12 samples that had been reported as triploid were also excluded from analysis of mosaic and segmental abnormalities (Section 3.3.3).
3.2.6 Classification of chromosomal abnormalities

On receipt of the anonymised PGT-A results from all four laboratories, the data were compiled, and embryos were stratified using custom macros in Microsoft Excel Visual Basic for Applications based on the pre-defined criteria described in Table 3.1. Briefly, embryos that were identified as triploid, were categorised as having karyotype-wide abnormalities (≥7 whole chromosome aneuploidies) (Table 3.1; Category 10). Embryo abnormality categories 1-16 were consolidated into five risk groups for embryo transfer based on the type of abnormalities reported. Risk groups were assessed with relation to maternal age, testing laboratory and embryos available for transfer per cycle.
Table 3.1: Embryo stratification categories and proposed ‘traffic-light’ system for clinical risk classification of embryos for transfer.

<table>
<thead>
<tr>
<th>Cat.</th>
<th>Eup.</th>
<th>Single whole aneu.</th>
<th>2-6 whole aneu.</th>
<th>≥7 whole aneu.</th>
<th>≥1 Mosaic whole aneu.</th>
<th>≥1 Uniform seg. aneu.</th>
<th>≥1 Mosaic seg. aneu.</th>
<th>Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Green: Low</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Red: High</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Purple: High</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Grey: High</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Amber: Unknown</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(± ✓)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Meeting exclusion criteria</td>
<td>None</td>
</tr>
<tr>
<td>18</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No PGT-A result achieved</td>
<td>None</td>
</tr>
</tbody>
</table>

Cat = Category; Eup. = Euploid; Seg. = Segmental; Aneu. = Aneuploidy/ies

3.2.7 Statistical analysis

For statistical analysis of differences in the data between different sample groups, Fisher’s exact test with a 95% confidence interval was applied. A P-value of <0.05 was considered to be statistically significant.

3.3 Results

3.3.1 A dataset of 2,083 embryo biopsies after stringent quality control.

Patients may present for assisted reproduction with male- and/or female-factor infertility. Increased aneuploidy associated with advanced maternal age (>35 years) is a major cause of repeated IVF failure and miscarriage. In some cases, patients may receive oocytes from younger donors to mediate this issue. This dataset is limited by cycles containing donor oocytes that were not indicated by the reporting laboratories, where maternal age was reported in terms of the patient receiving treatment, not the oocyte donor. Therefore, we could see variability in the maternal age effect on aneuploidy because of donor cycles. To mitigate this, I used an upper age inclusion limit of 44 years (not rounded) to screen the cycles. With this strategy we will suffer data loss, but we won’t detect false negatives due to the low rate of aneuploidies with a higher
(patient) maternal age. Additionally, the dataset is large such that any donor cycles that pass through the screening should be averaged out of incidence calculations. To this end, of the 2,177 embryo PGT-A reports received from four different PGT-A laboratories using NGS-based comprehensive chromosome screening, 40 (1.8%) embryos were excluded for failing to give a PGT-A result by NGS. This included two samples from different cycles in which only one embryo was tested (Table 3.1: Category 18). A further 54 samples from 15 cycles with a maternal age of <24 years or ≥45 years at the time of oocyte retrieval and were also excluded (Table 3.1: Category 17). The complete sample stratification of included samples is described in Table 3.2.

Table 3.2: Stratification of PGT-A samples by four collaborating laboratories.

<table>
<thead>
<tr>
<th>Stratification</th>
<th>PGT-A Laboratory</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Patients</td>
<td>100</td>
<td>130</td>
</tr>
<tr>
<td>Average Age ± S.D (All patients)</td>
<td>38.9 ± 4.1</td>
<td>34.2 ± 4.3</td>
</tr>
<tr>
<td>Total cycles</td>
<td>100</td>
<td>130</td>
</tr>
<tr>
<td>Total embryos</td>
<td>415</td>
<td>400</td>
</tr>
<tr>
<td>Age-excluded embryos</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Embryos with no PGT-A result</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Included embryos</td>
<td>393</td>
<td>399</td>
</tr>
<tr>
<td>Included cycles</td>
<td>94</td>
<td>129</td>
</tr>
<tr>
<td>Included patients</td>
<td>94</td>
<td>129</td>
</tr>
<tr>
<td>Included autosomes</td>
<td>18,078</td>
<td>18,354</td>
</tr>
<tr>
<td>Average Age ± S.D of included patients (24-44 years)</td>
<td>38.4 ± 3.7</td>
<td>34.3 ± 4.2</td>
</tr>
</tbody>
</table>

**S.D = Standard deviation**

### 3.3.2 Whole chromosome aneuploidies increase with maternal age, reducing the incidence of embryos affected by only mosaic or segmental abnormalities.

The age-related incidence of whole chromosome aneuploidies is a common feature in both natural and assisted conceptions. However, the incidence of mosaic and segmental aneuploidies is currently debated in the ART setting. In order to compare the incidence of chromosomal abnormalities reported following PGT-A with published clinical data, 2,083 samples from 532 cycles were stratified by clinical risk (Table 3.1). In total, 935 (45%) biopsies were reported as euploid (green: suitable for transfer), 377 (18%) as having mosaic and/or segmental aneuploidies only (amber: unknown clinical significance for transfer), 443 (21%) as having a uniform, single aneuploidy with or without mosaic and/or segmental aneuploidies (red: high risk, not suitable for transfer), 295 (14%) as having multiple chromosomal aneuploidies with or without mosaic and/or segmental aneuploidies (purple: high risk, not suitable for transfer), and 33 (2%) as
having karyotype-wide (≥7) aneuploidies (grey: high risk, not suitable for transfer), of which 12 were identified as triploid (Figure 3.1A). Triploid male samples can be identified due to the shift in the ratio of sequencing reads that align to the X and Y chromosomes compared to euploid autosomes. Triploidy is not possible to identify in female samples that do not contain aneuploidy when using copy number analysis, and therefore I expect a portion of euploid (46, XX) samples to also be triploid.

To compare this PGT-A dataset with the known maternal age effect on whole chromosome aneuploidy, the data were also divided into four age groups; 24-29, 30-34, 35-39 and 40-44 years (Figure 3.1C). As expected, there was a higher percentage of euploid embryos (green category) in younger patients (24-29 years; 56.0%; n=207), whereas both the ‘red’ (single aneuploidy ± mosaic/segmental abnormalities) and ‘purple’ (2-6 whole chromosome aneuploidies ± mosaic/segmental abnormalities) categories increased with maternal age, up to 44 years (30-44 years; 60.6%; n=391). The presence of mosaic or segmental aneuploidies only (amber category) was found to decrease with maternal age (24-29 years, 26.1%, n=54; 33-34 years, 22.4%, n=148; 35-39 years, 16.1%, n=133; 40-44 years, 10.7%, n=42). This observation is expected in a model where mosaic and segmental aneuploidies occur independently of whole chromosome aneuploidies, which increase with maternal age. Therefore, the incidence of biopsies reported as having both a whole chromosome aneuploidy and a mosaic or segmental event should also increase with maternal age; and conversely the incidence of biopsies reported and having only mosaic or segmental abnormalities should decrease with maternal age (Figure 3.1C). Across all maternal age groups, only a very small percentage (0.5–2.3%, n=33) of embryos had karyotype-wide abnormalities, involving ≥7 whole chromosome aneuploidies (grey category).

Finally, to identify variation in the incidence of abnormalities reported between the four PGT-A laboratories, the data were divided by testing centre (Figure 3.1D). A lower proportion of euploid samples were reported by Lab A (28.0%; n=393) than Labs B, C and D, respectively (56.1%, n=224; 47.4%, n=185; 46.2%, n=416). This can be explained by the higher average maternal age of the women undergoing PGT-A in Lab A (Figure 3.1B). The lowest proportion of aneuploid samples (red, purple and grey categories combined), was reported by Lab B (26.6%; n=399). When patient age is similar (Labs B, C and D), differences in abnormality reporting between labs may be due to inherent differences in patient populations, differences in embryo biopsy and sample handling procedures (storage and shipping), or to different
Figure 3.1: Proportion of biopsies categorised by PGT-A result. (A) Total embryos. (B) Age density distribution by PGT-A laboratory. (C) PGT-A result by maternal age. (D) PGT-A results by laboratory. Error bars show Standard Error of Proportion. Data drawn from experiment IDs: E32a_RCB_FinalClinicalAneuploidyDataset; E32b_AneuploidybyAge&Lab.
reporting standards used by each laboratory. This is important to consider when comparing the incidence of chromosome abnormalities reported from multiple labs; and demonstrates the requirement for standardised reporting procedures, particularly for events that are challenging to call, such as mosaicism or segmental aneuploidies.

3.3.3 **Mosaic and segmental abnormalities do not share the chromosome bias profile of whole chromosome aneuploidy.**

In natural conceptions and miscarriages, the risk of aneuploidy differs amongst chromosomes. We therefore investigated the frequency of whole chromosome, segmental, and mosaic aneuploidies amongst the 95,266 chromosomes in 2,071 embryos (following exclusion of triploid samples). As shown in Figure 3.2A, the incidence of whole chromosome aneuploidy (dark red) varied amongst the chromosomes, however chromosomes 15, 16, 21, and 22 showed the highest incidence of aneuploidy. This is expected from other studies as well as natural conceptions (Handyside et al., 2012; Soler et al., 2017). In contrast, there did not appear to be an elevation of specific chromosomes that were affected for mosaic aneuploidy (Figure 3A, light red). This is consistent with the notion that the mitotic errors during the embryonic divisions occur randomly amongst chromosomes.

Segmental aneuploidies were less frequent than whole chromosome aneuploidies and were analysed separately (Figure 3.2B). The larger chromosomes (1 to 9) displayed an elevated incidence of uniform and mosaic segmental abnormalities. However, there was also chromosome-specific effects. Chromosomes 1 and 2 showed higher frequencies of mosaic segmental events than full segmental events, and chromosome 9 showed a 2-fold higher frequency for uniform segmental events than mosaic segmental events. These patterns suggest that the underlying chromosomal pathologies that give rise to segmental aneuploidies (both uniform and mosaic) are distinct from those that result in whole chromosomal aneuploidies.

Current data suggests that mosaic and segmental aneuploidies occur independently of maternal age. To confirm this, I calculated the proportion of chromosomes affected by each abnormality type, stratified by age group (Figure 3.2C and 3.2D). As expected, the proportion of chromosomes affected by whole chromosome aneuploidy increases with maternal age (dark red), but the proportion affected by whole chromosome mosaicism does not follow this trend (light red) (Figure 3.2C). This is consistent with the notion that whole chromosome mosaicism occurs at a similar rate across all chromosomes, as well as across maternal age groups.
Figure 3.2: Frequency of chromosomal abnormalities reported in 2,071 embryo biopsies following PGT-A.  
(A) Frequency of whole chromosome aneuploidies, mosaic whole chromosome aneuploidies and combined total per chromosome.  
(B) Frequency of segmental aneuploidies showing a uniform copy number change, mosaic segmental aneuploidies and combined total per chromosome.  
(C) Proportion of analysed chromosomes (number of biopsies × 46) reported as whole chromosome aneuploid or whole chromosome mosaic, by maternal age.  
(D) Proportion of analysed chromosomes reported as having a uniform segmental abnormality or mosaic segmental abnormality, by maternal age. Error bars show Standard Error of Proportion. Data drawn from experiment IDs: E32a_RCB_FinalClinicalAneuploidyDataset; E32c_RCB_IncidenceMosaic&SegmentalByChromosome&Age.
The proportion of chromosomes affected by uniform segmental events appear lowest with maternal age <30 years and similarly for ages 30-44 years (Figure 3.2D: dark blue). The difference between the 24-29 and 30-34 age groups does not reach significance (p = 0.13, exact test); given the low incidence of segmental events, more data would be required to elucidate any maternal-age effect. Interestingly, the proportion of chromosomes affected by mosaic segmental events appears to decrease with maternal age (Figure 3.2D: light blue). This suggests an inverse age-related association between mosaic segmental abnormalities and maternal age, when compared to that of whole chromosome aneuploidies. This result could be explained by the temporal separation of errors giving rise to aneuploidy (meiotic) and mosaic errors (mitotic), since the proportion of chromosomes affected by mosaic segmental abnormalities is dependent on the presence of a whole chromosome aneuploidy in the same embryo.

Taken together, neither the incidence of uniform or mosaic segmental events are positively correlated with maternal age, as seen for whole chromosome aneuploidies. Therefore, these events are likely to result from different mechanisms. The negative correlation between mosaic segmental events and maternal age requires further investigation, considering the low number of chromosomes screened for the 24-29 years group compared to the 30-34 and 35-39 years groups.

3.3.4 Segmental aneuploidies frequently affect an entire chromosome arm.

Segmental aneuploidies are age-independent, they occur at a lower rate, and have different chromosomal distribution than whole chromosome aneuploidies; suggesting that they arise from distinct mechanisms. Therefore, I further characterised the segmental abnormalities by size, chromosome arm and copy number variation (gain or loss). The fragment size ranged from 10 to 150 Mb (Figure 3.3A), and 64 out of 355 (18.0%) of events affected a whole chromosome arm (Figure 3.3B). The enrichment for segmental events affecting an entire arm is suggestive of (peri)centromeric instability. To this end, I expected to observe equal frequencies of gains and losses in chromosome arms. Interestingly, whilst this was true for p-arm segmental events, a 2-fold higher proportion of losses to gains was observed for q-arm events (p < 0.001, exact test; Figure 3.3C). Stratification of segmental abnormalities by chromosome revealed that the bias towards q-arm losses was caused by specific chromosomes (including 1, 4, 6, and 9), where the frequency of q-arm losses was much higher than gains (Figure 3.3D).
Figure 3.3: Characterisation of uniform and mosaic segmental abnormalities in 2,083 embryo biopsies, stratified by size, chromosome arm affected (p or q) and copy number variation (gain or loss). (A) Size distribution of segmental aneuploidies. (B) Proportion of segmental events affecting a whole arm ((peri)centromeric break), with Standard Error of Proportion. (C) Frequency of segmental events affecting a whole arm ((peri)centromeric break), with Fisher’s Exact Test. (D) Frequency of segmental events by chromosome; acrocentric chromosomes (q arm only) are indicated under horizontal bars. Data drawn from experiment IDs: E32a_RCB_FinalClinicalAneuploidyDataset; E32d_RCB_StratificationOfSegmentalEventsBySize&Arm
Taken together, the data show a higher incidence of segmental events involving a loss in the q-arm than any other category; but across both chromosome arms, a similar proportion of events (~20%) show (peri)centromeric instability (entire-arm gain or loss). In addition to strong (peri)centromeric instability, the size distribution of segmental events is expected with the variability in chromosome size. Therefore, the majority of q-arm events that do not occur in the pericentromeric region, are losses. The skew in frequency of segmental events toward the larger chromosomes may be explained by the physical properties of the chromosome, such that the more material the higher the probability of suffering a break.

### 3.3.5 Independent validation of chromosomal abnormalities reveals the reported incidence of mosaic abnormalities is affected by technical limitations.

The PGT-A labs use a validated method for assessing uniform aneuploidies for diagnostics. Scoring low level mosaicism and small segmental (<10 Mb) abnormalities in PGT-A data are not part of the standardized workflow within BlueFuse® Multi software. It is therefore unclear how concordant the detection and classification of mosaic and segmental aneuploidies are. There are several challenges in detecting mosaic and segmental aneuploidies due to the NGS resolution and the noise floor associated with content analysis of only 3-10 trophectoderm cells. To validate the segmental abnormalities in PGT-A records from Lab D, we independently reviewed a subset of 522 records, comprising 24,012 chromosomes. The data from the NGS protocol (VeriSeq PGS; Illumina Inc.) were exported from BlueFuse® Multi software in order to ensure that the raw data analysed were the same. We used the Jaccard similarity coefficient to assess concordance between the two assessors. The Jaccard similar coefficient is a value between 0 and 1, reflecting the overlap between the two assessors.

Copy number data is shown in a consistent format throughout this thesis, where copy number values for each 1 Mb sequencing read bin, are plotted against their molecular position, by chromosome. The data was plotted using an independent software (JMP; SAS Institute Inc.) to preserve image quality and allow further annotation. Data points (DLR) are centred around the diploid copy number of 2.0 (Y-axis). Copy number variations are represented by a shift on the Y-axis value away from 2.0, where the number of bins also represents the size of the affected region. Reference lines at copy number 3.0 (red) and 1.0 (blue) represent a gain or loss of a single chromosome copy, assuming a diploid ground state. Additional dashed, green reference lines show the 20% deviations from copy number 2.0, 3.0 and 1.0, respectively. As discussed previously, this shows the thresholds recommended by the PGDIS and CoGEN consortia, for
categorising copy number changes as mosaic. Finally, I have used vertical reference lines (black) and alternating black and grey bin markers to visually separate the chromosomes.

Figure 3.4A shows the expected NGS profiles for whole chromosome and segmental abnormalities in a sample with a low noise floor. On the other hand, Figure 3.4B shows the typical NGS profile of a sample with high technical noise. Technical noise is characterised by high variation in bin-to-bin copy number values for a given chromosome or genomic region, that makes identifying biological CNVs challenging.

![Figure 3.4](image.png)

**Figure 3.4:** NGS trace of trophectoderm biopsy 44128_2, showing copy number variations with low technical noise following PicoPLEX WGA. Data re-drawn from exported BlueFuse® Multi records from Lab D. The trace shows examples of whole chromosome gain (Chromosomes 4 & 7), uniform segmental loss (Chromosome 8; q-arm) and a 50% whole chromosome mosaic gain (Chromosome 19) (red arrows).

### 3.3.5.1 Uniform segmental abnormalities

First, I assessed uniform segmental abnormalities, where the copy number variation was an integer between 0 and 4 according to the PGDIS and CoGEN guidelines. After inspection of the 522 reports, we detected 65 of the 78 (83.3%) uniform segmental abnormalities initially reported in the PGT-A reports. We identified an additional five uniform segmental abnormalities amongst the 522 samples, giving a Jaccard similarity coefficient of 0.78. Of the five uniform, segmental abnormalities detected by me, three were classified differently by the PGT-A lab (e.g. Figure 3.5A) and two were not detected at all (e.g. Figure 3.5B). Of the 13 uniform, segmental aneuploidies called by the PGT-A lab, nine were not scored by us due to high background noise in the NGS data (Figure 3.5D), three were classified differently and one was not called (Fig. 3.5C). These results show the challenges in having non-standardised scoring of abnormalities in a clinical setting.
Figure 3.5: NGS traces of trophectoderm biopsies with discordant uniform segmental copy number variations, following PicoPLEX WGA, between PGT-A Lab D and an independent review. (A) In sample 40669_6, chromosome 13 was scored as a whole chromosome aneuploidy by Lab D, but scored as uniform segmental with interstitial breakpoints by me (red ellipse). (B) Sample 43081_9, a uniform segmental event for the p-arm of chromosome 4 (red ellipse) was not reported by Lab D but was scored by me. (C) Sample 44696_3, Lab D reported a uniform segmental event in the p-arm of chromosome 6 (red ellipse), that was not included by me because the breakpoint is not clearly defined and could be an artefact of technical noise. (D) Sample 48537_E1, contains six discordant uniform segmental events (Chromosomes 1, 4, 6, 13, 14 and 15) that were reported by Lab D but were not included in my analysis due to the high level of noise throughout the sample (red ellipses). Data re-drawn from exported BlueFuse® Multi records.

3.3.5.2 Mosaic segmental abnormalities

The similar correlation for mosaic segmental abnormalities was significantly lower than uniform segmentals (0.66). Both assessors detected a common group of 81 mosaic segmental abnormalities. In addition, the PGT-A lab scored another four events, whereas I detected an additional 37 mosaic, segmental aneuploidies. Of the four events scored as mosaic, segmental aneuploidies by the PGT-A lab but excluded by me, three events were removed entirely due to high background noise across the samples as seen in Figure 3.5C and Figure 3.5D; and one was classified differently by me (Figure 3.6A).
The major difference causing the low similarity coefficient between two assessments, however, originated from the 37 events classified as mosaic, segmental aneuploidies by me but not by the PGT-A lab. Five of these were classified differently, including two that were erroneously called as whole chromosome mosaic (e.g. Figure 3.6B) and three there were classified by the PGT-A lab as uniform segmental (e.g Figure 3.6C). Thirty-two were not detected by the PGT-A lab (e.g. Figure 3.6D). These observations are consistent with the notion that the non-standardised detection of unusual events, which is not part of the workflow in PGT-A labs, cause under-reporting of mosaic, segmental aneuploidies.

Figure 3.6: NGS traces of trophectoderm biopsies with discordant mosaic segmental copy number variations, following PicoPLEX WGA, between PGT-A Lab D and an independent review. (A) Sample 47959_6, chromosome 12 was scored as segmental, mosaic by PGT-A Lab D, and whole chromosome mosaic by me (red ellipse). (B) Sample 47586_4GA, chromosome 4, was scored as whole chromosome mosaic by PGT-A Lab D, and segmental mosaic by me (red ellipse). (C) Sample 45815_6, chromosome 2 was scored as uniform segmental by PGT-A Lab D, and mosaic segmental by me (red ellipse). (D) Sample 44242_7 containing mosaic, segmental CNVs for chromosomes 5p and 16p (red ellipses) that were not reported by PGT-A Lab D. Data re-drawn from exported BlueFuse® Multi records.
3.3.5.3 Detection of uniform, whole chromosome abnormalities is affected by background noise.

We were concerned about the low similarity coefficient, especially for the mosaic, segmental aneuploidies (Section 3.3.5.2). Therefore, as a control, we decided to assess uniform, whole chromosome aneuploidies that the PGT-A workflow is designed to detect for clinical diagnosis. Therefore, one would expect a high similarity in the detection of uniform, whole chromosome aneuploidies.

To this end, I examined samples selected from 20 random patients (first 20 patient IDs). This comprised 101 biopsy records, of which 30 were reported to contain whole chromosome aneuploidies. The PGT-A lab reported 59 uniform, whole chromosome aneuploidies of which I detected 51. Of the eight not detected by me, one was classified differently (Figure 3.7A) and seven were excluded due to noise (e.g. Figure 3.7B). These observations suggest that noise in the background trace causes significant discordance between assessors for uniform, whole chromosome aneuploidies. Clinically, this may cause euploid embryos to be classified as aneuploid and therefore given a lower preference for transfer.
Figure 3.7: NGS traces of trophectoderm biopsies with discordant whole chromosome aneuploidies, following PicoPLEX WGA, between PGT-A Lab D and an independent review. (A) Sample 44233_3, chromosome 6 was scored by Lab D as uniform, whole chromosome aneuploidy, but scored by me as uniform segmental due to putative interstitial breakpoints (red ellipses). (B) Sample 43404_18, Lab D reported uniform, whole chromosome aneuploidies for chromosomes 3 and 6 (gain), and 17 (loss) (red arrows), that were not scored by me due to the high level of noise throughout the sample. Data re-drawn from exported BlueFuse® Multi records.
3.3.5.4 Lower similarity coefficient for detection of mosaic, whole chromosome abnormalities

We next examined whether noise also affected the similarity in detection of mosaic, whole chromosome abnormalities. We used the same 101 samples in which we assessed similarity of detection of uniform, whole chromosome aneuploidies. The PGT-A lab reported 27 mosaic, whole chromosome aneuploidies. My independent assessment detected only 15 of these as well as three additional events that were not reported by the PGT-A lab. The similar coefficient was 0.50 (Jaccard). Of the three events that I detected, one was at the edge of the noise floor (Figure 3.8A), whereas the other two clearly exceed the PGDIS 20% CNV detection guidelines (Figure 3.8B). Of the 12 mosaic, whole chromosome aneuploidies detected by the PGT-A lab, but not scored by me, 11 were excluded due to a high noise resulting in an uneven diploid baseline (e.g. Figure 3.8C) and one was unconfirmed in my assessment. These observations reveal two major issues in the interpretation of NGS data. First, biological noise or noise from the whole genome amplification causes discordance in scoring, even for uniform, whole chromosome aneuploidies for which this NGS pipeline was designed. Second, abnormalities that are not part of the standard analysis pipeline show poor similarity coefficients as the clinical genomics assessors are not specifically focussed on detecting such events (i.e. mosaic, segmental aneuploidies).
Figure 3.8: NGS traces of trophectoderm biopsies with discordant mosaic, whole chromosome aneuploidies, following PicoPLEX WGA, between PGT-A Lab D and an independent review. (A) Sample 44679_3, chromosome 12 was not reported by Lab D but was scored by independent review, as all data points fall outside the suggested 20% calling limit for mosaicism (green, dashed line at copy number 1.8) (red ellipse). (B) Sample 41080_2 showing mosaic aneuploidies for chromosomes 7 and 8, that were not reported by Lab D but were scored by me. Both events fall outside the noise floor of copy number 1.8-2.2 (green, dashed lines). (C) Sample 44679_1, containing three putative mosaic, whole chromosome aneuploidies that were reported by Lab D (red arrows), that were not included by me due to the elevated noise floor throughout the sample.
3.3.5.5  Removal of samples with a high noise floor improves similarity coefficients.

Our observations suggest that noise in the NGS traces cause significant problems in the assessment and inference of the chromosome constitution, leading to a high degree of discordance between different assessors (Table 3.3). Therefore, I reassessed the similarity correlation after removal of samples with uneven NGS traces across all chromosomes (high noise) as seen in Figure 3.5D, Figure 3.7B and Figure 3.8C (Table 3.4).

**Table 3.3:** Independent analysis of chromosomal abnormalities reported by Lab D.

<table>
<thead>
<tr>
<th>Abnormality type</th>
<th>Events reported in PGT-A records by Lab D</th>
<th>Additional events reported by independent review</th>
<th>Total events reported</th>
<th>Concordant events</th>
<th>Events removed following independent review</th>
<th>Discordant (%)</th>
<th>Jaccard score (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform Segmental</td>
<td>78</td>
<td>5</td>
<td>83</td>
<td>65</td>
<td>13</td>
<td>13/83 (15.7)</td>
<td>0.78</td>
</tr>
<tr>
<td>Mosaic Segmental</td>
<td>85</td>
<td>37</td>
<td>122</td>
<td>81</td>
<td>4</td>
<td>4/122 (3.3)</td>
<td>0.66</td>
</tr>
<tr>
<td>Mosaic Whole Chromosome</td>
<td>27</td>
<td>3</td>
<td>30</td>
<td>15</td>
<td>12</td>
<td>12/30 (40.0)</td>
<td>0.50</td>
</tr>
<tr>
<td>Whole Chromosome</td>
<td>59</td>
<td>0</td>
<td>59</td>
<td>51</td>
<td>8</td>
<td>8/59 (13.6)</td>
<td>0.86</td>
</tr>
</tbody>
</table>

**Table 3.4:** Independent analysis of chromosomal abnormalities reported by Lab D, whilst excluding samples that were removed during independent review due to uneven NGS traces.

<table>
<thead>
<tr>
<th>Abnormality type</th>
<th>Events reported in PGT-A records by Lab D</th>
<th>Additional events reported by independent review</th>
<th>Total events reported</th>
<th>Concordant events</th>
<th>Events removed following independent review</th>
<th>Discordant (%)</th>
<th>Jaccard score (noise removed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform Segmental</td>
<td>69</td>
<td>5</td>
<td>74</td>
<td>65</td>
<td>4</td>
<td>4/74 (5.4)</td>
<td>0.88</td>
</tr>
<tr>
<td>Mosaic Segmental</td>
<td>83</td>
<td>37</td>
<td>120</td>
<td>81</td>
<td>2</td>
<td>2/120 (1.7)</td>
<td>0.68</td>
</tr>
<tr>
<td>Mosaic Whole Chromosome</td>
<td>16</td>
<td>3</td>
<td>19</td>
<td>15</td>
<td>1</td>
<td>1/19 (5.3)</td>
<td>0.79</td>
</tr>
<tr>
<td>Whole Chromosome</td>
<td>52</td>
<td>0</td>
<td>52</td>
<td>51</td>
<td>1</td>
<td>1/52 (1.9)</td>
<td>0.98</td>
</tr>
</tbody>
</table>

3.3.5.6  Summary

Taken together, discordance rates are higher than expected, with the lowest similarity coefficients for mosaic abnormalities (Table 3.3). However, 29 out of 37 (78%) discordant events were attributed to high noise within the NGS traces, and removing these samples improved Jaccard scores (Table 3.4). In the NGS workflow, noise is likely to be introduced by the WGA step, and may be exacerbated by sub-optimal biopsy or sample handling, however this requires further work to elucidate. The high proportion of additional mosaic segmental events that were scored by independent review (37 out of 120; 31%) suggests a higher calling stringency at Lab D and highlights the challenge in distinguishing between technical noise and true events based on
a single biopsy. Guidelines from two international consortia suggest a noise threshold of 20% CNV deviation from an expected copy number of 2.0, for calling mosaic abnormalities (PGDIS, 2016; CoGEN, 2018). In practice, the same expectation cannot be applied to all samples, since the number of cells collected in a single biopsy may vary from ~3-10 cells, and the process of WGA and sequencing from such limited material can be challenging to replicate.

The data is confounded by the lack of an outcome control, or paired embryo biopsy to provide supporting information to the validity of the PGT-A report. Without such data the discordance observed here supports the notion that standardised reporting procedures should be applied when interpreting CNV sequencing traces, particularly for excluding samples that are difficult to interpret. Furthermore, intra-observer variability could be avoided by implementing an automated aneuploidy calling algorithm to filter results. Recently, the use of machine learning to identify mosaic and segmental aneuploidies in NGS data has been reported (CooperSurgical Inc, 2018).

3.3.6 Embryos affected by only mosaic or segmental abnormalities are present in cycles with a maternal age of 30-44 years.

The clinical significance of a mosaic PGT-A result is currently being debated. Transfer of embryos with apparent autosomal mosaic-monosomy have been shown to give rise to healthy babies (Greco and Fiorentino, 2015; Munné et al., 2017), which raises the question as to whether mosaic segmental events should be treated similarly. However, such embryos are given lower priority for transfer and their clinical use depends upon how many euploid embryos are present in the cycle. I observed a maternal age effect for the proportion of euploid embryos per cycle, in which the total number of embryos per cycle also decreased with age (Age 24-44 years: \( R^2 = 0.63 \)) (Figure 3.9A). This finding is further summarised by Figure 3.9B, where the average number embryos per cycle decreased by one (24-29 years: 4.47 ± 2.75 S.D; 40-44 years: 3.52 ± 3.23 S.D), but the average number of euploid embryos was 2-fold lower in the 40-44 years group (0.90 ± 1.73 S.D) than the 24-29 years group (2.47 ± 1.68 S.D). Taken together, I expect the incidence of having no euploid embryos, but having at least one embryo with only mosaic or segmental abnormalities, to be highest in cycles of higher maternal age (40-44 years); where both the total number of embryos per cycle, and the proportion of euploid results, is lower.

To test this, I first analysed the abnormality rate per cycle, and observed an increased proportion of cycles without any euploid embryos as a function of maternal age. In the youngest maternal age group of 24-29 years, 2 out of 47 cycles (2.4% ± 1.5% S.E.P) had no euploid
Figure 3.9: Incidence of reported embryo abnormalities per cycle, as a function of maternal age. (A) Frequency of euploid embryos per cycle total. (B) Average number of embryos tested (grey bars) and average number of euploid results (green bars) per cycle. Error bars show Standard Deviation. (C) Proportion of cycles with 0, 1, 2 or >2 embryos returning a euploid result. (D) Proportion of cycles in which all embryos returned a euploid PGT-A result (number in parentheses), out of the number of cycles in which ≥2 embryos were tested per cycle (n in legend). (E) Proportion of cycles that contained no euploid embryos but included embryo(s) with only whole chromosome mosaic, or segmental abnormality types. Combined data for the three age groups are shown to the right of the vertical separation line. The number of cycles is denoted by "n". For figures C, D and E, error bars show Standard Error of Proportion. Data drawn from E32a_RCB_FinalClinicalAneuploidyDataset, E32i_RCB_Euploidpercycle and E32e_RCB_Segmental&MosaicImpactPerCycle.
embryos. The remaining embryos all contained whole chromosome aneuploidy, mosaic or segmental chromosome abnormalities. These women may be of elevated risk of aneuploidy, however both patients only had a single embryo tested by PGT-A. It is unknown whether these patients had untested embryos, or if the oocyte retrieval and fertilisation rates were also reduced. For women aged 30-34 years, 20 out of 156 cycles (12.8% ± 1.3% S.E.P) had no euploid embryos; for women aged 35-39 years, 57 out of 215 cycles (26.5% ± 1.5% S.E.P) had no euploid embryos; and for women aged 40-44 years, 64 out of 114 cycles (56.1% ± 2.3% S.E.P) had no euploid embryos for transfer (Figure 3.9C; “0” category). To observe the number of euploid embryos per cycle, which could also be interpreted as the number of potential transfers per cycle, I also calculated the proportion of cycles in which 1, 2 or >2 embryos returned euploid PGT-A results (Figure 3.9B). For cycles with a single euploid embryo, the proportion of cycles was 28-33% across all maternal age groups. For cycles with 2 or >2 euploid embryos, increasing maternal age was associated with a lower proportion of euploid embryos. This is consistent with the notion that whole chromosome aneuploidy and a decreased total embryos per cycle with maternal age, impact the total number of euploid embryos identified with PGT-A.

Conversely, I calculated the proportion of cycles in which all embryos returned a euploid PGT-A result, provided that ≥2 embryos were tested (Figure 3.9D). Ensuring at least 2 embryos were tested, allows for an informed choice to be made regarding which embryo to transfer, or to transfer any embryos. For 24-29 years, 8 out of 43 cycles that had ≥2 embryos tested (18.6% ± 3.0% S.E.P) returned a euploid result for all embryos in the cycle. In contrast, for the 40-44 years group, only 2 out of 80 cycles that had ≥2 embryos tested (2.5% ± 0.9% S.E.P) returned a euploid PGT-A result for all embryos in the cycle. This supports the expected maternal age effect for aneuploidy, but importantly demonstrates that 81.4% of cycles with ≥2 embryos tested at 24-29 years, contain at least one embryo with a chromosomal abnormality. This is important because the average number of tested embryos per cycle for a maternal age of 24-29 years is 4.47 ± 2.75 S.D (Figure 3.9B), where 43 out of those 47 (91.5%) cycles had ≥2 embryos analysed. Therefore the 81.4% statistic describes the majority of cycles in the 24-29 years group, and this percentage decreases with maternal age (Figure 3.9D). Taken together, this suggests a limited benefit to PGT-A testing of embryos from all women, rather than just those of advanced maternal age (>35 years); however, the chance of selecting a euploid embryo at random from cycles of 24-29 years is higher than that of increased maternal age. To this end, patients with cycles that do not contain euploid embryos, may be recommended for a further cycle of ovarian hyperstimulation or treatments involving donated oocytes. Alternatively, we can ask whether the proportion of embryos that are reported as abnormal but contain only a mosaic or segmental aneuploidies, are compatible with healthy birth or were otherwise misinterpreted
from technical noise during PGT-A. Transfer of such embryos, would require appropriate genetic counselling for the patients.

I further probed the data to investigate the incidence of embryos containing only mosaic or segmental abnormalities. In this study, 377 out of 2,083 (18.1% ± 0.4% S.E.P) biopsies contained only mosaic or segmental abnormalities and were classified as ‘unknown clinical significance for transfer’ (amber risk group; Figure 3.1A). When analysed separately, 180 (8.6% ± 0.3% S.E.P) samples had one or more whole chromosome mosaic aneuploidies only; and 155 (7.4% ± 0.3% S.E.P) samples had one or more uniform or mosaic segmental aneuploidies only, which we expect to occur independently of maternal age (Figures 3.2C; 3.2D). The incidence of these abnormalities only reaches clinical significance when no other embryos within the same cycle return a euploid result; or if the prior transfer of euploid embryos in the same cycle are unsuccessful. Only in these scenarios would an embryo with only a mosaic or segmental abnormality PGT-A result be considered for transfer.

To elucidate the clinical significance, I calculated the proportion of cycles with no euploid embryos, but at least one embryo with only mosaic or segmental abnormality types reported by PGT-A, as a function of maternal age (Figure 3.9E). In the two cycles in the 24-29 maternal age group that had no euploid embryos, only one embryo was tested per patient. Therefore, this group is not included in the analysis and likely resulted from the higher euploidy rate observed for the youngest women in this study. For the remaining groups I did not observe a maternal age effect for any abnormality type, however the frequency was low across all maternal ages (<12 instances per cycle classification) indicating a larger sample size may be required to establish accurate rates. When age groups were combined (30-44 years), the incidence of embryos with whole chromosome mosaic abnormalities only (14 out of 485; 2.9% ± 0.4% S.E.P) was 2-fold higher than segmental abnormalities (7 out of 485; 1.4% ± 0.3% S.E.P), as seen previously (Figure 3.2). The incidence of having no euploid embryos but at least one with any combination of mosaic or segmental abnormalities in the absence of whole chromosome aneuploidy was highest in the 40-44 years group at 9.6% ± 1.2% S.E.P (11 out of 114 cycles), and in combined data accounted for 23 out of 485 (4.7% ± 0.5% S.E.P) of cycles. Taken together, the incidence of having no euploid embryos but at least one with either a whole chromosome mosaic or segmental aneuploidy in any maternal age group is <5%, with segmental error types being the least frequent.

In conclusion, a small number of cycles may be affected by a clinical decision regarding the transfer of ‘amber’ embryos. However, it does not detract from the importance and life changing decisions that this statistic represents for 5% of patients. Conversely, it must also be
noted that this study is limited by the lack of clinical outcome data for embryos discussed here. There are guidelines, advising on the transfer of embryos that do not report as euploid following PGT-A (PGDIS, 2016; CoGEN, 2018). Following appropriate genetic counselling, only a proportion of embryos seen here with mosaic or segmental abnormalities would be considered for transfer, depending on the size of event, chromosome, and likelihood of congenital birth defect or miscarriage. Uniform segmental events may also be classified as ‘unsuitable for transfer’ considering the size of the event, and separation between the CNV and NGS noise floor. In all cases, the number of cycles in which an ‘amber’ embryo could be transferred is expected to be much lower than 5% and is ultimately limited by the resolution of the detection assay.

3.4 Study conclusions

I analysed clinical data from four laboratories using NGS-based PGT-A of trophectoderm biopsies, and characterised the reported incidence of chromosomal abnormalities in cycles from women aged 24-44 years. The incidence of whole chromosome aneuploidy, but not mosaic or segmental aberrations, are associated with the maternal age effect (Hassold and Hunt, 2001; Vanneste et al., 2009; Fransasiak et al., 2014; McCoy et al., 2015; Vera-rodríguez et al., 2016; Babariya et al., 2017; Vera-Rodriguez and Rubio, 2017). Our data are consistent with this expectation and the notion that embryo mosaicism and segmental aneuploidy are prevalent in preimplantation embryos (~10%), with segmental events occurring less frequently by an order of magnitude (Figure 3.2C; Figure 3.2D). The observation of the maternal age-affect for aneuploidy mirrors the expected chromosome distribution of whole chromosome aneuploidies, including enrichment for chromosomes 15, 16, 21 and 22 (Handyside et al., 2012; Fransasiak et al., 2014; Soler et al., 2017) (Figure 3.2A). However, the profiles for mosaic and segmental events suggest a distinct mechanism of origin (Rabinowitz et al., 2012) (Figure 3.2B). The quality of sperm used for assisted reproduction, and the incidence of segmental aneuploidy should also be carefully considered, particularly DNA degradation. However, in this study I did not have access to this information. We remain confident that chromosome breaks are present in embryos, as I later demonstrate that we observe them in oocytes.

The transition to NGS-based PGT-A means CNVs can be detected with higher dynamic range, and therefore the incidence of mosaic and segmental aneuploidies must be considered during embryo prioritisation for transfer. The significance of mosaic and segmental aneuploidy reports in the clinic is that ~5% of cycles may be affected, where no euploid embryos are identified. According to current practice, these embryos with only mosaic abnormalities may be
considered for transfer, with appropriate genetic council, although with a reduced chance of implantation (Greco and Fiorentino, 2015; Munné et al., 2017).

In addition, segmental aneuploidies frequently affect whole chromosome arms (18%) (Figure 3.3B); and single breakpoints result in complete or partial q-arm loss twice as frequently than any other type (p<0.001, exact test; Figure 3.3C). The bias against segmental gains was caused by specific chromosomes such as 1, 4, 6, and 9 (Figure 3.3D). Taken together this suggests that segmental aneuploidies frequently occur in human preimplantation blastocysts and that (peri)centromeric instability may be a common cause. Uniform and mosaic segmental aneuploidies were represented equally, suggesting around half occur either very early in mitosis, or originate from the germline.
Chapter 4: Gross chromosomal rearrangements in human oocytes

4.1 Overview

The incidence of uniform segmental aneuploidy in human preimplantation embryo trophectoderm biopsies may be indicative of meiotic or early mitotic chromosome transmission errors. Segmental chromosome abnormalities occur independently of maternal age and are likely to originate from a distinct mechanism to whole chromosome aneuploidy. In this chapter the aim was to screen human oocytes and their matching polar bodies, to characterise the type and incidence of gross chromosomal rearrangements in oocytes from unstimulated women.

Single cell genomics is challenging due to the requirement for whole genome amplification (WGA), to amplify a single genome copy (~7 pg) to the mass needed for microarray analysis (200-400 ng) or NGS (1-100 ng depending on the application). Simple copy number analysis by low-pass NGS is more sensitive to stochastic amplification bias, because chromosome copy number in inferred relative to the remaining chromosomes within the same sample. Therefore, a quasi-linear WGA method is preferred (e.g. PicoPLEX, Takara Bio; or MALBAC, Yikon Genomics), at the expense of product length and thus genome coverage. Alternatively, genotype analysis by SNP microarray requires high genome coverage to minimise the incidence of allele drop out (ADO), with high fidelity. The preferred WGA method for SNP coverage is multiple displacement amplification (MDA) (e.g. SureMDA, Illumina; or REPLI-g SC, QIAGEN), which has been independently clinically validated for Karyomapping (Natesan et al., 2014). MDA is less suitable for CNV analysis because of the stochastic amplification bias that results in typically noisier CNV traces than linear WGA methods. However, study of gross chromosomal rearrangements in female meiosis would benefit from concurrent SNP and copy number analysis. SNP analysis allows inferences to be made about chromosome segregation patterns in meiotic duos (MII-arrested oocyte + PB1) and trios (MII oocyte + PB1 + PB2), because all four chromatids are captured for each chromosome. On the other hand, segmental aneuploidies cannot be detected by SNP arrays due to the limited number of included features, and therefore an NGS workflow is preferred. In this work, I have focussed on segmental events because they are poorly understood in oocytes and preimplantation embryos. In addition, there is less analysis on the maternal-age factor, since segmental events are believed to occur independently of whole chromosome aneuploidy.

Since WGA is destructive of the single cell, I propose to analyse oocytes using both technologies from MDA products. The expected noise profiles produced following MDA and NGS will be mitigated by the fact that collection of oocytes and matching polar bodies allows
reciprocal chromosome aberrations to be observed across duos/trios. However, there is a requirement to optimise a single-cell NGS pipeline for this purpose.

In this chapter, I have developed an MDA-NGS pipeline that has been validated using four independent cell lines with known whole chromosome and segmental abnormalities. I then apply the pipeline, with concurrent SNP analysis to 159 oocytes with matching polar bodies to assess segmental aneuploidies in human female meiosis.

This work constitutes my contribution to the following manuscripts, where we describe the current landscape of gross chromosomal rearrangements in human female meiosis.

- *Contributed equally.

4.2 Materials and Methods

4.2.1 Generating an MDA-specific reference database for NGS

For creation of a reference database of euploid samples for CNV analysis, male gDNA (Promega) was quantified for dsDNA using the fluorometric-based Quant-iT™ method described previously. Stock gDNA was diluted to 12.5 pg/µl in 1×PBS (provided in the SureMDA kit), in a three-step serial dilution (Stock → 2000 pg/µl → 200 pg/µl → 12.5 pg/µl). Ninety-six technical replicates were amplified using the SureMDA Amplification System (Illumina Inc., CA, USA) as described previously, where 4 µl of diluted gDNA provided an input of 50 pg dsDNA.

Sequencing libraries were prepared from SureMDA products using a custom protocol and sequenced as described in Chapter 2 (Section 2.10.1). Reads were aligned to the human genome.
genome hg19 using Burrows-Wheeler Aligner (BWA) within the MiSeq Reporter software. The resulting *.bam files were processed with bioinformatics support from Illumina Inc., to generate a custom reference database that was compatible with BlueFuse® Multi software for CNV analysis. In brief, the variation in chromosome medians for copy number per alignment bin were calculated, to ask whether specific regions of the genome are systematically called higher or lower copy number from the expected diploid state. The reference database was curated from the 96 technical replicates, to capture reproducible assay bias such as WGA noise or GC content. The reference database is therefore used to offset the bin counts of samples that are analysed against the reference. Aneuploidy calling in single cells for validation, and oocyte/polar body samples was performed as previously described in Chapter 2, using BlueFuse® Multi (Illumina Inc.) and the custom reference database described here.

4.2.2 NGS pipeline validation

4.2.2.1 Source of single cells and gDNA of known karyotype

The MDA-NGS pipeline for CNV detection in single cells was performed using cell lines obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, New Jersey, USA. Cell lines were cultured, and single cells were manually isolated as previously described; cells isolates were stored at 4 °C until WGA, performed on the same day. A summary of cell lines and samples collected are shown in Table 4.1. Refer to Table 2.3 for full karyotypes using the international system for human cytogenetic nomenclature (ISCN).

<table>
<thead>
<tr>
<th>Cell Line ID</th>
<th>Karyotype summary</th>
<th>Sample IDs</th>
<th>Isolation control ID</th>
<th>Amplification control ID</th>
<th>Date of isolation and MDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM00526</td>
<td>47, XY +13 114 Mb gain</td>
<td>V-15 to V-28 (n=14)</td>
<td>3C-10 BC-10</td>
<td>Mg⁺-3 NTC-6</td>
<td>24 Jan 2018</td>
</tr>
<tr>
<td>GM50121</td>
<td>47, XY Chr: 18 15.5 Mb p arm loss 59.3 Mb q arm gain</td>
<td>V-29 to V-42 (n=14)</td>
<td>3C-11 BC-11</td>
<td>Mg⁺-4 NTC-7</td>
<td>24 Jan 2018</td>
</tr>
<tr>
<td>GM04927</td>
<td>47, XY +21 47 Mb gain</td>
<td>V-43 to V-56 (n=14)</td>
<td>3C-12 BC-12</td>
<td>Mg⁺-5 NTC-8</td>
<td>25 Jan 2018</td>
</tr>
<tr>
<td>GM10985</td>
<td>46, XX Chr:3 10.3 Mb p arm loss</td>
<td>V-57 to V-70 (n=14)</td>
<td>3C-13 BC-13</td>
<td>Mg⁺-6 NTC-9</td>
<td>25 Jan 2018</td>
</tr>
</tbody>
</table>

*Chr* = Chromosome; *⁺* = Whole chromosome gain.
As a positive NGS control for each cell line in Table 4.1, gDNA was extracted from liquid culture using the QIAamp DNA Mini Kit (QIAGEN), and quality assessed as previously described.

4.2.2.2 Whole-genome amplification, library preparation and sequencing

Twelve single cell isolates and 12 gDNA technical replicates per cell line were whole genome amplified using the SureMDA DNA Amplification System as previously described. Genomic DNA was amplified immediately after dilution to 12.5 pg/µl in 1×PBS. SureMDA products were quality assessed by gel electrophoresis and quantitation of dsDNA using the fluorometric method described previously, prior to NGS library preparation. Only cells showing positive amplification were included for NGS.

Sequencing libraries from single cell and gDNA SureMDA products were prepared according to the custom protocol described previously, as used for creation of the reference database. Libraries were pooled for 24-plex sequencing on a MiSeq System (2×36 bp), and *.bam files were extracted for analysis using BlueFuse® Multi software with a custom, MDA-specific reference database.

4.2.3 Oocyte duos and trios

4.2.3.1 Sample collection

Human oocytes for use in this study were collected following ovarian stimulation, or from retrieval of small antral follicles directly from the unstimulated ovary (Donnez et al., 2013; Yin et al., 2016). In total, 41 oocytes were donated by 23 patients undergoing in vitro fertilisation (IVF) at the GENERA center for Reproductive Medicine, Rome, Italy with fully informed consent. Oocytes were collected, cultured, activated and biopsied according to the protocol described previously (Chapter 2) (Ottolini et al., 2016). Briefly, retrieved oocytes that had not completed meiosis I were matured in vitro until the PB1 was extruded. Oocytes that were arrested at metaphase II, had the PB1 removed prior to artificial activation with 100 µM calcium ionophore. Oocytes that completed meiosis II and extruded the PB2 were biopsied to yield oocyte-PB trios (PB1, PB2 and oocyte; tubed separately).

In addition, 118 oocytes were obtained from in vitro maturation (IVM) of germinal vesicle (GV)-stage oocytes, donated by 19 patients undergoing fertility preservation therapies at Juliane Marie Centre, Rigshospitalet, Copenhagen, Denmark with fully informed consent. Oocytes were matured in vitro, and those that completed meiosis I with extrusion of the PB1
were biopsied according to the protocol of Gruhn et al. (2018). A subset of 17 MII-arrested oocytes from 4 patients, were activated with calcium ionophore following PB1 biopsy, as described previously. Oocytes that then extruded the second polar body (PB2) were biopsied to yield oocyte-PB trios.

Oocytes that were not treated with calcium ionophore or failed to activate (no extrusion of the PB2), were classified as MI Duos (oocyte and PB1). Oocytes that were activated and extruded the PB2, but where one of the three cells (PB1, PB2 or oocyte) was lost during biopsy, were classified as MII Duos. Cells with no matching counterpart of individual meioses were classified as singles. A full stratification of samples is shown in Table 4.2. In total, 116 meiosis II-arrested oocytes and their matching PB1s (MI duos) were collected. An additional, 15 oocytes with both matching polar bodies (oocyte-PB trios) and 2 MII duos were collected following artificial activation of the oocyte with calcium ionophore (triggering the resumption and completion of meiosis II). Polar bodies were biopsied sequentially upon extrusion to differentiate between PB1 and PB2 samples. Following sample loss, degeneration or fragmentation during biopsy, 11 individual PB1s and 15 individual oocytes were also collected. All single cells were amplified using the SureMDA DNA Amplification System at the source laboratory and quality assessed by gel electrophoresis and quantitation of dsDNA using the Qubit dsDNA HS Assay Kit (Part: Q32851; Thermo Fisher Scientific Inc., MA, USA), prior to shipment of amplification products under dry ice for NGS and SNP microarray analysis at Illumina Ltd. (UK).

Table 4.2: Oocyte and polar body sample stratification for MDA-NGS pipeline.

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Clinic A</th>
<th>Clinic B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>23</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>Oocytes retrieved</td>
<td>41</td>
<td>118</td>
<td>159</td>
</tr>
<tr>
<td>Oocytes treated with calcium ionophore (Patients)</td>
<td>41 (23)</td>
<td>17 (4)</td>
<td>58 (27)</td>
</tr>
<tr>
<td>Activation rate (%)</td>
<td>16/41 (39.02)</td>
<td>2/17 (11.76)</td>
<td>18/58 (31.03)</td>
</tr>
<tr>
<td>MI duos</td>
<td>25</td>
<td>90</td>
<td>115</td>
</tr>
<tr>
<td>MII duos</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MII trios</td>
<td>13</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Individual PB1</td>
<td>0</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Individual oocyte</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Samples processed</td>
<td>95</td>
<td>212</td>
<td>307</td>
</tr>
</tbody>
</table>

4.2.3.2 SNP genotyping

For quality control of MDA yield and genome coverage of oocyte-PB duos/trios, MDA products were analysed by SNP microarray using the Infinium™ HumanKaryomap-12 DNA Analysis Kit as previously described. MDA products were not diluted prior to input. SNP genotype data was
used for assessment of WGA performance and provided supporting data when inferring aneuploidy and heterozygosity in single cells.

4.2.3.3 Library preparation and sequencing

Oocyte and polar body sample dsDNA was re-quantified upon arrival at the Illumina laboratory using the fluorometric method described previously. Samples were diluted prior to NGS using the same dilution plate used for quantification, to ensure the most accurate input concentration was achieved. Libraries were prepared for NGS using the custom protocol as described for validation of the pipeline. Samples were processed in 96-well plate format, with a maximum of two plates processed during a single library preparation event. Sample plates were divided by row for library pooling in a maximum of 24-plex configurations. The sequencing metrics used for sample and library QC were, total reads, mapped reads and reads passing filter. ‘Total reads’ refers to the sum of sequencing reads per sample, that include the sample-specific index barcodes that are added during library preparation. In a paired-end sequencing run (i.e. 2×36) one read is sequenced from each end of the fragment and therefore the total number of reads is expected to be twice the value of clusters on the flow cell surface. ‘Mapped reads’ are those which align to the reference genome (GRCh37 - hg19) and can be expressed as a percentage of total reads. ‘Reads passing filter’ (PF), is the number of reads that pass assay specific filtering. In this case, reads are filtered by the BlueFuse® Multi software using bash scripting, BEDtools (Quinlan & Hall, 2010), and SAMtools (Li et al., 2009) to remove unmapped reads, duplicate reads, reads with low mapping (mismatch threshold). The QC criteria for reads PF using the single index strategy for DNA libraries prepared from SurePlex WGA products, are >60% total reads or ≥250,000. Finally, the Derivative Log Ratio (DLR) metric is used to assess noise in the NGS genome coverage profile, displayed as a CNV chart. DLR measures the spread of the difference in copy number values, between all bins (aligned sequence read counts for a defined genomic region) of a chromosome, and is calculated within the BlueFuse® Multi software. The log ratio refers to the normalised ratio between the measured sample and the expected signal for each bin. The expected signal is defined by using a reference database. As the bins are a fixed variable size with a median of 1 Mb, regions with poor sequence coverage (due to GC-content or repetitive regions) typically have larger bins. This means the differences in spread that would be expected to alter the DLR score, are absorbed by the increased bin size. For DNA libraries prepared from SurePlex WGA products, DLR values >0.4 indicate low quality sample DNA or problems in amplification. These metrics are used as a guide to assess the performance of the MDA-NGS assay.
4.2.4 Exclusion criteria

All oocyte and polar body samples were quantified for dsDNA using the fluorometric-based method described previously and processed by SNP microarray. Thirty-eight samples were excluded following SNP QC, which involved assessment of internal Infinium™ BeadChip array controls in GenomeStudio software, and genotype call rates. Sample-dependent failures were identified by <5000 SNP calls in the green channel (G/C bases; B-genotype). This failure mode is characteristic of low sample input, where the genotype calling occurs at the noise floor, giving rise to a total call rate >60% (QC criteria for Karyomapping), but with a bias towards the red channel (A/T bases; A-genotype). A further 16 samples were excluded from NGS library preparation due to low MDA yield (<3 µg mass in 50 µl), and which also failed the Infinium™ assay sample-dependent controls. This is because poor library preparation results in an abundance of short ‘junk’ adapter sequences. During NGS, shorter fragments preferentially bind to the flow cell surface when competing with longer fragments generated from high quality libraries in the same library pool. The result is reduced read yield from high quality libraries.

Following NGS, 37 out of 253 (14.62% ± 1.11% S.E.P) samples were excluded from CNV analysis due to noise in NGS traces. Manual calling of CNVs is limited when variation between normalised copy number values (noise) for each alignment bin is high; either within or between chromosomes. High noise reduces the confidence of the operator when inferring CNVs, and in these cases, samples were excluded from analysis as ‘inconclusive for aneuploidy’. To this end, CNV charts were reviewed by two scorers, and CNVs were scored according to the PGDIS and COGEN international consortia guidelines for uniform aneuploidy (PGDIS, 2016; CoGEN, 2018) (only aneuploidies with >80% copy number change were reported). Agreement between scorers was reached for all sample exclusions. Finally, 56 out of 216 (25.93% ± 1.49% S.E.P) samples that passed NGS QC, had no included matching cell(s) of their respective meiotic division(s). These samples were excluded from CNV analysis, where reciprocal evidence of aneuploidy could not be obtained, i.e. where a chromosome gain in the MII oocyte is matched with a corresponding loss in the PB1.

4.2.4.1 Abnormality scoring and database curation

Chromosomal abnormalities in single cells from cell lines, and oocyte-PB duos/trios were manually scored by two operators, though observation of CNV charts generated by BlueFuse® Multi software. CNV charts presented here have been re-drawn using alternate software (JMP 13, SAS Institute, NC, USA) from exported BlueFuse® Multi data, to preserve image quality. For validation cell line samples, false positives and false negatives were counted using the expected
karyotype as a reference. For oocyte-PB samples, abnormalities were scored on a per chromosome basis and recorded in Microsoft Excel. B-allele frequencies and log intensity ratios in SNP microarray data were reviewed for evidence supporting chromosome imbalance. In this study, I focussed only on the incidence of segmental abnormalities in oocyte-PB duos and trios in which reciprocal were observed between two cells.

4.3 Results

4.3.1 Reference database curation and pipeline validation using cell lines

4.3.1.1 MDA is highly efficient for WGA of gDNA and single cells isolated from cell lines

MDA products were resolved by gel electrophoresis, and representative images are shown in Figure 4.1A. All 96 male genomic DNA (Promega) MDA products showed high molecular weight DNA fragments ~5 Kb, in addition to three DNA-negative NTC samples from two MDA events. All 48 samples prepared from genomic DNA extracted from cell lines showed high molecular weight DNA fragments, in addition to two DNA-negative NTC samples from a single MDA event. All 56 single cell samples from four different cell lines showed high molecular weight DNA fragments, in addition to DNA-negative isolation buffer and NTC controls, and DNA-positive Male gDNA (Promega) and 3-cell controls for each isolation/MDA event.

MDA dsDNA yield for each sample type, including isolation and amplification controls were assessed by fluorometric quantitation (Figure 4.1B). The average yield (± S.D.) in 50 µl for male genomic DNA (Promega), over two amplification events was 23.0 ± 3.5. The average yield (µg ± S.D.) for genomic DNA extracted from cell lines was 30.2 ± 5.7. The average yield for single cells from four different cell lines over four MDA events was 30.4 ± 3.8. The average yield (µg ± S.D.) for isolation and amplification controls were: 3-cell isolation positive, 33.9 ± 2.8; isolation buffer negative, 0.3 ± 0.002; male gDNA amplification positive, 31.0 ± 2.9; NTC amplification negative, 0.3 ± 0.1.

4.3.1.2 Sequencing libraries can be prepared from MDA products of gDNA and single cells isolated from cell lines

Quality control of PCR-amplified libraries was performed by quantitation of duplex DNA by fluorometric quantitation (Figure 4.1C). One gDNA and single cell sample for each cell line (n=8) were excluded due to high post-PCR yield (~30 ng/µl). These samples correlated with column 2 library preparation plate #2, suggesting that samples in this column had a higher input of MDA product compared to the other samples. Of the remaining samples, the average concentrations
(ng/µl ± S.D.) based on sample type, over two 96-plex library preparation plates were: Male gDNA (Promega), 10.5 ± 3.4; gDNA extracted from cell lines, 11.2 ± 3.1; single cells, 11.0 ± 4.2.

Figure 4.1: Quality control of WGA and library preparation for genomic DNA and cell lines of known karyotype. (A) Resolution of MDA products by gel electrophoresis for Promega male gDNA (M74, M75, M76), gDNA extracted from cell line GM04927 (G13, G19, G14) and single cells isolated from cell line GM04927 (V43, V51, V44) against a 1 Kb extension ladder “L”. In this case, negative controls were included in fluorometric analysis shown in Figure 4.1B. (B) Duplex DNA mass (µg) in 50 µl MDA product, for all samples by type and cell line, inclusive of positive and negative SCI and WGA controls; Promega Ref. (n=96); Cell Line gDNA (n=12 per cell line); Cell Line SC, (n=14 per cell line). (C) Duplex DNA concentration (ng/µl) of cleaned-up, post-PCR sequencing libraries by type and cell line; Promega Ref. (n=96); Cell Line gDNA (n=11 per cell line); Cell Line SC (n=11 per cell line). SC = single cell, SCI = Single cell isolation, NTC = no template control. Data drawn from E24_RCB_MDA_NGS_RefDB&CellLineLP.
In addition, library fragment size was assessed for a cross-section of gDNA and single cell samples in the library preparation plate using the Agilent Bioanalyzer. The average fragment length for gDNA samples \((n = 6)\) was 374 bp, 43% CV. The average fragment length for single cell samples \((n = 5)\) was 373 bp, 42% CV; inclusive of index and adapter sequences for NGS (135 bp). For both sample types the highest fragment peak appeared at ~220 bp. Whilst concentrations were variable between samples, the fragment distribution profile was comparable between all samples.

Sequenced libraries prepared from pooled Male genomic DNA (Promega), gave an average \((\pm\text{ S.D})\) of \(2.12 \times 10^6 \pm 0.31 \times 10^6\) total reads per sample (Figure 4.2; Blue) of which 98.70% \(\pm 0.47\%\) aligned to the reference genome (GRCh37 - hg19) (Figure 4.2; Red), and 89.80% \(\pm 0.52\%\) passed the filtering criteria for CNV detection with BlueFuse® Multi software (Figure 4.2; Green), with an average DLR score of 0.22 \(\pm 0.02\) (Figure 4.2; Purple). Libraries prepared from gDNA extracted from cell lines were comparable to Promega samples, with an average \((\pm\text{ S.D})\) of \(2.32 \times 10^6 \pm 0.33 \times 10^6\) total reads per sample (Figure 4.2; Blue) of which 98.44% \(\pm 0.87\%\) aligned to the reference genome (Figure 4.2; Red), and 89.49% \(\pm 0.94\%\) passed the filtering criteria for CNV detection with BlueFuse® Multi software (Figure 4.2; Green), with an average DLR score of 0.35 \(\pm 0.10\) (Figure 4.2; Purple). Whilst these QC metrics pass the criteria for CNV analysis from SurePlex-amplified samples (VeriSeq® PGS; Illumina Inc., CA, USA), the data are not directly comparable with this study, which used MDA products instead of SurePlex, and paired-end sequencing instead of a single-index strategy (Illumina Inc., 2014).

Sequenced libraries prepared from single cell isolates from cell lines, gave an average \((\pm\text{ S.D})\) of \(2.14 \times 10^6 \pm 0.34 \times 10^6\) total reads per sample (Figure 4.2; Blue) of which 98.46% \(\pm 0.93\%\) aligned to the reference genome (Figure 4.2; Red), and 89.94% \(\pm 0.84\%\) passed the filtering criteria for CNV detection with BlueFuse® Multi software (Figure 4.2; Green), with an average DLR score of 0.52 \(\pm 0.10\) (Figure 4.2; Purple). These metrics, with the exception of DLR, meet the criteria expected for CNV detection from SurePlex products, when using a single-index sequencing strategy. The elevated DLR value (>0.4) its likely due to the non-reproducible amplification bias profile inherent to MDA. In addition, a bimodal distribution was seen in the percent of reads mapping to the human genome (Figure 4.2; Red), and passing filter (Figure 4.2; Green), for cell line GM04927 and GM50121. The lower values in these data groups all correlate to samples from a single sequencing run (L01) from library plate #2. As no other sample in the library plate were affected, the reduction in reads available for CNV analysis is attributable to a single sequencing run, that had a cluster density of 1814 K/mm² and exceeded the acceptance value for MiSeq v3 (>1600 K/mm²). In both cases, the increased DLR value for cell line gDNA from
Figure 4.2: Box plot for primary low-pass, next-generation sequencing metrics for bulk genomic DNA and single cell isolates from four cell lines. Promega male gDNA samples (n=96) were used to build a CNV reference database. Genomic DNA (n=12 per cell line) and single cell isolates (n=12 per cell line) were used to validate CNV calls. Reads mapping to GRCh37 - hg19 (Red) and passing BlueFuse Multi pipeline filters (Green) are shown as a proportion of total (paired-end) reads (Blue). The derivative log ratio (DLR) metric describes overall, per chromosome noise (Purple).

Data drawn from E24_RCB_MDA_NGS_RefDB&CellLineLP and E31a_RCB_AllLibrariesPrimaryOutput.
GM00526, and the reduction in qualified reads from cell line GM04927 and GM50121, did not affect the ability to detect large structural chromosome abnormalities. This is likely due to the paired-end sequencing strategy for libraries generated from MDA products, which doubles the total read count per sequencing run and improves alignment efficiency to the reference genome compared to the single-index workflow.

4.3.1.3 High CNV concordance between NGS output and cell line karyotype information

Twelve genomic DNA replicates and twelve single cells from four cell lines of known karyotype (Coriell Institute) were analysed for copy number variation. A summary of copy number analysis is shown in Table 4.3. The expected karyotype was observed in all 12 gDNA replicates for each cell line, and a representative example of each karyotype is shown in Figure 4.3. The cell line GM00526 karyotype (47, XY +13) is clearly visible as the largest CNV used for validation. Interestingly, genomic DNA extracted from GM00526 gave higher DLR values (0.53 ± 0.03) than gDNA samples from the other cell lines (GM04927, 0.31 ± 0.01; GM10985, 0.29 ± 0.01; GM50121, 0.28 ± 0.01). This difference was not observed in single cell samples (GM00526, 0.54 ± 0.07; GM04927, 0.55 ± 0.11; GM10985, 0.51 ± 0.11; GM50121, 0.48 ± 0.10), indicating a possible problem during dilution of the GM00526 gDNA MDA input during library preparation. However, the condition did not affect the ability to identify trisomy 13 in these samples (Sample G-31; Figure 4.3A). The GM04927 karyotype (47, XY +21) was also detected, which is important as chromosome 21 is the smallest chromosome, and therefore trisomy 21 represents the smallest whole chromosome aneuploidy in the genome (47 Mb) (Sample G-16; Figure 4.3B). The GM10985 karyotype (46, XX -3p) (Sample G-56; Figure 4.3C) and the GM50121 karyotype (46, XY -18p +18q) (Sample G-44; Figure 4.3D) show detection of sub chromosomal CNVs (loss) in gDNA samples of 10.3 Mb and 15.5 Mb, respectively. The data confirm the CNVs are present in the cell lines used for validation and serve as a positive control.

Table 4.3: Validation of CNVs in genomic DNA and single cells of known karyotype

<table>
<thead>
<tr>
<th>Cell Line ID</th>
<th>Sample Type</th>
<th>n</th>
<th>Expected CNV detected</th>
<th>False Negative</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM00526</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM04927</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM10985</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM50121</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM00526</td>
<td>Single cell</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GM04927</td>
<td>Single cell</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>2*</td>
</tr>
<tr>
<td>GM10985</td>
<td>Single cell</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GM50121</td>
<td>Single cell</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Segmental abnormalities that are known to persist in a sub population of this cell line.
Figure 4.3: CNV validation in MDA products from gDNA samples of known karyotype. The expected karyotype is also shown above each chart. (A) GM00526; 47, XY +13. (B) GM04927; 47, XY +21. (C) GM10985; 46, XX -3p. (D) GM50121, 46, XY -18p +18q. Red arrows highlight CNVs of interest. Data drawn from E31c_RGB_Herlev_BFM_Exports.
For single cells, 12 out of 12 GM04927 samples showed a gain of chromosome 21 (3C) (Sample V-48; Figure 4.4). Sample V-48 previously showed a reduced rate of mapped reads and reads passing filter (Figure 4.2; Red and Green) due to sequencing run variation. The ability to detect trisomy 21 was not affected.

**Figure 4.4:** CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM04927 (47, XY, +21). Copy number is derived from the proportion of sequence read counts, where the increase observed for chromosome 21 (red arrow in cut-out) is used to infer trisomy. The copy number (1.0) of sex chromosomes in this male donor serve as an internal control.

One out of 12 cells from GM04927 showed segmental aneuploidies for chromosome 1 (q-arm; 4C), and chromosome 16 (q-arm; 1C) (Sample V-47; Figure 4.5). Both segmental aneuploidies have been previously observed in this cell line through personal experience and have been confirmed by karyotype analysis at Coriell (personal correspondence). Taken together this further demonstrates the ability of this pipeline to detect sub-chromosomal abnormalities.
Figure 4.5: CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM04927 (47, XY, +21), with additional CNVs detected in chromosome 1 (+1q; left cut-out) and chromosome 16 (-16q; right cut-out) (red ellipses). The expected CNV for Trisomy 21 was also detected (red arrow in cut-out; right).

For cell line GM00526 (47, XY +13), 11 out of 12 single cells showed a gain of chromosome 13 (3C) (Sample V-23; Figure 4.6A), and one out of 12 cells was copy number neutral for chromosome 13 (2C) (Sample V-20; Figure 4.6B). Two separate cells each showed an additional segmental aneuploidy, affecting chromosome 11 (q arm; 1C) (Sample V-19; Figure 4.6C) and chromosome 20 (p arm; 1C) (Sample V-22; Figure 4.6D).
Figure 4.6: CNV validation following MDA whole genome amplification and NGS, in single cells from cell line GM00526 (47, XY, +13). (A) NGS trace cut-out for sample V-23, showing the expected autosomal copy number deviation for chromosome 13 (red arrow). (B) NGS trace cut-out for sample V-20, showing copy number neutral chromosome 13 (red arrow). (C) NGS trace cut-out for sample V-19, showing the expected autosomal copy number deviation for chromosome 13 (red arrow) and an additional segmental loss of chromosome 11q (1C) (red ellipse). (D) NGS trace cut-out for sample V-22, showing the expected autosomal copy number deviation for chromosome 13 (red arrow) and an additional segmental loss of chromosome 20p (1C) (red ellipse).

For cell line GM10985 (46, XX -3p), 9 out of 12 single cells showed a loss of a 10.3 Mb region in the p-arm of chromosome 3 (1C) (Sample V-57; Figure 4.7A), and 3 of 12 cells were copy number neutral for chromosome 1 (2C) (Sample V-58; Figure 4.7B). One other cell showed an additional segmental aneuploidy, affecting chromosome 6 (q-arm; 1C) (Sample V-62; Figure 4.7C), and seven drops in telomeric signal, consistent with that seen in clinical trophoderm biopsies (Figure 3.5C).
Figure 4.7: CNV validation following MDA whole genome amplification and NGS, in single cells from cell line GM10985 (46, XX, -3p). (A) NGS trace cut-out for sample V-57, showing the expected autosomal copy number deviation for chromosome 3 (red ellipse). (B) NGS trace cut-out for sample V-58, showing copy number neutral chromosome 3, where the expected CNV is not observed within the noise floor and not at copy number 1.0 (red arrow). (C) Complete NGS trace for sample V-62, showing the expected autosomal copy number deviation for chromosome 3 (red arrow), and an additional segmental loss of chromosome 6q (1C) (red ellipse). Seven drops in telomeric signal (blue arrows) were also observed (suspected noise). Sex chromosome copy number serves as an internal control.

All 12 single cells of cell line GM50121 showed the expected karyotype (46, XY -18p +18q) (Sample V-36; Figure 4.8), with no additional aneuploidy events observed. Sample V-36 previously showed a reduced rate of mapped reads and reads passing filter (Figure 4.2; Red and Green) due to sequencing run variation. The ability to detect sub-chromosomal abnormalities in chromosome 18 was not affected. In addition, all gDNA and single cell samples showed the expected copy number for both X and Y chromosomes according to the Coriell karyotype descriptions.
Figure 4.8: CNV validation following MDA whole genome amplification and NGS, in a single cell from cell line GM50121 (46, XY, -18p +18q). This NGS trace cut-out shows the expected loss (1C) of chromosome 18p (red ellipse), and expected gain (3C) of chromosome 18q (red arrow). The copy number (1.0) of sex chromosomes in this male donor serve as an internal control.

In summary, the expected karyotype was detected in 48 out of 48 gDNA samples, and no false positive abnormalities were observed. For single cell isolates, 4 out of 48 (8.3% ± 2.0% S.E.P) samples were discordant with the expected karyotype. Whole chromosome false negatives that were seen in samples that do not show excessive noise in CNV traces (e.g. Figure 4.6B), may be due to chromosome segregation events that ultimately restore the diploid state in daughter cells during cell culture. On the other hand, CNVs of small sub-chromosomal regions (<11 Mb) may be supressed during normalisation processes, resulting in false negative results. This may indicate the limit of detection for such events with this assay, however further replicates would be required to elucidate this. In addition, 5 chromosomes gave false positive results. However, two that occurred in sample V-47 are known to be present in a sub population within the cell lineage. The remaining 3 out of 2208 (0.14% ± 0.04% S.E.P) chromosomes were false positive for sub-chromosomal abnormalities with a full copy number change. It is likely these events are due to de novo chromosome breakages during cell culture, however we cannot exclude the possibility that these CNVs are due to technical artefacts or biological phenomena, such as S-phase replication in unsynchronised cell cultures.

Taken together, NGS-based CNV detection from single cells has been demonstrated in a limited number of samples from four different cell lines, with abnormalities ranging from 10.3 Mb to 114 Mb (Chromosome 13). The approach is limited by the non-reproducible amplification bias that is characteristic of MDA. However, its use for detection of chromosome abnormalities in single cell products of meiosis, is supported because we can observe reciprocal CNV events in matched cells. This adds confidence to CNV discovery; however, we may underestimate the
incidence of small sub-chromosomal aberrations due to masking by technical noise. To this end, CNV detection from MDA products by low-pass NGS is suitable for analysing structural chromosome changes in individual meioses.

4.3.2 Structural CNV detection by MDA and low-pass NGS identifies reciprocal segmental abnormalities in meiotic duos and trios and allows extraction of their breakpoints. All MDA products from single cell samples were quantified and hybridised to SNP microarrays to assess human genome coverage. In total, 38 out of 307 (12.38% ± 0.94% S.E.P) single cell samples were excluded following SNP array analysis. Thirty-five samples were excluded due to absence of genotype calls in the green channel (C/G bases; B-type). These samples had a total call rate >60%, but an imbalance in the proportion of AA and BB calls such that <5000 out of 293,869 features were called BB genotype by GenomeStudio software (GenCall threshold: 0.15). This is consistent with the negative control (sample-dependent failure), in which no sample was hybridised to the array prior to washing, staining and scanning (Figure 4.9A). Of these 35 samples, 33 had an MDA yield of <15 µg per 50 µl reaction (Figure 4.9A) with an average of 5.40 µg ± 5.37 S.D. The remaining two samples gave higher MDA yields of >20 µg, however also showed the same failure mode due to BB call rate. These two samples were prepared for NGS but subsequently failed to give interpretable CNV results; due to either inaccurate quantification of the MDA product, or non-human contamination during WGA. An additional three samples were excluded following SNP array QC due to suspected human DNA contamination, with call rates that were more consistent with gDNA (total call rate >87% and AB call rate >21%). All remaining samples passed Infinium™ sample-independent controls, indicating no failures during array processing. Sample-dependent Infinium™ controls were not used for QC of MDA products since values are likely to be affected by aneuploidy in the samples. An additional 16 out of 307 samples, which had passed SNP array QC, were excluded from NGS library preparation due to low MDA yield <3 µg (Figure 4.9B), with an average of 1.53 µg ± 0.60 S.D. Finally, 37 out of 253 (14.62% ± 1.11% S.E.P) sequenced samples were excluded due to high noise/uninterpretable CNV traces generated using BlueFuse® Multi software, with a custom MDA reference database.

Taken together 91 out of 307 (29.64% ± 1.30% S.E.P) single cell samples failed to give interpretable CNV traces following DNA content analysis. MDA yield is not representative of genome coverage, and therefore cannot solely indicate sample quality. However, differences in the average MDA yield (µg ± S.D) were observed between oocytes and polar bodies from stimulated cycles (Clinic A; 24.18 ± 8.22) and IVM of GV oocytes from small antral follicles (Clinic B; 13.93 ± 7.59) (Figure 4.9B). The difference in pipeline efficiency between clinics was also
Figure 4.9: Quality control of MDA products derived from oocytes and polar bodies. (A) Correlation between AA and BB genotype calls for 307 single cell samples. Samples with <5000 BB calls were excluded from NGS (red dashed line; n=38). Genotype call frequencies from blank arrays show negative control (n=3; green diamond). (B) Duplex DNA mass (µg) in 50 µl MDA product, by cell type and collaborating clinic. Yield from single cell isolates from cell lines (as shown previously) is included for positive control (n=56). Samples passing array QC but with an MDA yield of <3 µg (red dashed line) were not included in library preparation for NGS (n=16). (C) Box plot for primary NGS metrics for samples with high noise in CNV profiles (n=37). (D) Primary NGS metrics for samples with a matched meiotic cell that also passed QC (n=160), and unmatched samples that had no reciprocal cell from the same meiotic division after exclusions (n=56). For (C) and (D), DLR values of 0.2 and 0.4 are shown with red solid and dashed lines, respectively. Oocytes from Clinic A were retrieved following ovarian stimulation, oocytes from Clinic B were retrieved from small antral follicles (SAFs) and matured in vitro. SC = single cell, SAF = small antral follicle. Data drawn from E31e_RCB_OocytesIncomingQC.
mirrored in samples failing NGS due to CNV trace noise profiles, with samples from Clinic A typically producing charts with low bin-to-bin variation within chromosomes, and high variation in representation between chromosomes (Oocyte: GEN123_4_36-3; Figure 4.10A) and samples from Clinic B typically produced charts with stochastic CNV values between bins (Oocyte: PT1063-27.3; Figure 4.10B). The difference in perceived noise is supported by the difference in average DLR values between clinics for samples that were excluded due to high noise (Figure 4.9C; purple; t-test, p<0.01). This is consistent with the notion that oocytes that are matured in vitro from the germinal vesicle stage to completion of meiosis I, are of reduced quality compared to MI-stage oocytes retrieved from stimulated cycles.

**Figure 4.10:** Copy number variation NGS traces from oocytes collected from two different clinics, re-drawn from exported BlueFuse® Multi data. On the Y-axis, the expected copy number for diploid cells (2.0; black line), single copy number gain (3.0; red line) and single copy number loss (1.0; blue line) are shown, in addition to reference lines (dashed; green) showing the 20% deviations from copy number 2.0, 3.0 and 1.0. Vertical reference lines (black) and alternating black and grey bin markers separate the chromosomes. (A) Sample GEN123_4_36-3 from Clinic A showing variation in genome representation between all chromosomes. (B) Sample PT1063-27.3 from Clinic B, showing stochastic noise between normalised copy number bins within and between chromosomes. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.
In total, 216 samples passed QC criteria for CNV analysis, with comparable metrics to single cells isolated from cell lines (Figure 4.9D). For all included oocyte/PB samples, the combined mean (± S.D) for percent mapped reads was 98.85 ± 0.47 compared to 98.46 ± 0.93 cell line isolates (t-test, p<0.01). As discussed previously, 12 out of 56 single cells from cell lines were associated with a single sequencing run with a cluster density >1400 K/mm² (L01). When excluding these samples, the mean (± S.D) for percent mapped reads for cell line samples was 98.96 ± 0.30 (t-test, p>0.05 n.s). However, for all included oocyte and polar body samples the combined mean (± S.D) for percent reads passing filter was 80.53 ± 11.68 compared to 89.94 ± 0.84 for all cell line isolates (n=56) (t-test, p<0.001). The high variability in reads passing filter (%) for oocyte/PB samples may be improved by lowering the cluster density during sequencing. A higher cluster density is likely to be caused by a shorter average fragment size within the library pool, when loading the same volume of library onto the reagent cartridge following bead-based normalisation. Whilst further work would be required to optimise this value, the present number of reads available after filtering is sufficient to perform simple content analysis due to the paired-end sequencing strategy. The DLR values for included samples were typically higher in oocytes following IVM from GVs (Clinic B) compared to in vivo maturation and ovarian stimulation (Clinic A), as seen for excluded samples (Figure 4.9C and 4.9D). However, for samples passing NGS QC, the average (± S.D) DLR value for cell line isolates was 0.52 ± 0.10 compared to 0.35 ± 0.14 for oocyte and polar body samples from clinic A (t-test, p<0.001). Taken together, oocytes and polar body samples show lower MDA bias, and therefore more linear genome coverage compared to single cells isolated from cell line cultures. This is consistent with the notion that oocytes and polar bodies are not in S-phase at the time of biopsy, compared to unsynchronised lymphoblasts in culture, and therefore we expect to see a higher variation in genome representation in these traces (Figure 4.9D).

Prior to CNV analysis, a further 56 out of 307 (18.24% ± 1.10% S.E.P) samples did not have a matched meiotic cell passing QC to allow reciprocal CNV detection and were also excluded from CNV classification. Taken together the pipeline efficiency for oocytes from Clinic A (stimulated cycles) per sample was 83.16% (79 out of 95 single cells), and the matched-cell inclusion efficiency was 80.49% (33 out of 41 oocytes). For Clinic B (GV IVM), the pipeline efficiency per sample was 64.62% (137 out of 212 single cells), and the matched-cell inclusion efficiency was 40.19% (43 out of 107 oocytes).

4.3.2.1 Chromosome breakage occurs in the germ line, and both meiotic divisions

In total, 160 single cell samples passed QC criteria for CNV analysis. The data set comprised 25 MI duos, 6 MII Trios and 2 MII Duos from Clinic A (stimulated cycles); and 41 MI duos and 2 MII

116
Trios from Clinic B (IVM oocytes from small antral follicles). We detected 12 reciprocal segmental events, occurring in 7 meioses (Clinic A: n=4; Clinic B; n=3) (Table 4.4).

One out of 12 events were detected in a complete oocyte-PB trio and affected the whole arm of chromosome 2q (Sample 942.44; Figure 4.11). The 150 Mb q-arm is detected at a 2:2:0 ratio across the PB1:PB2:oocyte trio, which is consistent with an MII abnormality. An additional whole chromosome, PSSC event was also detected for chromosome 16, resulting in a 3:1 loss in the oocyte at MII. This observation provides evidence supporting a 1N, 4C content of the complete meiotic trio.

The remaining 10 out of 12 reciprocal segmental events were detected in meiosis I oocyte-PB1 duos. The smallest segmental CNVs were detected in sample PT1053-1 and were <10 Mb; a 3.5 Mb event in chromosome 6q and an 8.4 Mb event in chromosome 18p, with both events resulting in a 3:1 loss in the oocyte (Figure 4.12).
Figure 4.12: Copy number variation NGS traces for oocyte-polar body duo PT1053-1, re-drawn from exported BlueFuse® Multi data. Traces show detection of two reciprocal segmental events; a 3.5 Mb event in chromosome 6q (red arrows) and an 8.4 Mb event in chromosome 18p (blue arrows); both with loss in the oocyte. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.

Sample 933-21 had an 18.0 Mb event affecting chromosome 17p and a second 16.4 Mb reciprocal segmental event for chromosome 19q that had copy number values of ± 0.3-0.7 (Figure 4.13A). Interestingly sample 933-21 also contained two non-reciprocal CNVs for chromosome 2 (whole chromosome event) and chromosome 17q (shared breakpoint with 17p) in the oocyte, with partial copy number changes (± 0.6). Taken together, the copy number values associated with abnormalities in sample 933-21 are more consistent with a triploid sample. This is because, in the current MI-duo pipeline the CNV ground state for the PB1 and Oocyte is expected to be diploid (1N, 2C), with the neutral copy number 2.0 (Y-axis centre line). If the cell is triploid (1N, 3C), but assigned a ground state CNV value of 2.0, then a 4C chromosome would have a CNV value of 2.66. Similarly, a 1C chromosome would have a CNV value of 0.66 and a 2C chromosome would have a CNV value of 1.33; these CNV values are consistent with those seen in Figure 4.13B, where the copy number value per bin is adjusted using the formula:

\[
\text{Triploid copy number value} = \left( \frac{\text{BlueFuse bin copy number value}}{2} \right) \times 3
\]

When combining the two cells, seven chromatids can be counted for chromosome 2 and only 5 chromatids are counted for the 17q fragment. Taken together, this suggests the chromosome imbalance occurred prior to the first meiotic division and occurred either during meiotic S-phase or was inherited from the germ line.
Figure 4.13: Copy number variation NGS traces for oocyte-polar body duo 933.21, aligned with the assumptions of diploid and triploid ground states. (A) Data re-drawn from exported BlueFuse® Multi data, centred on the assumption of a diploid copy number ground state 2.0. Traces show reciprocal segmental events for chromosome 17p (red arrows) and chromosome 20q (blue arrows). Non-reciprocal events are seen for chromosome 2 (red ellipse) and chromosome 17q (blue ellipse). Copy number values for non-reciprocal events are similar to those of putative mosaic events seen in multi-cell embryo biopsies, as they fall outside the 20% deviation from whole copy number states (green, dashed lines). (B) Data of figure A, re-drawn from exported BlueFuse® Multi data, with Y-axis copy number value adjustments that assume a triploid ground state 3.0. The same events are detected, however non-reciprocal events now align with the whole copy number states 4.0 (chromosome 2, red ellipse) and 2.0 (chromosome 17q, blue ellipse). This adjustment is made upon the notion that single cell analysis cannot give rise to mosaic copy number changes.
Sample PT1052-9 contained a 16.8 Mb segmental abnormality in chromosome Xq, with a 3:1 loss in the oocyte (Figure 4.14).

![Copy number variation NGS traces for oocyte-polar body duo PT1052-9, re-drawn from exported BlueFuse® Multi data. Traces show detection of a reciprocal 16.8 Mb segmental event in chromosome Xq (red arrows) with loss in the oocyte. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.](image)

**Figure 4.14:** Copy number variation NGS traces for oocyte-polar body duo PT1052-9, re-drawn from exported BlueFuse® Multi data. Traces show detection of a reciprocal 16.8 Mb segmental event in chromosome Xq (red arrows) with loss in the oocyte. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.

Sample GEN128_3_50 contained a segmental abnormality in chromosome 9q, that affected both chromatids of the chromosome, such that the MII-arrested oocyte contained a 4:0 gain of chromosome 9q (Figure 4.15). In addition, a 3:1 ratio was observed for whole chromosome aneuploidies in chromosomes 1 and 21. Taken together this data supports the diploid ground state of both cells and suggests that two independent chromosome breaks occurred approximately 10 Mb from the centromere on the q-arms of sister chromatids of chromosome 9. One additional oocyte also had a segmental event in chromosome 9q, that affected the whole arm, occurring approximately 1.6 Mb from the centromere and resulting in 3:1 loss in the oocyte (PT1060-16; Figure 4.16).
Figure 4.15: Copy number variation NGS traces for oocyte-polar body duo GEN128-50, re-drawn from exported BlueFuse® Multi data. Traces show detection of a reciprocal, 4:0 whole arm segmental event in chromosome 9q (red ellipses) with gain in the oocyte, and whole chromosome aneuploidies for chromosome 1 (red arrows) and chromosome 21 (blue arrows); both with loss in the oocyte. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.

Figure 4.16: Copy number variation NGS traces for oocyte-polar body duo PT1050-16, re-drawn from exported BlueFuse® Multi data. Traces show detection of a reciprocal, 3:1 whole arm segmental event in chromosome 9q (red ellipses) with loss in the oocyte, and 4:0 whole chromosome aneuploidies for chromosomes 1, 4, 8, 12, 13, 14, 16, 17 and X (blue arrows); all with loss in the oocyte. Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.
Finally, sample 860.2 contained four segmental events in three chromosomes with gain in the oocyte (Figure 4.17). Chromosome 1 showed two 4:0 abnormalities (oocyte gain) in the p- and q-arms, 39.3 and 16.1 Mb from the centromere, respectively. The interstitial fragment that included the centromere, was 2C in both cells of the MI duo. In addition, chromosome 7q showed a 4:0 30.9 Mb event and chromosome 12q showed a 3:1 33.1 Mb event. Sample 860.2 was successfully activated using calcium ionophore, but the PB2 was lost during biopsy.

![Copy number variation NGS traces for oocyte-polar body duo 860.2](image)

**Figure 4.17:** Copy number variation NGS traces for oocyte-polar body duo 860.2, re-drawn from exported BlueFuse® Multi data. Traces show detection of two separate reciprocal 4:0 breaks in chromosome 1 (red boxes), in addition to reciprocal segmental events in chromosome 7q and chromosome 12q (red arrows). A further ten 4:0, whole chromosome aneuploidy events were also detected (blue arrows). Data drawn from E31b_RCB_HerlevBiopsiesAnalysis.

Of the 76 oocytes with matched meiotic cells, I also identified 23 non-reciprocal chromosome imbalances in 12 meioses, of which two were segmental events. 19 out of 23 events (in 10 meiotic duos) were associated with oocytes collected by Clinic B (IVM oocytes from small antral follicles). Biopsy reports were available for 8 oocytes from this clinic, all of which indicated fragmentation of the PB1 during biopsy. This may explain why we see chromosome loss, or elevated noise in these samples. An example of polar body fragmentation is shown in Figure 4.18. Two meioses from Clinic A (stimulated cycles) accounted for three non-reciprocal imbalances in one oocyte-PB trio and one oocyte-PB duo. Biopsy reports indicated fragmentation of the polar bodies for both oocytes. Taken together, oocyte-PB duos obtained from IVM of GV-stage oocytes and activation with calcium ionophore are associated with a higher incidence of non-reciprocal chromosome imbalance than oocytes sourced from
stimulated cycles that were matured *in vivo*. In both cases, non-reciprocal abnormalities in matched meiotic cells appear to be explained by polar body fragmentation.

**Figure 4.18**: Fragmentation of the first polar body following *in vitro* maturation of GV-stage oocytes. (A) Sample PT1053-31 following completion of meiosis I. The PB1 has been extruded with good morphology. (B) Sample PT1045-25 following completion of meiosis I, with fragmentation of the PB1. During biopsy, all visible fragments of the extruded cell were collected. PB1 = First polar body; ZP = Zona pellucida. Images located in E31_RCB_Analysis_E29-E30_HerlevBiopsies.
**Table 4.4:** Segmental breakpoints detected in oocytes and polar bodies. The molecular positions (bp) of the sequencing alignment bins adjacent to the detected breakpoint are recorded (Start/Stop) by cell. The break positions within each cell are averaged, and the chromosome break position (Mb) is an average of the averaged per cell positions. CNV size is estimated by subtracting the break position from the known length of the chromosome/arm, displayed in terms of the fragment that deviated from a neutral copy number state. The distance (Mb) between the estimated break position and the chromosome centromere is also recorded, with a Figure ID for reference.

<table>
<thead>
<tr>
<th>Oocyte ID</th>
<th>Chr.</th>
<th>PB1</th>
<th>PB2</th>
<th>Oocyte</th>
<th>Chr. break Position (Mb)</th>
<th>CNV Size (Mb)</th>
<th>Centromere Distance (Mb)</th>
<th>Figure ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>942.44</td>
<td>2q</td>
<td>-</td>
<td>-</td>
<td>87,296,457</td>
<td>89,272,784</td>
<td>93,278,793</td>
<td>97,141,867</td>
<td>91.7</td>
</tr>
<tr>
<td>PT1053-1</td>
<td>6q</td>
<td>166,891,660</td>
<td>168,049,494</td>
<td>-</td>
<td>-</td>
<td>166,891,660</td>
<td>168,049,494</td>
<td>167.5</td>
</tr>
<tr>
<td>PT1053-1</td>
<td>18p</td>
<td>7,265,809</td>
<td>8,367,323</td>
<td>-</td>
<td>-</td>
<td>8,367,323</td>
<td>9,476,128</td>
<td>8.4</td>
</tr>
<tr>
<td>933.21</td>
<td>17p</td>
<td>17,217,874</td>
<td>18,698,383</td>
<td>-</td>
<td>-</td>
<td>17,217,874</td>
<td>18,698,383</td>
<td>18.0</td>
</tr>
<tr>
<td>PT1052-9</td>
<td>Xq</td>
<td>139,886,205</td>
<td>141,158,183</td>
<td>-</td>
<td>-</td>
<td>135,218,644</td>
<td>136,406,858</td>
<td>138.2</td>
</tr>
<tr>
<td>GEN128_3_50</td>
<td>9q</td>
<td>57,481,956</td>
<td>71,863,399</td>
<td>-</td>
<td>-</td>
<td>40,804,248</td>
<td>57,481,956</td>
<td>56.9</td>
</tr>
<tr>
<td>PT1060-16</td>
<td>9q</td>
<td>31,554,078</td>
<td>32,686,642</td>
<td>-</td>
<td>-</td>
<td>74,098,848</td>
<td>75,206,595</td>
<td>53.4</td>
</tr>
<tr>
<td>860.2</td>
<td>1p</td>
<td>81,685,145</td>
<td>82,755,509</td>
<td>-</td>
<td>-</td>
<td>87,330,643</td>
<td>88,416,476</td>
<td>85.0</td>
</tr>
<tr>
<td>860.2</td>
<td>1q</td>
<td>133,475,616</td>
<td>147,423,487</td>
<td>-</td>
<td>-</td>
<td>133,475,616</td>
<td>147,423,487</td>
<td>140.4</td>
</tr>
<tr>
<td>860.2</td>
<td>7q</td>
<td>127,497,470</td>
<td>128,660,317</td>
<td>-</td>
<td>-</td>
<td>127,497,470</td>
<td>128,660,317</td>
<td>128.1</td>
</tr>
<tr>
<td>860.2</td>
<td>12q</td>
<td>98,356,505</td>
<td>99,462,663</td>
<td>-</td>
<td>-</td>
<td>98,356,505</td>
<td>99,462,663</td>
<td>98.9</td>
</tr>
</tbody>
</table>

Chr. = Chromosome; WA = Whole arm.
4.4 Study conclusions

I have developed an NGS pipeline for CNV detection that can be run concurrently with SNP microarray analysis from a single MDA product. Previous work has demonstrated the utility of SNP genotyping to fingerprint chromosomes in meiotic trios by maternal homolog (Capalbo et al., 2015a; Ottolini et al., 2015). NGS analysis provides improved resolution to detect small segmental aneuploidies, which I validated with a per chromosome concordance >99% for single cells from cell lines of known karyotype (n=2,208), and a per sample concordance of 91.7% (n=48) (Table 4.3), although with noise in the cell line model. I mitigated the noise by stringent QC and detected 12 reciprocal sub-chromosomal aneuploidies in 7 out of 76 (9%) meioses and extracted their molecular positions (Table 4.4). A similar rate was observed in 10% (11 out of 111) of meiotic trios derived from oocytes from stimulated cycles (Hou et al., 2013; Ottolini et al., 2015), which provides evidence to suggest the pipeline does not over-call segmental aneuploidies due to noise introduced by WGA. Application of the method to oocytes and polar bodies from two sources (stimulated cycles and unstimulated, in vitro matured GV oocytes), revealed a higher pipeline efficiency from oocytes matured in vivo, and a higher incidence of non-reciprocal chromosomal abnormalities in matched cells from IVM oocytes. Despite improvements that could be made to the pipelines for sourcing human oocytes, the analytical method has been demonstrated in two studies of gross chromosomal rearrangements (GCRs) in human oocytes and polar bodies, and accounts for noise through stringent QC (Appendix A1.3, A1.4 and A1.5; submitted).

The rate of GCRs that we observe in human eggs and embryos exceed those seen in cancer cell lines. Chromosomal rearrangements drive tumour evolution and are prevalent in clonal populations (McGranahan and Swanton, 2017); however, the estimation of their frequency is lower than seen in oocytes and embryos (10^7 per cell division) (Jackson and Loeb, 1998). These higher rates are more consistent with the treatment of cells with inhibitors of DNA replication or exogenous DNA damaging agents (Lindsley et al., 1972). It is currently unclear where the gross chromosomal rearrangements originate from in the germline, however, our lab has evidence that heterogeneity in the expression of genes that prevent GCRs may cause the chromosome instability (Blanshard et al. In preparation).
Chapter 5: Direct detection of gene conversions

5.1 Overview

Meiosis gives rise to recombinant chromosomes by first inflicting double strand breaks (DSBs) on the genome. The repair and resolution of DSBs either results in a crossover (CO), or non-crossover (NCO) event. The decision to resolve DSB as a CO or NCO occurs early in DSB repair, around the time of strand exchange. Whilst COs are important for chromosome segregation during meiosis, NCOs are predicted to occur at a 10-fold higher rate in mice and human, with specific roles in mediating accurate homologue pairing and distribution of COs (Baudat et al., 2013). In both CO and NCO events, the DSB is repaired in using the other homologue as a template due to bias against repair of the sister chromatid. In CO events, the heteroduplex is thought to be longer than NCO events, where the unidirectional transfer of a ~50-1000 bp tract persists and may result in a gene conversion (Figure 1.11). Gene conversions play an important role in genome diversification, by breaking down local linkage disequilibrium (LD). The frequency of NCOs has been inferred from pedigrees using SNP array and NGS data. The incidence of meiotic gene conversions, based on the number of informative heterozygous SNPs analysed was $5.9 \times 10^{-6}$ per bp, per meiosis (Williams et al., 2015). The majority of confirmed gene conversions co-localised within DSB hotspots, which is consistent with the notion that NCOs occur in meiosis.

Direct detection of gene conversions in individual meiosis by single cell analysis is challenging due to that all four DNA strands have to be recovered as well as the limitations of whole genome amplification (WGA) and the cost of NGS at high depth. Unlike sequencing of bulk DNA, single cell sequencing requires WGA to generate sufficient material for array-based analyses (400 ng) or PCR-free NGS (1 µg) (Gawad et al., 2016). Since we can recover all four DNA strands in oocyte-PB1-PB2 trios (Ottolini et al., 2015) the rate limiting step is the WGA. In particular, detecting a gene conversion requires the same SNP to be mapped in all three cells with high precision. Systematic and stochastic amplification bias during WGA greatly affect the efficacy of genotyping. Specifically, both allele drop in (ADI) and allele drop out (ADO), the latter which occurs in up to 30% of genotyped SNPs in single cells (Blanshard et al., 2018), causes an false positive gene conversion (3:1 or 1:3). Thus, the noise associated with WGA causes a staggering false positive rate.

Sequencing is also critical when genotyping with high precision. PCR-free sequencing is preferred to prevent further confounding of replication errors by polymerases used for cluster generation and base extension. SNP microarrays offer a cost-effective solution for high confidence genotyping, particularly when analysing three cells per meiosis, as is required for
MeioMapping of maternal haplotypes (Ottolini et al., 2015) and detection of de novo mutations. In principal, genotyping from scanned SNP arrays is performed from two-fluorophore intensity data that represent alternate homozygous allele modes (AA or BB). Heterozygosity is inferred from detection of both fluorescent signals (AB). A number of algorithms are available for interpreting the intensity thresholds, normalising and calling genotypes from bulk genomic DNA (gDNA) with high call rates and accuracy (Ritchie et al., 2011; Li et al., 2012). However, as these algorithms are typically trained on intensity data generated from gDNA, it is unclear how they handle the amplification biases introduced by WGA of single cells. Simply excluding poorly performing SNPs results in substantial data loss, which greatly reduces the probability of detecting small gene conversion tracts when using an array with a limited number of features. Moreover, it's unclear how the SNPs that do call are affected by the unequal ratios of A and B allele input after various normalization procedures.

The lack of SNPs genotyped from single cells has led to a paucity in the development of algorithms that specifically characterize noise from single cells. To this end, I generated a comprehensive dataset of 28 million high quality SNPs from 104 single cell from two different cell lines. Single cells were genotyped using BeadChip SNP arrays under stringent quality control and compared to their reference genotypes. In collaboration with Ivan Vogel (Bioinformatician, University of Copenhagen), we have developed and validated a two-layered algorithm to identify the erroneous WGA signals in the single cell data. We then assign a probability score that a SNP has been correctly genotyped. To date, we have applied the algorithm to genotype data derived from one meiotic oocyte-PB trio and identified 10 putative gene conversion events. I independently verified four (40%) events with PCR-free NGS at high depth. The development of the algorithm (SureTypeSC) is available on BioRxiv (Vogel et al., 2018), and has been submitted for peer review elsewhere. My contribution to the paper constituted preparation of high quality, single cell reference libraries and processing of SNP arrays; design of the informatics pipeline and troubleshooting was performed in collaboration, and the script was written and implemented by Ivan Vogel. Using the SureTypeSC allowed us to explore gene conversions in human oocyte-PB1-PB2 trios. I prepared PCR-free sequencing libraries for eight oocyte-PB trios for which we had pre-existing SNP data. I sequenced four trios at 50x depth, of which we have used one to validate the SNP calling to date. Thus, we have detected directly human gene conversions.
5.2  Materials and Methods

5.2.1  Cell lines

Euploid cell lines were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (NJ, USA): GM07228, GM07224, GM07225 and GM12878. Cells were cultured according to the supplier’s recommendations, as described previously in Chapter 2. We generated two single cell reference databases, for cell lines GM07228 and GM12878 (Table 5.1). Single cells were manually isolated as described previously, and stored at -20 °C until WGA (Blanshard et al., 2018; Chapter 2). In addition, genomic DNA for each proband and parental cell lines (GM07224 and GM07225) was extracted from cell cultures for generation of high-confidence reference genotypes (QIAamp DNA Mini Kit; QIAGEN, Hilden, Germany). Cell line GM12878 was selected because genotype information for this cell line and both parental cell lines (NA12891 and NA12892) is publicly available following PCR-free NGS at 50× depth (Eberle et al., 2017); the trio was also sequenced using the 1000 Genome Project (Abecasis et al., 2010).

Table 5.1: Stratification of cell line samples used for algorithm training and validation

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Pedigree</th>
<th>Single Cell</th>
<th>gDNA</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM07224</td>
<td>Mother</td>
<td>-</td>
<td>9</td>
<td>Training</td>
</tr>
<tr>
<td>GM07225</td>
<td>Father</td>
<td>-</td>
<td>6</td>
<td>Training</td>
</tr>
<tr>
<td>GM07228</td>
<td>Proband</td>
<td>58</td>
<td>9</td>
<td>Training</td>
</tr>
<tr>
<td>GM12878</td>
<td>Proband</td>
<td>46</td>
<td>5</td>
<td>Validation</td>
</tr>
</tbody>
</table>

5.2.2  Whole genome amplification

Single cells were amplified by multiple displacement amplification (SureMDA, Illumina Inc.) within 48 hours of collection, using a 2-hour incubation with phi-29 DNA polymerase (Infinium™ Karyomapping Assay Protocol Guide: 15052710 RevB; Illumina Inc.). Positive amplification was confirmed by gel electrophoresis, and duplex DNA was quantified using the Quant-iT dsDNA Assay Kit, High Sensitivity (ThermoFisher Scientific Inc.).

5.2.3  BeadChip microarrays and SNP genotyping using GenCall

Single cell and gDNA samples for each of the cell lines obtained from Coriell were genotyped using the Infinium™ Karyomapping Assay Kit (Illumina Inc.), according to the manufacturer’s
instructions as described previously (Blanshard et al., 2018; Chapter 2). Samples were hybridised to HumanKaryomap-12 BeadChip arrays and scanned using either an iScan System or NextSeq550 System (Illumina Inc.). Intensity data were analysed using GenomeStudio v2.0.2 software with Genotyping Module v1.9 (Illumina Inc., California, USA) for genotype calling with the Illumina GenCall algorithm. A confidence score, known as the GenCall score (GC), is assigned to every SNP and is used as a measure of quality. As a default, genotypes with GC values lower than 0.15 are considered false positive.

5.2.4 Generation of reference genotypes on the HumanKaryomap-12 v1.0 BeadChip Arrays

High quality reference genotypes were generated for bulk genomic DNA extracted from cell lines. Nine replicates of maternal DNA from GM07224, six replicates of paternal DNA from GM07225 and nine replicates of proband DNA GM07228 were analysed by SNP array. Consensus genotypes for the mother, father and proband were created by only accepting SNP loci with 100% genotype concordance between the replicates. Parental inheritance patterns and those which did not show a mendelian inheritance of A and B genotypes in the proband (GM07228) were also excluded. In total 1.1% of autosomal SNPs were excluded. An additional reference dataset for GM12878 was also created using five gDNA replicates extracted from the cell line.

5.2.5 Genotyping in the single cell environment

5.2.5.1 The reduction in signal from single cells following WGA is characterised by features of microarray signal intensity

The SureTypeSC algorithm was developed by characterising SNP calls based on normalised intensity outputs in the red and green fluorescence channels; where A = Red fluorophore = X and B genotype = Green fluorophore = Y. Intensity data files (*.idat) were generated by the iScan and NextSeq550 Systems and imported into GenomeStudio software. The GenCall algorithm was executed with a GC threshold of 0.01, to include SNPs with potentially sub-optimal WGA, and allowing the full spectrum of single cell WGA noise to be captured. In total, 14,805,232 SNPs were typed for training (GM07228) and 11,799,864 SNPs were typed for validation (GM12878). An MA (linear-log) transformation was also applied to intensity data for each SNP loci for each sample (Vogel et al., 2018), whereby:
M = Log difference between X and Y intensities

A = Average log of X and Y intensities

The M-feature can powerfully discriminate between AA, AB and BB SNP clusters, whilst the A-feature is an indicator of signal quality. The M-feature and A-feature values serve as input into the machine learning workflow, and the match/mismatch status of each SNP loci to the consensus genotype serves as an output for machine learning. The error pattern associated with genotyping single cell data (with a GC threshold of 0.01) is shown in Figure 5.1.

![Figure 5.1](image)

**Figure 5.1:** Signal-noise detection in whole-genome amplified DNA from single cells. (A) Contour MA plot of all SNPs from one bulk DNA sample (gDNA-01) from GM07228; AA, BB and AB clusters are labelled accordingly. Correctly typed SNPs are rendered in blue, whereas incorrectly typed SNPs are shown in red. (B) Contour MA plot of all SNPs from one single-cell sample (SC-21) from GM07228; AA, BB and AB clusters are labelled accordingly. Miscalls are shown in red clusters, attributed to allele drop out (ADO) and allele drop in (ADI). For each cluster, the proportion of total call rate is shown.

5.2.5.2 A two-stage machine learning algorithm was trained to assign probabilities of correct genotype calls following GenCall with a minimal QC threshold

The supervised algorithm training using Random Forest (RF) was implemented by Ivan Vogel and has been described in detail (Vogel et al., 2018). Briefly, RF is an ensemble supervised training method that is built from the collection (forest) of classification (decision) trees (Breiman, 2001). We used the implementation of RF from the scikit package (Pedregosa et al., 2012) for fitting
the training data from cell line GM07228. The training data was randomly divided into subsets and each tree is trained on a different subset. The input features were m and a values derived from normalised SNP array intensity data. The model was limited to 30 trees, and predictions were evaluated using 10-fold cross validation and an independent validation on unseen data. The metrics used for validation are shown in Table 5.2, in addition to a posterior probability matrix that was used to calculate the probability of correctly calling a single cell genotype as well as ADO and ADI, when compared to reference genotypes derived from bulk DNA (Vogel et al., 2018).

Table 5.2: Metrics used to assess validation of the SureTypeSC genotyping tool

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>( \frac{TP}{TP + FN} )</td>
</tr>
<tr>
<td>Precision</td>
<td>( \frac{TP}{TP + FP} )</td>
</tr>
<tr>
<td>Specificity</td>
<td>( \frac{TN}{TN + FP} )</td>
</tr>
<tr>
<td>F1 score</td>
<td>Harmonic mean of precision and recall</td>
</tr>
<tr>
<td>Receiver Operating Characteristic (ROC) curve</td>
<td>Sensitivity (TP rate) as a function of 1-Specificity (FP rate)</td>
</tr>
<tr>
<td>ROC-AUC (area under curve) score</td>
<td>Area under the ROC</td>
</tr>
<tr>
<td>Precision-Recall (PR) curve</td>
<td>Recall as function of precision</td>
</tr>
</tbody>
</table>

TP = True Positive; FN = False Negative; TN = True Negative; FP = False Positive.

The genotype predictions derived from the RF algorithm were screened further using Gaussian discriminant analysis (GDA), that attempts to formalise the genotype clusters and improve precision. The algorithm has been described previously (Vogel et al., 2018). Briefly, two Gaussian models were used to calculate membership weights (SNP-specific) and component weights (genotype-specific). For each SNP, an Expectation step assigned a probability that the m and a intensity values fit to the predicted genotype cluster (true/false) using default (initial) parameters. In a Maximisation step, new component weights, means and covariances were calculated to assign new parameters. A log-likelihood score was calculated, and if the relative change between the two sets of parameters is below a given threshold then the algorithm halted; otherwise a new iteration was begun with the Expectation step, using the updated parameters (Dempster et al., 1977). Both models were run in iterations until convergence was reached for all SNPs. At this point, the analysis was re-run using the estimated parameters for both classes of genotype calls (true/false), but the class prediction from RF was hidden and each
SNP was evaluated for both Gaussian discriminants. The SNPs were labelled according to the winning likelihood score (true/false) and highest membership weight to a component (genotype) of the winning class. The scoring function of the RF alone was the probability of a correct genotype call, as a proportion of the trees that voted TRUE; for SureTypeSC (Random Forest + Gaussian discriminant analysis) we applied the Bayes rule to express the class-conditional posterior probability of a genotype being called as TRUE.

5.2.6 Oocytes and polar bodies for mapping of gene conversions

Eight meiosis II oocytes were ethically sourced from four patients who underwent ICSI treatment in the Centre for Reproductive Medicine GENERA in Rome between September 2008 and January 2010. At the time of oocyte collection, a maximum of three oocytes could be inseminated per patient under Italian law. Surplus oocytes were vitrified and included in this study with fully informed consent. This study did not interfere with patient treatment and was approved by the Institutional Review Board of the Valle Giulia Clinic. Oocytes were activated with 100 µM calcium ionophore for 40 min, using the previously described and published protocol (Ottolini et al., 2016). Both matching polar bodies were biopsied, and the oocyte was isolated. The three single cells (comprising an oocyte-PB trio) were amplified by SureMDA (Illumina Inc.) as described here. MDA products for four oocyte-PB trios were previously analysed by SNP array (HumanKaryomap-12; Illumina Inc.) and MeioMapping, and have been included in previously published work (Ottolini et al., 2015). SNP array intensity data files for these 12 samples were existing, and MDA DNA products were available for inclusion in the present work. Four additional oocyte-PB trios from one patient were amplified by SureMDA at the source laboratory, and I then performed SNP microarray analysis according to the Infinium™ Karyomapping Assay Kit (Illumina Inc.), according to the manufacturer’s instructions as described previously (Blanshard et al., 2018; Chapter 2). All four trios were confirmed to be euploid by MeioMapping (Blanshard, unpublished). To this end, 24 single cell MDA products and corresponding intensity data file from SNP microarray analysis were included in this study. The SureTypeSC algorithm was executed on one trio (666_5) to identify putative gene conversion events. The events were validated by two direct sequencing approaches.

5.2.7 PCR-free deep sequencing

PCR-free sequencing libraries were prepared from MDA products for all eight oocyte-PB trios trios, of which four trios were analysed by NGS. MDA products were quantified using the Quant-
iT dsDNA Assay Kit, High Sensitivity (ThermoFisher Scientific Inc.), and 1 µg duplex DNA was used per sample input. Libraries were prepared according to the TruSeq® DNA PCR-Free Library Prep Reference Guide and associated reagents (Illumina Inc.) as described in Chapter 2. Briefly, MDA products were fragmented by sonication and cleaned-up by Solid Phase Reversible Immobilisation (SPRI). Fragment end repair, A-tailing and adapter ligation were performed without modification. Fragment size was checked for quality control using the Agilent High Sensitivity DNA Kit on a 2100 Bioanalyzer instrument. Final sequencing libraries were quantified by qPCR (KAPA Biosystems), and insert size was assessed by low-pass 2×36 bp NGS on a MiSeq System as previously described. Library concentrations were re-calculated according to the fragment insert size determined by MiSeq analysis. Pooled libraries from four oocyte-PB trios were sequenced using a NovaSeq 6000 instrument with a NovaSeq 6000 S4 Reagent Kit (2 × 151 bp) (Illumina Inc.) as described previously (Chapter 2).

Sequencing output folders were uploaded to BaseSpace Sequence Hub (www.basespace.illumina.com) for primary analysis. BCL file demultiplexing and FASTQ generation was performed using the FASTQ Generation application (version 1.0.0). Alignment of sequence reads to the Homo sapiens hg19 reference genome (University of California, Santa Cruz) was performed using the BWA Aligner application (version 1.1.4). The resulting *.bam files were downloaded for variant calling using the GATK pipeline v4.0 (McKenna et al., 2010) as previously described.

5.3 Results

5.3.1 Generation of two single cell datasets from using MDA and SNP microarrays with stringent quality control

In total, 58 single cells from cell line GM07228 were isolated and underwent WGA over four independent collection events. An additional 46 single cell were collected and amplified for cell line GM12878 from one population. All single cells showed positive amplification, with no background amplification in isolation or WGA negative controls. Samples for cell line GM07228 were resolved by gel electrophoresis (Figure 5.2A). Single cells gave an expected MDA yield (µg ± S.D) of 27.7 ± 5.4 for GM07228 and 30.4 ± 2.4 for GM12878 (Figure 5.2B).
Figure 5.2: Quality control of single cell MDA. (A) MDA products from four SCI and WGA events were resolved by gel electrophoresis, together with collection (SCI) and amplification (MDA) controls. Original images have been cropped to exclude samples that were out of scope, and not used during this study. Image intensity and scale have been preserved. Sample run in 1.5% agarose against a 1 Kb extension ladder (L).

(B) MDA mass yield (µg) in 50 µl reactions for cell lines GM07228 and GM12878 including collection and amplification controls, and human oocyte-polar body trios. For single cells, GM07228 n=58; GM12878 n=46; Trios n=8 for PB1, PB2 and Oocyte, each. MDA = Multiple displacement amplification; SCI = Single cell isolation; WGA = Whole genome amplification. Data drawn from E05_RCB_HKM_SC_Validation_GM07228 and E13_RCB_HKM_SC_Validation_GM12878.
Undiluted single cell MDA products and technical replicates of 400 ng of each bulk genomic DNA were processed by SNP array. All samples performed as expected, with single cell average total call rates (± S.D) of 78.09% ± 6.46% for GM07228 and 82.37% ± 3.00% for GM12878. Genomic DNA replicates for both cell lines GM07228 and GM12878, and parental cell lines GM07224 and GM07225 gave a combined average total call rate of 98.03% ± 0.39% as expected (Figure 5.3). Heterozygous AB call rates were also within expectations, with an average of 17.06% ± 3.78% for GM07228 and 12.15% ± 2.86% for GM12878 single cells, and a combined average of 31.06% ± 1.34% for genomic DNA samples (Figure 5.3).

Figure 5.3: Genotype call rates for single cells and bulk genomic DNA hybridised to HumanKaryomap-12 BeadChip arrays and analysed in GenomeStudio with a GenCall threshold of 0.15. Total call rate describes the proportion of typed SNPs out of all features on the array. AB call rate describes the proportion of heterozygous genotypes out of the total number of typed SNPs. Data drawn from E33_RCB_AnalysisOfE05&E13/ArrayCallRateQC.

As an additional QC, Infinium™ control probes were assessed for efficiency of the microarray assay. A full description of the Infinium™ control probes is available in the training guide ‘Evaluation of Infinium Genotyping Assay Controls Training Guide’ (Illumina Inc., 2012). The internal control probes are used to assess the assay performance and robustness. Sample input quality as well as each step of the single base extension and staining process are evaluated, such that the controls can be used to troubleshoot assay performance. Depending on the probe design, each control is measured during array scanning, using the red or green channels, or both.
Briefly, staining (biotin and dinitrophenyl), hybridisation, extension and target removal controls are sample independent and verify the efficacy of the microarray workflow. Stringency (perfect match vs mismatch), non-specific binding and non-polymorphic site controls are sample-dependent and reflect quality of the input material as well as assay performance. When assessing the control charts (Figure 5.4 and Figure 5.5; generated for two different scanning systems, iScan and NextSeq550, respectively), each control probe has an expected signal intensity for the red and green channels, either above or at background levels, as described in the training guide (Illumina Inc., 2012). Figure 5.4 shows clear separation of the signal above the background for staining, hybridisation and extension controls, and an expected background signal for target removal for arrays scanned using the iScan System. Figure 5.5 shows up to a 3-fold reduced signal intensity for arrays scanned with the NextSeq550 system (Y-axis). This difference is inherent to the different imaging systems of the instruments, and the dual functionality of the NextSeq550 as a sequencing and array-scanning system. However, the positive control signals are clearly distinguishable from the background for all but one control. The hybridisation stringency control signal (mismatch probes) is not always distinguishable from the background in the red channel only, for gDNA and single cells scanned with the NextSeq550 System. This is an expected result and highlights a difference between the instruments. Mismatch oligonucleotide probes contain 25% mismatched nucleotides to the target sequence, and therefore we expect to see a reduced signal compared to the ‘perfect match’ probes. As such, the NextSeq 550 is more stringent towards the reduced signal that results from poor hybridisation than when the array is scanned on the iScan system.

On both scanning instruments, the response to the positive ‘non-polymorphic’ control is highly heterogenous in single cell MDA products when compared to bulk gDNA samples. In some cases, the positive signal overlaps the spread of the background signal. The non-polymorphic control works by targeting non-polymorphic sites in the genome, such that we expect A/T sites (red channel) not to be detected as C/G bases (green channel), and vice versa. For both SC and gDNA samples, the expected background signal is observed in all cases. However, the spread in positive signal in SC samples shows poor representation of non-polymorphic sites in this sample type. To this end, genotyping single cell MDA products by SNP microarray using a GenCall threshold of 0.15 (Illumina default), reveals noise that we believe to be coming from the WGA step. Not only do we see a reduced call rate in single cell samples, but the accuracy and/or coverage is likely to be affected by ADO and ADI due to inefficient WGA.
Figure 5.4: SNP microarray assay controls exported from GenomeStudio (Illumina Inc.) for arrays scanned using an iScan System. Data is derived from 14 BeadChip arrays and comprises genomic DNA samples from cell lines: GM07224, \( n = 3 \); GM07225, \( n = 3 \); GM07228, \( n = 9 \); and GM12878, \( n = 5 \); in addition to single cell MDA samples from cell lines: GM07228, \( n = 30 \) and GM12878, \( n = 46 \). Red and green box plots represent the signal intensity for the red and green scanning channels, for each control probe. Data drawn from E33_RCB_AnalysisOfE05&E13.
Figure 5.5: SNP microarray assay controls exported from GenomeStudio (Illumina Inc.) for arrays scanned using a NextSeq550 System. Data is derived from 8 BeadChip arrays and comprises genomic DNA samples from cell lines: GM07224, \( n=6 \); and GM07225, \( n=6 \); in addition to single cell MDA samples from cell line GM07228, \( n=28 \). Red and green box plots represent the signal intensity for the red and green scanning channels, for each control probe. Data drawn from E33_RCB_AnalysisOfE05&E13.
5.3.2 Improved genotype precision in single cell array data using two-stage machine learning

Genotyping in the single cell environment is challenging due to the noise introduced by WGA. We used two independent single cell datasets to characterise the noise profiles, train and then validate an algorithm to improve genotype precision over GenCall with standard QC. To this end, I typed nearly 28.7 million SNPs from 104 cells from two individuals (GM12878 and GM07228) using multiple displacement amplification and the HumanKaryoMap-12 SNP BeadChip array (Illumina Inc.). In addition, high-confidence reference genotypes from bulk genomic DNA were prepared for each cell line using either the full parental information (GM07228) or technical replicates from SNP array and sequence data (GM12878 and Eberle et al., 2017). Of the 276,528 autosomal features on the HumanKaryoMap-12 SNP array, 272,640 SNPs (98.6% autosomal SNPs) called correctly in every replicate from bulk DNA. Single cell genotypes were inferred using the GenCall algorithm (GenomeStudio; Illumina Inc.), with the default QC threshold (0.15) and a minimal QC threshold (0.01). Genotypes were compared to the cell line reference and the match/mismatch status was recorded (Table 5.3).

Table 5.3: Summary of genotype calls from single cells

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Region</th>
<th>Match</th>
<th>Mismatch</th>
<th>GC &lt;0.01</th>
<th>Match</th>
<th>Mismatch</th>
<th>GC &lt;0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>2,507,149</td>
<td>218,684</td>
<td>1,762,134</td>
<td>108,607</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM07228</td>
<td>AA, BB</td>
<td>10,139,650</td>
<td>1,939,749</td>
<td>1,186,470</td>
<td>1,496,791</td>
<td>2,830,096</td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>12,646,799</td>
<td>2,158,433</td>
<td>11,556,208</td>
<td>1,605,398</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>1,835,293</td>
<td>108,269</td>
<td>2,219,787</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM12878</td>
<td>AA, BB</td>
<td>8,443,004</td>
<td>1,413,298</td>
<td>920,424</td>
<td>8,106,969</td>
<td>1,090,869</td>
<td>2,219,787</td>
</tr>
<tr>
<td>all</td>
<td>10,278,297</td>
<td>1,521,567</td>
<td>9,365,395</td>
<td>1,135,106</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using the default threshold (0.15), 20.9 million SNPs from the two single-cell datasets called correctly, whereas 2.74 million generated false positives in GenCall, resulting in an incorrect genotype. 5.05 million SNPs gave ‘no calls’, having failed to fall within genotype clusters defined by bulk, gDNA genotypes. The true positive rate was higher when using a minimal QC (0.01) compared to the standard QC (44% and 40%, respectively, for cell line GM07228 and 36% to 33% for GM12878; Table 5.3). This suggests that the GenCall algorithm rejects about 3-4% of correct genotypes from single cell DNA, MDA products.

By transforming the X and Y fluorescent signal intensities (green and red, respectively) into the logarithmic difference $m$ and logarithmic average $a$, the noise profiles in single cell...
Genotyping can be observed. The distribution of correctly called SNPs (true positives; blue shading) is similar between bulk genomic DNA and amplified single cells, such that three distinct clusters represent AA, BB and AB genotypes (Figure 5.1A and Figure 5.1B). However, in single cell data three clusters of false positive calls are visible (red shading) in the transition zones between the true positive clusters (blue shading). False positive calls falling between heterozygous and homozygous clusters (AB to AA and AB to BB), are likely to represent allele drop out (ADO) events where suboptimal WGA leads to the underrepresentation of one allele. The third, false positive, cluster with low average intensity but with similar representation in both X and Y channels was also observed (Figure 5.1B). This is likely to represent a portion of allele drop in (ADI) events where a heterozygous genotype call is inferred at low intensity (near the signal to background floor), where GenCall is more sensitive to uneven WGA between alleles.

Taken together, the observations suggest good separability of the correct calls from miscalls, since the centres of the clusters are non-overlapping.

The characterisation of the patterns of noise in a comprehensive dataset allowed us to employ a supervised machine learning method to classify and separate high-quality genotypes from miscalls in the amplified DNA from single cells. The algorithm was implemented by Ivan Vogel, and is described in the attached submission for publication, and summarised here (Vogel et al., 2018). The scoring method comprises two-layered filtering using a non-parametric (Random Forest) and parametric method (Gaussian mixture model) to assign probabilities that a genotype has been called correctly, following initial GenCall calling with a minimal QC threshold of 0.01. The two layers, collectively referred to as SureTypeSC, were trained and validated using two different methods.

5.3.2.1 Cross-validation ensures precision without overtraining to limited data

Overtraining describes the optimisation of an algorithm on a limited dataset, that subsequently fails to generalise to independent, unseen data. To minimise this risk, we employed a 10-fold cross-validation using the GM12878 dataset. The data were randomly divided such that 90% was used for training and 10% was used for validation. Critically, the proportions of miscalls and correct calls were equally represented between folds, and each training sub-dataset comprised 11,448,258-11,448,260 SNPs. Whether employing the RF stage alone, or RF combined with GDA (SureTypeSC), we only observed minimal variation between validation experiments, showing that SNP selection does not influence algorithm performance. The implementation of SureTypeSC showed high precision at the expense of recall (Table 5.4).
Table 5.4: 10-fold cross-validation using dataset GM12878

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision ± S.D.</th>
<th>Recall ± S.D.</th>
<th>Accuracy ± S.D.</th>
<th>F1 score ± S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.92 ± 0.001</td>
<td>0.95 ± 0.003</td>
<td>0.88 ± 0.002</td>
<td>0.93 ± 0.001</td>
</tr>
<tr>
<td>SureTypeSC(^a)</td>
<td>0.95 ± 0.001</td>
<td>0.87 ± 0.01</td>
<td>0.85 ± 0.007</td>
<td>0.90 ± 0.005</td>
</tr>
</tbody>
</table>

\(^a\)SureTypeSC combines Random Forest and Gaussian discriminant analysis.

5.3.2.2 Independent validation shows high precision on unseen data

We next addressed the performance of RF alone and SureTypeSC on an independent MDA-amplified single cell dataset. To this end, we used the SNP genotypes obtained from 58 single cell MDA products from cell line GM07228 for training (training set), and the SNP genotypes from 46 single cell MDA products from cell line GM12878 for validation (validation set). Performance was assessed for heterozygous and homozygous SNPs separately, and combined (Table 5.5). For heterozygous sites, we observed a 22% increase in recall with similar precision (96-97%), when using the RF over GenCall with a default threshold of 0.15. Using SureTypeSC, precision was increased to 99%, with a 4% increase in recall over GenCall. The step-wise increase in genotyping accuracy is further shown by a shift to the left for the ROC curve and an increase in ROC-AUC values after implementation of each layer of the algorithm (Random Forest alone, then SureTypeSC) (Figure 5.6A; Table 5.5).

For homozygous genotypes, implementation of the RF and SureTypeSC also shifted the ROC curve to the left (Figure 5.6B) and increased the ROC-AUC score (Table 5.5). The SureTypeSC improved genotyping precision when compared to both GenCall and the RF alone (93%), but with nearly 7% fewer typed SNPs (recall, 88%). Taken together, we have shown that improved precision can be obtained when genotyping both heterozygous and homozygous SNPs from whole-genome amplified DNA, by using the filtering algorithms trained on single cell array data.
Figure 5.6: SureTypeSC improves the performance for single cell genotyping. ROC curve for heterozygous (A) and homozygous SNPs (B). P and N are the numbers of correctly (P = positive) and incorrectly (N = negative) typed SNPs. The dotted line in the ROC curves is the diagonal (random classification). RF = Random forest; ROC = Receiver operating characteristic.

Table 5.5: Performance of tested classifiers using cell line GM12878*

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>GenCall**</th>
<th>Random Forest</th>
<th>SureTypeSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.97</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>Recall</td>
<td>0.69</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.80</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>ROC-AUC score</td>
<td>0.71</td>
<td>0.66</td>
<td>0.64</td>
</tr>
</tbody>
</table>

*Data subjected to minimal QC with GenCall cutoff 0.01; values are proportions
**GenCall score cutoff 0.15

5.3.2.3 Confidence scores calculated from posterior probability

The improved precision (at the expense of recall) observed in Table 5.4 and Table 5.5, can be used to assign probabilities that a genotype call is correct when using SureTypeSC filtering on array data from amplified single cell DNA. The transition matrices in Table 5.6 show the posterior probabilities, calculated using the GM12878 single cell dataset, that a given single cell genotype call (Column: SC) matches each of the four possible reference consensus genotypes (Row: Ref). We calculated the probabilities depending on the type and thresholds of the algorithms used; including minimal QC only (GenCall threshold 0.01), default QC only (GenCall threshold 0.15), or SureTypeSC (GenCall threshold 0.01, and two-stage SureTypeSC filtering). Taken together, we improved the confidence in single-cell AA and BB calls by 7% (or by 5% compared with standard
GenCall genotyping), respectively. The confidence in single-cell AB calls is concordant with the precision obtained in the validation analysis in Table 5.3 (99%).

Table 5.6: Proportion of correct genotype calls when using GenCall and SureTypeSC on single cell array data from cell line GM12878. Call rate shows the proportion of SNPs called in single cell data.

<table>
<thead>
<tr>
<th>SC Ref</th>
<th>Call rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0.85</td>
</tr>
<tr>
<td>AB</td>
<td>0.15</td>
</tr>
<tr>
<td>BB</td>
<td>0.0001</td>
</tr>
<tr>
<td>NC</td>
<td>0.004</td>
</tr>
<tr>
<td>SC</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>1.85×10⁻⁵</td>
</tr>
<tr>
<td></td>
<td>0.003</td>
</tr>
</tbody>
</table>

5.3.2.4 SureTypeSC filters allele drop out (ADO) and allele drop in (ADI) in amplified single cell genotype data

In the context of screening oocyte-polar body trios for putative gene conversion events, the incidence of ADO and ADI are extremely important. Gene conversions and de novo mutations can be detected in meiotic trios by mapping the mendelian inheritance of alleles. In diploid organisms we expect two ‘A’ alleles and two ‘B’ alleles per SNP locus. Divergence from the 2:2 ratio of A and B is indicative of genome changes and is why we have designed the SureTypeSC algorithm to favor precision at the expense of recall. ADO and ADI will both manifest as divergence from the expected inheritance pattern, if occurring in isolation per locus (barring any aneuploidy). The more common ADO event results from non-amplification of one allele at heterozygous loci, whereas ADI results from genotype calls in low intensity array data, or possibly from replication errors early in the WGA reaction. The noise characteristics of ADO ad ADI are expected to be distinct from genuine genome changes, and therefore screening them
out will reduce the number of putative SNP conversions to investigate. To this end, we estimated the ADO and ADI rates before and after implementation of the RF and SureTypeSC algorithm using the transition matrices of the posterior probabilities (Table 5.6). Table 5.7 shows the reduction in both ADO and ADI rates when implementing the SureTypeSC algorithm, over the default GenCall or RF alone. The fact that ADO and ADI rates are similar between default GenCall (0.15) and the RF (3% and 10-12%, respectively), demonstrates that the GDA layer of the SureTypeSC algorithm is responsible for filtering these erroneous calls. The limitation of the Gaussian filter is the reduction in call rate, from 83% using default GenCall to 75% with SureTypeSC. Thus, the improvement in correct genotype detection in the single-cell environment comes with a certain data loss.

Table 5.7: Allele drop-in, allele drop-out and call rate with GenCall and SureTypeSC

<table>
<thead>
<tr>
<th></th>
<th>Minimal QC (GenCall 0.01)</th>
<th>Default GenCall (0.15)</th>
<th>Random Forest (Cutoff 0.5)</th>
<th>SureTypeSC (Cutoff 0.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI</td>
<td>0.05</td>
<td>0.03</td>
<td>0.03</td>
<td>0.007</td>
</tr>
<tr>
<td>ADO</td>
<td>0.14</td>
<td>0.12</td>
<td>0.1</td>
<td>0.07</td>
</tr>
<tr>
<td>Call rate</td>
<td>0.93</td>
<td>0.83</td>
<td>0.85</td>
<td>0.75</td>
</tr>
</tbody>
</table>

5.3.3 PCR-free NGS yields high genome coverage for single nucleotide variant calling in single cell MDA products

Sequencing libraries were prepared from MDA products of eight oocyte-PB trios (24 single cells), with an average MDA yield (µg) ± S.D of 30.09 ± 9.33 which is in line with amplified single cells from cell lines (Figure 5.2B). An additional 2 single cell MDA products from cell line GM12878, and a single bulk gDNA sample also from cell line GM12878 (as used for algorithm validation) were also included in NGS library preparation. Following mechanical DNA shearing and 1.6× SPRI clean-up, bioanalyzer analysis was performed on a cross-section of samples. An average fragment size of 508 bp (38.50% CV) was observed for single cell MDA products, and 469 bp (39.7% CV) for a single GM12878 gDNA sample following 55 seconds of sonication Figure 5.7. The initial fragmentation of input material is a crucial step in preparation of functional libraries for sequencing on Illumina platforms. The importance of in-process quality control using bioanalyzer is to check the input material, and later, size selection of functional libraries (following addition of adapters). If the fragment length is too long, we see poor bridge amplification of clusters. If the library pool contains a population of short fragments, these will bind with higher affinity to the flow cell at the nominal concentration, and therefore result in
over-clustering. Over-clustering leads to lower quality scores (Q-score) indicating a higher error rate per base call, as the NGS imaging systems struggle to identify individual clusters.

Following end repair, A-tailing and adapter ligation, a double-SPRI size selection enriched for a fragment length of 485 bp (350 bp fragment target + 135 bp adapter sequences) in the library pools. The samples were consistent, showing an average of 678 bp (23.77% CV) for single cells and 618 bp (30.10% CV) for gDNA when assessed by bioanalyzer (which tends to overestimate fragment length) (Figure 5.8). Following qPCR with an estimated median fragment length of 485 bp, all 27 libraries were normalised by dilution to 4 nM and pooled. MiSeq analysis of a 10 pM library pool was assessed for sample quality control, all results are shown as average ± S.D. Demultiplexing showed even sample representation 3.7% ± 0.6%, with 1.09 ± 0.16 million clusters per sample, of which 87.2% ± 0.6% passed filtering. A high alignment rate to human genome hg19 was observed with 97.7% ± 0.4% and 93.9% ± 2.5% aligned reads for Read 1 and Read 2, respectively. The base quality was summarised by 93.83% ≥ Q30. The median insert size was 428 ± 6.7 bp with a range of 149 ± 7.6 to 807 ± 18.2. Taken together the sequencing libraries showed expected performance, with sufficient fragment length to achieve 2×151 bp sequencing for maximum coverage.

As the actual average fragment size was higher than expected (428 bp + 135 bp adapters = 563 bp), the concentrations of each library were re-calculated using the sample-specific insert sizes obtained by MiSeq QC. A fresh 4 nM dilution of each library was prepared, and 14 samples were pooled for deep sequencing; oocyte-PB trios 666-5 and 1315-2, GM12878-SC137 and GM12878-gDNA. Overall, NovaSeq analysis of 1 nM pooled libraries generated 2898.65 Gbp of sequence data with >60% clusters passing filter and 87.12% >Q30. Detailed analysis was performed per sample, following demultiplexing and sequence alignment using BaseSpace tools (Illumina Inc.). Initially, I analysed single cells of trio 666-5, GM12878-SC137 and GM12878-gDNA. For single cells, 1.31 ± 0.13 billion reads passed filter per sample, with an alignment rate of 99.47% ± 0.05%, of which 16.34% ± 2.01% reads were marked as duplicates. For genomic DNA, 1.10 billion reads passed filter with 99.27% alignment rate, of which 16.52% were duplicate reads. For single cells this equates to a theoretical read depth of 50.96×, assuming equal coverage of 3.23 billion bases in the human genome for WGA and sequencing. WGA however does not achieve 100% genome coverage, and amplification bias is expected to result in heterogeneous read depths between regions. The actual sequencing depth was assessed directly at target SNPs.
Figure 5.7: Post-fragmentation DNA library quality control using bioanalyzer. Fragment size distribution for two oocyte-PB trios (666-5 and 1316-2), one single cell and one genomic DNA control from cell line GM12878. Data drawn from E17_RCB_NGS_TSTrios_LibraryPrep/E17d_PostFragmentation_Bioanalyzer.
Figure 5.8: Post-library preparation quality control using bioanalyzer. Fragment size distribution for two oocyte-PB trios (666-5 and 1316-2), one single cell and one genomic DNA control from cell line GM12878. At this stage, fragments are comprised of original insert size plus 135 bp of sequencing adapters. Data drawn from E17_RCB_NGS_TSTrios_LibraryPrep/E17e_PostLP_Bioanalyzer.
5.3.4 SureTypeSC accurately predicts putative gene conversions in individual human female meioses

One application of SureTypeSC would be to identify *de novo* mutation or single nucleotide variants in amplified single cells, using the cost-effective SNP microarray platform. The HumanKaryomap-12 array used in this study, contains approximately 300,000 SNP markers with even distribution throughout the genome (Figure 5.9). Single base conversions that deviate from the expected allele inheritance patterns in female meiosis may be indicative of larger structural genome changes such as gene conversions. To assess whether putative base conversions that are identified by SureTypeSC are true positives we validated the events using high depth NGS. To this end, we identified putative SNP conversions following genotype calling with minimal QC (GenCall 0.01) in one oocyte-PB trio (666-5) (Figure 5.10A). In Figure 5.10A, putative gene conversions appear to cluster to distinct blocks throughout the genome. However, this is inherent to the analysis pipeline, since only heterozygous loci are used to infer Mendelian representation of alleles, and therefore the detected signal is limited to the specific recombination pattern of this meiotic trio. Application of SureTypeSC then filtered this down to 10 putative SNP conversion events, of which three co-localised to the q10.22 region (Figure 5.10B). In Figure 5.10B, the putative gene conversions appear to occur close to chromosome centromeres. This may be due to technical issues, or a result *de novo* gene conversions. For example, there may be elevated mutation rates with WGA at centromeres, although there is no evidence for this. The hypothesis that recombination rates are lower near centromeres is derived from looking at live borns. However, we can speculate that recombination at centromeres is associated with chromosome breaks, and therefore chromosome instability. Therefore, these events could represent breaks that were repaired and did not lead to any major instability. Finally, this unexpected result may be due to gene conversions, since preliminary evidence suggests that crossovers occur near centromeres in oocytes, although centromeric crossovers are reduced in live borns (Ottolini et al., 2015). The SNPs were validated from whole-genome, PCR-free NGS data at a target depth of 50× (Table 5.8). Of 10 putative events, four were confirmed by NGS. To our knowledge, this preliminary data demonstrates the first direct detection of gene conversions in human.
Figure 5.9: Distribution pattern and frequency of genome-wide markers on the HumanKaryomap-12 SNP array. Each chromosome is represented by an ideogram, with labels (right). The frequency of SNP markers is represented by blue lines, using a 10 Kb sliding window. ‘Chr’ = Chromosome.

Figure 5.10: Direct detection of gene conversions in an oocyte-PB1-PB2 trio. Green lines represent heterozygous SNP loci that did not return an expected mendelian inheritance pattern of alleles. (A) Detection of SNP conversions in trio 666-5 prior to filtering with SureTypeSC. (B) Filtering with SureTypeSC revealed 10 SNP conversions, indicating putative gene conversions. Note that three co-converting SNPs on chromosome 10 are overlapping.
Table 5.8: Validation of SNP conversions detected using SureTypeSC by NGS. Concordant results in bold.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Chr</th>
<th>Position</th>
<th>SNP array (SureTypeSC)</th>
<th>NGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{sDNA}$</td>
<td>PB1</td>
</tr>
<tr>
<td>rs4659015</td>
<td>1</td>
<td>120145883</td>
<td>AC</td>
<td>AC</td>
</tr>
<tr>
<td>rs12682402</td>
<td>8</td>
<td>38381884</td>
<td>TC</td>
<td>TC</td>
</tr>
<tr>
<td>rs17835873</td>
<td>10</td>
<td>47605782</td>
<td>AG</td>
<td>AG</td>
</tr>
<tr>
<td>rs7074244</td>
<td>10</td>
<td>47608158</td>
<td>AC</td>
<td>AC</td>
</tr>
<tr>
<td>rs11259756</td>
<td>10</td>
<td>47634190</td>
<td>TC</td>
<td>TC</td>
</tr>
<tr>
<td>rs11530438</td>
<td>11</td>
<td>49836893</td>
<td>TC</td>
<td>TC</td>
</tr>
<tr>
<td>rs1045162</td>
<td>17</td>
<td>20903080</td>
<td>AC</td>
<td>AC</td>
</tr>
<tr>
<td>rs9303700</td>
<td>17</td>
<td>34450463</td>
<td>TG</td>
<td>TG</td>
</tr>
<tr>
<td>rs448372</td>
<td>19</td>
<td>22993128</td>
<td>TC</td>
<td>TC</td>
</tr>
<tr>
<td>rs2005736</td>
<td>22</td>
<td>18912119</td>
<td>TC</td>
<td>TC</td>
</tr>
</tbody>
</table>

‘Chr’ = Chromosome

5.4 Study conclusions

In this chapter I have typed over 27 million SNPs from 104 single cells from two independent cell lines and generated high-quality reference genotypes from bulk genomic DNA. A direct comparison of single cell genotypes and bulk DNA genotypes using the default GenCall algorithm revealed a high level of noise in the single cell environment, that would correlate to a high level of false discovery for meiotic genome changes. To address this, I developed an algorithm in collaboration with Ivan Vogel, to distinguish signal from noise in whole-genome amplified DNA when analysed by SNP array. The two-stage algorithm (SureTypeSC) uses a minimal GenCall QC threshold and employs the Random Forest to learn the error pattern from the single cell data. This stage greatly increases recall for heterozygous SNPs (that are required for linkage analysis in meiotic trios), from 69% to 91% over the default GenCall (0.15) alone. The second stage uses GDA to maximise precision at the cost of data loss, resulting in 2-fold reduction in ADO (7%) and a 4-fold reduction in ADI (0.7%). Having high precision makes it feasible to explore rare events across populations of cells. This includes assessing clonal expansion in tumour evolution, lineage tracing, or detecting rare de novo mutations in single cells that are averaged out and lost in bulk analyses (Lu et al., 2012b; Cooper et al., 2015; Chen et al., 2017).

I next tested the application of SureTypeSC for the study of de novo mutation and gene conversions. In the current work, I prepared sequencing libraries from eight oocyte-PB trios with stringent quality control, of which four have been sequenced to an estimated depth of 50x, sufficient for detection of de novo mutation (Besenbacher et al., 2015). To this end, we identified ten putative SNP conversion events in one trio using SureTypeSC, and independently validated
the loci using whole-genome NGS at high depth. Four SNP conversion events were confirmed, showing an early demonstration of the utility of SureTypeSC for detection of genome changes at the single variant level. The limited data available at the time of writing shows promise for single cell analysis, although several improvements should be considered. The probability of identifying SNP conversions in SNP microarray data is limited by the number of features on the array. Therefore, it is not possible to directly compare the incidence of gene conversions with that in the literature, where inheritance patterns were observed between individuals rather than products of single meioses (Williams et al., 2015). The high precision of SureTypeSC genotyping comes at the cost of recall, and therefore we expect to underestimate the incidence of putative gene conversions in our data. To mitigate this, the use of microarrays with higher SNP density would increase the probability of identifying conversion events; the Multi-Ethnic Global-8 array (Illumina Inc.) can analyse 1.7 million features per sample compared to the 300,000 included on the HumanKaryomap-12 array. In addition, confirmation of a single SNP conversion by NGS is not indicative of a larger gene conversion and may be a de novo base mutation. To elucidate this, analysis of the NGS data for oocyte-PB trios should be expanded at the sites of SNP conversions, to look for other co-converting SNPs within the same gene conversion tract. In the single trio analysed here, we have identified three SNPs that are co-converting within a 4 Kb tract on chromosome 10 (Table 5.6), which provides further confidence that we can detect gene conversion tracts.

Taken together SNP arrays are a cost-effective way of screening a large number of single cells when studying heterogenous populations. SureTypeSC offers high-precision screening of microarray data without the need for cluster files trained specifically on single cell data requiring hundreds of individuals to endure genotype diversity at each locus. Instead, the machine learning approach simplifies the training of the algorithm based on features of the WGA and microarray assay rather than the population. The robust algorithm has applications in both basic biomedical research as well as clinical settings such as in preimplantation genetic diagnosis, where reliable genotype information is required from a limited number of cells. The algorithm shows promise for the detection of gene conversion events but requires further validation.
Chapter 6: Discussion

5.5 Addressing the research proposal

The objective of this thesis work was to improve the single cell DNA sequencing capacities in order to investigate the origins of human aneuploidies in eggs and embryos. In Chapters 3 and 4, I investigated segmental aneuploidies and gross structural rearrangements to chromosomes in >2000 trophectoderm biopsies and developed single cell DNA sequencing assay that facilitates both SNP genotyping and low-coverage sequencing on the same MDA sample (Chapter 4). This allowed me to probe the incidence of segmental aneuploidies in human eggs, which is unexpectedly high. Finally, I have developed new technologies to type SNPs with high precision which has allowed us to detect gene conversions directly in human oocytes and their polar bodies. To our knowledge, this is the first direct detection of gene conversion events. Collectively, this thesis demonstrates that technology-driven research allows the exploration of basic features of human genome evolution and genetic health directly in our germline.

5.6 Solving the noise floor in single cell genomics

All the methods described here share a common goal of discerning true biological phenomena from external noise introduced by the handling, amplification and analysis of single cells. In the clinical setting this is critical, since interpretation of the signal may impact treatment decisions. The increased dynamic range of NGS technologies to detect CNVs raises questions over the clinical significance of (low-level) mosaicism, with some groups reporting successful pregnancy outcomes following the transfer of mosaic aneuploid embryos (Greco and Fiorentino, 2015; Munné et al., 2017). Data from mouse models suggests that embryo mosaicism at the blastocyst stage may be corrected during development, whereby aneuploid cells are out-competed by proliferation of euploid cell lines depending on their ratio (Bolton et al., 2016). The current debate suggests that mosaicism may be overestimated in human embryos due to technical noise introduced by the analysis system. This may lead to the discard of potentially viable embryos, thus reducing the selection pool for transfer and reducing the per cycle pregnancy rate. To this end, I independently validated 522 PGT-A reports from one lab, by redrawing and reviewing CNV traces following NGS. I observed a 44% false positive rate for whole chromosome mosaicism per event, of which 91% was accounted for by high noise (Table 3.3). Similarly, I saw a 31% false negative rate for mosaic segmental per event, of which 81% were explained by high noise. These discordances may be explained by the calling criteria used to report mosaic and sub-chromosomal abnormalities. Two international consortia have published guidelines for the
reporting of embryos following PGT-A (PGDIS, 2016; CoGEN, 2018), yet local differences may also exist between labs. In addition, these guidelines can only be implemented if the signal-background ratio is sufficiently high. My study was limited by the lack of ‘truth’ data, secondary biopsy or whole embryo analysis to confirm mosaicism, however the variability in the interpretation of CNV charts was striking; where assessment of the noise floor, and ability to make any call, is somewhat subjective. Current NGS technology has been shown to detect segmental aneuploidy at 10 Mb (Vera-rodríguez et al., 2016) and low level mosaicism (as determined using cell mixture models) (Greco and Fiorentino, 2015). However, to my knowledge no manufacturer has developed an assay specifically claiming to detect such events, because single (or low number) cell analysis is a sensitive procedure that may be influenced by the sample quality. For analytical methods that are distributed for Research Use Only, clinical laboratories are required to conduct their own independent validation of a protocol, which they may choose to modify to suit local needs. Therefore, caution should be taken to compare the reported incidence of mosaic and segmental abnormalities between research groups, where high resolution and low noise are required, yet genome coverage may vary depending on the local implementation. By extension, the same expectation may be applied to other assisted reproductive technologies (ART). Protocols for ovarian stimulation, embryo culture (media, supplements and handling), embryo biopsy and vitrification may all introduce noise to PGT-A, which typically manifests as an uninterpretable CNV traces. Elucidating this is challenging because embryo research is limited by the availability of material and lack of standardised procedures. To this end, I decided to further investigate the incidence of segmental aneuploidies directly in oocytes and polar bodies, where a single cell pipeline allows control external aspects of noise but is not without its own challenges.

New pipelines in single cell genomics are required when multiple analytical methods are needed to make inferences about the incidence and origin of discreet genome changes. I have previously described the benefits and limitations of quasi-linear and MDA-based WGA methods, with respect to the trade-off between amplification fidelity, genome coverage and representation (Blanshard et al., 2018). Novel methodologies are making impressive ground on improving uniformity and fidelity of genome amplification from single cells (Chen et al., 2017). At the time of this research, MDA was widely considered the gold-standard for SNP amplification fidelity, and has been independently validated for clinical utility (although labelled as Research Use Only – not for diagnostic purposes) (Natesan et al., 2014). For this work I used the MDA methodology for both SNP genotype analysis and copy number assessment of single cells, with a stringent reduction in the length of the amplification step to 2 hours. I have demonstrated
here that a combination of reduced MDA amplification and increased input into NGS library preparation (100 ng; Nextera) results in a tolerable noise profile for reciprocal-CNV detection in single cell products of meiosis. This is in contrast to comparison studies that benchmark the performance of MDA based on the manufacturer’s recommendation of 8-hour amplification (Huang et al., 2015; Babayan et al., 2016).

The MDA-NGS method offers flexible, cost-effective molecular analysis of single cell DNA, where low-pass NGS provides the resolution to detect (~10 Mb) CNVs (Vera-rodríguez et al., 2016) and SNP genotyping by microarray allows inferences to be made about the origin of chromosomes by maternal homolog (Ottolini et al., 2015). I used a paired-end (2×36 bp) sequencing approach to improve sequence alignment and to mitigate the effect of telomeric CNV ‘drop-off’ due to flagging of highly repetitive sequences as PCR duplicates. In theory, the resolution of NGS could be improved by higher depth sequencing (e.g. 2× average coverage), with increased read length (e.g. 2×100 bp). However, this development would increase the cost per sample, and require additional bioinformatics support to classify reads for inclusion into bin counts. Following validation of the method using cell lines, the per sample concordance was 91.7% (n=48), which was lower than I expected. I detected 4 false negatives for expected chromosome abnormalities and 3 segmental abnormalities that were false positive against the expected karyotype (Table 4.3). I believe that further validation is required to elucidate the true resolution of the assay. Three of the false negatives were found in one cell line with a 10.3 Mb segmental loss, which may indicate a limit of detection in this region of the genome. With only 12 single cell replicates per cell line, the sensitivity may be underestimated. On the other hand, the specificity of the pipeline is limited by the use of unsynchronised cell lines for validation. The effect of S-phase replication may give rise to sub-chromosomal CNVs of mitotic origin, which would manifest as higher noise in CNV chart traces after WGA and NGS. Oocyte and polar body samples (from stimulated cycles; Clinic A) show lower noise (DLR) following NGS, when compared to cell lines (Figure 4.9). This is likely because meiotic cells are analysed after bulk replication has completed and will not show replication-associated noise seen in mitotic cells. Alternatively, segmental aneuploidies may result from de novo chromosome breakage as a factor of cell culture and manipulation. In my analysis of meiotic duos and trios, I mitigated the effect of elevated noise by stringent QC, to only include events where I saw a reciprocal abnormality in a matched cell. Whilst this approach may result in an underestimation of incidence of chromosome abnormalities, we can be confident of true events that can in turn provide further information regarding their origin. To this end, MDA is prone to stochastic amplification bias that introduces noise for CNV detection in single cells. By using limited
amplification, paired-end sequencing and appropriate QC, my data show that sub-chromosomal abnormalities can be identified and accurately characterised, although with some data loss. The benefit of MDA however is the high precision genotype information that can be analysed in parallel.

SNP microarrays allow cost-effective generation of high-confidence genotypes, limited only by the number of features included in array design. Single-base extension of oligonucleotide probes using a two-dye chemistry with high redundancy (~15-30×) (Infinium®, Illumina Inc.) allows accurate genotyping, provided that the signal-background ratio remains high. Our SureTypeSC application is another example of how we used technology-driven research to mitigate the single cell noise effect of WGA, namely allele drop out (ADO) and allele drop in (ADI). Single cell WGA noise profiles are typically handled by algorithms that detect large scale genome changes (chromosome recombination), where hundreds of SNPs are used to infer the haplotype phase within trios (Handyside et al., 2010; Ottolini et al., 2015). On the other hand, gene conversions are characterised by the non-reciprocal exchange of ~1000 bp tracts between homologs, with a frequency that is one order of magnitude higher than crossovers (Williams et al., 2015). These tracts may manifest as single SNP conversions that do not follow the Mendelian inheritance pattern of alleles across meiotic duos/trios. SureTypeSC has significant benefit for screening large populations of cells, in addition to improved genotyping when analysing precious material, such as for preimplantation genetic diagnosis of embryos.

5.7 Limitations of the work
The occurrence of mosaicism within preimplantation embryos, and the challenges associated with clinical interpretation have been well reviewed (Taylor et al., 2014; Capalbo et al., 2016; Vera-Rodriguez and Rubio, 2017). Whilst the trophectoderm is typically representative of the inner cell mass (Capalbo et al., 2013b), on a per embryo basis and PGT-A result is still only an interpretation of the biopsy, not the whole embryo. Whilst some groups indicate that the level of detected mosaic correlated with the probability of clinical pregnancy (Munné et al., 2017), differences in the number of cells biopsied (3-10 cells) between embryologists may alter this detected proportion of overall mosaicism significantly. The incidence of mosaicism is therefore challenging to interpret between different clinics and testing labs. In this work, the incidence of mosaicism and segmental aneuploidy in preimplantation embryos cannot be separated from noise introduced from the pipeline, because we lack secondary molecular analysis and pregnancy outcome data for each embryo. In addition, PGT-A labs typically receive samples from
multiple clinics, and I am blind to some aspects of sample handling that may influence NGS noise; such as number of clinics involved, number of embryologists, shipment and storage conditions of the samples prior to analysis. Therefore, the data shown here is the incidence of abnormality reporting, not necessarily the biological incidence. In turn, PGT-A reports may influence the decision to transfer or discard embryos and from this I have drawn conclusions on clinical significance (affecting 5% of cycles).

The use of MDA to make copy number assessments is limited by the non-linear amplification of DNA, which is exaggerated with extended incubation time; however, my work has mitigated this with good effect, and the value of concurrent SNP analysis at reduced cost is sufficient to warrant its use. Further validation is required however, to elucidate the limit of detection of the assay with regards to Mb resolution. This work was limited by the availability of bioinformatics support to optimise the read filtering and alignment of longer reads (without the use of BlueFuse® Multi software). In addition, the MDA reaction could be further improved to reduce the effect of stochastic WGA bias by reducing the reaction volume by emulsion (Hosokawa et al., 2017). The collection of oocytes from unstimulated ovaries, appeared to produce biopsied single cells with higher NGS noise than oocytes from stimulated cycles and cell lines (Figure 4.9D). The pipeline is still new (Gruhn et al., 2018), and therefore may require further optimisation to ensure sample quality.

5.8 Conclusions, wider implications and future work

Following the development of novel tools for single cell genomics, our next steps are to apply them to explore genome changes in oocytes and polar bodies. I have demonstrated the CNV analysis of segmental aneuploidies in a subset of meiotic duo/trios, from which I extracted breakpoint positions. Our group are currently working to map these locations to known fragile sites (Mrasek et al., 2010; Fungtammasan et al., 2012) and recombination hotspots (Kong et al., 2010; Casper et al., 2018). The incidence of GCRs in preimplantation embryos and oocytes (~10%) is higher than expected in live births (<1 in 20,000), and typically only seen in transformed cell lines or following induction by DNA damaging agents or X rays (Lindsley et al., 1972). Ongoing work with human oocytes has shown heterogenous expression of DNA repair genes, which may be a cause of genome instability that gives rise to GCRs. Indeed, reduction of SMC6 dosage leads to elevation in GCRs in mouse oocytes (Hoffmann Lab; submitted). Therefore, we conclude that human oocytes are highly heterogenous and intrinsically unstable with respect to sub-chromosomal aneuploidy.
To date we have applied the SureTypeSC tool to one meiotic trio. Preliminary data has shown that single SNP conversions can be detected and validated using high depth NGS. During this work I prepared eight meiotic trios (24 single cell MDA products) for PCR-free sequencing, of which four trios have been sequenced. Our next steps are to further identify and validate SNP conversions in the remaining trios. The work of Williams et al. (2015) involved the use of higher density SNP microarrays (660K to 1 million features) that will have increased the discovery rate considerably over the 300K array used here. I also processed single cells and genomic DNA samples from cell line GM12878 using the Multi-Ethnic Global-8 array (Illumina Inc.) that includes 1.7 million markers (not included here). These data may be used for cross validation of shared SNP loci, or if necessary, used to retain the SureTypeSC tool for this platform. Higher density SNP arrays increase the likelihood of discovery of co-converting SNPs within close proximity, thus increasing the confidence that gene conversion has occurred.
References


Bartholin T. *Historiarum anatomicarum rariorum centuria III et IV. Ujusdem cura accessare observationes anatomicae*. 1656;


Berridge MJ. Inositol trisphosphate and calcium signalling mechanisms. *Biochim Biophys Acta* [Internet] 2009;**1793**:933–940. Elsevier B.V.


Capalbo A, Ottolini CS, Griffin DK, Ubaldi FM, Handyside AH, Rienzi L. Artificial oocyte activation with calcium ionophore does not cause a widespread increase in chromosome segregation errors in the second meiotic division of the oocyte. *Fertil Steril* 2015a;103:807–814.


Capalbo A, Wright G, Elliott T, Ubaldi FM, Rienzi L, Nagy ZP. FISH reanalysis of inner cell mass and trophectoderm samples of previously array-CGH screened blastocysts shows high accuracy of diagnosis and no major diagnostic impact of mosaicism at the blastocyst stage. 2013a;28:2298–2307.


United States.


Hassold T, Hall H, Hunt P. The origin of human aneuploidy: where we have been, where we are going. *Hum Mol Genet* [Internet] 2007;16 Spec No:R203-8.


164


Heindryckx B, Elst J Van der, Sutter P De, Dhont M. Treatment option for sperm- or oocyte-related fertilization failure: Assisted oocyte activation following diagnostic heterologous ICSI. *Hum Reprod* 2005;**20**:2237–2241.


Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and


Illumina Technical Note. Genome-Wide DNA Analysis BeadChips. 2009; Illumina, Inc., CA, USA.


Lamb N, Freeman S, Savage-Austin A. Susceptible chiasmate configurations of chromosome 21 predispose to non–disjunction in both maternal meiosis I and meiosis II. *Nat Genet* [Internet] 1996;14:400–405.


Li H, Handsaker B, Wyssker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The


McCoy RC, Demko ZP, Ryan A, Banjovic M, Hill M, Sigurjonsson S, Rabinowitz M, Petrov DA.


Nagaoka SI, Hassold TJ, Hunt P. Human aneuploidy: mechanisms and new insights into an age-


PGDIS. PGDIS position statement on chromosome mosaicism and preimplantation aneuploidy testing at the blastocyst stage. 2016;Available from:


Vera-Rodriguez M, Rubio C. Assessing the true incidence of mosaicism in preimplantation


Vogel I, Blanshard RC, Hoffmann ER. SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data. *bioRxiv* [Internet] 2018; Cold Spring Harbor LaboratoryAvailable from: https://www.biorxiv.org/content/early/2018/08/20/393256.


Appendix A1.1

Single cell genomics to study DNA and chromosome changes in human gametes and embryos.

Blanshard, R.C., Chen, C, Xie, X.S. and Hoffmann, E.R.

Methods in Cell Biology, 2018;144:441–457.
Single cell genomics to study DNA and chromosome changes in human gametes and embryos

Robert C. Blanshard*1, Chongyi Chen2, Xiaoliang Sunney Xie1,4,1
Eva R. Hoffmann*1,1

*Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, United Kingdom
1Clinical Genomics Group, Illumina Inc., Fulbourn, United Kingdom
1Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, United States
3Beijing Advanced Innovation Center for Genomics, Beijing, China
5Center for Chromosome Stability, University of Copenhagen, Copenhagen, Denmark
1Corresponding authors: e-mail address: xie@chemistry.harvard.edu; eva@sund.ku.dk

CHAPTER OUTLINE

1 Introduction.........................................................................................................................442
2 Methods .............................................................................................................................443
2.1 Design of a Single Cell Genomics Experiment .........................................................443
2.2 WGA Methods Comparison .......................................................................................446
2.3 Sample Collection ........................................................................................................448
2.3.1 Isolating Single Cells From Cell Lines ..................................................................448
2.3.2 Isolating All Three Products of Female Meiosis: Human Oocytes,
Embryos, and Polar Bodies.................................................................................................450
2.4 Whole Genome Amplification......................................................................................450
2.4.1 WGA by SureMDA.................................................................................................450
2.4.2 WGA by LIANTI......................................................................................................451
2.4.3 QC by Gel Electrophoresis.........................................................................................451
2.4.4 QC by Quantitation..................................................................................................451
2.5 NGS and Microarray Analysis.....................................................................................452
2.5.1 NGS for CNV..........................................................................................................452
2.5.2 NGS for Variant Calling.........................................................................................452
2.5.3 Whole-genome Genotyping by SNP Array............................................................453
Abstract
Genomic and chromosomal changes occur at a high rate in the germline and preimplantation embryos. To study such changes directly in the germline of mammals requires access to material as well as single cell genomics. Recent improvements in embryology and single-cell DNA amplification make it possible to study the genomic changes directly in human oocytes, sperm, and preimplantation embryos. This is particularly important for the study of chromosome segregation directly in human oocytes and preimplantation embryos. Here, we present a practical approach how to obtain high quality DNA sequences and genotypes from single cells, using manual handling of the material that makes it possible to detect genomic changes in meiosis and mitosis spanning the entire range from single nucleotide changes to whole chromosome aneuploidies.

1 INTRODUCTION
Genomic changes in the germline and somatic cells provide the genetic variability that drives immune diversity, tumor evolution, and germline evolution. Being able to study genome changes in single cells allows the deconvolution of ensemble population studies and therefore exploration of the true genetic states and heterogeneity of individual cells. The last decade has seen tremendous technological advances in probing DNA changes in single cells, referred to as single cell genomics (Gawad, Koh, & Quake, 2016; Huang, Ma, Chapman, Lu, & Xie, 2015; Zhang et al., 1992). Current technologies now make it possible to detect genomic changes ranging from single base mutations, to whole chromosome gains and losses (aneuploidy). With statistical methods to phase haplotypes, it is also possible to assess chromosome recombination and infer segregation patterns in meiosis, the specialized cell division process that generates gametes (Hou et al., 2013; Lu et al., 2012; Ottolini et al., 2015; Wang, Fan, Behr, & Quake, 2012).
Being able to explore genetic diversity of single cells together with advances in embryology has revealed new facets of chromosomal changes in meiosis that can explain genetic features of the human population, as well as those that impinge on reproductive health in women as they age (Capalbo, Hoffmann, Cimadomo, Maria Ubaldi, & Rienzi, 2017). Two studies of adult human oocytes have revealed the importance of recombination for chromosome segregation and thus prevention of aneuploidy (Hou et al., 2013; Ottolini et al., 2015). Detecting both sequence and copy number variation (CNV) in the same cell also allowed the identification of a new aberrant segregation pattern termed reverse segregation (Ottolini et al., 2015). Single cell technologies have also enabled de novo detection of complex chromosomal arrangements (chromothripsis) (Zhang et al., 2015), replication fork timing in individual cells (Chen et al., 2017), as well as mutation signatures.

In human preimplantation embryos, biopsies of one blastomere or 5–10 cells of the trophoderm lineage, which gives rise to the placenta, are being used to assess chromosome loss or gain (aneuploidy) as well as mosaicism and structural gains and losses, to improve overall clinical outcomes for patients (Vermeesch, Voet, & Devriendt, 2016). However, this also provides us with an opportunity to assess genome changes and cellular fate in early embryos (Bolton et al., 2016). Common to single cell (or low input) genome sequencing is the overall requirement to distinguish real genetic changes from noise that is introduced by the technology, especially during the whole genome amplification (WGA) step. Technological and statistical advances continue to improve and push the limits of detection (Chen et al., 2017). Here, we discuss the considerations to be taken into account when designing a single cell genomics experiment, the use of linear amplification via transposon insertion (LIANTI), a novel WGA method, and human female meiosis as a special focus.

2 METHODS

2.1 DESIGN OF A SINGLE CELL GENOMICS EXPERIMENT

The design of a single cell genomics experiment is critical and depends on the genomic changes that are being detected. In general, the work flow consists of four stages that influence subsequent processing: single cell isolation, WGA, sequencing, and data analysis (Fig. 1). Single cell sequencing for the purpose of detecting aneuploidy requires high reproducibility of WGA, but low coverage (0.01 ×) is sufficient to detect relative changes in chromosome numbers. In contrast, genotyping requires a high read depth for calling variants (30 × is preferred), which is expensive. An alternative to next-generation sequencing (NGS)-based variant calling, microarray approaches allow a cost-effective approach to detect highly characterized single-nucleotide polymorphisms (SNPs). However, the overall detection is limited by the number of features on the chip (e.g., 1.7 million on the Multi-Ethnic Global array (Illumina Inc.)). Both sequencing and microarrays have proven useful for detecting recombination events and inferring chromosome segregation patterns in human female meiosis (Hou et al., 2013; Ottolini et al., 2015) but require expert data
FIG. 1
Single cell genomics work flow.
analyses and informatics. Below, we focus on the two single cell pipelines used in our labs that utilize LIANTi and multiple displacement amplification (MDA)-based WGA, followed by NGS or microarray detection. We emphasize that other technologies can of course also be used and to illustrate this, we have made the methods described below as generic as possible.

Single cell isolation depends on the cell type and ranges from manual handling, such as for human oocytes and embryos, to microfluidics and emulsion approaches for cell suspensions (Fig. 1). Manual handling is time consuming and requires expert skills when obtaining single cells from rare material such as embryos and oocytes. In contrast, fluorescence-activated cell sorting (FACS), emulsion, and microfluidics approaches of cell suspensions are usually quick and relatively straightforward, allowing the parallel processing of thousands of cells. However, capturing multiple cells is a problem (doublets), and particularly the emulsion approach requires a large number of cells, because capture rates mostly follow poisson distributions (fewer than 10% of cells). “Dead” cells are another issue that influences the quality of the experiment, and we describe methods to exclude such cells both from the initial experiment and from the subsequent data analyses.

The original WGA method was PCR based (Zhang et al., 1992) and continues to provide a cost-effective solution to copy number detection, although genome coverage is limited (Blagodatskikh et al., 2017). Quasi-linear methods such as PicoPlex (Rubicon Genomics) or SurePlex (Illumina Inc.) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC; Yikon Genomics) have a “linear” phase, followed by a limited number of PCR-based cycles. Isothermal reactions such as MDA use relatively high fidelity polymerase (Φ29) with strand displacement activity, resulting in branched structures and high yields. When carried out in low volumes in microfluidics devices or in emulsion, MDA can detect de novo mutations with high precision and gives rise to an even coverage across the genome (Fu et al., 2015; Wang et al., 2012). MDA is widely used and has the advantage that the polymerase Φ29 has a low error rate and possesses strand displacement activity. Quasi-linear approaches such as PicoPlex or MALBAC are also available and used in several clinical applications, especially to detect CNV (Yan et al., 2015).

The most recent development in WGA is the use of a modified Tn5 transposon to fragment the single-cell genome and insert T7 promoters, from which T7 RNA polymerase can transcribe in vitro (LIANTI) (Chen et al., 2017). Transcription, not PCR or other DNA polymerases, thus amplify the genome in a linear fashion. LIANTI yields the best coverage yet in single-reaction amplification. Whereas LIANTI performs better than MDA and MALBAC, the yield is lower (3 μg compared to 25 μg), which may affect downstream processing. Quality control (QC) for WGA can be done directly after the reaction for MDA, where yields are high. For LIANTI, the QC is performed directly after the reaction, based on the yield, and again after library preparation based on quantitation of functional libraries by qPCR results, before loading to sequencers. Prior to sequencing or genotyping by microarrays, the amplified DNA must be processed to obtain an optimal length.

For NGS, adapters/linker sequences (known as indexes) are added for library multiplexing and flow cell attachment. There are several different technologies with some
that rely on mechanical shearing of the DNA (e.g., TruSeq DNA PCR-Free Library Preparation; Illumina Inc.) and others that use additional PCR-based amplification steps (e.g., Nextera DNA Library Preparation; Illumina Inc.). Microarray-based libraries are generated by enzymatic fragmentation. For both NGS and microarray-based genotyping, the data analyses associated with single-cell genome sequencing and genotyping are challenging since some algorithms need to be adapted for single cell applications. There are a variety of variant callers currently available, and the development in this area is moving very fast. Below, we give examples of two workflows from human female meiosis using microarray-based technologies and single cell sequencing using LIANTI to detect recombination and infer chromosome segregation.

2.2 WGA METHODS COMPARISON

Fig. 2 shows the comparison of single-cell amplification evenness between MDA, PicoPlex or MALBAC, degenerate oligonucleotide-primed (DOP)-PCR, LIANTI, and bulk sequencing. MDA relies on φ29 DNA polymerase, which has high replication fidelity generating accurate single-cell single nucleotide variant (SNV) detection, but CNV detection is limited due to its exponential amplification noise. PicoPlex or MALBAC uses a quasi-linear amplification scheme to achieve better amplification evenness than MDA, which, together with normalization of the systematic genome-wide amplification noise, enables single-cell CNV detection with megabase resolution. DOP-PCR has a similar level of amplification noise, but at the cost of a very low genome coverage. The linear amplification scheme of LIANTI offers the highest amplification evenness and fidelity of the whole genome, enabling single-cell micro-CNV detection with kilobase resolution, as well as accurate single-cell SNV.

![Amplification evenness comparison between single-cell WGA methods. Plotted are read depths across the genome with 10-kb bin size (only a 10-Mb region in Chromosome 1 shown in the plot).](image-url)
detection. LIANTTI has been used to detect the stochastic DNA replication origin firing and replicon formation in single human cells in S-phase (Fig. 3) (Chen et al., 2017), and to detect micro-CNVs (microdeletions) and trisomy in human embryos with a variety of genetic disorders (Fig. 4) (Wapner, et al., 2012).

**FIG. 3**
DNA replication origin firing and replicon formation. Plotted are genome-wide CNVs detected by LIANTTI in 11 single S-phase BJ cells with 10-kb bin size (only X chromosome shown in the plot).

**FIG. 4**
Frequencies of microdeletions and trisomy in human embryos.
The needs of different applications are often met by different single-cell WGA methods. A comprehensive comparison of DOP-PCR, MDA, PicoPlex or MALBAC, and LIANITI has been provided (Table 1).

2.3 SAMPLE COLLECTION

2.3.1 Isolating single cells from cell lines

1. Cells may be isolated manually by micropipetting, FACS analysis, or other methods. Please note that the cell cycle stage may affect results, and synchronization may be needed. Make two buffers 1 × PBS + 0.1% PVP (Origio, 10905000) and 1 × PBS (Cell Signalling). Expose the buffers and the collection tubes (0.2-ml thin-walled PCR tubes with closed lids in a rack; handle with metal forceps) to UV-C for ≥30 min. For FACS, use a skirted 0.2-ml plate (UV treated) and ensure that wells are sealed promptly after collection (in a Class II Microbiological Safety Cabinet).

2. Harvest 1.0 × 10^6 cells and perform an Annexin V-based Dead Cell Removal (Miltenyi Biotec, 130-090-101) according to the manufacturer’s instructions. Perform a viability assessment by staining with 0.4% Trypan blue. Viability of ≥98.0% is recommended. Centrifuge cell suspension at 100 × g in a microfuge for 5 min and wash pellet twice with 1 ml 1 × PBS + 0.1% PVP. The cells can be stored at 4°C for up to 5 h.

3. For manual isolation, use a stripper pipette (e.g., Origio) set at the WGA volume to handle individual cells and use a stereomicroscope in a hood. Use a sterile dish (e.g., Nunc, ICSI dish) and prepare 2–3 drops of 10 μL 1 × PBS + 0.1% PVP for diluting the cells. Isolate individual cells into 6 μL 1 × PBS + 0.1% PVP. Isolate 3–5 cells (positive control) and an empty drop (negative control). Wash the tip in a fresh drop of 1 × PBS between each cell. Include negative controls at several points and at the end. Check that all drops contain single cells and then remove cells in a fresh drop of 6 μL 1 × PBS + 0.1% PVP, pipetting the cell up and down. Dispense entire volume; move the tip away from the cell; reverse pipette the transfer volume (0.5–1 μL, pending WGA); pick up the cell at the very tip. This allows efficient flushing of the cell when dispensing into the WGA tube (e.g., the final volume for MDA is 4 μL). Dispense the cell directly into the solution into the collection tube without touching the sides. Rinse the tip under the microscope—no cell should be observed if the transfer is successful.

4. For FACS, wash the cell pellet in 1 × PBS and label the cells with Molecular Probes™ Dead Cell Apoptosis Kit (Fisher Scientific, 10257392) immediately prior to sorting to collect viable cells. Select appropriate gating for positive, negative, and single cell collection and prime the 0.2-ml tubes with the required volume of 1 × PBS. Determine the rate of two more cells if using a 96-well plate. These should not be used. The failure rate is about 10–15% (no cells).

5. Centrifuge the sample tubes/plates at 280 × g for 1 min, then proceed immediately to WGA. Alternatively, store samples at −80°C for no more than 24 h.
<table>
<thead>
<tr>
<th>WGA Method</th>
<th>Genome Coverage (%)</th>
<th>SNV Accuracy</th>
<th>SNV FPR</th>
<th>CNV Accuracy</th>
<th>CNV Spatial Resolution</th>
<th>Amplification Principle</th>
<th>Amplification Enzyme</th>
<th>Kit Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOP-PCR</td>
<td>45</td>
<td>Low</td>
<td>$2 \times 10^{-4}$</td>
<td>High</td>
<td>Megabase</td>
<td>Exponential</td>
<td>PCR DNA polymerase</td>
<td>Yes</td>
</tr>
<tr>
<td>MDA</td>
<td>87</td>
<td>High</td>
<td>$(1-2) \times 10^{-5}$</td>
<td>Low</td>
<td>Chromosome level</td>
<td>Exponential</td>
<td>Phi29 DNA polymerase</td>
<td>Yes</td>
</tr>
<tr>
<td>PicoPlex or MALBAC</td>
<td>73</td>
<td>Low</td>
<td>$(1-4) \times 10^{-4}$</td>
<td>High</td>
<td>Megabase</td>
<td>Quasilinear</td>
<td>PCR DNA polymerase</td>
<td>Yes</td>
</tr>
<tr>
<td>LIANTI</td>
<td>95</td>
<td>High</td>
<td>$(2-5) \times 10^{-6}$</td>
<td>High</td>
<td>Kilobase</td>
<td>Linear</td>
<td>T7 RNA polymerase</td>
<td>No</td>
</tr>
</tbody>
</table>
2.3.2 Isolating all three products of female meiosis: Human oocytes, embryos, and polar bodies

To obtain the genetic information from female meiosis, human oocytes and matching polar bodies (PB) are biopsied (Ottolini et al., 2016, 2015). The PBs should be biopsied sequentially to ensure that PB1 and PB2 are distinguishable a priori, and to prevent or limit degradation of in the PB1 prior to DNA amplification. The PB1 is removed from the meiosis II (MII)-arrested oocyte using a laser-assisted microscope and immediately tubed. The MII oocytes are activated by artificial activation using 40 min exposure to 100 μM A23187, a Ca²⁺ ionophore that triggers the second meiotic division and extrusion of the PB2. A single female pronucleus should form. Alternatively, the MII oocytes can be fertilized by sperm using intracytoplasmic sperm injection (ICSI). This generates a zygote, and the female pronucleus can be removed (Hou et al., 2013). All of these procedures require extensive handling skills of embryo and oocytes, and should be carried out by trained embryologists under appropriate ethics permissions and informed consent.

The biopsied cells are transferred into an 0.2-mL tube for WGA using a stripper pipette into the appropriate volume of 1 × PBS. The PBs are very small and washing them in PBS or media does not improve performance; however, the volume of media transferred with the cells should be minimized.

2.4 WHOLE GENOME AMPLIFICATION

WGA reactions should be assembled in a clean, dedicated pre-PCR work area, preferably under HEPA-filtered air flow via a Class II Microbiological Safety Cabinet (https://www.labogene.com) to minimize the risk of external contamination. Clean the work area, pipettes, and tube racks with a nucleic acid decontamination reagent and allow to dry, prior to exposure of equipment and consumables to UV-C for ≥30 min.

Alongside the cell isolation/biopsy controls, prepare replicates of 16–30 pg genomic DNA-positive controls in 1 × PBS, and 1 × PBS no template controls in 0.2-mL thin-walled PCR tubes, with a final volume matching that of the cell/biopsy collection. Immediately before use, remove the cell or biopsy samples from fresh or frozen storage and centrifuge at 300 × g for 3 min; immediate lysis is critical for the quality of the WGA. When adding reagents, it is critical to dispense the reagent above the sample droplet, onto the inner wall of the tube, to mitigate the risk of removing the sample cell or DNA when withdrawing the pipette tip. Change pipette tips between samples. Centrifuge sample and control tubes at 300 × g for 5 s to collect reagents at the bottom of the tube. Exponential amplification on the thermal cycler must be performed in a dedicated post-PCR area. Store amplified products at −20°C or −80°C. For best results, process samples within 1 week.

2.4.1 WGA by SureMDA

A number of MDA kits are commercially available. The SureMDA Amplification System (SureMDA; Illumina Inc.) has been QC tested by Illumina for the purpose of Karyomapping. We use a shortened amplification protocol than specified with
the REPLI-g® Single Cell Kit (QIAGEN) for our research. In brief, the reaction should be assembled according to the manufacturer’s instructions with the following modifications.

1. Handle samples on liquid-sealed –20°C cold blocks. Do not use wet ice, as this increases the risk of external contamination.
2. Following cell lysis, addition of the stop solution should be performed swiftly. When processing large sample sizes, dispense 3 μL stop solution onto the inner wall of the tube or well, and tap the plate periodically to collect the reagents at the bottom of the tubes or wells, prior to the final centrifugation.
3. Critically, we reduce the amplification time at 30°C for 2 h with SureMDA. This generates a yield of 20–25 μg and reduces artifacts.

2.4.2 WGA by LIANTI
The LIANTI assay should be assembled according to published protocols (Chen et al., 2017). Several critical points worthy of note are:

1. Transposome is made by mixing equal volume of 1.5 μM annealed LIANTI transposon DNA and ~1 μM Tn5 transposase, which is prepared in house or purchased from Epicentre (Illumina Inc. via Lucigen Corp.).
2. LIANTI transposome is made freshly every time before use, since long time storage of transposome may lead to degradation.
3. The steps before in vitro transcription should be carried out in a dedicated clean setting, to minimize the risk of contamination.

2.4.3 QC by gel electrophoresis
Successful amplification can be determined by gel electrophoresis. The long fragments generated by MDA should be resolved on a 0.8% agarose gel against a 1-kb extension ladder. A bright smear should be seen around 3–7 kb for our MDA protocol. The shorter 100–1000 bp products generated by PicoPlex and LIANTI should be resolved on a 1.5% agarose gel against a 1-kb ladder.

2.4.4 QC by quantitation
Accurate quantitation of input DNA for NGS and microarray analysis is essential for achieving optimal fragment size distribution during library preparation. A fluorometric-based assessment of MDA products for duplex DNA should be performed (Quant-it™ dsDNA High-Sensitivity (HS) Assay Kit, Invitrogen™). Perform a serial dilution of the WGA products in Molecular Biology Grade (MBG) dH2O to within the specified range, using at least 5 μL DNA for each step of the dilution. Quantify the final diluted WGA products according to the manufacturers’ instructions, using a microplate reader (FLUOstar® Omega; BMG Labtech) and appropriate software (MARS Data Analysis Software; BMG Labtech).
2.5 NGS AND MICROARRAY ANALYSIS

Selection of NGS vs microarray methodology should be based on output. This likewise determines the library preparation approach used. For high precision and variant calling, mechanical shearing and adaptor ligation is preferred, whereas CNV can readily be detected from low input volumes from PCR-based library preps. The hands-on time and required equipment needed to generate PCR-free libraries vs PCR-based libraries differ dramatically. High precision genotype calling at a lower density can be achieved by SNP microarrays at a much lower cost compared to the sequencing depth required for untargeted NGS-based genotyping.

2.5.1 NGS for CNV

CNV analysis can be performed on products from all WGA methods described here; however, the library preparation method for sequencing may vary depending on the input configuration. The main principles of library preparation for NGS comprise sample DNA fragmentation, library specific indexing for multiplexing, fragment size selection, library quantification, normalization, and pooling. CNV detection is performed by binning aligned sequence reads across the genome, and comparing the read counts against a euploid reference. Accurate copy number analysis requires a complete understanding of the noise and bias profiles of the WGA system. Amplification bias is well tolerated provided that it is systematic and not stochastic. There are several pipelines for CNV analysis (e.g., VeriSeq® from Illumina Inc. is a commercially available pipeline that uses SurePlex). For LIANTi, in addition to conventional CNV callers relying on read depth (e.g., Control-FREEC, BoevaLab), specific algorithms are also available to call single-cell CNV from LIANTi data based on the digital counting principle where sequencing reads with the same alignment positions can be grouped (e.g., the “LIANTi” toolkit from https://gitlab.com/lh3/liant).

QC measures for NGS library preparations include quantitation of functional libraries by qPCR (e.g., KAPA Library Quantification Kit for Illumina® Platforms (KAPA Biosystems)) and fragment size distribution (e.g., 2100 Bioanalyzer (Agilent Technologies)).

2.5.2 NGS for variant calling

Whole-genome sequencing at high depth >30 x is required for de novo mutation discovery and comprehensive genotyping. This may be an underestimate for de novo assembly. De novo mutation detection is sensitive to the fidelity of the polymerases used for NGS, and a PCR-free library preparation should be used. The advantage of PCR-based library preparation is a considerably shorter protocol, with input requirements as low as 1 ng DNA, where sequencing indexes are incorporated during an enzymatic fragmentation and reduced-cycle PCR reaction. In contrast, the PCR-free library preparation requires higher input (1–2 μg DNA depending on the desired average NGS insert size, which ranges from 350 to 550bp), mechanical fragmentation, end repair, and adapter ligation prior to sequencing.
The NGS insert size determines the optimal read length, and therefore is crucial in assay design, for ensuring maximum genome coverage with paired-end sequencing.

2.5.3 Whole-genome genotyping by SNP array

The microarray workflow uses a unique on-array, two-color, single base extension (SBE) biochemistry to identify A/T and C/G base incorporation at targeted SNP loci (Infinium BeadChip arrays (Illumina Inc.)). Briefly, the sample genomic DNA or WGA product is further amplified, fragmented, and hybridized to the microarray overnight. Following SBE, the incorporated bases are detected by a two-dye fluorescent signal amplification step, prior to scanning and analysis of intensity data from both red and green channels. Genome coverage is determined by the number of features in the array design, ranging from ~300,000 (e.g., HumanCytoSNP-12; Illumina Inc.) to 1.7 million (e.g., Multi-Ethnic Global-8; Illumina Inc.). For single cell analysis using the HumanCytoSNP-12 array (Illumina Inc.), we follow the Infinium Karyomapping Protocol Guide (15052710; Illumina Inc.) without modification.

2.6 DATA PROCESSING AND INFORMATICS PIPELINES

The two major limitations for single cell genomics are the random nonamplification of one of the alleles (allele drop out) and mutation due to the technology used. To QC a specific single-cell method, we recommend using a cell line that is well characterized to obtain the noise parameters (Eberle et al., 2017). Below, we discuss different pipelines for data analysis.

2.6.1 Data analysis pipelines for NGS

Sequencing the whole human genome at high depth generates a significant amount of data, and therefore the requirement for data processing and informatics should be considered. For LIANTI, after alignment of Illumina sequencing reads using BWA-MEM, a dedicated pipeline (“lianti” toolkit) is available to carry out subsequent analysis (https://gitlab.com/lh3/lianti).

2.6.2 Data analysis pipelines for microarrays

SNPs are detected using laser scanning, and intensity data are normalized and reported using dedicated software (GenomeStudio, Genotyping module; Illumina Inc.). An additional set of control probes that are included in the Infinium assay (Illumina Inc.) can be assessed for each stage of the process for QC and troubleshooting. The expected performance of each SNP locus is based on genomic DNA from a reference population with high heterogeneity (288 HapMap individuals). However, the limitation of such an approach is that it does not account for the noise within a single cell system; increasing the quality score in GenomeStudio improves precision, although at a cost of recall (Zamani Esteki et al., 2015).
2.6.3 Phasing algorithms

Once genotypes have been obtained, there are several phasing algorithms that do not rely on the information from three generations. We use the sibling approach that assumes that crossovers at the same position in different meioses are rare. Thus, when siblings share a specific crossover compared to an assumed reference (a 1N, 1C sample from an activated oocyte or PB2), the reference is likely to have experienced a crossover in that position (Hou et al., 2013; Ottolini et al., 2015). The assumed ancestor’s chromosome can thus be rearranged according to the positions of common crossovers among its siblings and the genotype of the parent can be deduced. The sibling approach requires haploid products from at least three independent meioses. Genomic DNA from parents are required to resolve maternal and paternal SNPs, if using embryos. For oocytes, the positions of heterozygous SNPs can be inferred from multiple haploid, meiotic products, if maternal DNA is not available.

2.7 MATERIALS

2.7.1 Sources of human oocytes and embryos

Research with human oocytes and embryos is conducted only with appropriate ethics, licenses, and informed consent as well as personal data protection that covers extensive genome sequencing.

2.7.2 Single cell isolation

Dead Cell Removal Kit and MS Columns (Miltenyi Biotech, 130-090-101, 130-042-201)
Molecular Probes™ Dead Cell Apoptosis Kit (Thermo Fisher Scientific, 10257392)
The STRIPPER® micropipetter and 125 μm tips (Origio, MXL3-STR, MXL3-IND-125)
PVP Clinical Grade, without phenol red (Origio, 10905000)
Phosphate Buffered Saline 20 × (Cell Signaling Technology, 9808)
Nunc™ IVF ICSI Dish 51 × 9 mm (Thermo Fisher Scientific, 150265)

2.7.3 WGA

SureMDA Amplification System (Illumina Inc., PR-40-405102-00)
REPLi-g® Single Cell Kit (QIAGEN, 150345)
SurePlex DNA Amplification System (Illumina Inc., PR-40-415101-00)
LIANTi (No commercial kit available at present)
Quant-IT™ dsDNA Assay Kit, high sensitivity (Invitrogen, Q33120)
FLUOstar® Omega microplate reader and MARS Data Analysis Software
(BMG Labtech)
2.7.4 NGS library preparation

VeriSeq™ PGS Kit and BlueFuse® Multi Software (Illumina Inc., RH-101-1001)
Nextera DNA Library Prep Kit (Illumina Inc., FC-121-1031)
Nextera Index Kit (Illumina Inc., FC-121-1012)
TruSeq® DNA PCR-Free HT Library Prep Kit (Illumina Inc., 20015963)
TruSeq® DNA CD Indexes (Illumina Inc., 20015949)
KAPA Library Quantification Kit for Illumina® Platforms (KAPA Bio systems, KK4828)
2100 Bioanalyzer High Sensitivity DNA Kit (Agilent, 5067-4626)

2.7.5 SNP microarray library preparation and arrays

HumanCytoSNP-12 v2.1 BeadChip Kit (Illumina Inc., WG-320-2103)
Multi-Ethnic Global-8 v1.0 BeadChip Kit (Illumina Inc., WG-316-1002)
GenomeStudio Software with Genotyping Module (Illumina Inc.)

3 CONCLUSIONS

We have discussed various considerations and methods for probing genome changes in single cells that allow new insights into meiosis and mitosis. We focused on two methods. LIANTI is a new WGA method that allows unprecedented coverage of the genome and resolution, which has allowed detection of replication origins in single cells. In contrast, MDA is an older generation WGA, but provides high fidelity and robustness, when studying female meiosis in precious material. Trouble shooting for mapping recombination and chromosome segregation in human female meiosis, termed MeioMapping, is further discussed elsewhere (Ottolini et al., 2016).

Single cell genomic technologies are rapidly evolving and successful integration into meiotic and mitotic studies require a combination of cellular approaches as well as computational skills. Several other single cell technologies have emerged, including sc-transcriptomics, sc-chromatin immunoprecipitation followed by sequencing, sc-methylation sequencing, and hybrid procedures that detect two or more features. Single cell proteomics is also available for human oocytes. Single cell genomics, however, remains one of the most challenging technologies due to the nature of detecting single copies at high precision and uniformly (“linear” amplification). Artifacts along the pipeline are amplified and may lead to erroneous conclusions, especially with regards to single variant calling (see Discussion in Chen et al., 2017).

The genome undergoes considerable changes, particularly in meiotic cells. Being able to explore these with high precision will enable future studies that allow new discoveries. The methods we have described are adaptable and should be useful for exploring genome changes.
ACKNOWLEDGMENTS

We thank our colleagues for technical discussions, in particular Alan Handyside, Antonio Capalbo, Sijia Lu, and Lei Huang. The authors receive research support from Illumina Inc., the Novo Nordisk Foundation (NNF15OC0016662), ERC (724718), MRC (G0902043, MR/M000664/1), the Danish National Research Foundation (DNRF115), NIH Director’s Pioneer Award (5DP1CA186693) and National Cancer Institute Grant (5R33CA174560), Beijing Municipal Science and Technology Commission (Z131100005213006), National Key Technology Research and Development Program (2016YFC0900100), National Basic Research Program of China (2015AA020407).

REFERENCES


Appendix A1.2

SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data.

Vogel, I, Blanshard, R.C. and Hoffmann, E.R.

Submitted.
Genetic and population analysis

SureTypeSC - A Random Forest and Gaussian Mixture predictor of high confidence genotypes in single cell data

Ivan Vogel\textsuperscript{1,2}, Robert C. Blanshard\textsuperscript{3,4} and Eva R. Hoffmann\textsuperscript{1,4} *

\textsuperscript{1}DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark.

\textsuperscript{2}Faculty of Information Technology, Brno University of Technology, Bozletechova 1/2, 616 66 Brno, Czech Republic.

\textsuperscript{3}Illumina Inc., Capital Park, Fulbourn, Cambridge, CB21 5XE, UK.

\textsuperscript{4}Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, BN1 9RQ, UK.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

\textbf{Motivation:} Accurate genotyping of DNA from a single cell is required for applications such as \textit{de novo} mutation detection, linkage analysis and lineage tracing. However, achieving high precision genotyping in the single cell environment is challenging due to the errors caused by whole genome amplification. Two factors make genotyping from single cells using single nucleotide polymorphism (SNP) arrays challenging. The lack of a comprehensive single cell dataset with a reference genotype and the absence of genotyping tools specifically designed to detect noise from the whole genome amplification step. Algorithms designed for bulk DNA genotyping cause significant data loss when used for single cell applications.

\textbf{Results:} In this study, we have created a resource of 28.7 million SNPs, typed at high confidence from whole genome amplified DNA from single cells using the Illumina SNP bead array technology. The resource is generated from 104 single cells from two cell lines that are available from the Coriell repository. We used mother-father-proband (trio) information from multiple technical replicates of bulk DNA to establish a high quality reference genotype for the two cell lines on the SNP array. This enabled us to develop SureTypeSC - a two-stage machine learning algorithm that filters a substantial part of the noise, thereby retaining the majority of the high quality SNPs. SureTypeSC also provides a simple statistical output to show the confidence of a particular single cell genotype using Bayesian statistics.

\textbf{Availability:} The implementation of SureTypeSC in Python and sample data are available in the GitHub repository: https://github.com/puko818/SureTypeSC

\textbf{Contact:} eva@sund.ku.dk

\textbf{Supplementary information:} Supplementary data are available at Bioinformatics online.

1 Introduction

Single cell genomics is an umbrella term for genotyping of individual cells from a heterogeneous population. The deconvolution of mixed populations allows detection of genetic diversity within a population of cells. Applications cover many disciplines from sequencing the complete genomes of microorganisms that are challenging to culture in the laboratory to \textit{de novo} mutation detection in tumour cells (Huang et al., 2015). Detecting genomic changes in single cells is a sensitive procedure,
complicated by the often rare, unique and precious nature of the starting material, such as during genetic testing of human embryos for diagnostic purposes. Unlike sequencing of bulk DNA, single cell sequencing requires a whole genome amplification (WGA) step to generate sufficient material for genotyping by next-generation sequencing (NGS) or single-nucleotide polymorphism (SNP) array (Gawad et al., 2016). A typical human cell contains 8–16 pg nuclear DNA that must be amplified to meet the input requirements for PCR-free sequencing (1 µg) or SNP array analysis (400 ng). The efficacy of genotyping from a single cell is critically dependent on the WGA method. Genome coverage, replication fidelity and the level of technical noise, such as systematic or stochastic amplification bias, are the main features considered when choosing the WGA method. However, all WGA methods deteriorate the signal from single cells. The signal deterioration potentially carries two risks: (a) sub-optimally amplified signal can lead to a complete loss of information about a particular locus, and (b) uneven signal amplification of two alleles at a heterozygous locus can result in an erroneous homozygous genotype call. The latter is called allele drop out (ADO) and affects up to 30% of SNPs from a single cell (Blanshard et al., 2018).

After WGA, the amplified DNA can either be sequenced or analysed by SNP arrays. There are several tools for genotype calling from WGA amplified DNA after sequencing (Zafar et al., 2016; Bohrson et al., 2017, bioRxiv; Lodato et al., 2015; Bae et al., 2018). In theory, sequencing can detect genomic changes from single base mutations to whole chromosome imbalances (Huang et al., 2015; Gawad et al., 2016). However, the sequencing depth required for applications such as detection of rare variants in a population of cells is limited by the number of cells that can be assessed, making sequencing impractical. In contrast, although SNP arrays are limited by assessing only a sample of the genome, the technology allows the analysis of a wide range of genetic variants with good coverage in a fast and cost-efficient manner. SNP arrays have been successfully applied to WGA DNA for discovery of new mutations, especially larger deletions that can cause loss of heterozygosity in cancers (Wong et al., 2004, Leung et al., 2002). They are also used in linkage analysis to screen preimplantation embryos for the presence or absence of monogenic variants that underlie serious genetic disorders. This is referred to as preimplantation genetic diagnosis or preimplantation genetic testing for monogenic disorder, PGT-M. This makes it critical that genome coverage is high and that the SNPs are typed with high precision (Handyside et al. 2010; Natesan et al., 2014; Zamani et al., 2015). SNP arrays have also been used to infer the parental origin of chromosome imbalances in human preimplantation embryos as well as recombination and segregation patterns in meiosis after WGA (McCoy et al., 2015; Ottolini et al., 2015). The unique genomic arrangement of SNPs that occur naturally as haplotype blocks, or can be induced experimentally, also allow lineage tracing in a variety of organisms, including by SNP array analysis (Woodworth et al., 2017; Ottolini et al., 2017).

There is a plethora of tools and algorithms currently available for genotyping bulk DNA from SNP arrays (Ritchie et al., 2011; Li et al., 2012). These algorithms are optimized for SNPs on the array and perform very well in terms of both call rates and sensitivity. However, an algorithm that is specifically designed for single cell variant calling using SNP arrays is currently missing. This is important because it is unclear how well the genotyping platforms deal with the biases introduced by whole-genome amplification of DNA from single cells. One solution is to include only SNP calls that are similar in properties to those from bulk DNA. This, however, causes a substantive loss of data (Zamani et al., 2015). It is also unclear how accurate genotyping is after the whole-genome amplification. Genotyping from SNP arrays relies on the detection of emission intensities (X and Y). Thus, when both X and Y are above a certain threshold, the genotype is inferred as heterozygous (AB). In contrast, when only X or Y is detected above a certain threshold, a homozygous genotype is assigned (AA or BB). Current genotyping algorithms are based on two distinct approaches. Model-based algorithms do not require a training data set and assume that every SNP can be modeled from a linear combination of multivariate components (Teo et al., 2007, Giannoulatou et al., 2008). Reference-based algorithms perform genotyping based on a comprehensive database of reference variants. Parameters of these algorithms are inferred from a training dataset (e.g. the HapMap population, International HapMap Consortium, 2005) and are used for normalization of the raw data (Ritchie et al., 2009) or provide a confidence measure of the genotype (Kermani, 2008). The training of the parameters can be performed via supervised machine learning methods, in particular neural networks (Kermani, 2008).

Here, we present a comprehensive database of 104 single cell samples from two different cell lines that we SNP-typed and compared with their reference genotype. This allowed us to create a database with two classes of calls: (a) high quality single cell calls and (b) misclassified single cell calls caused by deteriorated signal. We used both classes to develop a two-layered algorithm that combines a supervised machine learning method with a model-based algorithm. We refer to this as SureTypeSC, which is able to identify the noise in the single cell data coming from erroneous whole genome amplification and then assign a probability score of a SNP being correctly genotyped.

### 2 Materials and Methods

#### 2.1 Cell lines and molecular methods

We generated genotypes from whole genome amplified DNA (from single cells) or genomic DNA from bulk extraction using the Infinium Karyomapping Assay Kit (Illumina Inc., California, US). We obtained EBV-lymphoblastoid cell lines GM07228 and GM12878 from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, New Jersey, USA, and cultured these according to the supplier’s recommendation. All of the molecular methods and genotyping using GenCall for obtaining the SNP genotypes are provided in the Supplemental Methods.

#### 2.2 MA transformation

The MA transformation is an application of the Bland-Altman transformation (Bland and Altman, 1999) that has been used extensively in the analyses of gene expression data when intensity values for two channels are compared using microarrays (red and green, referred to as X and Y, respectively).

Formally, we apply a linear-log transformation for every SNP, i carrying a tuple of intensities \((x_i, y_i)\) by calculating the values \(m_i\) and \(a_i\), as follows:

\[
\begin{align*}
m_i &= \log_{10}(x_i) - \log_{10}(y_i) \\
a_i &= \frac{1}{2} \left[ \log_{10}(x_i) + \log_{10}(y_i) \right]
\end{align*}
\]

The M-feature has powerful discriminative ability to separate the three genotype clusters and is able to reduce variability between experiments and SNPs (Carvalho et al., 2007). The A-feature is a good general indicator of the signal quality (Ritchie et al., 2011).
2.3 Bioinformatics workflow for the machine learning algorithm

We developed a bioinformatics workflow with a supervised machine learning core that filters out the noise from the single cell data. The reference training intensities as well as validation intensities are first extracted from the intensity data (*.idat) files and subsequently genotype using the GenCall algorithm implemented in GenomeStudio. The training data are then exported from GenomeStudio, transformed using the MA transformation and fitted to a two-layered machine learning model. The results are subsequently tested on a cross-validated data set as well as on a set of independent single cell samples. Details of the workflow are shown in Fig. S1.

2.4 Training and validation datasets

We created a reference genotype for both single cell lines (GM07228 and GM12878) using parental information and multiple technical replicates from bulk DNA (Supplemental Methods). We subsequently compared our single cell datasets to the reference genotype. More specifically, for every candidate single cell call for SNP i and sample s we assigned a label: \( I_s \in \{\text{True, False} \} \), depending on the match or mis-match with the corresponding reference genotype call. The training dataset is then a triplet \((m, a, I)\), where \(m, a, I\) are input features and \(I\) is the output feature. Note that we omit sample index \(s\) in further explanation, as we do not distinguish between the origins of SNPs in the training data set. We included all autosomal single cell calls with GenCall score above 0.01 (QC001) totaling 14,403,139 SNPs for training (GM07228) and 11,737,508 SNPs for validation (GM12878). Lowering the GenCall score threshold for accepting a SNP allowed us to include potentially poorly amplified SNPs and to capture the full error pattern. Table 1 and Table S2 give a detailed overview of the datasets used.

2.5 Supervised training using Random Forest

Random Forest is an ensemble supervised training method that is built from the collection (forest) of classification (decision) trees (Breiman, 2001). Each tree is trained on a different random subset of data and different subsets of input features. Although the training data only contains two input features (\(M\) and \(A\)) the preliminary analysis (Fig. 1B) suggests that the function that separates the erroneous clusters (red areas) from the correct calls (blue areas) is non-linear. Random Forest (RF) has therefore approximate a non-linear separating function resulting in increased classification accuracy. The class of miscalls (\(I = \text{False}\)) is non-linear. The class of correct calls to tackle the class imbalance. We used the implementation of Random Forest from the scikit package (Pedregosa et al., 2011) for fitting the training data. We adjusted the following parameters of the algorithm:

- the number of features to consider when looking for the best split was set to two.
- the number of trees was increased from 10 to 30; according to Oshiro et al., 2012. A theoretical upper limit is 128 trees and further increase in number of trees does not contribute to higher accuracy. However our data suggest that forests with more than 30 trees contribute minimally to the accuracy of the model but increase the size of the model substantially (data not shown);
- the prediction was evaluated in two ways - by stratified 10-fold cross-validation and with an independent single cell dataset. We used metrics that are commonly used in classifier evaluation as well as metrics that are specific for the single cell environment (Supplemental Methods)

2.6 Cluster correction using Gaussian Discriminant Analysis

The second stage of the algorithm is a Gaussian Discriminant Analysis (GDA) that normalizes the genotype clusters obtained from the RF step and potentially improves the precision and recall.

Let \(D = \{x_j, y_j \mid 1 \leq j \leq N, x_j \in \mathbb{R}^4 \times \mathbb{G} \times \mathbb{L} \} \) denote a set of \(N\) validation SNPs that were classified by the trained Random Forest, where \( \mathbb{G} = \{ AA, AB, BB \} \), \( \mathbb{L} = \{ T, F \} \). Therefore, \( x_j = (m_j, a_j, I_j) \) is a quartet of the logarithmic difference, logarithmic average, genotype predicted by GenCall (QC 0.01) and class prediction by RF at the \(j\)-th SNP. We assume that both the positive (\(T\)) and negative (\(F\)) classes, which are represented by pairs \(d_i = (m_i, a_i)\), come from mixtures of multivariate normal distributions. Based on this, we define the following system of Gaussian discriminants:

\[
L \sim \text{Bernoulli}(\lambda) \quad \text{(1)}
\]

\[
p(I = T) = p(d_i | \theta_T) = \prod_{k=1}^{3} x_k \phi(d_i | \mu_k, \Sigma_k) \quad \text{(2)}
\]

\[
p(I = F) = p(d_i | \theta_F) = \prod_{k=1}^{3} x_k \phi(d_i | \mu_k, \Sigma_k) \quad \text{(3)}
\]

\[
\ln F(\theta) = \sum_{i=1}^{n} \ln p(d_i | \theta) \quad \text{(5)}
\]

We use an Expectation Maximization algorithm (Dempster et al., 1977) to estimate the parameters \(\theta_T\) of the positive and negative class that maximize their log-likelihood function (Eq. 6). The EM algorithm is divided into an Expectation-Step (E-Step) and a Maximization-Step (M-Step). These are run in iterations separately for the positive and negative classes until convergence is reached. The detailed description of the algorithm is provided in Supplemental Methods.

After the parameters of both classes have been estimated by the EM algorithm, they are subjected to a second run. Here, the class member...
ship \( \hat{t} \) is hidden from the algorithm and every SNP is evaluated for both Gaussian discriminants using the following formula:

\[
(s_{\text{RF}}, s_{\text{GDA}}) = \left[ \ln p(d|\theta_1), \ln p(d|\theta_2) \right]
\]

(6)

The final classification (membership to a positive or a negative class) is determined by higher value from the pair \((s_{\text{RF}}, s_{\text{GDA}})\).

2.7 Scoring function

The key role of a genotyping algorithm is to report the likelihood of a certain genotype in form of a score or a posterior probability. Besides GenCall having its own scoring scheme, we used the following equations to estimate the probability of a certain SNP being correctly genotyped:

1. Random Forest: the score of a genotype of the 4th SNP is given as a proportion of the trees in the forest that voted for a particular genotype being correct:

\[
s_{\text{RF}} = P(t_i = T|d_i)
\]

(7)

2. The scoring strategy of SureTypeSC is inferred from its second layer (GDA) as the class-conditional posterior probability of a genotype falling into positive class T:

\[
s_{\text{GDA}} = \frac{e^{s_{\text{RF}}} \times P(T)}{\sum_{T} e^{s_{\text{RF}}} \times P(T)}
\]

(8)

3 Results

3.1 Generation of 28.7 million high confidence SNPs from single cells

We typed nearly 28.7 million SNPs from 104 cells from two individuals (GM12878 and GM07228, Table 1 and Table S1) using the HumanKaryomap-12 array (Illumina Inc., California, USA). To amplify the DNA from the single cells, we used multiple displacement amplification (MDA), a commonly used first-generation WGA method that relies on Phi (\( \Phi \)) 29 polymerase. Its 3’→5’ activity allows proofreading and therefore improves the fidelity of amplification. This allows high precision genotyping with a mutation rate of \( 10^{-9} \) – \( 10^{-8} \). Furthermore, the ability to displace secondary DNA structures, such as hairpin loops that would cause other polymerases to stall or dissociate from the template DNA, allows the amplification of long DNA fragments (2-10 kb) (Blanshard et al., 2018; Dean et al., 2002).

Table 1. Summary of genotype calls from single cells

<table>
<thead>
<tr>
<th>Cell line</th>
<th>GenCall ≤ 0.001</th>
<th>GenCall ≤ 0.01</th>
<th>GenCall ≤ 0.05</th>
<th>GenCall ≤ 0.10</th>
<th>GenCall &gt; 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM07228</td>
<td>0.39 0.02 0.05 0.01 0.05 0.01 0.36 0.02 0.04 0.01 0.1 0.01</td>
<td>M SD M SD M SD M SD M SD M SD M SD</td>
<td>0.4 0.02 0.06 0.02 0.04 0.01 0.37 0.03 0.04 0.01 0.09 0.02</td>
<td>Total</td>
<td>0.8 0.11 0.09 0.73 0.08 0.19</td>
</tr>
<tr>
<td>GM12878</td>
<td>0.39 0.02 0.05 0.01 0.05 0.01 0.36 0.02 0.04 0.01 0.1 0.01</td>
<td>M SD M SD M SD M SD M SD M SD M SD</td>
<td>0.4 0.02 0.06 0.02 0.04 0.01 0.37 0.03 0.04 0.01 0.09 0.02</td>
<td>Total</td>
<td>0.8 0.11 0.09 0.73 0.08 0.19</td>
</tr>
<tr>
<td>Total</td>
<td>28.7 million SNPs, 104 cells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values are proportions. *GenCall cutoff 0.001, **GenCall cutoff 0.01, ***GenCall cutoff 0.1, ****concordant with reference genotype, ⋅ disconcordant with reference genotype, NC: no calls; M: mean; SD: standard deviation.

3.2 Noise characterization of genotypes from single cells

To characterize the noise associated with genotyping from whole-genome amplified DNA from single cells, we compared the 28.7 million SNP genotypes from the two single cell datasets to their reference genotypes obtained from bulk, genomic DNA. To this end, we created high confidence reference genotypes from bulk DNA using nine independent bulk DNA samples hybridized against the HumanKaryomap-12 array and inferred genotypes using either the full parental information (GM07228, Table S2) or multiple technical replicates of bulk DNA and sequence data (GM12878, Supplemental Methods and Eberle et al., 2017). This allowed us to identify 264,269 SNPs for GM07228 and 270,681 for GM12878 (95.8% and 97.9 % of autosomal SNPs, respectively) on the HumanKaryomap-12 array that called correctly in every replicate from bulk DNA. From these, we generated high confidence reference genotypes.

Using the standard QC cutoff from GenCall (0.15), 73% SNPs (20.9 million) from the two single cell datasets called correctly, whereas 8% SNPs (2.36 million) were not concordant with the reference genotype (Table 1, Table S1). 19% SNPs gave ‘no calls’ (5.05 million; Table 1, Table S1), having failed to fall within the genotype clusters defined by bulk, DNA genotypes (Fig. 1A). The true positive rate was higher when we used a minimal QC (0.01) compared to the standard QC of GenCall (39%, SD=0.02% and 36%, SD=0.02%, respectively, for cell line GM07228 and 40%, SD=0.02% to 37%, SD=0.03% for GM12878, Table 1). These differences in true positive rates are statistically significant (p<0.0001, Fig. S2). In total for both datasets, the GenCall algorithm rejects about 7% of correctly genotyped SNPs from WGA DNA and increases precision by 2% (Table S1). We also listed call rates and error rates of individual cells and chromosomes from GM12878 and GM07228 (Table S3).

We displayed the pattern of the noise from the genotyping of SNPs from WGA DNA from single cells by first transforming the fluorescence intensities (X and Y) of each SNP into the logarithmic difference \( \log(5^X/5^Y) \) of the fluorescent intensities \( X \) and \( Y \) of each SNP from WGA DNA from single cells by first transforming the fluorescence intensities (X and Y) of each SNP into the logarithmic difference \( \log(5^X/5^Y) \) of the fluorescent intensities \( X \) and \( Y \). At this stage of the workflow (“Building training dataset”, Fig. S1) we were able to observe the error pattern in the single cell data and display it in the form of contour plots (Fig. 1B). Three clusters of miscalls (false positives) became apparent in the single cell data. Two clusters were from allele drop out (ADO), where AB genotypes were incorrectly genotyped as AA or BB. A smaller cluster of allele drop in (ADI) also appeared. The ADI cluster was clearly separated from the true AB genotypes. Most of the errors, however, occur in the transition area between AB to AA or AB to BB (ADO) but nevertheless suggest good separability of the correct calls from miscalls, since the centers of the clusters are non-overlapping (Fig. 1B).

3.3 Design and implementation of the SureTypeSC algorithm

The characterization of the patterns of noise in a comprehensive dataset allowed us to employ a supervised machine learning method to classify and separate high quality genotypes from miscalls in the WGA DNA from single cells (Fig. S1). We combined a non-parametric (Random Forest) and parametric method (Gaussian Discriminant Analysis) and developed a scoring strategy that assigns probabilities that a specific SNP from a single cell dataset has been correctly genotyped (Methods, Eq. 8). Using a Random Forest prevents over-fitting of the data and provides good estimates of the positive and negative classes for the
Gaussian discriminant analysis (Methods). Both the RF and GDA can be implemented on their own. We implemented the RF-GDA and the testing procedures in Python using the scikit library (Pedregosa et al., 2014) and pandas (McKinney, 2010). An example of a division of the feature space consisting of $M$ and $A$ by the RF-GDA algorithm is shown in Fig. 1D. Collectively, we refer the single layers and the combined layer (RF-GDA) as SureTypeSC. The output from SureTypeSC is compatible with GenomeStudio and allows the user to import the results of the analysis back to GenomeStudio for further investigation.

3.4 Cross-validation of SureTypeSC

To assess whether our algorithm captures noise from the WGA and to exclude the possibility of overtraining, we first ran stratified 10-fold cross-validation on the single cell dataset from cell line GM07228. The dataset is imbalanced and mistyped SNPs are the minority class. We therefore used stratification to ensure that every fold contains both correctly genotyped and mistyped SNPs. In every iteration, we trained the algorithm on nine folds (27,445-27,772 SNPs) and used the 10th fold for testing. To tackle the imbalance problem, we always balanced the training fold by downsampling the correctly genotyped SNPs. We evaluated the performance of every testing fold with the single layers individually (RF or GDA) as well as with a combination of them (RF-GDA). We scored the genotypes of all algorithms using the GenCall score or equations 7 or 8 (Methods). Consistent with random sampling of the SNPs, the mean performances of all algorithms have narrow confidence intervals (at 95%), which suggests that the algorithms are invariant to SNP selection (Table S4). Pairwise comparison of the algorithms using paired t-test shows that both, GDA and RF-GDA outperform GenCall in precision at similar recall (mean difference 4%, $p < 0.0001$, Table S5). Note that the RF outperforms GenCall in precision as well (mean difference 4.7%, $p < 0.0001$, Table S5), but has a lower performance in other metrics (Table S4). RF-GDA is more accurate than GenCall (mean difference 3.5%, $p < 0.0001$) and has a higher f1-score (mean difference 2%, $p < 0.0001$, Table S5).

3.5 Validation of SureTypeSC on an independent dataset

We next addressed how well our algorithms and GenCall performed on an independent dataset. To this end, we used the SNP genotypes obtained from 58 single cells after WGA from cell line GM07228 for training and the SNP genotypes obtained from WGA DNA from 46 single cells from a different cell line, GM12878 (Table 1, Table S1), for testing (‘tester’ set). The genotyping data from the tester set were obtained at an independent time, with different batches of WGA reactions and genotyping arrays. This avoids systematic errors introduced by the chemistry used to obtain the genotypes.

We first evaluated the performance of GenCall, RF, GDA and SureTypeSC (RF-GDA) separately for heterozygous and homozygous regions using ROC and Precision-Recall curves (Fig. 2, Fig. S3). These metrics gave us visual insight into overall performance of the classifiers, invariant to the score cutoffs used. For the heterozygous calls, RF-GDA outperforms all tested algorithms, which is also quantified by the ROC-AUC score (Table 2). Whereas GenCall achieves a 74% ROC-AUC score on average, this is increased to 86%, 87% and 92% for RF, GDA and RF-GDA, respectively ($p < 0.0001$, Table S6). For the homozygous regions, the RF outperforms GenCall at all points of the ROC and Precision-Recall curves, which is supported by the increase in the ROC-AUC score from an average of 67% (GenCall) to 81% for the RF (Table 2). This is further increased with the GDA or RF-GDA (both 83%). Interestingly, at a precision of 93%, the RF curve crosses that of the GDA and RF-GDA and recalls more true positive homozygous calls (Fig. S3A). This suggests that the RF alone might be a good option if higher recall is required at the costs of lower precision, which is nevertheless higher than GenCall in the homozygous regions. GenCall crosses the Precision-Recall curve of the RF-GDA at a precision around 88% and recalls more true positives (Fig. S3A). This is, however, very close to a recall of 100%, which also means accepting all calls without any significant filtration.
Next, we were interested in how our methods perform compared to GenCall with standard settings (QC 0.15). Standard GenCall recalls 68% of the true positive heterozygous genotypes with a precision of 97%. The RF-GDA has 84% recall, achieves 99% precision on average, and thus outperforms the standard GenCall in both precision and recall. At similar precision, the RF and GDA on their own recall fewer true positive heterozygous genotypes (Table 2). As expected, high precision and recall is reflected in a high harmonic mean of precision and recall (f1-score) for the RF-GDA (Table 2) and a high rate of correctly classified SNPs (accuracy, Table 2). GenCall recalls 96% of the true positive homozygous genotypes on average at precision 89%. At similar recall, the RF alone increases precision by 2.5% (p < 0.001; Table S6). GDA and RF-GDA further improve precision, but at the cost of recall. Both methods achieve an average precision of 92% at 90% recall for the homozygous calls (Table 2). Recalling fewer true positives at higher precision causes a drop in the f1-score for GDA and RF-GDA. This is because recall declines much quicker than the precision increases (Fig. S3A). The effect of lower recall from the GDA and RF-GDA is also mirrored in the lower accuracy. As GDA and RF-GDA have higher precision, they are also more likely to reject correct SNPs, thereby decreasing the number of true positives.

Table 2. Performance of the genotyping algorithms on independent dataset 12878

<table>
<thead>
<tr>
<th>Alg.</th>
<th>GenCall</th>
<th>RF</th>
<th>GDA</th>
<th>RF-GDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>het</td>
<td>homo</td>
<td>het</td>
<td>homo</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.68±0.01</td>
<td>0.86±0.012</td>
<td>0.71±0.013</td>
<td>0.88±0.008</td>
</tr>
<tr>
<td>f1-score</td>
<td>0.8±0.01</td>
<td>0.92±0.007</td>
<td>0.82±0.012</td>
<td>0.93±0.005</td>
</tr>
<tr>
<td>precision</td>
<td>0.97±0.01</td>
<td>0.89±0.009</td>
<td>0.99±0.001</td>
<td>0.91±0.008</td>
</tr>
<tr>
<td>recall</td>
<td>0.68±0.01</td>
<td>0.96±0.005</td>
<td>0.7±0.017</td>
<td>0.96±0.001</td>
</tr>
<tr>
<td>ROC-</td>
<td>AUC</td>
<td>score</td>
<td>0.74±0.01</td>
<td>0.67±0.015</td>
</tr>
</tbody>
</table>

The two-layered architecture, RF-GDA, generally outperforms its constituent single layers (RF or GDA alone). Combining the RF and GDA together is particularly advantageous in the heterozygous regions, where the RF-GDA performs better in all metrics (Fig. S4). For the homozgyous calls, the RF-GDA performs better than single RF and GDA in precision (mean difference 1.7% and 0.2% for RF and GDA, respectively, p < 0.0001, Table S6). However, the single GDA has better ROC-AUC score, which is 0.2% higher in the GDA than in RF-GDA (p < 0.001, Table S6). The ROC curve in Fig. 2A and Precision-Recall curve in Fig. S3A confirm that the difference is minor, since the RF-GDA and GDA largely overlap. Collectively, the benefits of the two layered RF-GDA compared to its single layers is the maximized precision and recall for the heterozygous calls. There is a further benefit in the maximized precision in the homozygous calls at the relatively modest loss of true positive calls.

3.6 Genotyping confidence in the single-cell environment.

Our observations suggest that SureTypeSC can effectively improve precision of both homozygous and heterozygous SNPs (on average, 99% for heterozygous calls and 92% for homozygous calls, Table 2). Precision can be further improved at the cost of recall, particularly for homozygous SNPs, as Fig. S3 suggests. We therefore adjusted both SureTypeSC and GenCall for high precision, recalling ~47% of the true positive SNPs. To compare their performance, we developed a simple statistical tool that shows a detailed view of confidence in AA, BB or AB calls using a transition matrix of posterior probabilities (Table S7). The posterior probabilities show the probability that a certain genotype from the single cell application is genotyped correctly compared to the reference genotype. Table S7 shows that compared to GenCall, SureTypeSC achieves major improvements of 8% and 7% confidence of AA and BB, respectively, and an improvement of 0.3% in confidence of an AB genotype.

3.7 Allele-drop out and allele-drop in rates are reduced using SureTypeSC

Incorrect genotype calls arise predominantly from imbalances in the allele frequencies generated during the chemical reaction when the whole genome is amplified. The deviation from a 1:1 allele ratio of heterozygous SNPs can lead to allele drop out (ADO). Analogously, mistyping of a homozygous SNP results in allele drop in (ADI). We calculated the ADO and ADI rates for GenCall and SureTypeSC at high precision using the transition matrices (Table 3; performances of single layers RF and GDA are shown in Table S8). At a call rate of 42% for GenCall and 39% for SureTypeSC, GenCall is able to decrease ADI 7 times and SureTypeSC 12.5 times compared to minimal filtering (GenCall QC 0.01). The ADO rate is decreased 1.5 times by GenCall and 5.6 times and SureTypeSC 12.5 times compared to minimal filtering (GenCall QC 0.01). The ADO rate is decreased 1.5 times by GenCall and 5.6 times by SureTypeSC (Table 3). Although SureTypeSC outperforms GenCall and minimizes the error incidence, the loss of data is inevitable (call rate 39%, Table 3).

Table 3. Allele drop-in and allele drop-out

<table>
<thead>
<tr>
<th></th>
<th>Min. QC</th>
<th>GenCall</th>
<th>RF-GDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI</td>
<td>0.05±0.01</td>
<td>0.007±0.003</td>
<td>0.004±0.005</td>
</tr>
<tr>
<td>ADO</td>
<td>0.14±0.009</td>
<td>0.096±0.01</td>
<td>0.025±0.004</td>
</tr>
<tr>
<td>Call rate</td>
<td>0.92±0.003</td>
<td>0.42±0.01</td>
<td>0.39±0.01</td>
</tr>
</tbody>
</table>

*GenCall QC threshold 0.01; *GenCall QC threshold 0.87; * RF-GDA score threshold 0.75. Proportions and confidence intervals at 95% are shown.

3.8 Proof-of-concept of biological inference

To provide proof-of-concept that SureTypeSC would improve biological insight when used for high precision (RF-GDA), we assessed copy number variants (CNVs) in human oocytes (Fig. S5; Ottolini et al., 2015). The loss of a chromosome or chromosome segment results in one
cell with only A and B calls (no heterozygous SNPs). The loss, however, is obscured by ADI when using the standard GenCall algorithm (Fig. S5). SureTypeSC removes the ADIs (erroneous AB), increasing the certainty of the inference (loss of hetSNPs).

We were interested in whether we could use SureTypeSC with high precision to reveal biological variability within the tested cell line GM12878. We assumed that all variants in the cell line that do not match the reference genome (‘erroneous variants’) are allele drop outs or allele drop ins (Fig. S6A). We were curious whether some of these variants, however, could be real and therefore used to detect heterogeneity within a cell population. We chose SureTypeSC as this documented the best overall performance in terms of precision (Table 2, Table 3, Table S7, Table S8) and compared it to GenCall. We performed hierarchical clustering (Supplemental Methods) on raw data with minimal filtering (QC 0.01), data processed by GenCall (GenCall QC 0.87), and RF-GDA at high precision (Table 3). The hierarchical clustering reveals there are potentially four subpopulations in cells in GM12878 cell line that are invariant to the type of filtration used (Fig. S6B, C, D). The bootstrap analysis (Supplemental Methods), however, reveals that only the RF-GDA consistently gives four stable subpopulations (Jaccard mean bootstrap value for a cluster > 0.75, Hennig 2007). The unstable clusters present in the trees from the minimal filter (QC 0.01) and ‘high precision’ genotyping using GenCall suggest non-reproducible noise being transferred to the bootstrapped replicates that is removed by SureTypeSC.

Discussion

Whereas there are specialised tools for single cell genotyping from next-generation sequencing data (Zafar et al., 2016; Bohrson et al., 2017, bioRxiv; Lodato et al., 2015; Bae et al., 2018), no such algorithms exist for genotyping data from SNP arrays. Instead, genotyping WGA DNA has relied on increasing the threshold of the genotyping algorithms, which causes a substantial data loss (Zamani et al., 2015).

In this study, we have typed nearly 30 million SNPs from 104 single cells from two independent cell lines and developed an algorithm to distinguish signal from noise in whole-genome amplified DNA. SureTypeSC consists of two layers – a Random Forest (RF) and a Gaussian Discriminant Analysis (GDA) that work singly or in a cascade. The cascade approach is particularly beneficial for heterozygous SNPs, as it improves both precision and recall compared to the single layers and GenCall. We observed an increase in recall from 68% in GenCall at standard QC (0.15) to 84% in SureTypeSC. Resolving most of the heterozygous SNPs makes SureTypeSC highly relevant and applicable when heterozygosities are needed, such as tag SNPs during linkage analysis of transmission of monogenic diseases and aneuploidy detection (Handyside et al., 2010; Natesan et al., 2014; Zamani et al., 2015). At the same time, SureTypeSC improves precision for both homozygous and heterozygous SNPs by 3% and 2%, respectively (Table 2, Table S6).

Having high precision makes it feasible to explore rare events across populations of cells. This includes assessing clonal expansion in tumour evolution, lineage tracing, or detecting rare de novo mutations such as large genomic rearrangements in single cells that are averaged out and lost in bulk analyses (Cooper et al., 2015; Lu et al., 2012; Chen et al., 2017; Wong et al., 2004, Leung et al., 2002). High precision is also needed to obtain high resolution at haplotype breakpoints, which is particularly important in diagnostics (Handyside et al., 2010; Natesan et al., 2014; Zamani et al., 2015). Our proof-of-concepts show that SureTypeSC is likely to improve diagnostics as well as biological inferences (Fig. S6).

As running the single layers of SureTypeSC could be potentially beneficial, such as high recall in homozygous regions at lower precision by the RF alone, SureTypeSC always operates in both modes (cascade and single layers), simultaneously, and scores the genotypes with RF, GDA, as well as RF-GDA.

Analysing a large number of single cells allows the decomposition of heterogeneous populations. Understanding how single cells in a population contribute towards a ‘mosaic’, mixed SNP call is particularly important as use of SNP arrays increases in cytogenetics. Furthermore, having a robust algorithm of genotyping from WGA DNA from single cells improves the certainty of genotype calling when only few cells are available. This is important in both basic biomedical research as well as clinical settings such as in preimplantation genetic testing. We show that SureTypeSC can be used for both (Fig. S6).

We have implemented SureTypeSC in two modes in terms of performance. Using the standard mode where precision and recall are balanced, SureTypeSC was able to successfully identify a chromosomal loss in the single cell oocyte data, where GenCall failed to reject SNP calls from the chromosomal region (Section 3.8, Fig. S6). Using a high precision mode, SureTypeSC, but not GenCall, was able to stably detect four subpopulations in the reference GM12878 cell line. Thus, SureTypeSC most likely revealed true heterogeneity within the single cell population. This allows the use of SNP arrays in the single cell environment to explore fine differences between closely related cells. This was previously not possible due to the low resolution of the SNP array in combination with the noise coming from WGA.

Although genotyping from SNP arrays cover only a fraction of the genome compared to next-generation sequencing, the cost of de novo genome assembly is prohibitive even for bulk, genomic DNA when assessing a large number of cells or samples. The sequencing depth or coverage, needed in one recent reference genome assembly for the detection of de novo mutations was nearly 50x (Besenbacher et al., 2015). For single-cell applications, the coverage to accurately identify new mutations from the noise and bias introduced by the whole-genome amplification step is in excess of this (Beljatia et al., 2014). Thus, SureTypeSC allows a cost-effective approach to improve genotype accuracy using SNP arrays.

Acknowledgements

The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM07228, GM07224, GM07225 and GM12878. We thank Alex Bladon from Illumina Inc. for critical discussion and members of Hoffmann lab for comments on the manuscript.

Funding

I.V. was supported by the Danish National Research Foundation Center grant (DNRF115, 6110-00344B). Work in the ERH lab was supported by a NNF Young Investigator Award (NNF15OC0016662).
Appendix A1.3

Chromosome errors in human eggs shape natural fertility


Submitted.
“Chromosome errors in human eggs shape natural fertility”

Jennifer R Gruhn¹,†, Agata Zielinska²,†, Vallari Shukla¹,†, Robert Blanshard³,⁴,†, Antonio Capalbo⁵, Danilo Cimadomo⁶, Dmitry Nikoforov⁷,⁸, Andrew Chi-ho Chan¹, Louise J. Newnham³, Ivan Vogel¹, Catello Scarica⁹, Marta Krapchev¹⁰, Deborah Taylor¹¹, Stine Gry Kristensen⁷, Erik Ernst¹², Anne-Mette Bay Bjørn¹², Lotte Berdiin Colmorn¹³, Joanna Liss¹⁰, Geraldine Hartshorne¹¹, Marie Louise Groendahl¹⁴, Laura Rienzi⁶, Filippo M. Ubaldi⁶, Rajiv McCoy¹⁵, Krzysztof Lucaczezk¹⁰,¹⁶,¹⁷, Claus Yding Andersen⁷, Melina Schuh² and Eva R. Hoffmann¹,³,*

¹DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

²Max Planck Institute for Biophysical Chemistry, Department of Meiosis, Göttingen Germany

³Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, UK

⁴Illumina Inc., Fulbourn, UK

⁵iGenomix, Marostica, Italy

⁶GENERA, Clinica Valle Giulia, Rome, Italy

⁷Laboratory of Reproductive Biology, The Juliane Marie Centre for Women, Children and Reproduction, Copenhagen University Hospital and Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

⁸Unit of Basic and Applied Biosciences, Università degli studi di Teramo, Teramo, Italy

⁹DAHFMO, Unit of Histology and Medical Embryology, Sapienza, University of Rome, Italy.

¹⁰INVICTA Fertility Clinic, Gdańsk, Poland

¹¹Warwick Medical School, University of Warwick and Centre for Reproductive Medicine, University Hospital Coventry, UK

¹²Department of Obstetrics and Gynaecology, University Hospital of Aarhus, Skejby Sygehus, Aarhus, Denmark
13 The Fertility Clinic, The Juliane Marie Centre for Women, Children and Reproduction, Copenhagen University Hospital, Denmark

14 Fertility Clinic, Copenhagen University Hospital, Herlev Hospital, Herlev, Denmark

15 Department of Biology, Johns Hopkins University, Baltimore, MD, USA

16 Department of Obstetrics and Gynecological Nursing, Faculty of Health Sciences, Medical University of Gdansk, Gdansk, Poland

17 Department of Gynecological Endocrinology, Medical University of Warsaw, Warsaw, Poland

† These authors contributed equally to the work.

* Corresponding author: eva@sund.ku.dk
Abstract.

Human conceptions are affected by an exceptionally high incidence of chromosome errors (aneuploidy), which cause genomic imbalances resulting in pregnancy loss and congenital disorders. To understand why maternal age is the most important risk factor for aneuploidy, we collected >3,000 oocytes from 268 girls and women and obtained chromosome error rates from 218 oocytes, covering 5,014 segregation events in females aged 9 to 42 years. Chromosome errors in oocytes follow a convex J or U curve according to maternal age, consistent with natural pregnancies. The curve is a composite of different error types with distinct age-dependencies, indicating that separate molecular mechanisms cause the high incidence of aneuploid conceptions at the two extremes of the curve. Thus, aneuploid conceptions are a major component in determining natural fertility in humans.

One sentence summary.

Chromosome errors in female meiosis shape the J curve of aneuploid conceptions and natural fertility.
Main text.

Aneuploidy is the gain or loss of chromosomes and affects 20 to 90% of human conceptions, resulting in preclinical and clinical pregnancy losses as well as congenital disorders in live births (1-3). The wide range is due to the maternal age effect, where aneuploid conceptions increase in incidence with advancing maternal age such that by 42 years, an estimated 90% of human conceptions are aneuploid (1-3). The maternal age effect is thought to be due to the prolonged arrest during meiosis, the unique cell division that ensures that only one chromosome set is transmitted from each parent. Indeed, the majority of these chromosome errors are inferred to occur during the reductional division (meiosis I), when the oocyte resumes meiosis and homologous chromosomes segregate after a decades-long arrest (3, 4).

Given the socioeconomic changes causing women to delay childbirth (5), most of our focus has been on understanding the exponential increase that occurs as women advance in maternal age (mid-30s and above) (2, 6). What is less appreciated is that aneuploidy rates in clinically-recognized pregnancies are also elevated in teenagers, resulting in a slight biphasic J or U shape, reminiscent of a convex J curve (Fig. S1) (3). However, since studies of female meiosis to date have relied predominantly on observations in oocytes donated by women undergoing fertility treatment, and therefore by default exclude younger girls, we currently do not understand whether the elevation in teenagers is due to errors in oocytes, sperm (7), or embryonic mitoses in preimplantation embryos (8). We reasoned that understanding meiotic chromosome errors directly in human oocytes across the entire spectrum of female ages would provide insight into the forces that shape the J curve, the mechanisms that give rise to errors in chromosome segregation, and natural fertility in human females.
To obtain oocytes from a wide spectrum of female ages, we collected 3,253 GV oocytes (dictyate or G2/M arrested) directly from the ovarian tissue of 145 unstimulated girls and women aged 2.5 to 39 years under informed consent and ethical regulation (Cohort 1; small antral follicles; **Fig. 1A-D, Table S1, Methods**). The GV oocytes are from small antral follicles (0.5-3 mm), which are more numerous in the ovary compared to large antral follicles (**Fig. 1D**) (9), and are independent of high circulating levels of follicle stimulation hormone (FSH). These girls and women were due to undergo chemotherapy for blood disorders and a range of cancers (excluding ovarian cancer; **Fig. S2**) and therefore had ovarian removed for fertility preservation (10-12). More than 95% of such women re-establish their hypothalamus-pituitary-ovarian axes after autotransplantation of the cortical tissue, indicating that their ovarian tissue functions normally (10-12).

Homologous chromosomes segregate during the first meiotic division concomitant with oocyte maturation and ovulation. We matured 2,642 GV oocytes enclosed in their cumulus cells (cumulus-oocyte-complexes) from 119 patients using an established *in vitro* maturation protocol (**Fig. 1E-G**) (13). Nearly 30% of the oocytes completed meiosis I and arrested at metaphase II (mature MII oocytes; **Fig. 1E**), a rate that was highly dependent upon the morphology of the cumulus-oocyte-complexes, but not clinical diagnosis or center where the surgery took place (**Fig. S2-S4**). Maturation rates improved with age (**Fig. 1F, G; β = 0.045 ± 0.016, 95% CI, p < 0.01**), and this was not due to an increase in cumulus-oocyte-complexes with good morphology (**Fig. S4**). These findings suggest that GV oocytes obtained from small antral follicles in younger girls (2.5 to 15 years) have intrinsically reduced meiotic potential. This is consistent with irregular and lower ovulation rates, when GV oocytes complete meiosis I *in vivo*, amongst teenagers after the onset of menarche (14-17).
Only a fraction of oocytes from small antral follicles will eventually ovulate *in vivo*. We were therefore interested in whether they were inherently different in their potential to complete the meiotic divisions and in their ability to segregate chromosomes compared to oocytes from large antral follicles (16-22 mm). 123 women undergoing controlled ovarian stimulation as part of their fertility treatment in three clinics across Europe provided informed consent to participate in our study. Of these, 78 participants donated 238 GV oocytes and 86 *in vivo* matured MII (MII*) oocytes not needed for fertility treatment (Cohort 2; large antral follicles; age range: 20 to 42 years; **Fig. 1A, C, and D; Table S2**). The GV oocytes were induced to undergo the first meiotic division (maturation) after removal of cumulus and granulosa cells. We only used the fast-maturing GV oocytes that enter and complete meiosis I within 24 hours after removal of the cumulus cells, since this population is similar to *in vivo* matured MII oocytes (18, 19). The maturation potential for fully-grown GV oocytes obtained from large antral follicles (Cohort 2) in IVF clinics where women underwent gonadotrophin stimulation was not dissimilar from those recovered from small antral follicles (Cohort 1), although the rates varied between IVF centers (**Fig. 1H, Fig. S5**) (9). Both populations of oocytes that matured *in vitro* showed lower activation potential (completion of meiosis II) in response to treatment with a calcium ionophore compared to *in vivo* matured MII oocytes (MII*; **Fig. 1I, Table S3**) (20, 21). Collectively, our observations suggest that GV oocytes obtained from small antral follicles are capable of entering and completing both meiotic divisions *in vitro* with similar efficiency as those from large antral follicles.

We next assessed the efficiency of chromosome segregation at the first meiotic division, which is particularly error-prone and contributes the majority of chromosome errors in natural conceptions (3). If mis-segregation events in oocytes shape the J curve according to
maternal age, then one would expect an elevation in both teenagers and women of advancing maternal age compared to women in their twenties.

To address whether this was the case, we inferred chromosome segregation in meiosis I by obtaining the chromosome content in both the MII oocyte and PB1 (Fig. 2). We analysed genome content in 178 MII-PB1 duos (or 356 single cells) using a custom next-generation sequencing pipeline validated on single cells (Fig. 2A; Methods). After stringent quality control, we obtained data from 218 meiosis I divisions covering 5,014 segregation events. Using generalized linear models, we found no significant effects of the source (small versus large antral follicles), the maturation method, or the technology used to assess aneuploidy (Fig. S6, S7; Table S5). Maternal age was the only significant effect and the best fit model was provided by a quadratic equation, suggestive of the convex J or U curve ($\chi^2$ of deviance, $p < 0.001$ for both oocytes and chromosomes; pseudo-$R^2_{oocyte} = 0.10$, pseudo-$R^2_{chromosome} = 0.12$; Fig. 2B, C). Consistent with this, both the slope from the mid-point (25.4 years) towards the younger oocytes as well as upwards towards the oocytes from women of advancing age contributed to the fit (pseudo-$R^2 = 0.84$, $p < 0.05$). To further explore the elevation in the youngest and oldest oocytes, we divided them into three age groups based on the curvature of the J shape in natural conceptions (Fig. S1) (3, 23, 24): ‘teenagers’ (< 20 years), ‘mid’ (20-32 years), and ‘advancing maternal age’ (33-43 years), where we used a conservative cut-off of 33 years. Both teenagers and women of advancing maternal age experienced elevated levels of aneuploidy compared to women aged 20-32 years (Fig. 2B, C; $p < 0.025$, Fisher exact test). This is consistent with observations from natural conceptions and supports the notion that errors in female meiosis underlie the shape of the convex J curve, especially the elevation in teenagers. We noted that a relatively modest rate of chromosome errors (4%; Fig. 2C) affected a large proportion of oocytes (>50%; Fig. 2B), and
consistent with this, the majority of aneuploid oocytes were affected by a single event (Fig. 2F, G). This suggests a relatively uniform risk amongst oocytes as opposed to a minority of oocytes having more complex outcomes, although these do occur in a subset of oocytes (Fig. 2F, G) (25).

To understand the convex nature of the J curve of chromosome segregation errors in human oocytes, we assessed the error type as a function of maternal age. Meiosis I nondisjunction (MI NDJ), the gain or loss of a whole chromosome, was noted in early cytological studies of rare, mature MII oocytes obtained directly from ovarian tissue (26-28). Unexpectedly, the prevalence of meiosis I nondisjunction was higher in teenagers compared to the older age groups (p < 0.025, Fisher exact test; Fig. 2D, E; Table S6). In contrast, precocious separation of sister chromatids (PSSC) or predivision (29, 30), in which sister chromatids of one of the homologs separate in meiosis I, increased linearly with maternal age ($\beta_{\text{oocytes}} = 0.14$ and $\beta_{\text{chromosomes}} = 0.11$, p < 0.001; Fig. 2D, E). We observed the elevation in PSSC compared to meiosis I nondisjunction in the mid- and advancing maternal age groups in both cohorts (Fig. S7). Thus, the type of chromosome error changes with maternal age suggesting that the shape of the overall incidence of aneuploidy is a composite of at least two distinct curves.

Our findings suggest that at both ends of the convex J curve, oocytes are of lower genetic quality due to the elevation of two distinct segregation errors in meiosis I, the predominant segregation errors in clinically-recognized pregnancies (3). This predicts meiotic errors should also follow a convex J curve in preimplantation embryos, which would be consistent with a high incidence of preclinical pregnancy losses. To test this prediction, we leveraged a previously published dataset consisting of 36,786 preimplantation embryos biopsies obtained during 5,819 cases of genetic testing for aneuploidy (31, 32). In this, and other large datasets of biopsies from cleavage or blastocyst stage preimplantation embryos, total aneuploidy rates amongst younger
women appear elevated, supporting the convex J curve (8, 31, 33). When we assessed meiotic errors from the mother that result in a trisomic embryo containing both maternal haplotypes (BPH–‘both parental homologs’; Methods), we found that a quadratic model fits the maternal meiotic errors better than a reduced-sloped model ($\chi^2$ of deviance, $p < 3 \times 10^{-43}$; pseudo-$R^2 = 0.240$, **Fig. 2H**). To assess the maternal age effect in the lower age range, we truncated the dataset at a maximum maternal age of 27.1 years. This threshold was chosen to maximize power based on simulations informed by the quadratic model (power = 0.179; **Fig. S8**; Methods). Furthermore, this is also the estimated local minimum of predicted aneuploid conceptions from our human oocyte data (Methods). Consistent with our observations in human oocytes, we observed a significant decline in meiotic errors with maternal age within this age range ($\beta = -0.082$, 95% CI [-0.157 to 0.059], $p < 0.035$).

The elevation in aneuploidy rates in teenagers and women of advancing maternal age compared to the mid group appear to be due to different types of chromosome errors. In young teenagers, meiosis I nondisjunction is the predominant error type, whereas PSSC is the major error type in women of advancing maternal age. This suggests that different underlying molecular mechanisms cause the higher incidences in the two groups. We first considered the age-dependent increase in PSSC (**Fig. 3A**, “PSSC”). For sister chromatids to separate in meiosis I, both cohesion between sister centromeres and co-orientation of the sister kinetochores must be disrupted. How co-orientation is achieved in human oocytes is unclear given that the sister kinetochores are unfused (34, 35). However, an age-dependent decline in sister chromatid cohesion between sister centromeres in both meiosis I and meiosis II has previously been reported (34-36). The cohesion fatigue hypothesis proposes that since the major phase of cohesion establishment occurs during S-phase in mammalian fetal oocytes decades earlier, decreased maintenance or loss during dictyate
arrest causes cohesion deterioration that can result in mis-segregation (37-44). The extent of cohesion between centromeres can be inferred from the distance between sister kinetochores in either meiosis I or meiosis II oocytes (34-36). Thus, as cohesion is lessened, the distance between sister kinetochores increases. We measured 986 inter-sister kinetochore (iKT) distances in 44 MII oocytes and found that centromeric cohesion decreased with age (15 to 40 years; $R^2 = 0.24$, $p < 0.001$; **Fig. 3B**). In particular, the inter-sister kinetochore distances was not elevated in the younger oocytes (< 25 years) and appeared to follow a linear increase with maternal age (**Fig. 3B**). Maternal age explained 24% of the variance, however, an equal amount was attributed substantial variability between oocytes from the same participant. Thus, lessened centromere cohesion is positively correlated with PSSC as oocytes age. However, even in teenagers, centromeric cohesion might be affected, since oocytes from this age group display a low incidence of PSSC (**Fig. 2D, E**). This implies that meiosis I nondisjunction events may be mediated by other mechanisms in younger women.

If lessened centromeric cohesion is a predisposing factor of PSSC, then reverse segregation, when *both* homologous chromosomes segregate their sister chromatids at meiosis I, should also increase with female age (**Fig. 3A”,”RS”) (21). Since reverse segregation results in an MII oocyte and first polar body that each contain normal chromosome content but two different maternal chromatids, our next-generation sequencing pipeline would not have detected this event. However, reverse segregation is distinguishable from a normal meiosis I in that the two chromatids have different genotypes around their centromeres, reflecting their distinct parental origins (**Fig. 3C**). We obtained genotype data from 79 MII-PB1 duos around their centromeres to determine the parental origin of each chromatid (**Table S7**). As shown in **Fig. 3C**, we observed an age-dependent elevation in reverse segregation that was particularly pronounced in the advanced maternal age
submitted Manuscript: Confidential

group (≥ 33 years; p < 0.001, Fisher exact test). This may explain why some (21, 22), but not other studies (45) observed reverse segregation. The pronounced elevation in oocytes from women of advancing maternal age is also consistent with the hypothesis that reverse segregation is not simply a function of two independent PSSC events (p < 0.001, Fisher exact test,) (21). We used two datasets that cover 37 meiosis II events after a reverse segregation occurred in meiosis I (21, 22).

As seen in Fig. 3D, the segregation of the two non-sister chromatids is better-than-expected (75%) compared to random segregation, where only 50% of the activated oocytes would be expected to be euploid (p < 0.01, Fisher exact test). We therefore propose that reverse segregation may be caused by a single event that predisposes both homologs to separate sister chromatids at meiosis I, followed by non-random segregation at meiosis II. It is unclear to what extent this occurs due to loss of centromeric cohesion alone and/or an age-dependent elevation in univalents (34, 46), which may be permissive of bioriented (amphitelic) attachments and cause equational segregation of both homologs at meiosis I (Fig. 3A, “RS”) (34, 47). Taken together, our findings suggest that as oocytes age, an elevated proportion show lessened cohesion that result in aberrant segregation, specifically PSSC and reverse segregation (Fig. 3E). PSSC seems to be modelled well by a linear response to age (Fig. 2D, E), whereas reverse segregation might be explained by an exponential increase as women reach their mid-30s. Thus, the three types of meiosis I errors follow different age-dependent curves.

Having established that PSSC and reverse segregation increase with maternal age, we next considered the age-dependent decline in meiosis I nondisjunction, which is elevated in teenagers compared to the two older age groups (Fig. 3E). Unexpectedly, this was in part due to an elevation in errors that affected large chromosomes (chr.1-6; p < 0.001, Fisher exact test; Fig. 3F). Aneuploidies of chromosomes 1-6 are observed at low rates in clinical pregnancy losses, as
they are largely incompatible with implantation and fetal development (48). The convex J curve is therefore dependent on chromosome-specific vulnerabilities to the type of segregation errors that occur as a function of maternal age. I.e. in teenagers, the chromosome-specific vulnerabilities to meiosis I nondisjunction of chromosomes 1-6 may cause an elevation in chromosomally abnormal eggs and hence aneuploid conceptions, but these may only occasionally be detected clinically, since they would mostly cause preclinical pregnancy loss. In contrast, with advancing maternal age, the incidence of PSSC and reverse segregation of chromosomes 15, 16, 21 and 22 (trisomies compatible with implantation) would shape the J curve. This is consistent with their higher incidence in pregnancy losses compared to chromosomes 1-6 (48). For chromosome 21, the trisomy which underlies Down Syndrome (49, 50), our findings are consistent with observations in live births, where the incidence is low and flat for teenagers and women in their early 20s before increasing around 30-35 years (1, 23, 51, 52). Finally, some chromosomes follow the average, composite convex J curve, with initial higher incidences of meiosis I nondisjunction and then a shift towards PSSC and reverse segregation as women age. This trend was most pronounced for chromosome 13, which causes Trisomy 13 or Patau Syndrome (Fig. 3F). Taken together, our data suggest that the J curve is composed of chromosome-specific responses to maternal age, such that some chromosomes are vulnerable in teenagers (chr. 1-6), others missegregate later in life (e.g. chr. 21), and some missegregate differently – and therefore with variable impact – depending on maternal age (e.g. chr. 13).

Although a single chromosome mis-segregation event is sufficient to result in embryonic aneuploidy and pregnancy loss, multiple chromosome segregation errors cumulatively reduce the chances of a viable pregnancy. Interestingly, our data show that while aneuploidy affects a large proportion of oocytes ( > 50%, Fig. 2B), the majority of chromosomes segregate as
expected, even in aneuploid eggs, across all ages. In fact, 94% of chromosomes segregate accurately during meiosis I after ≥ 33 years of dictyate arrest (6% mis-segregate; Fig. 3E). This is true even in situations where chromosomes would be predicted to missegregate such as during reverse segregation, where the two unpaired non-sister chromatids would be expected to segregate independently of each other. Yet, in 75% of meiosis II, the two non-sister chromatids segregate normally, resulting in a euploid conception (Fig. 3D).

To understand this, we conducted high resolution imaging of the chromosomes on intact metaphase II spindles, an arrangement where chromosome architecture can be assessed in detail. We frequently detected chromatin threads between paired chromatids in meiosis II (Fig. 4A; Supplemental Movie S1). In fact, a striking 40% of meiosis II oocytes had at least one chromosome with a prominent chromatin thread bridging the two chromatids (Fig. 4B). Surprisingly, we found that 24 out of 90 meiosis II oocytes analysed exhibit at least one chromosome in which the two sister chromatids are separated by a prominent gap with no apparent DNA mass between them, as judged by Hoechst staining (Fig. 4C, D). Following such pronounced cohesion weakening, one would expect single chromatids to separate prematurely and migrate towards opposite spindle poles, generating two functionally and spatially independent single chromatids. This is because conventionally, centromeric cohesion is the sole factor promoting sister chromatid linkage during meiosis II (40). The fact that the two chromatids remained correctly aligned at the spindle equator, suggests that despite pronounced spacing they still act as a single, functional pair. This prompted us to evaluate in detail the arrangement of the two chromatids of each pair in over 1900 human meiosis II chromosomes. While the average chromatid separation in our oocyte cohort was 1.7 µm, 5% of chromosomes had two chromatids separated by a pronounced gap ranging between 2.5 and 6.6 µm (Fig. 4E). Chromatids with such prominent gaps
were nevertheless co-aligned at the spindle equator, with the two kinetochores facing opposite spindle poles. Interestingly, we frequently detected the presence of thin chromatin threads between the two chromatids (Fig. 4F). The chromatin threads frequently bridged not only the pericentromeric regions, but also the distal portions of the chromosome arms (Fig. 4F).

It is currently unclear whether the threads are present between sister- or non-sister segments (or both) of the two chromatids. Nevertheless their presence implies an unexpected role of chromatin threads in linking chromatids in meiosis II. These abundant threads may augment cohesin complexes in promoting linkage between chromatids. Their role may become more prominent once cohesin complexes are lost, such as after degradation of arm cohesin at anaphase I, which is required to release homologous chromosomes from each other (chiasma dissolution) (53, 54). Therefore, chromatin threads may explain the better-than-expected likelihood of an accurate segregation of two non-sister chromatids during reverse segregation (Fig. 3D). These threads may have a more ubiquitous function in linking chromosomes throughout female meiosis, since they may also hold homologous chromosomes together during meiosis I (Fig. S9). This may be particularly important if cohesion deteriorates as women age (37-44). More widely, chromatin threads have been reported to physically connect non-exchange and even heterologous chromosomes in meiosis of Drosophila (55-57) and Luzula (58). However, the pathological accumulation and lack of resolution of chromatin threads impede segregation of chromosomes at both meiotic divisions in a variety of organisms (59-61). Our observations suggest that in human meiosis, where cohesion appears to be predominantly suboptimal, chromatin threads provide physical connections that may improve the likelihood of correct chromosome orientation on the metaphase spindles and, ultimately, during segregation. On the flip side, failure to resolve these threads could also result in mis-segregation.
In this work, we have explored meiotic progression and chromosome segregation in human oocytes across a wide spectrum of ages (2.5 to 42 years) and found that aneuploidies follow a convex J curve. Our data suggest that in young teenagers, increasing age is protective against aneuploidy, whereas the risk of aneuploidy increase in women of advancing maternal age. Intriguingly, we found that in teenagers meiosis I nondisjunction drives aneuploidy; the incidence of PSSC appears to become elevated in women aged 20 and over and increases linearly over time; and reverse segregation shows a major increase in the advanced maternal age group. Thus, three distinct age-dependent errors shape the overall J curve. The prevalence of different segregation errors reflect chromosome-specific responses to maternal age. This complex interplay may allow the fine-tuning of error rates and ensure that distinct mechanisms determine reproductive outcomes. Thus, cohesion loss and other chromosomal aging features that predispose to PSSC and reverse segregation may act as a molecular clock that limits reproductive success as women advance in age. However, such an intrinsic mechanism would not allow elevated aneuploidy rates in teenagers, indicating that the elevation in meiosis I nondisjunction in this group is caused by other mechanisms that we currently do not understand. We note that the incidence in aneuploidy appears to be inversely related with the concave shape of natural fertility in human females (24). We speculate that the incidence of aneuploidy defines the effective reproductive lifespan and shapes the natural fertility curve in humans. High aneuploidy rates would limit pregnancies that are high-risk such as those due to insufficient physical development in teenagers (62-64) and those that may interfere with child rearing in later life (65-67).
References and Notes:


41. L. M. Lister et al., Age-related meiotic segregation errors in mammalian oocytes are preceded by depletion of cohesin and Sgo2. Curr Biol 20, 1511-1521 (2010).
43. R. Jessberger, Age-related aneuploidy through cohesion exhaustion. EMBO Rep 13, 539-546 (2012).


R. C. Team. (Vienna, Austria, 2018).


**Acknowledgments**: We are grateful to the many girls and women who participated in this study and the multitude of colleagues who helped with informed consent, fertility treatment, and ensured that the ethical framework were adhered to. Full acknowledgements are shown in the Methods.

**Funding**: ERH was funded by a NNF Young Investigator Award (NNF15OC0016662) and ERC Consolidator Grant (724718-ReCAP). JG was funded by a project grant from the DFF-FSS. ERH gratefully acknowledges general support from the DNRF Center Grant (6110-00344B). LJN and the oocytes at Warwick were collected as part of the MRC project grant MR/M000664/1. GH was funded in part from The University of Warwick and University Hospitals Coventry and
Warwickshire NHS Trust. APZ and MS have received financial support from the Max Planck Society, the European Research Council under grant agreement no. 337415 and the Lister Institute for Preventive Medicine to conduct research on human oocytes. AZ is further supported by a Rosetrees Trust PhD Fellowship.

**Author contributions:** JRG, APZ, VS and RB conducted the maturation and aneuploidy analyses. JRG and VS were responsible for Fig. 1-3, with assistance as described below. RB developed the MDA-NGS technology (Fig. 2). APZ collected and analysed data in Fig. 4, where intact spindles were preserved. AC, DC, DN, LJN, CS, MK, DT, SGK, EE, AMBB, LBC, JL, GH, MLG, KL, LR, FMU, and CYA conducted the clinical coordination and collection of human oocytes. CAC, IV helped with the genomic analyses of the data. RMC analysed the human preimplantation dataset. APZ and MS developed the high resolution imaging, analysed the data, and discovered the chromosome threads. ERH conceived the study, analysed the data, and wrote the manuscript. All authors commented and approved the manuscript. Responsibilities for ethics, informed consent and personal data protection are described in the Supplemental Methods.

**Competing interests:** The authors declare no competing interests.

**Data and materials availability:** All data are available under appropriate Material Transfer Agreements.

**Supplementary Materials:**
Materials and Methods
Figures S1-S9
Tables S1-S7 (Excel file)
Movies S1
**Fig. 1. Developmental competence of human oocytes from small- and large antral follicles.**

(A) Schematic of follicular growth in the human ovary from the small antral to mature stage. GV oocytes from Cohort 1 were obtained from small antral follicles, whereas GV and *in vivo* matured MII oocytes were recovered from large antral follicles (Cohort 2) from women undergoing ovarian stimulation.

(B) Recovery of oocytes from small antral follicles from ovarian tissue. (i) Human ovaries contain small antral follicles in the medulla (inside the dotted circle; (ii)) from which cumulus-oocyte-complexes are released when the ovarian cortex is prepared for cryopreservation (iii-v). (vi) Autotransplantation of ovarian cortex. Bar (iv): 250 µm.

(C) The age range of patients included from each fertility clinic. The LRB clinic sourced all small antral follicles. Warwick (WAR), INVICTA (VIC), and GENERA (GEN) sourced large antral follicles from gonadotrophin-stimulated women. The number of participants providing informed consent from each clinic is shown in brackets. Red boxes represent the mean age of patients from each clinic.

(D) The number of oocytes collected per patient for each cohort population. For the small antral follicles (Cohort 1), all of the 3,253 oocytes were GV stage at the time of collection (red dots). Of the 1,655 oocytes obtained from large antral follicles from gonadotrophin-stimulated women (Cohort 2), 1,201 were MII, 186 MI and 238 GV at the time of collection (grey dots). The majority of *in vivo* matured MIIs were used for fertility treatment in the IVF clinics, however 86 of the *in vivo* matured MIIs from INVICTA (VIC) were donated for this project.

(E) Maturation of cumulus-oocyte-complexes from small antral follicles (Cohort 1). After 48 hours, the oocyte was denuded and scored as mature MII by the extrusion of the first polar body (green), MI or failed polar body extrusion by the lack of both a polar body and GV (light grey), GV stage (dark grey), or degenerate, as determined by dark and/or grainy cytoplasm (black). PB1 - polar body.

(F-I) Maturation rates of oocytes obtained from small antral follicles (F, G) and large antral follicles (H). (I) Activation rates for both cohorts compared to *in vivo* mature MII (MII*) oocytes. The numbers in parentheses show the number of patients (F) or the number of oocytes (H, I). Artificial activation was carried out using a calcium ionophore (100 µM A23187), which we previously showed to be as efficient as ICSI in causing the formation
of a female pronucleus and extrusion of PB2 when used to activate \textit{in vivo} matured MII oocytes \cite{20, 21, 68}. The A23187 treatment regime does not perturb meiosis II errors \cite{20, 21, 68}. Clinic-specific \textit{in vitro} maturation and artificial activation rates are shown in Fig. S5.
Figure 2

A

PB1 Biopsy needle

MI oocyte

1N, 2C

B

Oocytes (meiosis I)

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>prop. aneuploid MI oocytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20</td>
<td></td>
</tr>
<tr>
<td>20-32</td>
<td></td>
</tr>
<tr>
<td>≥33</td>
<td></td>
</tr>
</tbody>
</table>

D

Error rate (per oocyte)

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>prop. aneuploid chr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20</td>
<td></td>
</tr>
<tr>
<td>20-32</td>
<td></td>
</tr>
<tr>
<td>≥33</td>
<td></td>
</tr>
</tbody>
</table>

E

Fertilized embryos (maternal trisomies)

H

Day 3

Day 5

36,786 biopses from 5,819 genetic testing cases

Observed data

fitted values ± SE

Day 3 Blastomere

Day 5 TE Biopsy

Prop. of embryos w. maternal BPH

fitted values ± SE

Day 3 Blastomere

Day 5 TE Biopsy

Complexity (meiosis I)

F

G

Chromated content

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

No. chromosomes affected

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>No. chromosomes affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 2. Meiotic segregation errors follow opposing trajectories in human oocytes.

(A) Next-generation sequencing (NGS) based analysis of chromosome segregation in meiosis I. The mature MII oocyte and corresponding PB1 were biopsied and subjected to whole-genome amplification and NGS. Copy number variant analysis reveals reciprocal gains and losses of single chromatids (3C and 1C) and whole chromosomes (4C and 0C), consistent with meiotic errors.

(B-E) Aneuploidy rates and error types in human oocytes after meiosis I. The total number of oocytes (B, D) or chromosomes (C, E) analysed from small and large antral follicles are shown in parentheses. The dataset includes 90 meiosis I were assessed via NGS (MII-PB1), 55 from cytological assessment of MII oocytes, and 73 from in vivo mature MII (MII*) from previously published data (21, 22). The fitted curve is shown in B and C. Age groups were binned as described in the text and are significantly different in Fig. 2B and C (Table S6). Stars denote p-values <0.05 (*) and <0.025 (**) for two-sided Fisher exact tests. Beta and the p-value are shown in red for PSSC.

(F) Scatterplot of the number of chromatids or chromosomes affected in each oocyte according to maternal age. Blue: simple aneuploidies. Red: complex aneuploidies where multiple chromosomes are affected.

(G) Oocyte counts of the number of affected chromosomes. The embedded graph shows zoomed in view of the 2-10 range of affected chromosomes.

(H) Generalized linear models fit to the PGT-A data. Lines indicate fitted values and standard errors across the maternal age range of the data, while colors distinguish sample types (day-3 blastomeres versus day-5 TE biopsies). Upper panel: A quadratic model was fit to the full dataset, and provided significantly improved fit compared to a model with a linear maternal age term ($\chi^2$ of deviance, $p = 1 \times 10^{-43}$; pseudo-$R^2 = 0.240$). Observed per-case proportions of embryos with one or more maternal meiotic chromosome gains are overlaid to visualize the underlying data. Lower panel: Focusing on the lower portion of the age distribution, we fit a linear model to the subset of data from patients younger than 27.1 years old (dotted line; Methods), revealing a negative relationship with maternal age ($\beta = -0.082, 95\% \ CI [-0.157, 5.88 \times 10^{-3}], p = 0.034$).
Fig. 3. Age-dependent decline in centromeric cohesion and increase in reverse segregation.

(A) Schematic of kinetochore attachments, their potential chromosome segregation patterns and final MII-PB1 chromosome content. Meiosis I nondisjunction (MI NDJ) is associated with co-orientation of sister kinetochores and well-established cohesion. Precocious separation of sister chromatids (PSSC) events may arise from bi-oriented (amphitelic)
kinetochore attachments and decreased centromeric cohesion. Reverse segregation (RS) is potentially associated with high rates of bioriented attachments, which cause an equational segregation pattern of both homologs at meiosis I.

(B) Sister kinetochore distances in MII oocytes according to female age. The average distance between kinetochore signals is shown and the error bars demonstrate the st. dev. of distances in the same oocyte. The data were obtained from 31 MII oocytes from small antral follicles (Cohort 1) and 13 from large antral follicles (Cohort 2). Example of spread chromosome from MII oocyte and the measurement of inter-kinetochore distances (CREST, magenta). The p-value of adj. R² are shown for a quadratic fit using all 986 data points.

(C) Schematic of reverse segregation and the incidence according to maternal age. Stars denote p-values (* < 0.05, ** < 0.025, *** < 0.01).

(D) Chromosome segregation errors at meiosis II in reverse segregation. Data from 37 events, including 26 from Ottolini et al., 2015 and 11 events from Ottolini et al. (2017). Stars denote p-values as in (C).

(E) Error rates in meiosis I. The numbers of chromosomes assessed were 552, 2530 and 1932 across the three age groups for cytological and NGS-based detection of MI NDJ (blue) and PSSC (red). 345, 713, and 759 chromosomes were assessed for reverse segregation (yellow) using SNP arrays.

(F) Chromosome errors in meiosis I broken down by chromosome and group (A to G), by maternal age. MI NDJ (blue), PSSC (red), and reverse segregation (yellow). Error bars: standard error of a proportion.
Fig. 4. Chromatin threads in human meiosis may provide physical connections for chromosomes.

(A) Chromatin threads between MII chromatids. Scale bars: 2 µm in overview and 0.4 µm in insets.

(B) Proportion of MII oocytes with chromatin threads. All analysis was performed on intact spindles.

(C) Meiosis II chromosomes contain prominent gaps but are nevertheless aligned correctly on the spindle. Arrows illustrate a chromatin thread. Scale bars: 4 µm in overview and 2 µm in insets.

(D, E) Proportion of oocytes having at least one chromosome with a pronounced gap (D) and a quantitative analysis of separation of the two chromatids across the entire oocyte cohort (E).

(F) Chromatin threads between chromatids with intact (left panel) and weakened centromeric cohesion. All chromosomes were captured prior to anaphase resumption Scale bars: 2 µm in overview and 0.5 µm in insets.
Appendix A1.4

Improving quality and quantity of human immature oocytes collected relating to freezing of cortical tissue for fertility preservation.


Submitted.
Improving quality and quantity of human immature oocytes collected relating to freezing of cortical tissue for fertility preservation

Dmitry Nikiforov*1,8, Cheng Junping*1,9, Vallari Shukla2, Robert Blanshard3, Susanne Elisabeth Pors1, Stine Gry Kristensen1, Kirsten Tryde Macklon4, Lotte Colmorn4, Erik Ernst5,6, Anne-Mette Bay-Bjørn5, Marie Louise Grøndahl7, Eva Hoffmann2, Claus Yding Andersen1

*Contributed equally

1 Laboratory of Reproductive Biology, The Juliane Marie Centre for Women, Children and Reproduction, Copenhagen University Hospital and Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark

2 DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

3 Clinical Genomics Group, Illumina UK Ltd., Fulbourn, UK. & Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Brighton, BN1 9RQ, UK

4 The Fertility Clinic, Copenhagen University Hospital, Rigshospitalet, 2100 Copenhagen, Denmark

5 The Fertility Clinic, Aarhus University Hospital, Skejby, 8200 Aarhus, Denmark

6 The Fertility Clinic, Horsens Regional Hospital, 8700 Horsens, Denmark (currently)

7 The Fertility Clinic, Herlev Gentofte Hospital, 2740 Herlev, Denmark

8 Università degli studi di Teramo, 64100 Teramo, Italy

9 People’s Hospital of Guangxi Autonomous Region, 530000 Nanning, China
Abstract:

Study question: Can immature oocytes released from small antral follicles during cortical tissue preparation for fertility preservation constitute a reliable source of mature eggs?

Summary answer: Immature oocytes released during cortical tissue preparation mature to the MII stage with a 31% success rate, resulting in an average of 11 MII oocytes per patient. Nearly two of three MII oocytes are shown to be euploid.

What is known already: Immature oocytes are released from small antral follicles (0.5-3 mm) during ovarian tissue cryopreservation for fertility restoration. Recent studies of oocytes obtained from unstimulated women suggest that they can be used for fertility treatment. The aim of the present study was to improve the procedure of immature human oocytes collection from the surplus tissue and to analyze whether these oocytes might be used clinically after in-vitro maturation, thereby augmenting the fertility potential.

Participants/materials, setting, methods: A total of 25 patients aged 17-37 years (mean age 28) who had one ovary excised for fertility preservation were included. After collection GV oocytes were divided into one of three groups: cumulus oocyte complexes (COCs) with large amount of cumulus cells, small amount of cumulus cells and naked oocytes. After maturation measurements of oocyte diameter, diameter of egg’s cytoplasm and zona pellucida thickness were taken and related to the maturation status of the oocyte. The MII oocytes were also assessed for aneuploidy using a next-generation sequencing method.

Main results and the role of chance: On average, 36 immature oocytes per patient were collected (range 7-90, N=895). 31% (N=274) matured to the MII stage within 48 hours, resulting in an average of 11 MII oocytes per patient based on improvement of media and collection procedures. The diameter of MII oocytes were significantly larger than MI and GV ones (p<0,001) and maturation rate for the oocytes with large COCs was 52%, small OCCs 26% and naked ones 12% (p<0,001). Aneuploidy rates amongst 53 oocytes from 13 patients aged 19-33 was 36% (+3%).

Limitations, reasons for caution: More patients are needed to evaluate the developmental potential of oocytes obtained by the described method. Danish legislation does not consider IVM standard practice and does not allow fertilization to generate embryos. Wider implications of the findings: The unexpected high number of immature oocytes collected from surplus ovarian tissue while maintaining a relatively good capacity for maturation allowed generation of MII oocytes in numbers similar to normal
ovarian stimulation. Provided that these MII oocytes prove suitable for IVF this will provide a substantial improvement in fertility preservation for these patients and advance IVM and very small antral follicles as an interesting platform for further improvement of assisted reproduction.

**Key words:** Human immature oocytes, fertility preservation, ovarian cryopreservation, in-vitro maturation
Introduction.
The use of immature human oocytes for infertility treatment is not widely used – only a few centers maintain the procedure in which the patient receives a few days of exogenous FSH stimulation followed by a bolus hCG trigger before aspiration of follicles 10 – 14 millimeters in diameter. Despite the introduction of the agonist trigger to avoid potential ovarian hyper stimulation syndrome the milder stimulation used in connection with the in-vitro maturation (IVM) procedure remains attractive, but the significant lower pregnancy outcome has resulted in most clinics have abandoned the procedure. However, the introduction of fertility preservation where one whole ovary is normally excised have changed this picture. The cortical tissue containing the clear majority of resting primordial follicles is isolated and frozen, while the medulla tissue containing the growing follicles which are too big and their architecture is too complicated to sustain freezing is normally discharged. Several studies have shown that immature oocytes can be aspirated from small antral follicles, normally exceeding a diameter of 3-5 millimeters, visible on the surface of the ovary after excision (Cavilla et al., 2008; Fasano et al., 2011; Abir et al., 2016). These oocytes may be able to augment the fertility potential in addition to the cortical tissue.

Four pregnancies resulting from cryopreserved embryos obtained from IVM of such oocytes have been reported from cancer survivors (Kedem et al., 2018; Segers et al., 2015, Uzelac et al., 2015; Prasath et al., 2014). However, the small antral follicles (hSAFs) in the medulla may span diameters from less than one millimeter to around 10 millimeters and beyond, and several studies have attempted to locate all immature oocytes released from small antral follicles during the preparation of the cortical tissue (Durinzi et al., 1995; Imesch et al., 2013; Wilken-Jensen et al., 2014, Yin et al., 2016). The developmental competence of oocytes is normally considered to be inversely associated to the diameter of the follicle from where it originates and oocytes from small antral follicles with a diameter of just a few millimeters or even smaller is likely to be reduced compared to that of follicles exceeding maybe 10 millimeters (Rosen et al., 2008; Wirleitner et al., 2018). Nonetheless, OTC now provides new opportunities for developing fertility treatments based on oocytes from hSAFs understanding the underlaying physiological mechanisms. In a recent study we collected an average of 11 immature oocytes from the surplus medulla tissue per patient having one ovary removed for fertility preservation with an average maturation rate to the MII stage of around 30% (Yin et al., 2016). It is unclear, however, what the aneuploidy incidence is amongst oocytes obtained from hSAFs and matured in vitro.

Classification of human oocytes are not well defined yet, often oocytes will merely be classified as either mature (i.e. MII stage with various classifications of this stage) or immature (oocytes showing a germinal vesicle). However, immature oocytes require a further classification as oocytes are collected from follicles with a range of diameters and their developmental competence evaluated with the aim
of being used clinically. Thus, it is warranted to determine characteristics which may predict the developmental competence of immature oocytes and facilitate a more precise categorization of human immature oocytes. One key parameter to evaluate is the aneuploidy rate of those oocytes from hSAFs that sustain MII transition - information which is currently lacking.

The aim of the present study was to improve the process of immature oocytes collection from surplus medulla tissue and evaluate their ability to mature to the MII stage in order to use as an additional measure for fertility preservation. We evaluated the morphometric characteristics of the oocytes including oocyte diameter, zona pellucida thickness and the importance of cumulus cell mass surrounding the oocyte and assessed the aneuploidy rate of the in vitro matured MII oocytes.

Materials and methods

Patients

A total of 25 consecutive patients with normal ovarian function (mean age 28 years, range: 17 –37) who underwent fertility preservation at the Laboratory of Reproductive Biology, University Hospital of Copenhagen, Denmark, from December 2017 to March 2018 and who consented to donate their surplus ovarian tissue for research purposes were included in the study. One patient with imminent POI was excluded from the study because her ovarian tissue a priori contained fewer oocytes than an average patient. The ethical committee of the municipalities of Region Hovedstaden approved the project (H-2-2011-044) and the patients signed an informed consent form prior to the surgical procedure. Assessment of aneuploidy and GDPR were approved by the ethical committee and data protection agency.

Collection of cumulus-oocyte complexes (COCs)

Immature oocytes were obtained from surplus medulla tissue from patients receiving freezing of ovarian tissue for fertility preservation. The medulla tissue was otherwise going to be discarded. One entire ovary was excised by unilateral oophorectomy and delivered to the laboratory from the local or collaborating hospitals. According to the Danish scheme “The patient stays - the ovary moves” the ovaries were delivered in IVF flushing medium (Origio A/S, Denmark) either at 37ºC within 10 minutes (the surgical procedure performed at the University Hospital of Copenhagen) or on ice within 2-5 hours (the surgical procedure performed at collaborating hospitals). The entire procedure of cortical tissue preparation was performed in a flow hood at the room temperature using saline or HEPES buffered HTF medium (Invitrogen, UK) at ambient temperature or pre-cooled at 4ºC, corresponding to the temperature during transportation. It was evaluated whether the maturation rate of immature oocytes depended on whether saline or HEPES-buffered medium was used to prepare the cortical tissue. Each ovary was cut into two halves and the cortical tissue was isolated using sterile scalpels and forceps as
previously described (Andersen et al. 2008; Schmidt et al. 2011). All petri dishes with saline or medium containing surplus tissue were meticulously examined for the presence of immature oocytes. In order to optimize harvest of immature oocytes, pieces of medulla tissue were finely dissected once by a tissue chopper (Mcllwain Tissue Chopper, Campden Instruments LTD, UK). Settings for the chopper insured that tissue was cut into pieces of approximately 2 mm\(^3\). Thereafter, pieces of the tissue were rinsed in saline or in the HEPES buffered HTF medium, which was also examined for the presence of immature oocytes. Collection was performed using a stereomicroscope (Leica MZ12, Germany) on 37ºC heating table in a flow hood within approximately one hour after extraction of the ovary. All immature oocytes were collected into the holding medium consisting of McCoy’s 5α-medium containing 25 mM HEPES (Invitrogen, UK), supplemented with 5 mg/ml human serum albumin (HSA - CSL Behring 20%, Germany), 2 mM Glutamax (Invitrogen, UK), 0.05 mg/ml penicillin/streptomycin (Invitrogen, UK), and 10 mg/ml of ITS solution (insulin, 5.5 mg/ml transferring, 6.7 ng/ml selenium, Invitrogen, UK). The immature oocytes (Figure 1) could be present either as cumulus-oocytes complexes or naked oocytes. After collection, immature oocytes were graded according to the size surrounding cumulus cell mass into three categories: grade 2 - more than 10 layers of cells, grade 1 - with 3-10 layers of cumulus cells surrounding the oocyte and grade 0 - with small clumps of cumulus cells or naked oocytes (Figure 2). Only healthy-looking oocytes, naked or surrounded by cumulus cells, were selected for the further culture, which corresponds to approximately 95% of all collected cells. Oocytes with obvious signs of atresia (dark or irregular ooplasm) or mechanical damage were not included in the study.

**In-vitro maturation and evaluation parameters**

For in-vitro maturation the MediCult IVM system (Origio A/S, Denmark) with modifications was used. Briefly, IVM medium was supplemented with 75 mIU/ml rFSH (Puregon, MSD, the Netherlands), 100 mIU/ml rLH (Luveris, Serono, Germany) and 10 mg/ml HSA. Immature oocytes were washed in IVM medium and placed into the 25 μL drops of media, covered by liquid paraffin (Origio A/S, Denmark). The culture dishes were equilibrated overnight prior to oocyte collection in 5% CO\(_2\) and ambient O\(_2\) at 37ºC, conditions which were maintained for the subsequent 48 hours culture period. Oocytes were cultured individually in drops and data were recorded for each oocyte separately.

After 48 hours all oocytes were denuded by repeated gentle aspiration using denudation glass pipettes with a diameter of 130–133 μm (Vitrolife, Sweden). Denuded oocytes were examined under inverted microscope (Carl Zeiss Axiovert 135, Germany) equipped with Hoffman interference optics with 20X objective lenses and classified as germ vesicle stage (GV), Metaphase-I (MI) or Metaphase-II (MII). GV stage oocytes were identified on the basis of their distinct GV, refractive nucleolus, darkened center and granular ooplasm; MI oocytes were identified as oocytes without either a GV or polar body; and MII oocytes were identified as oocytes with the first polar body by rotating oocyte manually (Figure
3). Oocytes which remained immature (GV or MI) after 48 hours were regarded as incapable of maturation.

All oocytes were imaged through the oocyte equator with the clear oolemma in the focus, using the Carl Zeiss microscope and the AxioVision software (SE64 Rel.4.9.1). Two perpendicular measurements were obtained from each oocyte and the mean diameter used – see figure 4. Data were stored for later analysis which included measurements of the external and internal zona pellucida diameters, ooplasm diameter and zona pellucida thickness. The value was obtained by calculating an average of the parameters measuring twice perpendicularly for each specific oocyte.

Oocytes vitrification

Mill oocytes were vitrified in accordance with the protocol described by Lucena et al., 2006 with modifications. Briefly, oocytes were washed in three drops of basic solution (BS), which contained 199-medium with 10% HSA. Then the last third drop of BS was merged with the first drop of equilibration solution (ES), contained 7.5% (v/v) of dimethyl sulfoxide (DMSO - Sigma, Germany) and 7.5% (v/v) of ethylene glycol (EG - Merk, Germany) in the BS for three minutes. Then previous drop was merged with the second ES drop for another 3 minutes and finally oocytes moved into the third ES drop for 6 minutes. During this process, three drops with vitrification solution (VS) were made, which contained 15% (v/v) EG, 15% (v/v) DMSO, 0.5 M sucrose (Sigma, Germany) and 10% HSA. After equilibration, oocytes were transferred into VS drops for 20 seconds in each drop, then loaded on the Cryolock carrier device (Biotech, USA) and immediately plunged into liquid nitrogen. The total time in VS was up to 1 minute. Volume of all drops was 30 µL and entire procedure was performed at room temperature.

Warming of oocytes after vitrification

One milliliter of thawing solution (BS with 1 M sucrose) was pre-warmed in a central well dish at 37°C. When the carrier was removed from the liquid nitrogen, the tip of the carrier was quickly immersed into TS1 for one minute. Then oocytes were moved into 1 mL of dilution solution 2 (DS) for 3 min, which contained 0.5 M sucrose in the BS and finally to 1 ml of BS for 5 min (last step was repeated twice). DS and BS were pre-warmed at the room temperature for 30 minutes and entire procedure was performed at the room temperature.

First polar body biopsy and tubing

Vitrified/warmed oocytes were kept for at least one hour prior to biopsy in a cleavage medium (Origio A/S, Denmark) in individual drops under the oil in an incubator with 5% CO₂, ambient O₂ at 37°C. Individual oocytes were fixed in a micro-manipulation system (Narishige, Japan) by a holding pipet. For the dissection of the zona pellucida a laser was employed (Research Instruments, UK). Polar body was aspirated in the manner in which it did not entirely enter the biopsy pipet (Origio A/S, Denmark) and
released in the same drop as oocyte. Then, with further assistance of the laser and biopsy pipet, zone pellucida was removed. The biopsied first polar body and the oocyte were tubed separately with the stripper pipet (Origio A/S, Denmark) in a DNA/RNA-free test tube in a volume of approximately 5 μl. Immediately thereafter test tubes were flash-frozen in a liquid nitrogen and stored at – 80°C prior to further procedures.

Whole genome amplification and next-generation sequencing
To amplify the DNA from the single MII oocyte, we used the SureMDA (Illumina Inc.) following the instructions by the manufacturer (User guide: 15052710, Revision B; Illumina Inc.)(Blanshard et al., 2018). Briefly, working in a pre-PCR area we centrifuged the tube at 300 × g for 3 min., followed by lysis immediately afterwards. After collection of the lysis buffer and single oocyte sample, we centrifuged the sample for 5 sec. at 300 × g followed by addition of the stop solution and amplification reagents, using a reduced amplification time of 2 hours at 30 °C. We assessed the quality of the SureMDA product using gel electrophoresis and generated sequencing libraries using a custom Nextera®-based library preparation workflow (Illumina Inc.). MDA products were quantified (Quant-IT™ dsDNA High-Sensitivity Assay Kit, Thermo Fisher Scientific Inc.), and 100 ng was tagmented according to the Nextera® DNA Library Prep Reference Guide (Part: 15027987 v01; Illumina Inc.). Tagmented DNA was cleaned up by 1.8× Solid Phase Reversible Immobilisation (SPRI) (Part: 15041032; Illumina Inc.), prior to limited-cycle PCR amplification for addition of indexed barcodes with a modified program: 72°C for 3 min, 98°C for 3 min; 10 cycles of 98°C for 30 sec, 60°C 30 sec, 72°C 30 sec; 72°C 5 min, 10°C hold. Amplified libraries were cleaned up by 1× SPRI and normalized according to the VeriSeq® PGS Library Prep Reference Guide (Part: 15052877 v03; Illumina Inc.). Normalized libraries were pooled for 24-plex sequencing using a MiSeq System using Reagent Kit V3 consumables (2 × 36 bp; Part: MS-102-3001; Illumina Inc.). The data were analysed for copy number variation by extracting the .bam files and importing them into BlueFuse Multi® software (Illumina Inc.) with a custom, MDA-specific reference database. The analysis pipeline was validated using single cells and genomic DNA from four cell lines obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, New Jersey, USA, of which two contained structural abnormalities: GM00526 (47, XY, +13), GM04927 (47, XY, +21), GM0121 (46, XY, -18p 15.5 Mb, +18q 59.3 Mb), GM10985 (46, XX, -3p 10.3 Mb). Cell lines were cultured according to the supplier’s recommendations, and single cells were isolated manually using the method described previously (Blanshard et al. 2018). Per chromosome the concordance was greater than 99% for single cells (n=2,208).

Statistical analysis
The frequency of maturation for immature oocyte to MII was modeled as a mixed logistic regression with maturation to MII (yes/no) as outcome and transport (yes/no), medium (saline/HEPES-buffer),
and diameter of oocyte (linear effect) as explanatory variables. Patient effect was included as a random intercept. A cut-off value for the oocyte diameter where a 50% maturation rate was achieved (MII = yes) was determined using the predicted values from the mixed logistic regression. Analysis of maturation rates was done using Fisher exact test and diameters with one-way analysis of variance (ANOVA).

Results
Oocyte retrieval
A total of 895 immature oocytes from 25 patients were collected from the surplus tissue remaining after isolation of the cortical tissue for cryopreservation. After the collection and prior to the start of the IVM procedure any immature oocytes recognized as having degenerated (approximately 5%) were excluded from the study. After the removal of initially degenerated oocytes the average number of oocytes collected per patient was 36 ±5 (mean ±SEM, range 7–90). Patient characteristics and outcome parameters regarding the rates of maturation in relation to whether the ovary was transported or originated form the local hospital prior to procurement are shown in table 1. No significant correlation was observed between the number of oocytes collected and the age of patients (R²=0.03), as well as between the number of oocytes collected and the ovarian volume (R²=0.26).

Maturation rate
The overall IVM rate was 31% (274 MII oocytes out of total 895), however with variability between participants (range 6–55%). The size of the cumulus cell mass was significantly associated with maturation rate, with oocytes surrounded by a large cumulus cell mass being more likely to mature (52%) than oocytes enclosed in small cumulus complexes (26%) or naked oocytes (12%) (p<0.001, table 2). The IVM rate from patients below 20 years of age (n=2, 40%) was not different with the patients both aged 20–30 years (n=14, 29%) and above 30 years old (n=9, 31%; p>0.1).

Impact of transportation on the oocyte maturation rate
Oocytes collected from ovaries originating from patients operated at the local hospital (n=7) matured with a significantly higher rate (42%) as compared to oocytes from ovaries transported on ice (n=18) prior to dissection for cryopreservation (27%; p<0.001, table 1). The number of large COSs were similar irrespective of whether the ovary was transported or not (33 versus 40 %). Maturation rate for oocytes collected from ovaries dissected in HEPES buffered medium (n=12) was significantly higher (36%) as compared to oocytes collected in saline (n=13), which matured with the rate 26% (p<0.001). Consequently, oocytes collected in HEPES buffered medium from the local hospital matured with significantly higher rate (49%) than oocytes from the local hospital collected in saline (30%) or than that of both HEPES and saline groups of transported ovaries (29% and 26 %, respectively) (p<0.05), (table
Oocytes diameter
The average diameter of MII stage oocyte (n=274) was 110 ± 7 μm, 105 ± 9 μm for MII stage cell (n=223) 96 ± 12 μm for GV (n=117) and 88 ± 19 μm for degenerated ones (n=281). The diameters for all stages were significantly different (p<0.001). In contrast, the thickness of the zona pellucida were not significantly different among the groups of oocytes after 48 hours of IVM culture (20 ± 4, 21 ± 5, 20 ± 5 and 21 ± 4, respectively) (p>0.1) (table 4).

To understand the effect of the factors such as collection site and medium on the diameter, we developed a linear model. The equivalent diameter of 50 % MII transition, (i.e. the diameter at which 50% of the oocytes will mature to MII stage), under different conditions was calculated on the basis of linear model of diameters for all oocytes included in the study (Figure 4). Clearly, oocytes originated from ovaries delivered from the local hospital and dissected in HEPES-buffered medium show the lowest ED50% value (104.3 μm) as compared to any other group (Figure 4). The average diameter of oocytes collected in connection with IVF treatment is around 110 – 112 μm (Romao et al., 2010).

Chromosomal error rates amongst MII oocytes after IVM of GV oocytes obtained from hSAFs.
Gain or loss of chromosomes (aneuploidy) during the meiotic divisions causes unbalanced genomes that are result in preclinical and clinical pregnancy losses. We used a next generation sequencing method to assess the aneuploidy rate of the mature MII oocytes after IVM (figure 5). To do this, we removed the PB1 and isolated the MII oocyte, which was subjected to whole-genome amplification. Analysis of the shallow next-generation sequencing profiles (0.01x) revealed that 36% ± 3.3% of the 53 oocytes sampled from 13 patients aged 19 to 33 years were aneuploid (Fig. 5,B and C). Thus, we estimate that for women below 35 years that around two-thirds of the MII oocytes obtained after IVM of GV oocytes from hSAFs are euploid with a normal chromosome constellation.

Discussion
The present study found an unexpected high number of immature oocytes from ovarian medulla tissue left over from isolation of the cortical tissue. On average 36 immature oocytes were obtained from one ovary of an unselected group of patients with a few patients that provided nearly 100 immature oocytes. The IVM procedure yielded on average 11 MII oocytes with the one patient having 30. Obtaining such a high number of oocytes from a single ovary is a result of meticulous recovery of oocytes after the dissection of ovary for the cortex cryopreservation. This also include oocytes from very small antral follicles, which are barely visible with the naked eye and most of them probably having diameters well below 2-3 millimeters. Therefore, this study is one of the first to demonstrate that oocytes
obtained from such small follicles may have potential for fertility treatment (Durinzi et al., 1995; Imesch et al., 2013; Wilken-Jensen et al., 2014, Yin et al., 2016).

The genomic analyses of the MII oocytes revealed that 63% were euploid, which is approximately 10% lower than estimates obtained from MII oocytes retrieved from large- and mature follicles after ovarian stimulation (Plachot et al., 2001). However, the wide range of aneuploidy rates reported for women below 35 years makes it difficult to compare our figure, although it seems elevated (Hoffmann et al., 2018; Capalbo et al., 2017). Despite the slight elevation in aneuploidy rates, the large number of oocytes retrieved from the ovarian tissue provides MII oocytes in numbers comparable to one cycle of ovarian stimulation, thus on average, eight euploid oocytes would be recovered that could be used for fertility treatment. The embryonic development from oocytes matured in vitro might be lower than those that mature in vivo in stimulated cycles, and the efficiency of embryo formation remains to be determined. Currently, the Danish health authorities consider IVM experimental and further ethics permission would have to be obtained before the MII oocytes can be fertilized and the constitution of embryos assessed.

Our study demonstrated that optimizing the collection procedure and by use of more appropriate media to support immature oocytes during the preparation of cortical tissue help to preserve better the developmental capacity of the immature oocytes. Surprisingly, immature oocytes that were cooled to around zero degrees during transportation for 2 to 5 hours showed a significant reduction in their developmental capacity compared to those from the local hospital. Nevertheless, the magnitude of the recovery of oocytes justifies their collection and putative use for fertility preservation. Collectively, this suggests that immature oocytes obtained from the surplus medulla tissue may provide a source of additional fertility to women receiving OTC to a much larger extent than previously thought.

Previous studies isolating immature oocytes from surplus medulla tissue in connection with OTC found an average of 11—15 oocytes per patient (Escribà et al., 2012; Wilken-Jensen et al., 2014; Segers et al., 2015; Yin et al., 2016). The present study more than doubled the harvest of immature oocytes and is the result of careful search for oocytes and introducing an additional chopping of the remaining medullar tissue which is likely to release more oocytes from very small antral follicles. Further, the present study demonstrated the potential of accumulating immature oocytes by optimizing the collection procedures and improved the maturation rate by using more appropriate media for preparing the cortical tissue and thereby enhancing the developmental capacity of the immature oocytes. It this context, it is interesting to notice that some of the oocytes appear in very large cumulus oocyte complexes with multiple layers of cumulus cells which is an appearance seldom observed in normal IVF. Surprisingly the category of large COC’s show the highest maturation rate of 52%. We hypothesize that many
cumulus cells sustain oocyte development better than lower number, but are currently unable to pinpoint to which specific follicle diameter they derive from.

It would be expected that the majority of oocytes harvested from very small antral follicles were of low quality, which only would be able to support MII transition in vitro to a limited extend. However, the MII rate was almost similar to previous studies (Revel et al., 2009; Fasano et al., 2017; Hourvitz et al., 2015) and therefore the number of resulting MII oocytes were also increased significantly. This is an important outcome of the current study and probably reflects the release of the immature oocytes to a culture medium rather than to saline. In addition, this measure obviously needs to be combined with the general accepted rules for performing ART procedures: fast handling of cells when out-side of incubator concomitantly with meticulous control of temperature, pH and medium osmolarity.

The present study was unable to evaluate directly the clinical benefit of these MII oocytes as the IVM procedure is not approved by Danish authorities, which means that it is not allowed to generate and transfer embryos derived from IVM oocytes. This is an obvious limitation of the present study since it MII oocytes originating in very small follicles may possess an attenuated developmental competence.

The average diameter of in-vitro matured MII oocytes from small antral follicles (110 ±7µm) is similar to the mean diameter of oocytes retrieved at IVF clinics after ovarian stimulation (112 ±3µm, Romao et al., 2010). At the same time, the diameter of oocytes originating from hSAFs appears to reflect their developmental capacity: the larger the better chance of supporting MII transition. This observation confirms and extends an earlier study that found a similar result where 75% of 49 oocytes collected from excised ovaries from six women underwent MII transition provided the diameter exceeded 105 µm, while only 25% matured when diameter was less than that (Durinzi et al., 1995). It is interesting that the different conditions employed to improve the conditions for oocyte collection and maturation become reflected in the equivalent diameter of 50 % MII transition (ED 50% MII transition). The ED 50% value is significantly reduced when comparing local ovaries at room temperature versus transported ovaries at zero degrees. The same result is observed when HEPES buffered media versus saline was compared. Collectively this suggests that the diameter of immature oocytes from small antral follicles reflect their developmental capacity and that the measure ED 50 % transition is a useful parameter that can be employed to optimize conditions for human immature oocytes including collection and culture. In addition, these results introduce oocyte diameter as a new first parameter to subdivide or categorize the large heterogenous group of human immature oocytes. In this context, it is interesting to notice that the MII rate of oocytes with a diameter of less than 100 µm was as modest as 3% (9
MII oocytes out of 289 oocytes). We therefore suggest distinguishing immature oocytes based on diameter and propose an initial two categories: human immature oocytes with diameters below or above 100 µm based on their developmental competence.

Follicular development from the primordial stage until the preovulatory stage has traditionally been divided into three phases, a gonadotrophin independent growth phase, a growth phase requiring gonadotropins at a tonic level and a phase requiring cyclic levels of gonadotropins. The transition of follicles from the tonic to the preovulatory growth phase is usually considered to take place when follicles reach a diameter of around 0.5 to 2 millimeters (Gougeon, 1996). The atresia rate of human follicles with diameters of around 0.5 to 2 millimeters has been estimated to be relatively small of only 15-24 % (Gougeon, 1996). In contrast the atresia rate of follicles with diameters of from 2 to 10 millimeters have been estimated to be between 58 and 77 %. The reason for this augmented atresia rate is unknown but it may be hypothesized that oocytes from very small antral follicles as obtained in the present study may circumvent a stage in follicular development where a high atresia rate exists and by being placed directly in an environment with sufficient FSH stimulation the oocytes escape and overcome the natural selection and achieve a developmental competence corresponding to oocytes from larger follicles.

Although collection of immature oocytes was intended to augment the fertility potential of women having one ovary excised for OTC, the unexpected high number of immature oocytes collected may pave the way for justification of the procedure itself. Women in whom transplantation of ovarian tissue is not warranted because of a risk of reintroducing the original disease may benefit from this procedure. This would include women with ovarian cancer, borderline ovarian cancer, granulosa cell tumor or cancers with a high risk of infiltration of malignant cells into the ovarian tissue, such as leukemia or colon cancer. In some of these cases, the intervention could be associated with the surgical removal of tumor tissue from the ovary. OTC is currently not offered to many women with ovarian cancer because the risk of reintroducing the malignancy due to transplantation. However, collecting immature oocytes would eliminate this risk since the oocyte will not contain malignant cells. Therefore, this is a new procedure, which for the first time in itself may provide a group of women, who previously were without any chances of saving their fertility, with a chance of having their own child.

The longer-term perspective of the present findings is potentially the dawn of fertility treatment without stimulation with exogenous hormones in certain groups of women. Up to 100 immature oocytes may be harvested from one ovary of women with for instance PCOS. It may be envisioned that collection of small antral follicles can be optimized by development of new procedures for aspiration of very small follicles. In women with PCOS and two ovaries it may become possible to collect a sufficient
number of immature oocytes without ovarian stimulation to create equal or even a higher number of MII oocytes for treatment as compared to conventional ovarian stimulation. If necessary, repeated aspirations of oocytes may be performed with stockpiling of frozen oocytes in between.

In conclusion, the present study demonstrated that an unexpectedly high number of immature oocytes from surplus medulla tissue can be obtained in connection with fertility preservation. Despite many of these oocytes derived from antral follicles with a diameter of just a few millimeters or maybe even less, the developmental competence in terms of MII transition and was maintained at an overall rate of around one in three oocyte showing an acceptable aneuploidy rate. It is hypothesized that oocytes from such small follicles undergo an accelerated developmental competence by a quick physical handling and placing them in an environment with high levels of FSH whereby they avoid degenerative changes that they could have encountered in vivo. The study suggests a sub-division of the large ill-defined category of immature human oocytes based on an oocyte diameter of 100 μm, which in the present study with high precision separates oocytes which poses the capacity to undergo MII transition from those which are unable to sustain this transition. It is noticeable that maturation of immature oocytes as described in the present study results in generation on an average of 11 MII oocytes, of which nearly two thirds are euploid, and which is similar to most ovarian stimulation regimes, which show an interesting developmental potential for infertility treatment in the future.

Acknowledgements

The following grants are gratefully acknowledged for financial support. They had no role in the study design, collection and analysis of data, data interpretation or in writing the report: ERH, VS and is grateful for the support from The Independent Research Fund Denmark (No.), the NNF Young Investigator Award (grant number 16662), the ERC (grant number 724718) and the Danish National Research Foundation (DNRF115, 6110-003448), the project ITN REP-BIOTECH 675526 funded by the European Union, European Joint Doctorate in Biology and Technology of the Reproductive Health is thanked for support to DN, The University Hospital of Copenhagen, the EU interregional project ReproUnion and Vera and Carl Johan Michaelsens Legat CYA.

All personnel including the clinical activities in fertility preservation are thanked for their passionate work.
References


Fig. 1. Immature oocytes of different categories (large COCs, small COCs and naked oocytes) after collection.

Fig. 2. Categories of immature oocytes: A - naked immature oocyte, B - small COCs, C - large COCs.
Fig. 3. Stages of oocytes after 2 days of IVM. A - MII, B - MI, C - GV, D - degenerated oocyte

![Image of oocyte stages]

100 μm

<table>
<thead>
<tr>
<th></th>
<th>ED-50 % value HEPES group</th>
<th>ED-50 % value saline group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transferred on ice</td>
<td>109,6 b</td>
<td>115,0 c</td>
</tr>
<tr>
<td>Not transported</td>
<td>104,3 s</td>
<td>110,0 b</td>
</tr>
</tbody>
</table>

Figure 4. Equivalent diameter of which 50% (ED 50%) of oocytes collected from surplus ovarian tissue will mature to MII stage - Suggested to become a new measure to optimize IVM media composition?

a,b: a,c; b,c: P<0.002 - b,b: P>0.10
Figure 5. Aneuploidy rates in GV oocytes from hSAFs after in vitro maturation (IVM). (A) The first polar body was removed and the MII oocyte was used as template for single-cell whole-genome amplification (WGA) and shallow next-generation sequencing (NGS, 0.01×). (B) The fraction of MII oocytes that were euploid is shown in green, whereas the fraction that were aneuploid is represented by gray bar. Error bars represent standard error of a proportion. The number above the graph is the number of oocytes. In total, 53 MII oocytes were assessed after IVM. (C) The frequencies of euploid (green) and aneuploid MII oocytes per participant, ordered according to chronological age.

Table 1. Patient characteristics and outcome of IVM. Data are presented by mean value per patient ±SEM (range)

<table>
<thead>
<tr>
<th></th>
<th>No. patients</th>
<th>Age (years)</th>
<th>Ovarian volume (mL)</th>
<th>No. oocytes per patient</th>
<th>No. MII oocytes per patient</th>
<th>Maturation rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovaries transported</td>
<td>18</td>
<td>28 ±1 (17-36)</td>
<td>9 ±1 (2,7-19,8)</td>
<td>38 ±6 (9-90)</td>
<td>10 ±2 (1-27)</td>
<td>27 ±2 (6-38)</td>
</tr>
<tr>
<td>Ovaries local hospital</td>
<td>7</td>
<td>28 ±3 (19-37)</td>
<td>9 ±2 (3,4-16,5)</td>
<td>31 ±7 (7-56)</td>
<td>13 ±4 (2-30)</td>
<td>42 ±5 (23-55)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>25</td>
<td>28 ±1 (17-37)</td>
<td>9 ±1 (2,7-19,8)</td>
<td>36 ±5 (7-90)</td>
<td>11 ±2 (1-30)</td>
<td>31 ±2 (6-55)</td>
</tr>
</tbody>
</table>

* 100 oocytes were obtained from one of the patients, however, ten of these was transferred to another study
### Table 2. Maturation and cumulus expansion for immature oocytes in relation to the initial morphology

<table>
<thead>
<tr>
<th>Category of oocyte</th>
<th>No.</th>
<th>No. MI oocytes (day 2)</th>
<th>Maturation rate (%)</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. COCs expanded cumulus</td>
<td>Cumulus expansion rate (%)</td>
</tr>
<tr>
<td>Naked oocytes</td>
<td>212</td>
<td>25</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Small COCs</td>
<td>399</td>
<td>102</td>
<td>26</td>
<td>54</td>
<td>14</td>
</tr>
<tr>
<td>Large COCs</td>
<td>284</td>
<td>147</td>
<td>52</td>
<td>56</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>895</td>
<td>274</td>
<td>31</td>
<td>110</td>
<td>16</td>
</tr>
</tbody>
</table>

### Table 3. Maturation rate in relation to the medium used during isolation of the cortical tissue. Data are mean per patient ±SEM (range).

<table>
<thead>
<tr>
<th>Dissection medium</th>
<th>No. patients</th>
<th>No. oocytes per patient</th>
<th>No. oocytes matured</th>
<th>Maturation rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saline</td>
<td>11</td>
<td>39 ±6 (13-88)</td>
<td>10 ±2 (1-27)</td>
<td>26 ±3 (6-34)</td>
</tr>
<tr>
<td>HTF HEPES</td>
<td>7</td>
<td>37 ±12 (9-100)</td>
<td>10 ±3 (2-24)</td>
<td>29 ±3 (13-38)</td>
</tr>
<tr>
<td>Saline</td>
<td>2</td>
<td>41 ±3 (38-44)</td>
<td>13 ±3 (10-15)</td>
<td>30 ±8 (23-39)</td>
</tr>
<tr>
<td>HTF HEPES</td>
<td>5</td>
<td>27 ±9 (7-56)</td>
<td>13 ±6 (2-30)</td>
<td>49 ±5 (29-55)</td>
</tr>
<tr>
<td>Saline</td>
<td>13</td>
<td>39 ±5 (13-88)</td>
<td>10 ±2 (1-27)</td>
<td>26 ±3 (6-39)</td>
</tr>
<tr>
<td>HTF HEPES</td>
<td>12</td>
<td>33 ±8 (7-100)</td>
<td>12 ±3 (2-30)</td>
<td>35 ±4 (13-55)</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>36 ±5 (7-100)</td>
<td>11 ±2 (1-30)</td>
<td>31 ±2 (6-55)</td>
</tr>
</tbody>
</table>

### Table 4. Oocyte Diameter and zona pellucida thickness in relation to the maturation stage after 48 hours of IVM. Data are mean ±SEM (range)

<table>
<thead>
<tr>
<th>Oocyte stage</th>
<th>No. oocytes</th>
<th>Diameter of oocyte (µm) (Mean ±SEM)</th>
<th>ZP thickness (µm) (Mean ±SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MII</td>
<td>274</td>
<td>110 ±7 (83-148)</td>
<td>20 ±4 (10-33)</td>
</tr>
<tr>
<td>MI</td>
<td>223</td>
<td>105 ±9 (75-153)</td>
<td>21 ±5 (10-51)</td>
</tr>
<tr>
<td>GV</td>
<td>117</td>
<td>96 ±12 (49-118)</td>
<td>20±5 (8-32)</td>
</tr>
<tr>
<td>DEG</td>
<td>281</td>
<td>88 ±19 (37-142)</td>
<td>21 ±4 (8-38)</td>
</tr>
</tbody>
</table>
Appendix A1.5

Gross chromosomal rearrangements in human oocytes and preimplantation embryos


*Contributed equally

Submitted.
“Gross chromosomal rearrangements in human oocytes and preimplantation embryos”.

R. Blanshard1,2†, V. Shukla3†, A. Zielinska4, L. J. Newnham1, J. Gruhn4, A. D. Widger1, A. Capalbo5,6, L. Ju1, M. Krapchev7, A. Mirecka7, S. G. Kristensen8, Danilo Cimadomo9, L. Xanthopoulou10, S. Newcombe, R. B. Helweg-Larsen4, G. Jakiel7,11, J. Liss6, L. Rienzi8, F. Ubaldi8, K. Lucaszuk6,12,13, T. Gordon9, M. Large10, C. Y. Andersen7, M. Schuh3, A. H. Handyside2,9, and E. R. Hoffmann1,4*

1Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, BN1 9RQ, UK.
2Clinical Genomics Group, Illumina Inc., Fulbourn, UK.
3DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark.
4Max Planck Institute for Biophysics, Department of Meiosis, Göttingen, Germany
5IGENOMIX, via Fermi 1, Marostica (VI), Italy
6University of Rome, Sapienza, Dipartimento di Scienze, Anatomiche, Istologiche, Medico-Legali e dell’Apparato Locomotore, Sezione Istologia ed Embriologia Medica, Rome, Italy.
7INVICTA Fertility Clinic, Gdańsk, Poland.
8Laboratory for Reproductive Biology, Rigshospitalet, Denmark
9Genera Center for Reproductive Medicine, Clinical Valle Giulia, via de Notaris 2b, Rome, Italy
10Genesis Genetics, London, UK
11Department of Obstetrics and Gynecology, Center of Postgraduate Education, Warsaw, Poland
12Department of Obstetrics and Gynecological Nursing, Faculty of Health Sciences, Medical University of Gdansk, Gdansk, Poland
13Department of Gynecological Endocrinology, Medical University of Warsaw, Warsaw, Poland
14The Bridge Clinic, London, UK

†These authors contributed equally to the work.

*corresponding author: eva@sund.ku.dk
Abstract.

Genome instability can result in gross chromosomal rearrangements (GCRs) such as translocations or deletions. GCRs occur de novo in the human germline and cause serious genetic disorders such as Wolf Hirschon and Jacobsen syndrome\(^1\textsuperscript{-3}\). The origin and incidence of GCRs in our germline are unclear due to the challenges associated with accessing and analyzing material from embryos and oocytes. Here we show that despite their low incidence in live births (5 in 10,000), large GCRs (>5Mb) occur at an unexpectedly high frequency in human eggs (10%) and preimplantation embryos (16-24%). The breakpoints are enriched for recombination hotspots (>30-fold) and to a lesser extent, common fragile sites (10-fold). Unlike live borns, human oocytes are enriched for centromeric recombination that may explain why 28% of breakpoints occur within 5 Mb of the centromere. Single cell mRNA sequencing reveals heterogenous expression of the GCR network genes\(^4\). Using a mouse model with reduced SMC6 dosage, we provide proof-of-concept that GCR genes may be rate limiting in preventing chromosome breaks in vivo. Our findings indicate that GCRs are common in the human germline and may in part by caused by heterogeneity in the expression of GCR genes. Thus, the differences in magnitude between the germline and live births suggest a high attrition rate of conceptions with GCRs, likely due to implantation failure and early pregnancy loss\(^5\textsuperscript{,}\textsuperscript{6}\).
Main text.

To understand genome instability in the human germline, we focussed on gross chromosomal rearrangements (GCRs), since their impact are significant given the size of the structural rearrangement (> 5 Mb). This may lead to genetic dysregulation, genomic imbalance and loss of heterozygosity. GCRs result in segmental aneuploidies and are associated with 6% of first trimester miscarriages\textsuperscript{6,7}. We analysed the incidence of GCRs in 901 embryo biopsies from 206 patients using a next-generation sequencing method \textsuperscript{8,9} (Fig. 1a). The embryos were generated for fertility treatment and the primary purpose was to screen for whole chromosome aneuploidy (e.g. Down Syndrome) to prioritise the transfer of euploid embryos for pregnancy and live births in IVF settings.

We detected 71 GCRs in 60 samples from 47 patients (6.0% of blastocyst embryos), with a copy number change indicative of the GCR being uniformly present throughout the cells in the biopsy sample (Fig. 1a). A further 118 segmental aneuploidies were ‘mosaic’, being present only in subset of cells from the sample\textsuperscript{10,11}. The latter is consistent with an origin of GCRs during the embryonic divisions or arising through technological issues\textsuperscript{12-14}. The combined incidence (18.1% of biopsies) is similar to that reported using different technologies\textsuperscript{14,15}. It is conceivable that a much higher proportion of preimplantation embryos contain GCRs, when considering their high incidence amongst cleavage stage \textsuperscript{15} and arrested embryos\textsuperscript{15-17}, which make up ≥ 30% of preimplantation embryos in IVF clinics. Thus, 16 to 24% of preimplantation embryos may contain a GCR. We used a conservative approach and focussed on uniform GCRs (n=71) to avoid overestimating GCRs.

Several observations suggest that GCRs arise independently of whole chromosome aneuploidies, which affect between 20 to 90% of human conceptions, depending on the maternal age\textsuperscript{18,19}. First, amongst the blastocyst embryos, the proportion of biopsies that
contained both a whole chromosome and segmental aneuploidy was not dissimilar from that expected from independent events (2.4% versus 1.4%; $\chi^2 1.8$, p > 0.17, power: 0.68). Secondly, the incidence of GCRs was unaffected by maternal age (Fig. 1b) and GCRs were not enriched for chromosomes 15, 16, 21 and 22, which are highly error-prone for whole chromosome aneuploidies. Rather, GCRs affect chromosomes uniformly (Fig. 1c; Wilcox, p < 5 × 10^{-5}). The size of the GCR ranged from 10 to 150 Mb (Fig. 1d) and we observed an equal number of gains and losses (39 and 40, respectively). 28% affected an entire chromosome arm, with a preponderance for loss (Fig. 1e), suggesting that (peri)centromeric regions are particularly prone to a chromosome break (OR: 2.48, C.I. 95% 1.2-5.3; p < 0.01, exact test). Importantly, we observed a range of chromosome arm deletions that have been reported to occur at low incidence in prenatal screening, miscarriage, and live births. Although in vitro propagation of mammalian preimplantation embryos may increase the rate of genome instability, our observations detect types of events that are seen in natural conceptions and support the conclusion that GCRs may contribute towards the substantial early pregnancy losses in human.

To understand the origin of GCRs that occur uniformly in the trophectoderm biopsies, we assessed GCRs in human oocytes. We developed a single-cell genomics pipeline to analyse GCRs in human oocytes and their corresponding polar bodies (Fig. 2a, Methods). By analysing both the oocyte and first polar body (MII-PB1 duo), we reduced the false discovery rate from 5% to less than 3 × 10^{-6}, since we detect reciprocal events, i.e. the loss of a segment in the egg was matched by the gain in the first polar body and vice versa. In total, we obtained data from 178 meioses and detected 29 segmental aneuploidies in 21 of the MII-PB1 duos (11.8%; Fig. 2b). The presence of the GCR in two cells is consistent with the interpretation that the segmental aneuploidy was generated or present at the first meiotic division (meiosis I).
To obtain further evidence of the high rate of GCRs in human female meiosis, we inspected three independent datasets where in vivo matured MII oocytes obtained from women undergoing controlled stimulation in IVF clinics were activated artificially or fertilized with sperm and where polar body analyses had been conducted as well. GCRs were reported in all of the three independent studies in 11 of 111 oocyte-PB1-PB2 trios (10%)\textsuperscript{17,26,27}. None of the oocyte donors contained a GCR in their blood-derived DNA or other oocytes, consistent with the interpretation that the GCR was generated de novo during meiosis (Fig. 2c). We conclude that up to 12% of human eggs display inherent genomic instability leading to the formation of a structurally abnormal chromosome, the vast majority of which are incompatible with fetal development and live birth.

Next, we asked whether human oocytes display cytological evidence of chromosome deletions or unbalanced translocations. Chromosomes are naturally highly compacted after the first meiotic division during metaphase II arrest. We either spread the chromosomes and evaluated their morphology (Fig. 2d, e) or examined chromosomes intact on the metaphase II spindle using high resolution imaging (Fig. 2f, g). 7.4% (5.8-9.1%, n=54) the MII oocytes contained a broken or atypical chromosome structure consistent with a GCR. The rate is slightly lower than the NGS-based assessments (Fig. 2b, c), likely due to smaller GCRs not being detected cytologically. In summary, we observe GCRs by single-cell NGS as well as cytological means, which support the conclusion that GCRs occur in up to 12% of human female meioses.

At least two mechanisms may result in the high, intrinsic rate of genome instability in human oocytes and preimplantation embryos. Common fragile sites (CFS) are expressed during DNA replication and up to 600 double-strand breaks are introduced throughout the genome to initiate meiotic recombination during prophase I of meiosis\textsuperscript{28,29}. We observed a high overlap of GCR breakpoints (1Mb resolution) with meiotic recombination
hotspots (30-fold enrichment over random; Fig. 3a)\textsuperscript{30} and a 10-fold enrichment for aphidicolin-induced CFSs\textsuperscript{31,32} (Fig. 3b). The enrichment for CFSs may be more modest (3.2% or 3-fold enrichment; Fig. 3c) since they correlate with historically-enriched recombination sites\textsuperscript{32}.

CFS are expressed in a cell-type specific fashion\textsuperscript{33} and are thought to be initiated by replication stress upon encountering RNA polymerase II. We did not observe an enrichment for breakpoints that mapped to FRA3B (\textit{FHIT}) and FRA16D (\textit{WWOX}), which are the highest expressed CFS in several cell lines\textsuperscript{33}. However, since many of the CFSs are located in genes that are specific to neurones, we analysed expressing on human oocytes from three donors. This is unlikely to be caused by oocytes not initiating the aCFS expressed highly in lymphoblastoid cells, since we detected transcripts for both genes, and others implicated in gap or break formation at aCFS, at levels equivalent to other tissues where aCFSs are expressed (Fig. 3c). Similarly, for \textit{FRAXA}, a rare fragile site, we observed expression of \textit{FMR1} at similar levels to brain and other tissue\textsuperscript{34}, yet we did not observe any GCRs that involved \textit{FRAXA} amongst the embryos or oocytes. Our observations suggest that CFS are intrinsically expressed in human oocytes.

The significant enrichment of meiotic recombination sites with GCRs is consistent with the hypothesis that the induction and aberrant repair of programmed double-strand breaks may lead to GCRs\textsuperscript{29}. It was therefore surprising that nearly 28% of breakpoints occur near centromeres, where recombination rates are suppressed in population-based assessments\textsuperscript{30,35,36}. One possibility is that GCRs with breakpoints near centromeres have a different origin. Alternatively, and not mutually exclusive, meiotic recombination may be associated with an elevated risk of GCRs and/or whole chromosome missegregation, both of which could result in preclinical and clinical pregnancy losses\textsuperscript{37-41}. Such events would therefore not be represented in genome projects that rely on genomically healthy individuals.
To assess whether recombination occurs close to centromeres in human oocytes, we examined meiotic recombination in 58 oocyte-PB1 duos after obtaining genotype data using single nucleotide polymorphism arrays (Fig. 3d). Recombination breakpoints are visible as a shift from homozygous (AA or BB) regions to heterozygous blocks (AB). This approach avoids the requirement of three or more oocytes per participant to phase the mother’s genome, however, up to 30% of breakpoints occur within heterozygous blocks and are not detected using this method. Nevertheless, we observed recombination rates similar to those in fetal oocytes as well as a high degree of heterogeneity even in oocytes from the same woman, suggesting that the reduced detection compared to oocyte-PB1-PB2 trios is not majorly affected by the lack of phasing within heterozygous AB blocks. We observed recombination events near centromeres in 44 oocytes (76%) similar to previous observations suggesting that recombination is not necessarily suppressed in meiosis. Instead, we suggest that recombination events are less likely to be detected in genome projects that rely on the birth of genomically healthy children, discounting the massive pregnancy loss that occur both preclinically and clinically (25%).

Our observations suggest that human oocytes and preimplantation embryos are display high degree of genome instability. Previous microarray-based expression studies have suggested that mRNA expression in MII oocytes differ between women. We were interested in whether the expression profiles in MII oocytes from the same female were heterogenous with regards to the GCR network genes and could explain why a certain proportion of MII oocytes display GCRs. To understand the endogenous genome instability in human oocytes, we assessed expression of genes that are implicated in preventing GCRs in human tumours. To this end, we sequenced 14 MII oocytes from three women at high read-depth. We observe heterogeneity in the expression of the 237 highly conserved genes associated with GCRs in cancer genomes (Fig. 4a-d). In contrast, in mouse oocytes, where chromosomes breaks
occurred with a lower incidence⁴⁵, the expression of the GCR network was more homogenous (Fig. 4e,f).

In this work, we have shown that the de novo generation of GCRs occurs in up to 12% of human oocytes. This is an exceptional rate. Human tumours often contain rearranged chromosomes that drive tumour evolution and become prevalent in clonal populations⁴⁶. However, their actual generation per cell division in vivo during the lifetime of an individual is infrequent (est. one in 10⁷ per cell per division) ⁴⁷. Even tumour-derived cell lines with abnormal karyotypes remain relatively stable when passaged in vitro⁴⁸. The high rate of genome instability in human oocytes is normally only seen when cells are treated with inhibitors of DNA replication, exogenous DNA damaging agents or cells with mutations in DNA repair genes⁴⁹. The GCRs could originate from meiotic S-phase and prophase I in fetal oocytes, when hundreds of programmed double-strand breaks are introduced into the genome to initiate meiotic recombination ²⁹. Consistent with this, we observed significant overlap with meiotic recombination hotspots and, to a lesser extent, common fragile sites. In mouse models, DNA breaks induced in fetal oocytes can persist into the meiotic divisions in adult oocytes ⁵⁰,⁵¹. If this is also true in human, then breaks in DNA persist for decades until the oocyte matures in the adult woman.

Acknowledgments.
We thank the donors and patients who contributed oocytes and embryos for this study, and our colleagues who obtained their informed consent for these studies. We are grateful to our many colleagues in the IVF and fertility clinics who obtained and collected material. The human
oocyte work in the ERH lab is funded by a Novo Nordisk Foundation Young Investigator Award (NNF15OC0016662) the MRC grant (MR/M000664/1), ERC grant (724718-ReCAP), and Danish National Research Foundation (DNRF115, 6110-00344B). The mouse model was funded by MRC grant (MR/M000664/1). RB was supported by Illumina Inc., who provided reagents for analysis of human oocyte-PB1 duos. JG was funded by a postdoctoral fellowship DFF-FSS. ERH is grateful to the DNRF (grant DNRF115) for infrastructure support. AZ and MS have received financial support from the Max Planck Society, the European Research Council under grant agreement no. 337415 and the Lister Institute for Preventive Medicine to conduct research on human oocytes. AZ is further supported by a Rosetrees Trust PhD Fellowship. We gratefully acknowledge Ian Hickson, Jo Murray, Tony Oliver, and Alan Lehmann for reagents and discussions of the work, as well as the initial breeding data from the Smc6<sup>+/GT</sup> female mice.
Figure legends.

**Fig. 1 Segmental aneuploidies (GCRs) in human preimplantation embryos.**

(A) Trophectoderm (TE) biopsies were analysed by the VeriSeq (NGS) protocol for aneuploidy. The proportion of biopsies with different types of aneuploidies. Uniform
calls were within 2 S.D. of the normalized NGS reads. The number of biopsies are shown in parentheses above the graph.

(B) Incidence of aneuploidies according to maternal age. * denote p-values of less than 0.01 compared to the 24-29 year old age group.

(C) Frequency of chromosomes affected by whole chromosome aneuploidy or segmental aneuploidies.

(D) Size distribution of segmental aneuploidies.

(E) Frequency of loss and gain of whole arms (p or q).

Fig. 2 Segmental aneuploidies and chromosome breaks in human oocytes.

(A) Strategy to recover the genetic information from polar bodies and oocytes. Sequential biopsy of the two polar bodies allows detection of maternal inheritance pattern from meiosis.

(B) The fraction of mature MII and PB1 that contained a segmental (seg.) or whole chromosome (whole) aneuploidy.

(C) The fraction of oocytes that were activated and contained a segmental or whole chromosome aneuploidy in the oocytes or polar bodies. Data are from Hou et al. (2013), Ottolini et al. (2015) and Ottolini et al. (2017).

(D) Example of a broken chromosome fragment without a CREST (kinetochore-centromere) in human MII oocyte. The broken chromosome is indicated by the white area in the inset to the right. Bar: 5 µm.

(E) The fraction of human MII oocytes with broken chromosome fragments. Images from all three are shown in Supplemental Fig. S2.

(F) Symmetrical (left) and atypical chromosome structures on MII spindles in human oocytes. Bar: 5 µm.

(G) The fraction of MII oocytes that contains an atypical chromosome structure. The fraction of spread mouse MII oocytes with a chromosome break compared to the rate in human oocytes (from E and G).

(H) The fraction of murine MII oocytes (spread) with chromosome breaks compared to human MII oocytes.

Fig. 3 Heterogenous mRNA expression profiles in human MII oocytes.

(A) mRNA expression profiles for 14 MII oocytes from three patients.

(B)
Fig. 4 Heterogeneity in expression of the GCR network in human oocytes.

(A) PCA plot of 14 MII oocytes from three patients based on 4,797 genes with transcripts per million, TPM > 0.
(B) Heatmap with two-way hierarchical cluster of expressed genes.
(C) PCA plot of the GCR network (222 genes with TPMs > 0.
(D) Heatmap of GCR network.
(E) T-SNE plot showing clustering of more homogeneous mouse oocytes compared to human for the SMC5/6 complex genes.
(F) Heatmap of SMC5/6 complex genes for mouse oocytes (green) and human.

References (up to 30)


Figure 1

(a) 3-10 cell TE biopsy

(b) Maternal age (years)

(c) Chromosome affected

(d) Size (Mb)

(e) chr. arm
Figure 2

(a) Schematic diagram showing the process of fertilization (ICSI) or A23187 on a mature MII oocyte, leading to the formation of a "zygote".

(b) Graph showing the fraction of MII-PB1 duos with CREST and DNA staining. The graph includes data points with n=178.

(c) Graph showing the fraction of activated oocytes with PB1 biopsy. The graph includes data points with n=111.

(d) Images of human meiosis II spread chromosomes stained with CREST and DNA. The images illustrate symmetrical and atypical chromosomes.

(e) Graph showing the type of chromosomal defect, either broken or whole, with n=54.

(f) Images of human meiosis II spindle showing symmetrical and atypical chromosomes.

(g) Graph showing the fraction of MII oocytes with atypical and whole types of chromosomal defects, with n=90.
Fig. 3

(a) Hotspots

(b) aCFS

(c) adjusted aCFS

(d) N=58

(e) Centromeric

Telomeric
PCA of 4,797 detected genes

TPM, 4797 detected genes

PCA, TPM of 222 GCR genes

TPM, 222 GCR genes

t-SNE MEMS genes
Methods.

Detection of GCRs in trophectoderm samples

The primary purposes of the trophectoderm samples from 963 embryos from 215 couples was to screen by whole chromosome aneuploidies using the SurePlex DNA Amplification System (Illumina Inc.) followed by next-generation sequencing (NGS). We used the VeriSeq protocol that has previously been validated for the detection of segmental aneuploidies as small as 1.8 Mb {Zheng, 2015 #1599; Vera-Rodriguez, 2016 #965}. The average age of women providing samples was 36.1 years (± 4.2, S.D., range: 24 to 44 years). We included 809 biopsies after removal of traces indicative of triploid, samples with an uneven noise floor and two independent assessors using different software both detected 189 segmental aneuploidies (≥5 Mb). Of these, 71 segmental aneuploidies had a whole copy number change indicative of a uniform distribution within the cells of the trophectoderm sample. The remaining 118 segmental aneuploidies showed an intermediate copy number change, suggestive of being present in only a subset of the sampled cells (mosaic){CoGEN, 2018 #1965; PGDIS, 2016 #1964}. To be conservative, we excluded the mosaic segmental aneuploidies from further analysis, since the detection of mosaic copy number changes are still poorly understood{CoGEN, 2018 #1965; PGDIS, 2016 #1964}. They are, however, consistent with that gross chromosomal rearrangements occur during the early embryonic mitosis.

MDA-NGS pipeline for GCR and aneuploidy detection in human eggs and polar bodies.

Whole genome amplification of single human eggs and polar bodies was performed by multiple displacement amplification (SureMDA; Illumina Inc.). Sequencing libraries were prepared from MDA products using Nextera® reagents (Illumina Inc.) in an optimised workflow. Briefly, DNA fragments were tagmented (enzymatic fragmentation and addition of Illumina-specific adapter sequences), prior to reduced-cycle, PCR amplification for sample multiplexing
by addition of unique barcodes. Amplified libraries were normalised using a bead-based workflow (VeriSeq® PGS; Illumina Inc.) and analysed on a MiSeq® instrument using paired-end, 2×36 bp sequencing (MiSeq® Reagent Kit v3; Illumina Inc.). Reads were aligned to the human genome hg19 using the Burrows-Wheeler Aligner (BWA) within the MiSeq Reporter software, and the resulting *.bam files were analysed for copy number variations (CNVs) using BlueFuse® Multi software (Illumina Inc.).

A custom reference database was created using ninety-six, male genomic DNA replicates (Promega) that were also amplified using MDA (with 50 pg input), to establish the euploid baseline NGS signal. MDA products were prepared for sequencing as described for oocytes and polar body samples. The euploid baseline was established from reference samples by measuring the variation between sequence read bins and the chromosome copy number medians, to address whether specific regions of the genome were systematically called higher or lower copy number from the expected diploid state. This method captures reproducible assay bias such as WGA noise or GC content. Reference database compatibility with BlueFuse® Multi software was provided under a research collaboration with Illumina Inc. Therefore, MDA-specific offsets were applied to NGS bin CNV values for all samples that were analysed against the reference using the standard software analysis pipeline.
**MDA-NGS pipeline validation**

The MDA-NGS pipeline for CNV detection was validated using cell lines of known karyotype, obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, New Jersey, USA. Cell lines were cultured, and single cells were manually isolated under a stereo microscope in 1×PBS prior to WGA using SureMDA (Illumina Inc.) A summary of cell lines and samples collected are shown in Table S1. For each round of single cell isolation, a 3-cell sample was isolated for positive control and isolation buffer (1×PBS) was sampled for negative control. During MDA, all single cell samples and isolation controls were amplified together with a 50 pg genomic DNA positive control and a ‘no template’ (kit-supplied 1×PBS) negative control. Genomic DNA was also extracted from each cell line using the QIAamp DNA Mini Kit (QIAGEN), as a positive control for CNV validation.

<table>
<thead>
<tr>
<th>Cell Line ID</th>
<th>Karyotype summary</th>
<th>Single cells</th>
<th>Isolation controls</th>
<th>Amplification controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM00526</td>
<td>47, XY +13</td>
<td>14</td>
<td>3-cell, positive</td>
<td>gDNA positive</td>
</tr>
<tr>
<td></td>
<td>114 Mb gain</td>
<td></td>
<td>Buffer, negative</td>
<td>No template, negative</td>
</tr>
<tr>
<td>GM50121</td>
<td>46, XY Chr: 18</td>
<td>14</td>
<td>3-cell, positive</td>
<td>gDNA positive</td>
</tr>
<tr>
<td></td>
<td>15.5 Mb p arm loss</td>
<td></td>
<td>Buffer, negative</td>
<td>No template, negative</td>
</tr>
<tr>
<td></td>
<td>59.3 Mb q arm gain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM04927</td>
<td>47, XY +21</td>
<td>14</td>
<td>3-cell, positive</td>
<td>gDNA positive</td>
</tr>
<tr>
<td></td>
<td>47 Mb gain</td>
<td></td>
<td>Buffer, negative</td>
<td>No template, negative</td>
</tr>
<tr>
<td>GM10985</td>
<td>46, XX Chr:3</td>
<td>14</td>
<td>3-cell, positive</td>
<td>gDNA positive</td>
</tr>
<tr>
<td></td>
<td>10.3 Mb p arm loss</td>
<td></td>
<td>Buffer, negative</td>
<td>No template, negative</td>
</tr>
</tbody>
</table>

‘Chr’ = Chromosome; ‘+’ = Whole chromosome gain.

Twelve single cell isolates and 12 gDNA technical replicates per cell line were amplified using SureMDA and analysed using the MDA-NGS workflow described previously. SureMDA products were quality assessed by gel electrophoresis and quantitation of dsDNA using a
fluorometric assay (Quant-iTM dsDNA Assay Kit, high sensitivity; ThermoFisher Scientific), prior to NGS library preparation. Sequencing libraries were pooled for 24-plex sequencing on a MiSeq System (2×36 bp), and *.bam files were extracted for analysis using BlueFuse® Multi software against the custom, MDA-specific reference database described here.

**CNV concordance between NGS output and cell line karyotype information**

Twelve genomic DNA replicates and twelve single cells from four cell lines of known karyotype (Coriell Institute) were analysed for copy number variation. A summary of copy number analysis is shown in Table S2. All gDNA samples showed the expected karyotype, serving as a positive control and confirming the CNVs are present in each of the cell lines used for validation. For single cells, 12 out of 12 GM04927 samples showed the expected gain of chromosome 21 (3C), and one out of 12 cells showed additional segmental aneuploidies for chromosome 1 (q-arm; 4C) and chromosome 16 (q-arm; 1C). Cell line GM04927 has since been confirmed to be mosaic for these abnormalities (personal communication with Coriell), further demonstrating the detection of sub-chromosomal abnormalities using this pipeline. For cell line GM00526, 11 out of 12 single cells showed the expected gain of chromosome 13 (3C), with the remaining cell being copy number neutral for chromosome 13 (2C). Two cells each showed an additional segmental aneuploidy, affecting chromosome 11 (q arm; 1C) and chromosome 20 (p arm; 1C). For cell line GM10985, 9 out of 12 single cells showed the expected loss of a 10.3 Mb region in the p-arm of chromosome 3 (1C), with the remaining three cells being copy number neutral for chromosome 3 (2C). One cell showed an additional segmental aneuploidy, affecting chromosome 6 (q-arm; 1C), and loss of NGS signal at the ends of seven chromosomes, consistent with poor WGA. All 12 single cells of cell line GM50121 showed the expected karyotype for chromosome 18, with no additional aneuploidy events
observed. In addition, all gDNA and single cell samples showed the expected copy number for both X and Y chromosomes according to the Coriell karyotype descriptions.

**Table S2:** Validation of CNVs in genomic DNA and single cells of known karyotype

<table>
<thead>
<tr>
<th>Cell Line ID</th>
<th>Sample Type</th>
<th>$n$</th>
<th>Expected CNV</th>
<th>False Negative</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM00526</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM04927</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM10985</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM50121</td>
<td>gDNA</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM00526</td>
<td>Single cell</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GM04927</td>
<td>Single cell</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>2*</td>
</tr>
<tr>
<td>GM10985</td>
<td>Single cell</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GM50121</td>
<td>Single cell</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Segmental abnormalities that are known to persist in a sub-population of this cell line.

In summary, the expected karyotype was detected in 48 out of 48 gDNA samples, and no additional abnormalities were observed. For single cell isolates, 4 out of 48 (8.3% ± 2.0% S.E.P) samples were discordant with the expected karyotype. Chromosomes that were copy number neutral for expected CNVs, could be explained by mitotic chromosome segregation events in cell culture, that restore the diploid state in daughter cells. On the other hand, CNVs of small sub-chromosomal regions (<11 Mb) may be suppressed during normalisation of NGS data. This may indicate the limit of detection for such events with this assay, however further replicates would be required to elucidate this. In addition, 5 chromosomes gave false positive results. However, two that occurred in cell line GM04927 are known to be present in a sub-population within the cell lineage. The remaining 3 out of 2208 (0.14% ± 0.04% S.E.P) chromosomes were false positive for sub-chromosomal abnormalities with a full copy number change. It is likely these events are due to *de novo* chromosome breakages during cell culture,
however we cannot exclude the possibility that these CNVs are due to technical artefacts or biological phenomena, such as S-phase replication in unsynchronised cell cultures.

Taken together, NGS-based CNV detection in MDA products from single cells has been validated using multiple cell lines with CNVs ranging from 10.3 Mb to 114 Mb. However, the sensitivity is limited by the low-level stochastic amplification bias that is inherent to MDA. In this study, we mitigate this by reporting reciprocal CNV events in matched cells, thus providing increased confidence to CNV detection. However, we may underestimate the incidence of small sub-chromosomal aberrations due to masking by technical noise. To this end, CNV detection from MDA products by low-pass NGS is suitable for analysing structural chromosome changes in individual meioses.

**Human oocytes**

Human oocytes for use in this study were collected following ovarian stimulation, or from retrieval of small antral follicles (SAFs) directly from the unstimulated ovary (Donnez et al., 2013; Yin et al., 2016). 41 oocytes were donated by 23 patients undergoing in vitro fertilisation (IVF) at the GENERA center for Reproductive Medicine, Rome, Italy with fully informed consent. Oocytes were collected, cultured, activated and biopsied according to the protocol of Ottolini et al. (2016). Briefly, retrieved oocytes that had not completed meiosis I were matured in vitro for 24 hours. Oocytes that were arrested at metaphase II, had the PB1 removed prior to artificial activation with 100 µM calcium ionophore. Oocytes that completed meiosis II and extruded the PB2 were biopsied to yield oocyte-PB trios (PB1, PB2 and oocyte; tubed separately).

In addition, 118 oocytes were obtained from in vitro maturation (IVM) of germinal vesicle (GV)-stage oocytes, donated by 19 patients undergoing fertility preservation therapies
at Juliane Marie Centre, Rigshospitalet, Copenhagen, Denmark with fully informed consent. Oocytes were matured *in vitro*, and those that completed meiosis I with extrusion of the PB1 were biopsied according to the protocol of Gruhn *et al.* (2018). A subset of 17 MII-arrested oocytes from 4 patients, were activated with calcium ionophore following PB1 biopsy, as described previously. Oocytes that then extruded the second polar body (PB2) were biopsied to yield oocyte-PB trios.

Oocytes that were not treated with calcium ionophore or failed to activate (no extrusion of the PB2), were classified as MII-PB1 duos (oocyte and PB1). Oocytes that were activated and extruded the PB2, but where one of the three cells (PB1, PB2 or oocyte) was lost during biopsy, were classified incomplete trios. Cells with no matching counterpart of individual meioses were classified as singles. A full stratification of samples is shown in Table S3. In total, 116 meiosis II-arrested oocytes and their matching PB1s (MI duos) were collected. An additional, 15 oocytes with both matching polar bodies (oocyte-PB trios) and 2 MII duos were collected following artificial activation of the oocyte with calcium ionophore (triggering the resumption and completion of meiosis II). Polar bodies were biopsied sequentially upon extrusion to differentiate between PB1 and PB2 samples. Following sample loss, degeneration or fragmentation during biopsy, 11 individual PB1s and 15 individual oocytes were also collected. All single cells were amplified using SureMDA (Illumina Inc.) and quantified (Quant-iT™) prior to library preparation for NGS analysis as described for validation of the pipeline.

**Table S3:** Oocyte and polar body sample stratification for MDA-NGS pipeline.

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Clinic A</th>
<th>Clinic B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>23</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>Oocytes retrieved</td>
<td>41</td>
<td>118</td>
<td>159</td>
</tr>
</tbody>
</table>
Oocytes treated with calcium ionophore (Patients) | 41 (23) | 17 (4) | 58 (27) |
---|---|---|---|
Activation rate (%) | 16/41 (39.02) | 2/17 (11.76) | 18/58 (31.03) |
MII-PB1 duos | 25 | 90 | 115 |
Incomplete trios | 6 | 0 | 6 |
Oocyte-PB1-PB2 trios | 13 | 2 | 15 |
Individual PB1 | 0 | 11 | 11 |
Individual oocyte | 0 | 15 | 0 |
Samples processed | 95 | 212 | 307 |

**SNP genotyping**

MDA products of oocyte and polar body samples were analysed by SNP microarray using the Infinium™ HumanKaryomap-12 DNA Analysis Kit, with scanning on the iScan System or NextSeq 550 System (Illumina Inc.). MDA products were not diluted prior to input. SNP genotype data was also used to examine meiotic recombination single cells. Internal assay controls (included on the BeadChip array) were analysed using GenomeStudio software (Illumina Inc.), and genotype call rates were reviewed as a quality assessment of genome coverage and MDA yield.

**Sample quality control and exclusion criteria**

Sample quality was assessed following consideration of MDA yield, SNP call rates and NGS read and alignment metrics. NGS ‘noise’ was assessed using the derivative log ratio of observed and expected CNV values per alignment bin, when compared to the euploid reference database. In total, 38 out of 307 (12.38% ± 0.94% S.E.P) oocyte or polar body single cell samples were excluded following SNP array QC as a consequence of suboptimal MDA yield or suspected gDNA contamination. A further 16 out of 307 samples were excluded following
NGS as they failed to produce sufficient reads for analysis. The remainder of the samples were also sequenced, of which 37 out of 253 (14.62% ± 1.11% S.E.P) samples were excluded due to high noise within CNV traces generated using BlueFuse® Multi software with our custom reference database. High noise reduces the confidence of the operator when inferring CNVs, and in these cases, samples were excluded from analysis as ‘inconclusive for aneuploidy’. Finally, 56 out of 216 (25.93% ± 1.49% S.E.P) samples that passed NGS QC, had no included matching cell(s) of their respective meiotic division(s). These samples were excluded from CNV analysis, where reciprocal evidence of aneuploidy could not be obtained, i.e. where a chromosome gain in the MII oocyte is matched with a corresponding loss in the PB1.

**Abnormality scoring and database curation**

Chromosomal abnormalities in single cells from cell lines, and oocyte-PB duos/trios were manually scored by two operators, though observation of CNV charts drawn from raw and normalised, unsmoothed data exported from BlueFuse® Multi software. Abnormalities were scored on a per chromosome basis and recorded in Microsoft Excel. B-allele frequencies and log intensity ratios in microarray data were reviewed for evidence supporting chromosome imbalance and meiotic recombination. In this study, the incidence of uniform segmental aneuploidies was calculated from events that showed a reciprocal imbalance between two matched cells, of which we detected 29 reciprocal uniform segmental events in 21 MII-PB1 duos.

**Statistical methods.**

We used R for statistical methods, as described throughout, including the libraries ‘boot’ {Canty, 2017 #1943; Davison, 1997 #1944} and ‘mosaic’ from R.
Appendix A2

Appendix A2 is provided electronically, and comprises:

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2.1</td>
<td>Coriell Material Transfer Agreement, Blank</td>
</tr>
<tr>
<td>A2.2</td>
<td>Coriell Material Transfer Agreement, Illumina</td>
</tr>
<tr>
<td>A2.3</td>
<td>Coriell Material Transfer Agreement, University of Copenhagen</td>
</tr>
<tr>
<td>A2.4</td>
<td>Coriell Statement of Research Intent, E.R. Hoffmann_01</td>
</tr>
<tr>
<td>A2.5</td>
<td>Coriell Statement of Research Intent, E.R. Hoffmann_02</td>
</tr>
</tbody>
</table>
NIGMS HUMAN GENETIC CELL REPOSITORY

Material Transfer Agreement (Assurance Form)
for Human Cell Lines, Somatic Cell Hybrids, and DNA Samples

This Material Transfer Agreement ("MTA") pertains to the samples ("NIGMS Repository Sample(s)") that are part of the NIGMS Human Genetic Cell Repository ("Repository"), and which are administered by the Coriell Institute for Medical Research, Camden, New Jersey ("Coriell"). The Institutional Official is the legal representative ("Institutional Official") of the Institution ("Institution") receiving the NIGMS Repository Samples.

The Principal Investigator is the person receiving the NIGMS Repository Sample(s) and is responsible for the conduct of the Statement of Research Intent, defined below. The Principal Investigator’s research team that is under the direct supervision of the Principal Investigator may have access to the NIGMS Repository Sample(s) only after they have been informed of and agreed to the provisions of this MTA.

To ensure compliance with the Office for Human Research Protections, Department of Health and Human Services ("DHHS") regulations for the protection of human subjects (45 CFR Part 46), before NIGMS Repository Sample(s) can be shipped from the Repository by Coriell, the Principal Investigator must provide to Coriell a written description of the purpose of the research to be done using the NIGMS Repository Sample(s) ("Statement of Research Intent"). The Statement of Research Intent and the signed MTA must be submitted to Coriell.

The Principal Investigator must acknowledge on the signature page of this MTA that he/she has read and understands the terms and conditions of this MTA. The Principal Investigator’s Institutional Official must also sign this MTA agreeing to adhere to the terms and conditions of this MTA. The Institutional Official acknowledges that the conditions for use of the NIGMS Repository Sample(s) are governed by the Coriell Institutional Review Board ("Coriell IRB") in accordance with DHHS regulations (45 CFR Part 46). The Institution agrees to comply fully with all such conditions and to report promptly to the Coriell IRB any proposed changes in the Statement of Research Intent. The Institution remains subject to all applicable state and local laws or regulations and Institution policies that provide additional protections for human subjects.

Coriell will under no circumstances provide information that will allow investigators to identify human subjects. Furthermore, the Institution and the Principal Investigator agree not to try to identify or contact the submitter of the sample or the donor subject from whom the cell line or DNA sample was derived. The Institution and the Principal Investigator agree not to name the population from whom the samples were obtained, if this information is not essential. (See Policy for the Responsible Collection, Storage, and Research Use of Samples from Named Populations for the NIGMS Human Genetic Cell Repository.)
WARRANTY AND LIABILITY

Warranty: THE REPOSITORY AND CORIELL MAKE NO REPRESENTATIONS AND EXTEND NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE MATERIAL WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS.

IN ADDITION TO THE TERMS CONTAINED IN THIS MTA, USE OF SOME OF THE NIGMS REPOSITORY SAMPLE(S) MAY BE SUBJECT TO CERTAIN RESTRICTIONS AND/OR LICENSE REQUIREMENTS. RESTRICTIONS AND/OR LICENSE REQUIREMENTS OF WHICH THE REPOSITORY IS AWARE HAVE BEEN NOTED IN THE REPOSITORY’S ONLINE CATALOG ON THE SAMPLE DETAIL PAGE IN THE “REMARKS” FIELD UNDER THE “OVERVIEW” TAB. LICENSE REQUIREMENTS RELEVANT TO NIGMS REPOSITORY SAMPLES MAY BE FOUND IN THE LIMITED USE LICENSES AVAILABLE HERE AND ARE INCORPORATED IN FULL INTO THIS AGREEMENT. THERE MAY EXIST ADDITIONAL RESTRICTIONS OR PROPRIETARY RIGHTS OF WHICH THE REPOSITORY IS UNAWARE; THE INSTITUTION IS RESPONSIBLE FOR COMPLIANCE WITH ALL CONDITIONS OF USE OF THE NIGMS REPOSITORY SAMPLE(S).

Liability Statement for State Institutions: The Institution agrees to be responsible for any claims, costs, damages, or expenses resulting from any injury (including death), damage, or loss that may arise from the use of the NIGMS Repository Sample(s) to the extent permitted under the laws of the Institution’s state. This provision shall also apply to any byproducts or derivatives of the NIGMS Repository Sample(s).

Liability Statement for U.S. Government Laboratories: The United States assumes the liability for any claims, damages, injuries, or expenses arising from the use of NIGMS Repository Sample(s) or any byproduct or derivative, but only to the extent provided under the Federal Tort Claims Act (28 U.S.C. Chapter 171).

Liability Statement for All Other Institutions: The Institution agrees to indemnify and hold harmless the United States Government, Coriell, and the submitter of the sample from any claims, costs, damages, or expenses resulting from any injury (including death), damage, or loss that may arise from the use of the NIGMS Repository Sample(s). This provision shall also apply to any byproducts or derivatives of the NIGMS Repository Sample(s).

HUMAN EXPERIMENTATION

Human experimentation utilizing the NIGMS Repository Sample(s) is strictly prohibited.
RESEARCH USE, COMMERCIAL USE, AND RESTRICTIONS ON REDISTRIBUTION AND PROHIBITIONS ON RESALE

The Repository provides biomaterials as a service to the research community. The purpose of the Repository is to stimulate and facilitate research in genetics and related fields, leading to a better understanding of normal genetic and cellular processes, to the identification and function of disease-related genes, and to the diagnosis and treatment of genetic disorders.

It is expressly understood that the NIGMS Repository Sample(s) delivered pursuant to this MTA are experimental in nature and are for use in research, in teaching, and as reference materials in clinical genetics laboratories. Institutions using NIGMS Repository Sample(s) for use as reference materials or controls are responsible for complying with all laws and regulations applicable to the intended use of the NIGMS Repository Sample(s), including any requirements for FDA approval.

The Repository number(s) of the cell line(s) or the DNA sample(s) must be cited as follows in publications or presentations that are based on the use of these materials: "The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: [list Repository ID numbers here]."

There is no restriction on development of commercial products resulting from the knowledge gained from studies using the NIGMS Repository Sample(s). However, with the exception of the NIGMS Repository Samples listed here, NIGMS Repository Samples, or material isolated from them, such as RNA, DNA, or protein, may not themselves be used in the manufacture of commercial products.

SHARED USE AND POSTING OF PERSONALLY IDENTIFYING GENETIC INFORMATION DERIVED USING NIGMS HUMAN GENETIC CELL REPOSITORY SAMPLES

NIGMS Repository samples, or material isolated from them, such as RNA, DNA, or protein, have the potential to generate data that could be used to identify an individual (e.g.: Combined DNA Index System (CODIS) reports, whole-genome microarray genotyping data, whole-exome or whole-genome DNA sequencing data). Investigators are strictly prohibited from posting or making available through open-access public websites and/or databases genetic data that might identify an individual. Investigators, however, are encouraged to submit such datasets to the Database of Genotypes and Phenotypes (dbGaP), which requires users of the data to abide by a Code of Conduct. For a limited subset of NIGMS Repository samples for which the donor-subject provided explicit informed consent for public data posting, open-access public posting of potentially individually identifying data is permitted (click here for a list of these samples).

Prior to posting or sharing with a collaborator personally identifying genetic information derived using NIGMS Repository biospecimens, the Principal Investigator must request permission to do so from the NIGMS Repository by submitting a completed, signed Statement of Research Intent.
Form describing the proposed data posting and/or sharing. The collaborator must also submit to
the NIGMS Repository a completed and signed Statement of Research Intent Form.

SECONDARY DISTRIBUTION AND SHARED USE OF CELL CULTURES AND DNA
SAMPLES FROM THE NIGMS HUMAN GENETIC CELL REPOSITORY

Genetic research often involves collaborations among several investigators or several
laboratories that share materials toward a common goal. Also, as a result of new genomic
technologies, data are often generated by multi-user core facilities. Many laboratories benefit
from using common biological reference materials for research or clinical purposes. Thus,
consistent with the mission to facilitate genetic research, the Repository will permit secondary
distribution to accommodate certain situations if it can be established that protection of human
subjects and quality control of the samples can be ensured. Secondary distribution, defined as the
sharing of NIGMS Repository Sample(s) from the Repository with members of laboratories
other than the Principal Investigator’s, is permitted only under certain clearly defined
circumstances.

Principal investigators who might wish to share NIGMS Repository Sample(s) with other
investigators should read the information below very carefully and must contact Coriell
before proceeding with a secondary distribution.

Permitted Uses:

1. **Single purpose collaboration**: Two or more investigators initiate a collaborative project
   that requires the use by each laboratory of the same NIGMS Repository Sample(s). One
   Principal Investigator obtains NIGMS Repository Sample(s) and explains in the
   Statement of Research Intent that the sample will be shared with specific, named
   collaborator(s) for a common research project. Secondary distribution to named
   collaborator(s) is permitted when the Statement of Research Intent is identical for all the
   named collaborator(s) and is consistent with this MTA. Each collaborating investigator
   and his or her Institutional Official must sign and submit a copy of this MTA.

2. **Multi-user core facility**: A core facility (for high-throughput genotyping, for example)
   purchases NIGMS Repository Sample(s) for use by the investigators within the facility to
   perform assays for investigators at his or her Institution or at a consortium of institutions.
   The Statement of Research Intent describes the range of studies that will be conducted
   using the NIGMS Repository Sample(s). In this situation, the use of these NIGMS
   Repository Sample(s) in the core facility may be permitted after the Coriell IRB assures
   that the use of these samples is consistent with the research subject’s informed consent.
   Since the NIGMS Repository Sample(s) will be used in the same facility by multiple
   investigators, quality can be ensured.

3. **Distribution of aliquots of samples for use as reference materials**: An Institution
   purchases a sample and describes in the Statement of Research Intent that the NIGMS
   Repository Sample(s) will be distributed for use as a reference material (for proficiency
testing, for example). The Statement of Research Intent may not be able to specify the laboratories that will receive the materials. Prior approval by the Coriell IRB for this use of the NIGMS Repository Sample(s) is required. The Coriell IRB will decide this type of request on a case-by-case basis with the advice of the NIGMS Repository’s Project Officer. The NIGMS Repository Sample(s) that are distributed must be accompanied by a disclaimer of the Repository’s responsibility regarding safety and quality. Furthermore, residual NIGMS Repository Sample(s) must be returned to the Principal Investigator or destroyed.

4. Development of a Highly Unique Resource: An Institution purchases a cell line from the Repository and develops it into a Highly Unique Resource that requires significant modification or specialized expertise to grow, characterize, and maintain (such as an induced pluripotent stem cell line). A Highly Unique Resource is substantially different from the original NIGMS Repository Sample obtained from the Repository. Simply modifying an NIGMS Repository Sample obtained from the Repository through the introduction of a gene (e.g., hTERT or green fluorescent protein) would not qualify as creating a Highly Unique Resource. The Principal Investigator may distribute aliquots of the Highly Unique Resource material by using an appropriate agreement between the Principal Investigator and/or the Principal Investigator’s Institution and the secondary institution receiving the Highly Unique Resource (“Secondary Recipient”). Often a material transfer agreement is used for transfers of research materials for this purpose. The agreement to transfer the Highly Unique Resource to a Secondary Recipient must include: (1) a statement naming the NIGMS Human Genetic Cell Repository and the Repository ID number of the cell line from which the Highly Unique Resource was derived; (2) a statement that the Secondary Recipient must acknowledge the Repository and the cell line number(s) in any publications or presentations based on the utilization of the NIGMS Repository Sample(s) as follows: "[Name of Highly Unique Resource] was derived from cell line [list NIGMS Repository ID number] from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research"; and (3) a statement that the Highly Unique Resource obtained from different sources will not have undergone the standard quality control of the Repository.

The terms of the agreement between the investigator who developed the Highly Unique Resource and the Secondary Recipient who obtains the Highly Unique Resource for research purposes must be consistent with NIH’s Simple Letter of Agreement for the Transfer of Materials or the Uniform Biological Material Transfer Agreement. Both of these documents can be found at: http://www.ott.nih.gov/forms-model-agreements#MTACTA.

An Institution that purchases the NIGMS Repository Sample(s) is encouraged to make available aliquots of the Highly Unique Resource derived from the NIGMS Repository Sample and appropriate protocols and training to the NIGMS Repository for the NIGMS Repository to expand, characterize and distribute the unique resource through the NIGMS Repository, should the NIGMS Repository wish to do so.
Prohibited Uses:

1. **Multi-purpose use** - An investigator working on a particular project submits a Statement of Research Intent describing that project and obtains NIGMS Repository Sample(s). At some time after obtaining the NIGMS Repository Sample(s), the Principal Investigator wishes to give a portion of the NIGMS Repository Sample(s) or a culture derived from the NIGMS Repository Sample(s) to an investigator who is working on another project. In this case, **secondary distribution is prohibited** because the use of the NIGMS Repository Sample(s) by the second investigator may not be consistent with this MTA and the Statement of Research Intent. In addition, errors in cell culture technique and identification of cultures or DNA samples can occur and could compromise the Repository's reputation.

2. **The Secondary Distribution or sale of NIGMS Repository Sample(s) for any purpose not specified above is prohibited.**

**BIOHAZARD**

All cultured animal and human cells have the potential for carrying viruses, latent viral genomes, and other infectious agents in a latent or inactive state. NIGMS Repository Samples shipped by the Repository should therefore not be treated as if they are free of contamination. These NIGMS Repository Sample(s) should always be handled carefully by trained persons under laboratory conditions which afford adequate biohazard containment following [MINIMUM SAFETY GUIDELINES RECOMMENDED FOR WORKING WITH HUMAN CELL CULTURES](#). By accepting NIGMS Repository Sample(s), the undersigned assume full responsibility for their safe and appropriate handling.
We, the undersigned, have read and understand this document and agree to adhere to the restrictions and warnings stated herein.

Name of Institution

Name of Institutional Official who can make legal commitments on behalf of the Institution (typed or printed)

Title of Institutional Official

Signature of Institutional Official Date

Read and Understood the terms and conditions of this Agreement:

Name of Principal Investigator (typed or printed)

Signature of Principal Investigator Date

To contact the CORIELL CELL REPOSITORIES:
Write: 403 Haddon Avenue. Camden, New Jersey 08103 USA
Call: 800-752-3805 in the United States; 856-757-4848 from other countries
Fax: 856-757-9737
E-mail: ccr@coriell.org
We, the undersigned, have read and understand this document and agree to adhere to the restrictions and warnings stated herein.

ILLUMINA, Fulborne, UK
Name of Institution

KAREN GUTEKUNST
Name of Institutional Official who can make legal commitments on behalf of the Institution (typed or printed)

VP, PRODUCT DEVELOPMENT
Title of Institutional Official

Karen Gutekunst 23 Aug 2017
Signature of Institutional Official  Date

Read and Understood the terms and conditions of this Agreement:

PHILIPPA BURNS
Name of Principal Investigator (typed or printed)

Signature of Principal Investigator  23 Aug 2017  Date

To contact the CORIELL CELL REPOSITORIES:
Write:  403 Haddon Avenue, Camden, New Jersey 08103 USA
Call:  800-752-3805 in the United States; 856-757-4848 from other countries
Fax:  856-757-9737
E-mail:  ccr@coriell.org

Form 1401-65 Rev D-121713: NIGMS HGCR MTA/Assurance Form_Humans
We, the undersigned, have read and understand this document and agree to adhere to the restrictions and warnings stated herein.

**UNIVERSITY OF COPENHAGEN**

Name of Institution

**OLE WILLIAM PETERSEN**

Name of Institutional Official who can make legal commitments on behalf of the Institution (typed or printed)

**HEAD OF DEPARTMENT**

Title of Institutional Official

Signature of Institutional Official 24 NOVEMBER 2015  

Read and Understood the terms and conditions of this Agreement:

**PROF. EVA R. HOFFMANN**

Name of Principal Investigator (typed or printed)

Signature of Principal Investigator 23-11-2015  

To contact the CORIELL CELL REPOSITORIES:

Write: 403 Haddon Avenue, Camden, New Jersey 08103 USA  
Call: 800-752-3805 in the United States; 856-757-4848 from other countries  
Fax: 856-757-9737  
E-mail: ccr@coriell.org
NIGMS HUMAN GENETIC CELL REPOSITORY

STATEMENT OF RESEARCH INTENT

For each research project, submit a separate Statement of Research Intent. Please fill out all parts of the form. Use additional sheets as necessary.

Part I

List the NIGMS Repository number for each cell culture, DNA sample, or DNA Panel you wish to order:

<table>
<thead>
<tr>
<th>NIGMS Repository Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM07228B</td>
<td>Coriell order #: 112012, 113673, 119130, 118888, 122216</td>
</tr>
<tr>
<td>GM010608</td>
<td>Coriell order #: 114071</td>
</tr>
<tr>
<td>GM010985</td>
<td>Coriell order #: 114071</td>
</tr>
<tr>
<td>GM016362</td>
<td>Coriell order #: 114071</td>
</tr>
<tr>
<td>GM07226, GM07224</td>
<td>Coriell order #: 107527</td>
</tr>
<tr>
<td>GM07225</td>
<td>Coriell order #: 107527</td>
</tr>
<tr>
<td>GM04927H</td>
<td>Coriell order #: 122216</td>
</tr>
</tbody>
</table>

Part II

These samples will be used in the following ways (check all that apply):

☐ Perform functional studies
☐ Develop or characterize induced pluripotent stem cell (iPSC) lines
☐ Serve as positive or negative controls for genetic testing
☐ Serve as positive or negative controls for assay development
☐ SNP discovery/genotyping/haplotypeing
☐ Sequence portions of the genome
☐ Map genes
☐ Identify novel genes
☐ Characterize genes and mutations
☐ Study gene expression
☐ Determine the ancestral state of a polymorphism/haplotype
☐ Conduct proteomic studies
☐ Other (please specify): Sequence assembly, SNP calling, genotyping validation

Part III

Form 1401-05 Rev H-120715: NIGMS HGCR Statement of Research Intent
Revised December 07, 2015
Describe more specifically the study or studies you will conduct using these samples. (You may type and attach the description, or include a copy of the abstract of your research grant that describes the project). If, in the future, you plan to use these samples for a purpose different from what you provide here, you must submit another Statement of Research Intent. There will be no additional charge.

The cell material will be processed for whole genome amplification and genomic DNA will be used for sequencing and microarray analysis. The aim is to develop and validate a new machine learning based method for accurate genotyping in the single cell environment, and to find limiting factors in training the algorithm. Specifically, the next-generation sequencing data will be assembled and processed by SNP calling algorithms and the microarray data (SNP) will be genotyped using existing software and newly developed algorithms.

Part IV

Provide information about proposed sharing of personally identifiable genetic information (PIGI)\(^1\) with individuals outside your laboratory:

- [ ] PIGI will not be generated.
- [ ] PIGI will not be shared with individuals outside my laboratory.
- [x] PIGI will be shared with another investigator\(^2\).
- [ ] PIGI will be deposited in a controlled-access database.
- [ ] PIGI will be deposited in an open-access, public database\(^3\).
- [ ] Other (please specify): ________________________________

---

\(^1\) PIGI includes, but is not limited to: whole-genome microarray genotyping data, whole-exome sequence data, whole-genome sequence data, and other genetic data that could potentially be used to identify an individual.

\(^2\) The investigator(s) with whom PIGI will be shared must submit a completed NIGMS Human Genetic Cell Repository Statement of Research Intent Form.

\(^3\) Please note that it is permitted to deposit PIGI in an open-access public database for only a small subset of NIGMS Human Genetic Cell Repository samples. Click here\(^4\) for a list of these samples.

Form 1401-05 Rev H-120715: NIGMS HGCR Statement of Research Intent
Revised December 07, 2015
Describe specifically intended PIGI sharing, indicating with whom you propose to share the data and what type of PIGI will be shared (e.g., whole-genome microarray data, whole-genome sequence data, etc.):

The cell material will be processed at Illumina, Fulborn by WGA, SNP microarray and NGS. The secondary SNP and NGS data will be shared between Illumina, Fulborn and Prof. Eva Hoffmann at the University of Copenhagen. Cell material and genomic DNA will not be shared outside Illumina, Fulborn. The data that will be shared will be whole genome.

Part V

Provide information about who will be using the samples:

☐ These samples will be used only in my laboratory.
☐ These samples will be shared with one or more investigators for a single research study.
☐ These samples will be shared within a multi-user core facility.
☐ These samples will be distributed as aliquots or derivatives for use as biological standards.
☐ These samples will be shared as a Highly Unique Resource.

Name(s) of any collaborating investigator(s):
Dr. Philippa Burns, Illumina Fulbourn (UK), pburns@illumina.com

Describe specifically the nature of any proposed secondary distribution:
The secondary distribution will be data only from SNP and NGS, not cell lines nor genomic DNA or any other physical material derived from the cell lines. A description of personal data protection at the Faculty of Health Sciences, University of Copenhagen is attached. The secondary data will be transferred via secure SSH.

All shared usage must conform to the Secondary Distribution Policy described in the NIGMS Human Genetic Cell Repository Material Transfer Agreement.

Form 1401-05 Rev H-120715: NIGMS HGCR Statement of Research Intent
Revised December 07, 2015
Part VI

Contact information:

Provide the following information for the Principal Investigator directly responsible for the use of the cell culture or DNA sample:

Name: Eva Hoffmann
Institution: University of Copenhagen, Denmark
E-mail: eva@sund.ku.dk
Signature: [Signature]
Date: 02-12-2016

To contact the CORIELL CELL REPOSITORIES:
403 Haddon Avenue
Camden, New Jersey 08103
Telephone: (800) 752-3805 or (856) 966-7377
Email: customerservice@coriell.org

https://catalog.coriell.org/0/Sections/Collections/NIGMS/pigi_consenteds.aspx?PgId=745&coll=GM
NIGMS HUMAN GENETIC CELL REPOSITORY

STATEMENT OF RESEARCH INTENT

For each research project, submit a separate Statement of Research Intent. Please fill out all parts of the form. Use additional sheets as necessary.

Part I

List the NIGMS Repository number for each cell culture, DNA sample, or DNA Panel you wish to order:

<table>
<thead>
<tr>
<th>NIGMS Repository Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM12877</td>
<td>CEPH/UTAH PEDIGREE 1463 ITNL HAPMAP PROJECT</td>
</tr>
<tr>
<td>GM12878</td>
<td>CEPH/UTAH PEDIGREE 1463 ITNL HAPMAP PROJECT</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part II

These samples will be used in the following ways (check all that apply):

- [ ] Perform functional studies
- [ ] Develop or characterize induced pluripotent stem cell (iPSC) lines
- [ ] Serve as positive or negative controls for genetic testing
- [ ] Serve as positive or negative controls for assay development
- [ ] SNP discovery/genotyping/haplotyping
- [ ] Sequence portions of the genome
- [ ] Map genes
- [ ] Identify novel genes
- [ ] Characterize genes and mutations
- [ ] Study gene expression
- [ ] Determine the ancestral state of a polymorphism/haplotype
- [ ] Conduct proteomic studies
- [ ] Other (please specify):  **Sequence assembly, SNP Calling, genotyping validation**

Part III

Form 1401-05 Rev H-120715: NIGMS HGCR Statement of Research Intent
Revised December 07, 2015
Describe more specifically the study or studies you will conduct using these samples. (You may type and attach the description, or include a copy of the abstract of your research grant that describes the project). If, in the future, you plan to use these samples for a purpose different from what you provide here, you must submit another Statement of Research Intent. There will be no additional charge.

The cell material will be processed for whole genome amplification and genomic DNA will be used for sequencing and microarray analysis. The aim is to develop and validate a new machine learning based method for accurate genotyping in the single cell environment, and to find limiting factors in training the algorithm. Specifically, the next-generation sequencing data will be assembled and processed by SNP calling algorithms ad the microarray (SNP) data will be genotyped using existing software and newly

Part IV

Provide information about proposed sharing of personally identifiable genetic information (PIGI)¹ with individuals outside your laboratory:

- PIGI will not be generated.
- PIGI will not be shared with individuals outside my laboratory.
- □ PIGI will be shared with another investigator².
- □ PIGI will be deposited in a controlled-access database.
- □ PIGI will be deposited in an open-access, public database³.
- □ Other (please specify): ______________________________________________________________________

¹ PIGI includes, but is not limited to: whole-genome microarray genotyping data, whole-exome sequence data, whole-genome sequence data, and other genetic data that could potentially be used to identify an individual.
² The investigator(s) with whom PIGI will be shared must submit a completed NIGMS Human Genetic Cell Repository Statement of Research Intent Form.
³ Please note that it is permitted to deposit PIGI in an open-access public database for only a small subset of NIGMS Human Genetic Cell Repository samples. Click here⁴ for a list of these samples.

---

¹ PIGI includes, but is not limited to: whole-genome microarray genotyping data, whole-exome sequence data, whole-genome sequence data, and other genetic data that could potentially be used to identify an individual.
² The investigator(s) with whom PIGI will be shared must submit a completed NIGMS Human Genetic Cell Repository Statement of Research Intent Form.
³ Please note that it is permitted to deposit PIGI in an open-access public database for only a small subset of NIGMS Human Genetic Cell Repository samples. Click here⁴ for a list of these samples.
Describe specifically intended PIGI sharing, indicating with whom you propose to share the data and what type of PIGI will be shared (e.g., whole-genome microarray data, whole-genome sequence data, etc.):

The cell material will be processed at Illumina Fulbourn by WGA, SNP microarray and NGS. The secondary SNP and NGS data will be shared between Illumina Fulbourn and Prof. Eva Hoffmann at the University of Copenhagen. Cell material and genomic DNA will not be shared outside Illumina Fulbourn. The data that will be shared will be whole genome.

Part V

Provide information about who will be using the samples:

☐ These samples will be used only in my laboratory.
☒ These samples will be shared with one or more investigators for a single research study.
☐ These samples will be shared within a multi-user core facility.
☐ These samples will be distributed as aliquots or derivatives for use as biological standards.
☐ These samples will be shared as a Highly Unique Resource.

Name(s) of any collaborating investigator(s):

Dr Philippa Burns, Illumina Fulbourn (UK), pburns@illumina.com

Describe specifically the nature of any proposed secondary distribution:

The secondary distribution will be data only from SNP and NGS, not cell lines nor genomic DNA, or any other physical material derived from the cell lines. The secondary data will be transferred via secure SSH.

All shared usage must conform to the Secondary Distribution Policy described in the NIGMS Human Genetic Cell Repository Material Transfer Agreement.
Part VI

Contact information:

Provide the following information for the Principal Investigator directly responsible for the use of the cell culture or DNA sample:

Name: Eva Hoffmann
Institution: University of Copenhagen, Denmark
E-mail: eva@sund.ku.dk
Signature: Eva Hoffmann
Date: 30 May 2017

To contact the CORIELL CELL REPOSITORIES:
403 Haddon Avenue
Camden, New Jersey 08103
Telephone: (800) 752-3805 or (856) 966-7377
Email: customerservice@coriell.org