The role of new genetic technologies for the analysis of early human embryos: clinical application and research

Nada Kubikova
Brasenose College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2020

Acknowledgements

I would like to express my greatest gratitude to my mentor and my supervisor Prof. Dagan Wells, who has always given me his full support, trust and tremendous opportunity to establish a valuable international network of collaboration.

I thank the members of the Wells lab who, with their dedication and commitment to excellence, pushed me to my limits and provided stimulating working environment.

I would also like to thank my dear Oxford and college friends, Vilija, Virginia, Eszter, Florence, Helena and Eddie for their continuous support and friendship. Their diverse stories and international backgrounds widened my horizons and served as true inspiration. I will never forget the fun times over college dinners and endless intellectual stimulation I received during my times as a DPhil student. Some of the friendships formed during this period will, undoubtfully, last a lifetime.

Furthermore, my gratitude goes to my lovely collaborators at the Francis Crick Institute, Kathy Niakan and James Turner (and their group members), for involving me in their projects. I learned a lot from them, and I greatly benefitted from this opportunity. I hope that the fruitful collaborations that were formed with other labs during my doctoral degree will carry on in the future.

Finally, this work would not have been possible without the generous financial support provided by the Clarendon Fund and Brasenose Joint Scholarship, to whom I will forever be grateful for giving me the opportunity to take up this degree.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5C-DNA	5-cells DNA
A	Adenine
ADO	Allele drop-out
ARMS	Amplification refractory mutation system
ATL	A-tailing Mix
ATP	Adenosine triphosphate
BAM	Binary version of SAM
C	Cytosine
CGH	Comparative genomic hybridization
CIP	Calf intestinal phosphatase
CRISPR	Clustered regularly interspaced short palindromic repeat
DNA	Deoxyribonucleic acid
DSB	Double-stranded break
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EGA	Embryonic genome activation
ERM	End Repair Mix
ESHRE	European Society of Human Reproduction and Embryology
FISH	fluorescent *in situ* hybridisation
G	Guanine
gDNA	Genomic DNA
GE	Genome editing
HBB	Hemoglobin beta

HbS	Hemoglobin S
HDR	Homology-directed repair
HEK	Human embryonic kidney
hESCs	Human embryonic stem cells
HFEA	Human Fertilisation and Embryology Authority
ICSI	Intracytoplasmic sperm injection
IGV	Integrative Genomics Viewer
IVF	*In vitro* fertilization
LIG	Ligation Mix
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MARSALA	Mutated Allele Revealed by Sequencing with Aneuploidy and Linkage Analyses
MDA	Multiple displacement amplification
mRNA	Messenger ribonucleic acid
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
ONT	Oxford Nanopore Technologies
PAM	Protospacer-adjacent motive
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PGT	Preimplantation genetic testing
PGT-A	Preimplantation genetic testing for aneuploidy
PGT-M	Preimplantation genetic testing for monogenic disease
PGT-SR	Preimplantation genetic testing for segmental rearrangements
PVA	Polyvinyl alcohol
RGEN	RNA-guided programmable endonuclease

RNP	Ribonucleoprotein
RSB	Resuspension buffer
RT	Room temperature
SAM	Sequence Alignment/Map
SC-DNA	Single-cell DNA
SCA	Sickle cell anaemia
SGD	Single gene disorder
sgRNA	Single guide ribonucleic acid
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
T	Thymine
TAF	Total amplification failure
TALENs	Transcription activator-like effector nucleases
TE	Tris/EDTA
tracrRNA	Trans-activating crispr RNA
Tris	Trisaminomethane
TSV	Tab-separated value
VCF	Variant call format
WGA	Whole genome amplification
WGS	Whole genome sequencing
ZFNs	Zinc finger nucleases

Abstract

Mutations of the beta-globin gene (*HBB*) are the most common cause of inherited disease in humans, causing ß-thalassaemia and sickle cell anaemia. Traditional preimplantation genetic testing (PGT) protocols for the detection of *HBB* mutations frequently involve labour intensive, patient-specific test designs owing to the wide diversity of disease-associated *HBB* mutations. Chapter 1 focuses on the development, validation and clinical implementation of a novel and universally applicable PGT method for the diagnosis of *HBB* gene mutations, utilising next generation sequencing (NGS). Employing this protocol *HBB* mutation status and associated single nucleotide polymorphism (SNP) haplotypes were successfully determined in all 21 embryos of three couples undergoing PGT for ß-thalassaemia. Analysis of 141 heterozygous sites showed no instances of allele dropout in the clinical samples and the test displayed 100% concordance compared with the data obtained using an established method (karyomapping). Taken together, the results suggest that the new method may deliver superior accuracy than typically achieved with traditional PGT methods. Furthermore, the test is streamlined and economical, which should improve patient access to PGT, reducing costs and waiting times. This will be especially important in less affluent parts of the world where diseases affecting hemoglobin synthesis are of high prevalence.

An alternative to PGT for reducing the burden of inherited disorders is the correction of disease-causing mutations in human embryos using genome editing methods. However, this possibility is subject to numerous ethical and technical concerns, especially as intervention in gametes or preimplantation embryos would inevitably involve modification of the germline and a high likelihood of transmission of edited genes to future generations. CRISPR-Cas9 is currently the leading technology for introducing specific and heritable modifications into the genome. Chapter 2 evaluated the CRISPR-Cas9 system from the perspective of technical feasibility. It involved development of a methodological framework and computational pipelines for evaluation of on-target mutagenesis as well as potential off-target consequences of the editing in OCT4 *(POU5F1)* CRISPR-Cas9-targeted human zygotes and controls. A high efficiency of editing was observed with a wide variety of indel mutations, characteristic of non-homologous end-joining (NHEJ). No evidence of off-target activity was recorded.

There is a possibility of unintended editing outcomes following the use of CRISPR-Cas9 that may have pathologic consequences, potentially exacerbated by insufficient DNA repair in early human embryos prior to embryonic genome activation (EGA). Chapter 3 investigated the repair outcomes of Cas9-induced double-strand breaks (DSBs) introduced in the *POU5F1* locus. Strikingly, the results showed that 37.5% of the targeted zygotes present with breaks that remain unrepaired or participate in complex genomic rearrangements, resulting in segmental aneuploidy with breakpoints within the targeted 6p21.3 region. Gains and losses of large regions, stretching from 6p21.3 to the end of the short arm of chromosome 6, as well evidence indicating a complexity of DNA sequence mutations at on-target sites after CRISPR-Cas9 editing, provide a cautionary note to those considering the technology for clinical use and underscore the importance of basic research into DNA repair pathways and genomic stability in human embryos. Such research is not only relevant in the context of genome editing, but also has importance for assisted reproductive treatments and stem cell research. Furthermore, collaborative work of which this thesis is a part emphasizes that CRISPR-Cas9 remains a powerful molecular biology tool for the study of gene function and the biology of early human development.

Introduction

Preimplantation genetic testing

Preimplantation genetic testing (PGT) encompasses a wide variety of methods for the analysis of genetic material sampled from early human embryos, prior to the establishment of a pregnancy (Ray and Handyside 1996). Thus far, PGT has been used exclusively in combination with *in vitro* fertilization (IVF) treatment, and is carried out either to avoid the transmission of inherited disorders or in an attempt to improve IVF treatment outcomes. In the former case, embryos from couples known to be at high-risk of passing on a familial condition are tested and only those found to be free of the disorder are considered for transfer to the mother's uterus (Handyside et al., 1990). Since genetically abnormal embryos are never transferred, any pregnancies established are expected to be free of the condition and pregnancy terminations and affected births are thus avoided. When applied to the detection and avoidance of autosomal dominant, recessive, and sex-linked disorders caused by mutation of a single gene, this strategy is referred to as PGT for monogenic disease (PGT-M). Genetic testing at the preimplantation stage can also be applied for the detection of unbalanced chromosomal rearrangements, resulting from translocations and inversions present in one or both parents, a strategy referred to as PGT-SR (PGT for segmental rearrangements) (Jiang et al., 2012). The use of PGT to enhance IVF outcomes has focused on the detection of chromosome abnormalities, which are common in human preimplantation embryos and are the principal cause of failed embryo implantation and miscarriage. The transfer of embryos found to be chromosomally normal, in preference to those that are aneuploid, has been reported to lead to a higher probability of a live birth per embryo transferred and lower rates of miscarriage (Forman et al., 2013; Scott et

al., 2013a; Yang et al., 2012), although the clinical utility of this approach (known as PGT for aneuploidy - PGT-A) is still debated.

Preimplantation genetic testing for monogenic disease

PGT-M was developed more than two decades ago as a result of the pioneering work of scientists such as Verlinsky, Handyside and others. The first application of PGT-M involved determination of sex of embryos in order to avoid transmission of X chromosome-linked diseases such as Duchenne muscular dystrophy and retinitis pigmentosa, using the polymerase chain reaction (PCR) for Y-chromosome sequences (Handyside et al., 1990). Embryos found to be healthy are prioritised for transfer to the mother's uterus, circumventing the need to consider pregnancy termination and eliminating the disease segregation through the family. Figure 1 summarises the main steps involved in a PGT-M cycle.

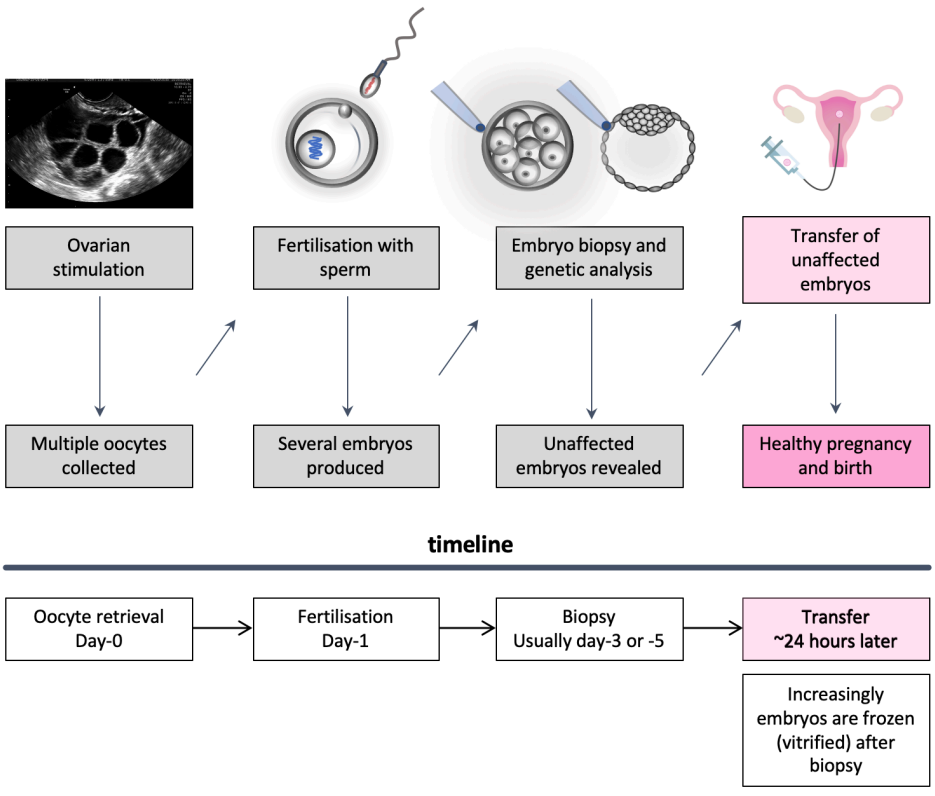


Figure 1: An overview of a PGT-M cycle.

Fundamentally, PGT-M can be divided into two constituent components: 1) embryological, which involves many aspects of routine IVF, but also a method of collecting DNA from the embryos through biopsy of one or more cells; 2) genetic, which involves the processing and analysis of the genetic material in order to yield a clinically useful result. Both of these aspects of PGT have seen considerable evolution in the 30 years since the first cases. Since the 90s, the utilisation of preimplantation genetic testing has grown immensely and continues to increase annually, with the birth of more than 10,000 babies and more than 50,000 cycles performed worldwide thus far (Handyside, 2010, De Rycke et al., 2017). Couples who undergo PGT-M are often influenced by their personal ethics, religious views and/or a complicated reproductive history that may have involved recurrent miscarriages or a history of affected conceptions and pregnancy termination (Van der Aa et al., 2013). Although technically PGT can be applied to virtually any genetic condition for which the causative mutation is known, regulatory authorities may choose to restrict the list of testable conditions to those that are life-threatening or have a significant impact in the quality of life. In the UK alone, 880 single gene conditions have been approved (or are awaiting approval) for testing by the Human Fertilisation and Embryology Authority (HFEA) to date. The code of practice excludes the selection of embryos for social (non-disease) reasons, such as sex determination for family balancing (Human Ferilisation and Embryology Authority, 2015). The most common indications for PGT-M are listed in Table 1. In contrast, PGT-A currently has blanket approval by the HFEA (Chen et al. 2014).

Table 1: Most common indicators for PGT-M. Collected in 2009 by PGD European Society of Human Reproduction and Embryology (ESHRE) Consortium (Moutou et al., 2014).

PGT-M indication	Autosomal Recessive	Autosomal Dominant	X-Linked
1.	B-thalassemia, sickle-cell anaemia	Huntington disease	Fragile X syndrome
2.	Cystic fibrosis	Myotonic dystrophy type 1	Duchenne and Becker muscular dystrophy
3.	Spinal muscular dystrophy	Neurofibromatosis	Haemophilia A and B

Overview of the PGT treatment cycle

For many years the dominant method for obtaining genetic material involved the biopsy of cells (blastomeres) from cleavage stage embryos (on Day-3 of embryo culture). In most cases, only one cell was removed for testing. However, the last decade has seen a dramatic shift in practice, with blastomere biopsy increasingly abandoned in favour of sampling several trophectoderm cells at the blastocyst stage (on Day-5/6) (Chen et al. 2014). The reason for the change in practice is related in part to the fact that genetic tests are more accurate and robust when several cells are available for analysis, rather than just one. Nonetheless, a more important explanation is the growing evidence that removal of a single blastomere at the relatively fragile cleavage stage has the potential to impair subsequent development, leading to lower pregnancy rates for biopsied embryos (Scott et al., 2013b). Breaching the zona pellucida using mechanical, chemical, or laser-assisted technology is indispensable to provide access to the cells within. Unlike the early days of PGT, when rapid genetic testing was essential in order to permit embryo transfer within the same cycle, today biopsied embryos are usually cryopreserved until the results of genetic diagnosis are available. Embryo(s) determined to be at low risk of the specific disorder under analysis are considered to be eligible for therapeutic use and may subsequently be thawed/warmed and transferred to the uterus. In

response to increasing clinical use of PGT-M over the past decade, the European Society of Human Reproduction and Embryology (ESHRE) PGD Consortium was formed and has established a set of practice guidelines, intended to provide information and support to existing PGT-M programmes (Harton et al., 2011).

Strategies for PGT-M of single gene disorders by DNA amplification

Historically, PGT-M has relied on polymerase chain reaction (PCR) to enrich a DNA sample for specific fragment(s) encompassing diagnostically relevant regions, such as mutation sites. DNA amplification-based methods and techniques for the analysis of DNA sequence have changed greatly since they were first introduced (Wells and Sherlock 1998; Chen et al. 2014). Amplification of individual sites in the genome evolved to assess additional loci of diagnostic relevance using multiplex-PCR approaches and, more recently, whole genome amplification techniques have been applied. The earliest methods utilised in order to reveal the presence of disease causing mutations involved heteroduplex formation, fluorescent PCR and amplification refractory mutation system (ARMS) (Wells and Sherlock 1998). The most recent PCR-based method used for PGT-M was quantitative real-time PCR, which combined both amplification and detection in a single step. Table 2 summarises the main strategies developed for PCR-based PGT-M. The DNA amplification methods adapted for PGT-M are significantly more susceptible to technical difficulties than conventional PCR. The technical problems arise because samples are assayed at the single cell level with only a few picograms of DNA available for amplification and therefore require an unusually large number of amplification cycles. Misdiagnosis can potentially occur as a result of contamination, amplification failure or allele drop out (Spits and Sermon 2009), discussed in more detail in the following sections.

Table 2: Strategies developed for PCR-based PGT-M.

PCR method used for genotyping of embryos	Principle	Disease	Examples of references
Heteroduplex formation	Detection of heterozygosity based on the restriction of migration during electrophoresis	cystic fibrosis	Handyside et al. 1992
Fluorescent PCR	Fluorochrome-labelled primer allows analysis of amplified DNA fragments on an automated sequencer with higher sensitivity. Differences related to small deletions and insertions are detected with high accuracy	cystic fibrosis	Hattori et al. 1992 Khosravi et al., 2016
Amplification refractory mutation system	Annealing of allele-specific primers that leads to preferential amplification of mutated/wild-type alleles	cystic fibrosis, spinal muscular atrophy	Sherlock et al. 1998
Restriction endonuclease digestion	Preferential endonuclease digestion allows recognition of mutated/wild-type DNA fragments based on the length of the product	cystic fibrosis, β-thalassemia, sickle-cell anaemia, retinitis pigmentosum	Strom et al. 1998
Minisequencing Multiplex PCR	Minisequencing primer anneals one base before the mutation site and undergoes single base extension, revealing the nature of the nucleotide at the mutation site Simultaneous amplification of multiple informative short tandem repeats (STR) and single nucleotide polymorphisms (SNP) markers linked with the mutation site. Often includes amplification and analysis of the mutation site too. Provides redundant testing, which reduces the risk of misdiagnosis due to ADO	cystic fibrosis, β-thalassemia, retinoblastoma, haemophilia A cystic fibrosis, β-thalassemia, sickle-cell anaemia	Fiorentino et al. 2003 Sherlock et al. 1998; Rechitsky et al. 1998

Quantitative real-time PCR	Quantification of amplicons present in the sample. Can be combined with sequence detection probes (e.g. TaqMan) to reveal the presence of specific mutations.	β-thalassemia, spinal muscular dystrophy, potential to detect wild-type/mutated mitochondrial genomes	Chen et al. 2014

Contamination

The limited amount of template DNA in a single-cell necessitates a large number of cycles of PCR amplification. This increases the risk that any contaminating exogenous DNA will be amplified to detectable levels, potentially compromising the diagnosis. Thus, strict laboratory practices and dedicated clean rooms for setting-up the PCR, physically separated from the laboratory where amplified products are analysed, are required in order to minimise the risk of "carry-over" of previously amplified DNA fragments and reduce the incidence of contamination (Wells and Sherlock 1998).

Allele drop-out and mutation analysis

One of the most significant hazards encountered during single cell amplification in protocols for single gene disorders (SGD) is allele drop out (ADO), defined as the successful amplification of only one of the two alleles present in a heterozygous cell (Ray and Handyside 1996). ADO can affect any of the two alleles and may result in misdiagnosis, since heterozygous embryos appear homozygous when ADO occurs. The consequences can be particularly severe with autosomal dominant disorders where affected embryos may potentially be misdiagnosed as healthy and selected for uterine transfer (Wells and Sherlock 1998; Harper et al. 2002). Imperfect cell lysis and DNA strand breaks have been suggested as possible

explanations for ADO, the latter possibility is supported by the observation that ADO frequency is higher for cells derived from embryos of lower morphology quality scores, which likely contain degraded DNA (Piyamongkol et al., 2003; Spits, 2009).

Efforts to reduce the incidence of ADO include multiplex PCR where multiple polymorphic loci, situated in close proximity to the affected gene, are co-amplified simultaneously using different combinations of primers in order to permit accurate and redundant diagnosis based upon genetic linkage. Although ADO is just as likely to affect the polymorphic marker as the mutation site, it is statistically improbable that both loci will suffer from the amplification failure in the same reaction (Piyamongkol et al., 2003). In order to apply this approach in clinical practice, family-specific single-cell PCR tests need to be designed, validated and optimised. Multiple polymorphic sites may need to be assessed in order to identify markers that are informative in the family undergoing PGT-M (i.e. the polymorphisms must permit differentiation of chromosome carrying a mutation from its homologue bearing the normal copy of the gene). This approach interrogates highly polymorphic short tandem repeats (STRs) or single nucleotide polymorphisms (SNPs). While SNPs have lower levels of heterozygosity and are therefore less often informative, they are much more numerous in the genome than STRs and therefore many may exist in relatively close proximity to the causative gene. This is an important consideration given that meiotic recombination occurring between the gene and the polymorphism will negate attempts to use the marker diagnostically and are more likely when the gene and polymorphism are distant from one another (Renwick et al., 2006).

A priori knowledge of which polymorphism alleles are to be expected in the embryo also provides a method to detect contamination in cases where extraneous genotypes appear during analysis. As the polymorphisms provide a simple form of DNA fingerprint for the sample, it is often possible to not only detect contamination, but also determine the potential source, such as cumulus cell DNA (i.e. maternal origin) or operator cells (Spits and Sermon 2009). Although

current methodologies have managed to reduce the frequency of ADO to approximately 5-10%, these frequencies still represent a significant risk for erroneous clinical results. As recommended by ESHRE PGD Consortium Best Practice Guidelines, the rates of ADO should not only fall below 10%, but require additional steps to be employed in order to reduce the risk of misdiagnosis, such as the use of tests examining multiple diagnostically relevant sites in parallel, as described above (Harton et al., 2011).

Considerations prior to PCR amplification

The utilisation of redundant tests, which analyse several amplified fragments and provide additional opportunities to detect the presence of the mutant gene, are of great importance. However, multiplex-PCR strategies used for PGT-M frequently require an extensive investigation, prior to clinical application, in order to identify polymorphic markers that are informative in the family seeking PGT. To develop a patient-tailored PCR protocol requires skilled personnel and is a labour-intensive and costly undertaking. It is often the case that a protocol developed is only applicable to a particular family. The need to develop optimised, family-specific tests, often leads to a delay in the initiation of patients' cycles and the high associated cost restricts access of patients to this reproductive strategy, particularly in regions where there is no provision of such service by national health care. Although the more established PGT laboratories have access to a considerable number of validated protocols accumulated over the years, it is not uncommon for PGT providers to offer a limited number of tests for monogenic disease, restricted to those conditions where there is a common mutation (Renwick et al., 2006). While understandable from economic and operational perspectives this policy risks discrimination against patients with rare inherited conditions or unusual mutations.

Strategies for PGT-M by whole genome amplification (WGA)

The scarcity of genetic material obtained from a blastomere or a trophectoderm biopsy, coupled with the desire to obtain ever greater quantities of diagnostically valuable information, encouraged the development of more comprehensive techniques for genetic testing in preimplantation embryos. Whole genome amplification (WGA) methods were developed to amplify the entire embryonic genome from the approximately 5-10 picograms present in a single cell to an amount sufficient for a more detailed investigation (Wells et al., 1999; Zheng et al., 2011). Historically, the majority of WGA methods have been PCR-based, where a universal set of self-inert degenerative primers is used to anneal to numerous loci throughout the genome and to then amplify the resulting fragments exponentially by PCR utilizing the universal priming sites (Deleye et al., 2015). An example of such a technology is Picoplex/Sureplex (Rubicon Genomics Inc.), which has been validated for use for PGT in the context of comprehensive chromosome screening by microarray comparative genomic hybridization (microarray-CGH) and, more recently, using next generation sequencing (NGS). Multiple displacement amplification (MDA) represents another important WGA technique, regularly applied for the purpose of PGT-M, usually in combinations with SNP-microarrays (as discussed below). MDA allows amplification of DNA from single cells up to microgram levels, and generates products of significantly higher fidelity, yield and genomic coverage compared to other techniques (up to 99% of the entire genome is amplified) (Martín et al., 2013). These features are valuable for accurate PGT, and offer the possibility of simultaneous analysis of mutation site, polymorphic markers, and chromosome copy number (Zheng et al., 2011).

Strategies for PGT-M by SNP-arrays

Improved WGA technologies together with the development of genome-wide SNP-arrays have managed to overcome some of the challenges with the delivery of PGT-M. In particular, they have permitted a move away from patient-specific tests, providing more generic PGT-M protocols (Handyside et al., 2010; Natesan et al., 2014). A SNP array-based method known as Karyomapping arose as a result of these advancements and combines the detection of single gene defects and cytogenetic abnormalities, using the principle of linkage analysis (Handyside et al., 2010). Parental DNA is genotyped at ~300,000 SNPs scattered throughout the genome using a microarray comprising probes for the different SNP alleles at each polymorphic locus. This provides a means of establishing all four parental haplotypes and their pattern of inheritance. An additional member of the family, such as an affected child of the couple undergoing PGT-M, is also required to be tested and allows setting of phase (i.e. the parental haplotypes associated with the mutant gene are revealed) (Natesan et al., 2014). Provided these conditions are met, genome-wide SNP genotyping serves to identify the unique combination of haplotypes an embryo inherits from its parents, including those specific to mutant and normal copies of the disease gene. Embryos found to carry a high-risk haplotype (inherited along with a gene mutation) are excluded from transfer. In terms of the cytogenetic status of embryo samples, the haplotype information from individual chromosomes reveals which parental chromosomes have been inherited (and how many), while quantitative data concerning the amount of DNA hybridised to the SNP probes on a given chromosome provides a complementary indication of its copy number. Furthermore, in some instances it can be possible to differentiate mitotic segregation errors from the meiotic ones.

A similar technique to Karyomapping, called Haplarithmisis, utilises a next-generation sequencing approach in order reveal the genotypes of polymorphisms spread across the genome and to deduce haplotypes, thus providing the linkage data needed to diagnose embryos (Zamani

Esteki et al., 2015). Additionally, Haplarithmisis employs an analytical pipeline that corrects for the potential whole-genome amplification artefacts to reconstruct the haplotype architecture (Zamani Esteki et al., 2015). Together, these techniques have become widely implemented for genetic testing for monogenic disease in human embryos, enabling diagnosis of disease alleles as well as numerical and structural chromosome abnormalities.

Strategies for PGT-M by next generation sequencing technology

In the context of PGT-M, genome-wide SNP-arrays and NGS-based haplotyping, despite their wide applicability, remain relatively expensive and require DNA samples from close relatives in addition to the male and female patients requesting diagnosis of their embryos. With genetic methodologies evolving rapidly, there is potential to develop PGT protocols of even higher accuracy and lower cost. Chief amongst the new methods used for research and diagnostics is next generation sequencing, an umbrella term that encompasses a range of different techniques which have in common the generation of high-throughput DNA sequence information. In the last decade, sequencing technology had undergone a continuous evolution and improvement in speed, accuracy, and cost efficiency. NGS has now almost entirely replaced the first commercialized method of DNA sequencing known as the Sanger sequencing, which was used for the sequencing of the first human genome (The Human Genome Project). The first generation of high-throughput DNA sequencers were developed and commercialized by Illumina (Solexa), Roche (454) and Applied Biosystems (Ion Torrent), each unique in the technology and chemistry they use (sequencing by synthesis, pyrosequencing, and the direct release of H+ ions from the incorporation of new nucleotides, respectively) (Martín et al. 2013). Fundamental to all three technologies is the principle of shotgun sequencing where a DNA sample is broken into small fragments, ligated with adapters and sequenced. Each fragment of

sequence produced is known as a ‘read’ and the number of times a specific piece of the genome is sequenced is represented by the ‘depth’ (or coverage) (Martín et al., 2013). As the fragments are sequenced in a massively parallel fashion, an NGS run takes a fraction of time required for the conventional sequencing methods (a few hours or days versus weeks). The major advantage of sequencing approaches over SNP genotyping is its scalability, where a user can fully control the throughput, and consequently the costs, depending on how many samples are multiplexed together in a single run. The major disadvantage of this generation of sequencers is their cost, as each requires a significant capital investment, making them inaccessible to many laboratories.

The last couple of years have seen continued technological advancement with the arrival of the “third” generation of sequencers, pioneered by the Oxford Nanopore Technologies (ONT). The technology behind the sequencing method is unique in that it uses a nanopore embedded in a synthetic membrane, through which a single DNA molecule passes and disrupts the electrical current. The change in the voltage is unique to each nucleotide and recorded as a basecall. Nanopore sequencing has numerous advantages over the other technologies, most importantly it confers significantly lower cost and does not require any capital investment. This is because the technology does not use any imaging for the detection of nucleotides, and was, therefore, able to be scaled it down to a portable level. Oxford Nanopore Technologies’ MinION weighs only 90 grams and connects to a laptop computer via a USB port, meaning that sequencing can be conducted anywhere, including in the field (Kono and Arakawa, 2019). The other major advantage of nanopore sequencing is the length of the sequencing reads it is able to generate. As nanopores detect DNA molecules without any intermediate amplification or synthesis step, there is no real limitation to the read length, other than the one represented by the library preparation step, thus enabling numerous previously challenging types of genomic and

transcriptomic study, such as ultra-long read sequencing, detection of structural variation, *de novo* assembly, and analysis of haplotypes (Kono and Arakawa, 2019).

Various NGS methods have already been adapted for use in assisted reproductive technology, providing a robust platform for the detection of aneuploidy in embryo biopsy specimens (Fiorentino et al., 2014a, 2014b; Wells et al., 2014; Zheng et al., 2015). Most NGS-based methods for aneuploidy detection have focused on a 'shotgun' strategy, randomly sequencing a small fraction of the genome (typically <1%) and then counting the number of reads per chromosome to gain an understanding of the number of chromosomal copies. Since so little of the genome is sequenced, the probability of obtaining data concerning any given mutant gene is negligible. While efforts to develop NGS-based methodology to screen for chromosomal imbalance have been successful, attempts to harness the technology for targeted mutation detection still presents a challenge (Fiorentino et al. 2014; Wells et al. 2014). Recently, the first reports of NGS having also been used for the diagnosis of single gene mutations in embryos appeared, although the protocols used clinically still require significant customisation in designing and optimizing multiplex-PCR for the amplification of mutation sites and polymorphisms (Chen et al., 2016; Ren et al., 2016; Treff et al., 2013a; Yan et al., 2015). Further studies should, therefore, be carried out to design, validate and optimize NGS protocols for SGD detection that are more generic and more comprehensive.

The first experimental chapter of this thesis, describes the design, validation and clinical implementation of a novel PGT-M protocol, based on NGS technology, for the detection of virtually all mutations responsible for β-thalassaemia and sickle cell anaemia in preimplantation embryos. This single cost-effective method is applicable to the vast majority of couples seeking PGT for these conditions and represents a valuable option for patients undergoing IVF coupled with PGT for β-thalassemia and sickle cell anemia (Kubikova et al., 2018).

Germline genome editing

Undoubtedly, preimplantation genetic testing for monogenic disease (PGT-M) continues to be a valuable reproductive option for couples who are at high risk of transmitting heritable genetic disorders to their children. Yet despite its growing clinical implementation over the last two decades, PGT-M has a number of limitations. Firstly, there is a finite number of embryos created in an IVF cycle. This is important because the probability of finding an unaffected embryo using PGT-M and the likelihood of achieving a viable pregnancy are strongly influenced by the number of embryos tested. The majority of these embryos will usually be discarded, either as a result of inadequate development or because they are diagnosed affected by the inherited condition for which they were being tested (Wells et al., 2019). When both parents are carriers of a recessive mutation, the proportion of embryos suitable for transfer to the uterus is expected to be reduced by 25% with respect to a routine IVF cycle. In the case of a dominant disease, where inheritance of a single copy of a mutant gene generates a disease phenotype, only 50% of the embryos are expected to be found free of the disease (Barrangou and Doudna, 2016). Published data indicates that 16-20% of PGT-M cycles do not reach the point of uterine transfer and that fewer than 25% of cycles result in a healthy ongoing pregnancy (De Rycke et al., 2017; Gutiérrez-Mateo et al., 2009). The relative infrequency of embryos that are both developmentally competent and unaffected by the familial condition are the principal reasons why the majority of PGT-M cycles fail to produce the much-wanted pregnancy after the physically, psychologically, and often financially demanding treatment. Thus, the development of novel strategies that prevent germline transmission of mutations, while avoiding the discard of affected but otherwise potentially viable embryos, is desirable. Such approaches could increase the number of embryos suitable for transfer, improving pregnancy outcomes and reducing the risk of inherited disease in patients who are opposed to pregnancy termination and are therefore unable to utilise conventional prenatal testing

strategies, and also those unwilling to discard affected preimplantation embryos. In the absence of a means of 'curing' affected embryos, reproductive options for such patients are extremely limited and currently include the use of donor oocytes, sperm or embryos and adoption, all of which sacrifice the possibility of having a genetically related child. The other option is to try and conceive naturally (assuming the couple is fertile) and undergo prenatal testing. Of course, if the prenatal test reveals the presence of genetic disease in the fetus, a choice will have to be made between pregnancy termination and the birth of an affected child (Kubikova and Wells, 2020).

The reproductive options discussed above have, for the past thirty years, been the only available strategies that couples have at their disposal when trying to avoid transmission of an inherited disorder. More recently, dramatic advances in genome analysis and engineering technologies have brought on another possibility: germline genome editing (GE). While heritable genome editing does not currently exist as a reproductive strategy, it has potential to become one in the future. The application of GE for the avoidance of inherited disease would involve the modification of a mutant gene in order to restore a wild-type sequence. Such an approach would shift the paradigm of the current reproductive strategies away from the diagnosis and exclusion and towards cure (Kubikova and Wells, 2020).

From a technical perspective, conducting genome editing interventions at the time of fertilization or during early preimplantation embryonic stages, can be viewed as ideal, since it is easier to ensure that the active components used for GE will be delivered to all cells. This is important because many inherited conditions affect multiple tissues or organs and a return to normal function may require the mutant gene to be corrected in most, if not all, cells. It can be extremely difficult to access the genomes of cells later in life, when the cells are potentially numbered in the millions and populate hard to reach internal organs. A fertilized oocyte or embryo that underwent GE would ultimately give rise to a whole individual carrying the edited

gene in all cells of their body. However, this would not only include all of the somatic cells, but the germ cells also, a fact which has been the basis of some ethical concerns. Genome editing interventions, performed on human embryos, are without a doubt ethically controversial; however, some might argue that the current practice of discarding affected but otherwise viable embryos is wasteful and perhaps no better from ethical, moral as well as certain religious perspectives (Wells et al., 2019). With GE, embryos previously diagnosed as affected could become eligible for transfer, thus rescuing them from being destroyed and increasing the number of IVF embryos available to the couple, which would likely lead to improved chances of a pregnancy during that treatment cycle. In an ideal scenario, if GE could be combined with natural cycle IVF, in which only a single oocyte is collected and fertilized with the partner's sperm, perhaps no viable embryo would need to be discarded (Kubikova and Wells, 2020).

Can germline genome editing be ethically justifiable?

Germline genome editing has received considerable interest from the scientific media and, to a lesser extent, the mainstream and social media in the recent years. As a result of these, there have been multiple initiatives to promote discussion and public engagement on the topic. Such inclusive societal debate might involve various stakeholders such as patient organisations, as well as the members of the general public and expert organisations to better understand the views and reasoning of the society as a whole. In view of supporting public debate, the Nuffield Council on Bioethics (2018) concludes in their Enquiry into genome editing and human reproduction that, on a national level, an independent body should be formed to promote and coordinate debate on genome editing and related technological innovations and help develop norms for governance.

The ethical considerations largely stem from how the reproductive desires and decisions of prospective parents are embedded in a context of knowledge and possibilities for actions that link different types of interests and responsibilities. In this case, the knowledge relates to the role of the genome and their own genetic status, and the possibilities represent the genome editing technology. In broad terms, when analysing the ethics of genome editing, the considerations can be split into three main categories: 1) those immediately involved (the prospective parents and the future child), 2) members of society indirectly affected and general public, and 3) future generations.

For people immediately involved, there may be various reasons for wanting to have a biologically related child, and this has been recognised and protected as a human right. In some cases, the use of assisted reproductive treatment provides the only solution, but its success is limited. It is the view of the Nuffield Council of Bioethics (2018) that any genetic modification affecting the germline should only be permissible if it protects the welfare of the future individual, and welfare beyond just good health but also social and psychological wellbeing should be considered. Furthermore, the difficulty in deciding which traits are justifiable for intervention lies in the complexity of deciding what constitutes disability. This depends on a person's particular circumstances as well as social environment such as assistance and access to healthcare. To complicate matters further, it might be impossible to fully predict the effect of certain interventions with confidence, as would be the case with traits controlled by polygenic inheritance as well as those affected by external and environmental factors. Finally, one must ensure the safety of the technique, and there are concerns relating to unintended consequences of editing that may persist in the future generations. Even though few medical interventions carry no risk, it is, therefore, imperative to also consider possible alternatives alongside genome editing.

Although the prospect of wide clinical implementation of germline genome editing to eliminate inherited disease remains largely speculative at this point, should it happen, there are potential effects for others in society and future generations. Firstly, there is the argument that heritable editing could contribute to reduction of population diversity, not only loss of variants associated with disease but also those with beneficial characteristics, such as resistance to pathogens. Furthermore, there is a concern about the attitudes towards disabled people and how does the existence of such reproductive technologies shift opinions about disability. Some believe it might reinforce negative messages about disability and propagate the view that life of a disabled person is not worth living. Finally, there is a concern for equity and justice. If access to genome editing is not equally distributed, because of the financial costs involved, then the potential benefit will not be shared equally in the society, which may exacerbate the existing division and inequality. To conclude, it is the view of the Nuffield Council (2018) that clinical use of genome editing should only be permissible if intended to secure welfare of the future individual and the interventions safe and consistent with social justice and solidarity so that it should not increase disadvantage, discrimination, or division in society.

How far are we from (safe) clinical application of genome editing?

Recent technological advancements in the area of genomics have enabled previously unprecedented levels of interrogation of DNA variation and have opened up the possibility of genome editing (GE) for research and, potentially, for eventual clinical application. CRISPR-Cas9 (clustered regularly interspaced short palindromic repeat-Cas9) is currently the leading approach for GE of cells, tissues and whole organisms. It has been successfully applied in microorganisms, plants, animals, and most recently in human embryos donated for research (Dever et al., 2016; Fogarty et al., 2017; Ma et al., 2017; Maddalo et al., 2014; Wang et al., 2014; Yin et al., 2014). Cas9 is an endonuclease of bacterial origin, which can be guided to

specific 'target' DNA sequences by a single-stranded guide RNA (sgRNA) containing a specified sequence (~12 bp in length) complimentary to the target DNA. Upon successful recognition of a protospacer-adjacent motive (PAM) sequence, Cas9 flanks the targeted region and cleaves each DNA strand, generating a double-stranded break (DSB) (Fogarty et al., 2017) (summarised in Figure 2).

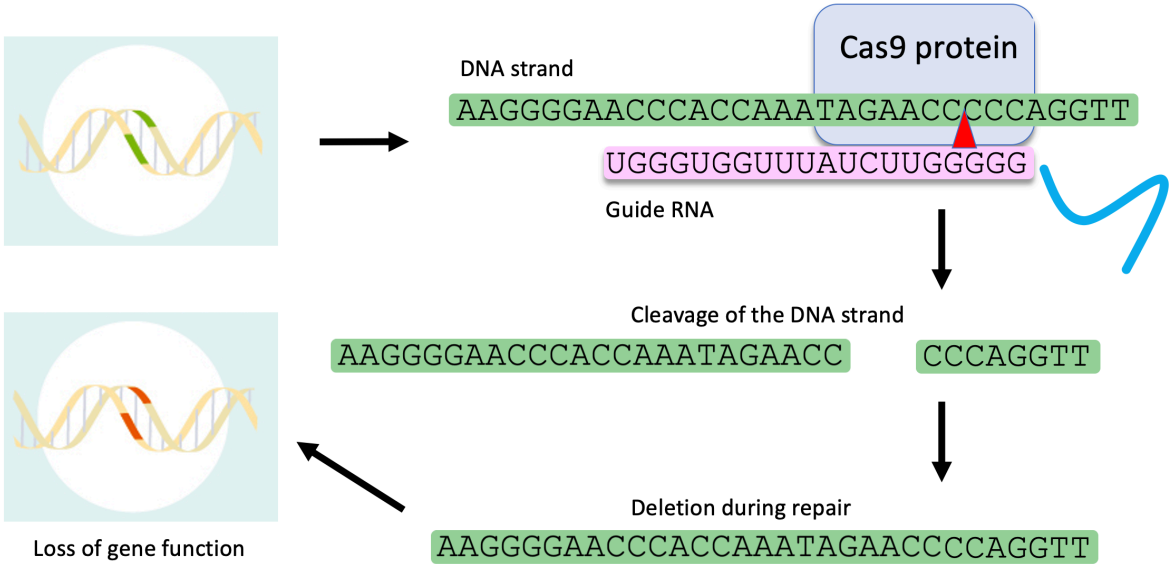


Figure 2: An Overview of the CRISPR-Cas9 system. Engineered CRISPR-Cas9 system contains two components: a guide RNA (sgRNA) and a CRISPR-associated endonuclease (Cas9 protein). The sgRNA is a short synthetic RNA composed of a scaffold sequence (shown in blue) necessary for Cas-binding and a user-defined ~20 nucleotide spacer (in pink) that defines the genomic target to be modified. Thus, one can change the genomic target of the Cas9 protein by changing the target sequence present in the gRNA. The DNA strand is then cleaved, and the resulting double-stranded break repaired using the predominant form of DNA repair, non-homologous end joining (NHEJ). NHEJ generates a frameshift due to the deletion present in the modified sequence and expression of the protein encoded by the targeted gene is subsequently lost.

Most cells resolve the induced DSB predominantly by two mechanisms: error-prone non-homologous end joining (NHEJ) and, less frequently, homology-directed repair (HDR). During the process of reconnecting the two ends of the broken DNA strand, NHEJ usually introduces

insertions and deletions (indels), which typically results in disruption of the targeted gene and a consequent loss of function. In contrast, HDR rebuilds the site of breakage using a homologous DNA molecule as a template (summarised in Figure 3). In diploid cells, the second, undamaged copy of the gene is usually employed as the template for HDR, leading to replacement and correction of the copy in which Cas9 has induced a DSB. However, it is also possible to supply an exogenous DNA sequence, with homology to the targeted site, which the cell can utilise as a template for repair (Fogarty et al., 2017).

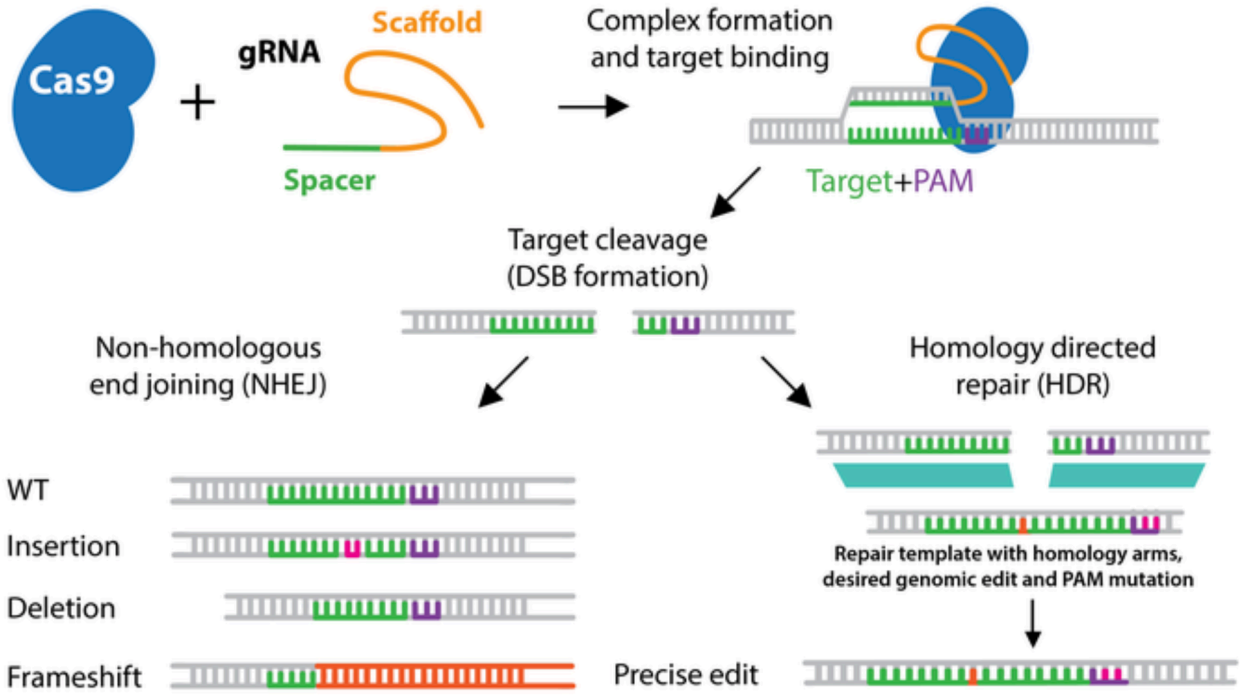


Figure 3: An overview of CRISPR-Cas9 and possible repair outcomes via NHEJ and HDR. The CRISPR-Cas9 ribonucleocomplex (RNP) produces a double stranded break upon the recognition and cleavage of the target DNA sequence. DNA repair will proceed in two possible routes: non-homologous end joining (NHEJ) producing indels (insertions, deletions) most commonly resulting in a frameshift and consequent loss of function or homology directed repair (HDR) repairing the break with a template DNA molecule (either present in the cell or supplied exogenously), allowing restoration of the wild-type genotype or precise modification of the allele (adopted from Adhikari and Poudel, 2020).

Genome editing technology has already been applied to correct pathogenic mutations such as those causing Duchenne muscular dystrophy and β-thalassemia in cellular and animal models (Long et al., 2014; Song et al., 2015). In 2017, a proof of concept study demonstrated the possibility of targeted correction of germline mutations in human preimplantation embryos for the first time, using a CRISPR-Cas9-based technology to remove a dominant heterozygous *MYBPC3* mutation, responsible for an inherited form of cardiomyopathy (Ma et al., 2017). In the same year, our group collaborated with researchers in the UK who edited a group of human zygotes donated for research in order to study the role of an important developmental regulator OCT4 using a similar CRISPR-Cas9-based approach, but in this case deployed to knockout gene function rather than correct a mutation (Fogarty et al., 2017). This research marks the beginning of a new era, in which modification of the human germline at the preimplantation stage is used with the intention to answer basic biological questions related to embryonic development. Together, the two studies of human embryos also lay down a framework for the assessment of the CRISPR-Cas9 system during early embryonic development in terms of its efficacy and safety of the potential clinical application.

Although it seems possible to achieve relatively high targeting efficiency, it appears that only a proportion of embryos might be able to resolve the DSBs by HDR, and therefore succeed in correcting the mutant gene. A significant proportion of the edited embryos might harbour additional indels, further disrupting the targeted gene. In addition, the induction of large deletions and/or complex structural rearrangements, extending over many kilobases of DNA from the cut site, may occur, potentially leading to genomic instability and mitotic arrest. Large deletions and chromosomal aberrations, as well as any form of mosaicism (i.e. where some cells are successfully edited and others are not, or where a proportion of the cells harbour additional deleterious indels produced by NHEJ) have been difficult to detect in preimplantation embryos subjected to GE, due to limited quantity of the DNA obtained from

single cells, and this would present a serious safety concern for clinical application. It remains to be determined whether mosaicism in gene editing can be reduced by modulating the cell cycle stage at which DSBs are induced and whether this approach can yield 100% uniformity in the desired genotype.

In addition to developing a precise GE tool, it is essential to examine the potential off-target consequences of GE technologies, particularly their capacity to induce mutations in unintended regions of the genome, which have close homology to the targeted locus. The approaches that investigate the off-target effects of GE have thus far relied on the use of *in silico* modelling as they are rapid, easy to use and inexpensive. However, they have only a predictive value and their estimates are far from definitive. With the prospect of clinical implementation, one might foresee that whole genome sequencing of the edited embryo might become an inevitable and essential accessory to exclude the possibility of any off-target mutagenesis and to ensure the delivery of safe and efficacious treatment (Kubikova and Wells, 2020).

The second experimental chapter of this thesis sets out to examine the technical feasibility of germline genome editing using the CRISPR-Cas9 system in terms of its potential clinical application. This investigation involved characterisation of the spectrum of mutagenesis induced by CRISPR-based editing in human preimplantation embryos donated for research and development of tools for the assessment of gene editing in terms of its efficiency and safety, the paramount considerations that preclude clinical application. The methodology described in this chapter will investigate the on-target and off-target activity of the CRISPR-Cas9 system utilising an sgRNA specifically targeting the *POU5F1* gene in human zygotes. The chapter concludes with the review of ethical issues concerning the therapeutic use of genome editing in the germline.

Prior to curing disease, do we need to think about DNA repair?

The ability to modify virtually any region of the genome with relatively little effort or expense has transformed biological research. With the advent of genome editing, correction and disruption of genes has been demonstrated in a variety of models, including cells, animals and the human. The expansion in the application of GE was largely facilitated by the development of the rapid, simple and efficient system utilising CRISPR-Cas9 technology. The ease with which it is possible to generate sgRNAs to target specific loci in the genome, as opposed to the earlier versions of programmable nucleases such as the ZFNs and TALENs, made the notion of human germline editing much seem a more realistic possibility, at least from a technical perspective.

In 2019, only seven years after CRISPR-Cas9-based genome editing was first described as molecular biology tool (Gasiunas et al., 2012; Jinek et al., 2012), the birth of the first "CRISPR babies" Lulu and Nana was announced at a summit in Hong Kong. Gene editing had been applied at an embryonic stage in order to disrupt healthy copies of *CCR5* gene, with the intention of conferring resistance to the HIV virus (the father of the children was HIV positive). The announcement was met with condemnation from many quarters, not least from the scientific community, and stimulated wide scientific, societal and ethical discussions, as well as triggering calls for a moratorium on human germline genome editing across the world. The news has also led to initiation of efforts to introduce a new regulatory framework, since rapid technological advancements had outpaced the ability to legislate and modify existing controls. Considering how little is understood about the underlying biology of early human development and the impact of introducing DNA damage in the form of DSBs on the developmental capacity of human embryos, it is clearly premature to consider techniques such as CRISPR-Cas9 for clinical application. This should not be attempted until the functionality of key pathways required for the resolution of DNA damage at early embryonic stages is better understood.

The first studies to have applied germline GE in human embryos used abnormally fertilised tripronuclear zygotes, as these abnormal embryos are not viable and would never be considered for uterine transfer, diminishing ethical objections (Kang et al., 2016; Liang et al., 2015; Tang et al., 2017). As proof-of-concepts, these studies, set out to characterise how CRISPR-based editing performs in terms of efficacy, on-target and off-target activity, mosaicism and continued compatibility with preimplantation embryonic development. While some of the former objectives may be achievable, it might be difficult to address the issue of continued development after editing when using material that is ultimately non-viable, discarded during routine IVF. Since these embryos develop abnormally, it is plausible that their DNA repair pathways may be dysfunctional (Lea and Niakan, 2019). Most of the studies published in the area of germline genome editing have focused on achieving "gene correction" facilitated by HDR as a proof-of-concept for future clinical application. However, the HDR rates reported from these studies are generally extremely low (in the range of 5-15%), and one can hypothesise that this is due in part to the technical difficulty of inducing HDR repair in preference to NHEJ and also due to the inherent biology of the early human embryo, which is transcriptionally quiescent for the first two days of life and potentially less able to adjust and respond to challenges such as DNA damage, which typically require dynamic gene expression. As the specificity and efficiency of genome editing steadily increases, some fundamental aspects of human developmental life, including DNA repair capacity, mechanisms of DNA damage response, cell cycle control and timing, will need to be more fully elucidated. Without a doubt, basic research into the fundamental questions concerning embryonic development will inform the germline GE debate and help establish when and how CRISPR-based editing can be considered appropriate and safe for clinical use, if ever.

Genomic instability and lack of checkpoint control is common in early human embryos

The success rates of human preimplantation development during assisted reproductive treatments remain relatively low, with less than 50% of all fertilised zygotes reaching the blastocyst stage (Hardy et al., 1989). The initial point that marks significant embryo loss is reaching of the cleavage stage, which, in the human, coincides with the activation of embryonic genome (EGA, between 4-8 cell stage) (Braude et al., 1988). Problems with EGA initiation, which may be associated with aneuploidy in some cases, most commonly manifest as developmental arrest. Of note, it is plausible that embryo culture and medium formulation exacerbate this issue and do not currently provide optimal conditions to adequately support and sustain *in vitro* development, in particular those conferring protection from DNA damage and chromosome malsegregation (Swain et al., 2016).

Embryonic mosaicism, where two or more cytogenetically distinct cell lines are present within one embryo, is, at its core, a failure of chromosomes to properly segregate during mitosis (Taylor et al., 2014). Mosaic embryos can be characterised by the presence of cells with different types of aneuploidy in the absence of any normal cells or a mixture of euploid and abnormal cells (Munné and Wells, 2017). Mosaicism arises as a consequence of errors in chromosome segregation occurring during mitotic divisions, and they occur at an appreciable frequency. For cleavage stage embryos, the reported frequency varies greatly, ranging from 15% to 90% (Baart et al., 2006; Harper et al., 1995; Taylor et al., 2014). There is a decrease in the proportion of abnormal cells from the cleavage stage towards the blastocyst stage, which may indicate selection against abnormal cells, either in terms of diminished survival of abnormal cells or reduced rate of cell division. Furthermore, it has been reported that mosaic embryos that do not arrest prior to blastocyst formation are less likely to implant, suggesting

that this form of genomic instability in human preimplantation development, even if confined to the cleavage stage, is deleterious to long-term survival (Munné et al., 2016).

While non-mosaic aneuploidy is most commonly a consequence of errors occurring during meiosis and its occurrence increases dramatically with advancing maternal age, the presence of segmental aneuploidy in human oocytes and preimplantation embryos is largely a result of post-zygotic mitotic errors (Babariya et al., 2017; Vanneste et al., 2009). In the study of Babariya et al. segmental aneuploidy, involving loss or gain of chromosomal fragments was found at an appreciable frequency of 10% in human oocytes but increased dramatically during the first three days of *in vitro* development, reaching almost 25%, before declining to about 15% at the blastocyst stage (Babariya et al., 2017). This suggests that the cell cycle is more relaxed during the first few mitotic divisions following fertilisation, permitting progression through cell division even in the presence of DSBs. Chromosome instability, a hallmark of tumorigenesis, could partly explain the low human fecundity as frequently observed segmental aneuploidy could directly lead to embryo loss. It could be assumed that the high rates of segmental post-zygotic aneuploidy that persist during preimplantation development are due to decreased checkpoint control from the time of fertilisation until the cleavage stage, with the stringency beginning to increase after EGA, allowing elimination of cells affected by segmental gains and losses (Lea and Niakan, 2019). It is, therefore, an imperative to consider that in the context of CRISPR-Cas9-based editing, human oocytes and zygotes may not recruit similar mechanisms for the repair of DSBs compared to other model organisms and cellular systems.

Consistent with the idea that cellular pathways responsible for correcting genetic errors/damage are relaxed until after EGA, the study of Bazrgar et al. (2014) found that Day-4 embryos that contain more than one aneuploid chromosome and exhibit abnormal morphological features as well as displaying atypical timing of cell division overexpress at

least five genes involved in DNA repair (*MSH3, XRCC1, RAD50, LIG1* and *CDK7*) compared to the controls. Conversely, genes involved in checkpoint control and apoptosis are downregulated in these embryos, implying that DNA repair may be attempted without cell cycle arrest. A separate study detected association between altered gene expression of several genes involved in cell cycle checkpoint control and DNA repair, including *TP53* and *BRCA1*, and certain abnormal morphological features such as granular cytoplasm, condensed organelles and multinucleation in 4-10-cell human embryos (Wells et al., 2005). Kiessling and colleagues also found that blastomeres of cleavage stage embryos are under unique cell cycle control, overexpressing cell cycle progression drivers and downregulating checkpoints. In their comparative array-based analysis of cleavage stage embryos, hESCs and fibroblasts, canonical checkpoint genes such as *RB1* and *WEE1* were downregulated in cleavage stage embryos, while cell cycle progression genes such as those encoding cyclins E and D, *CDC25B* and *MYC* were enriched. Interestingly, they observed enrichment of genes encoding elements of circadian clock in the cleavage stage embryos, indicating a certain type of periodical control of division, irrespective of the status of DNA (in)stability (Kiessling et al., 2010, 2009). The absence of checkpoint control in the early human pre-EGA embryo may be advantageous for embryo survival through escaping the p53-dependent arrest and apoptosis pathways present in somatic cells. On the other hand, downregulation of genes involved in checkpoint control may also favour accumulation of DNA damage, and lead to embryo loss once the blastocyst stage is reached and apoptotic pathways becomes more active (Lea and Niakan, 2019).

DNA repair after DSB formation in preimplantation human embryos

The above studies, focused on the expression of DNA repair and cell cycle factors, raise an interesting question of how exactly early human embryos carry out DNA repair after CRISPR-Cas9-induced DSBs, if indeed they are able to carry out repair at all. Since successful DNA

repair is a dynamic process that requires active transcription, it is likely that embryos prior to EGA exist in a state of vulnerability during which they are highly susceptible to genotoxic damage, and their ability to successfully repair DNA is compromised. Two studies have attempted to manipulate whatever endogenous DNA repair capacity exists, attempting to favour the HDR pathway - necessary for replacement of a mutation with wildtype DNA sequence. The rates of HDR increased in human pluripotent stem cells and mouse zygotes after providing external factors that favour this type of repair and inhibit others, such as by overexpressing RAD51 and inhibiting 53BP1 (Canny et al., 2018; Takayama et al., 2017). Techniques to improve the utilisation of HDR may be necessary if CRISPR-Cas9 is ever to be used to correct a mutation, since other studies have shown that DSBs induced by GE methods are predominantly repaired by NHEJ, resulting in the generation of indel mutations, with only low frequencies of HDR even when an exogenous repair template is provided (Kang et al., 2016; Liang et al., 2017; Ma et al., 2017; Tang et al., 2017). Even though some factors required for HDR are present in early human embryos (such as RAD50), what is not clear is whether these factors alone are sufficient for functional repair. Furthermore, the CRISPR-Cas9 components are usually microinjected at the fertilisation or pronuclear stage, several days before EGA, and it is therefore plausible that any repair that takes place is reliant upon maternally deposited factors.

Poor control of timing when CRISPR-Cas9 components enter the cell and induce a DSB offer an alternative explanation for low frequencies of the HDR observed in preimplantation embryos. The reagents can be delivered via microinjection into the cytoplasm or into the two pronuclei, or through electroporation. Furthermore, the CRISPR components can either be injected during fertilisation by ICSI, or at the zygote stage. The appearance and fading of pronuclei typically mark the end of the zygotic S-phase and progression to G2 (Balakier et al., 1993; Capmany et al., 1996). It has been reported that HDR is restricted to late S and G2 phases

when DNA replication has been completed and sister chromatids are available as repair templates. Conversely, HDR has been shown to be downregulated at the M and early G1 phases, thus favouring NHEJ-induced indel generation (Lin et al., 2014; Orthwein et al., 2014). Furthermore, the timing in which the reagents act will differ depending on the form in which they are delivered, i.e. whether using Cas9 mRNA or an RNP complex, with mRNA requiring transcription and translation causing a delay of 6-12 hours in comparison to RNP. Similarly, the rate of protein degradation differs based on the form of delivery, with mRNA taking approximately 72 hours post-injection compared to ~24 hours for the RNP. If chromatids condense in preparation for mitosis, the process of DSBs formation may be further delayed due to the target DNA sequence being inaccessible (Horlbeck et al., 2016; Isaac et al., 2016). Taken together, the above considerations indicate that it may be difficult to experimentally control when exactly the DSB formation occurs, and in turn make it difficult to dictate which kind of DSB repair takes place. Gu et al. (2018) coordinated the microinjection of CRISPR-Cas9 reagents with the zygotic genome activation in the mouse (taking place at the 2-cell stage) and found that this greatly enhanced the rates of HDR through which they were introducing a large insertion to generate a knock-in mouse. The authors concluded that this approach took advantage of the prolonged G2 phase with likely increase in homologous recombination during which the chromatin structure is open and accessible. It remains to be established whether this technique could be replicated successfully in human embryos since it is currently unknown whether and when there is a prolonged G2 phase. It is worth noting, however, that it might not be technically feasible to carry out successful editing of all cells of the human post-EGA embryo using microinjection technique, since EGA takes place at the 4-8-cell stage and this would greatly increase the difficulty of the microinjection procedure. Even if this limitation could be overcome with electroporation or another method of delivery, the technique might generate embryos that are mosaic for various indels/mutations as well as HDR signatures, and

this would, almost certainly, be incompatible with the clinical use of CRISPR-Cas9 technology in the human germline. The various potential options for timing of CRISPR-based editing in human preimplantation embryos are summarised in Figure 4.

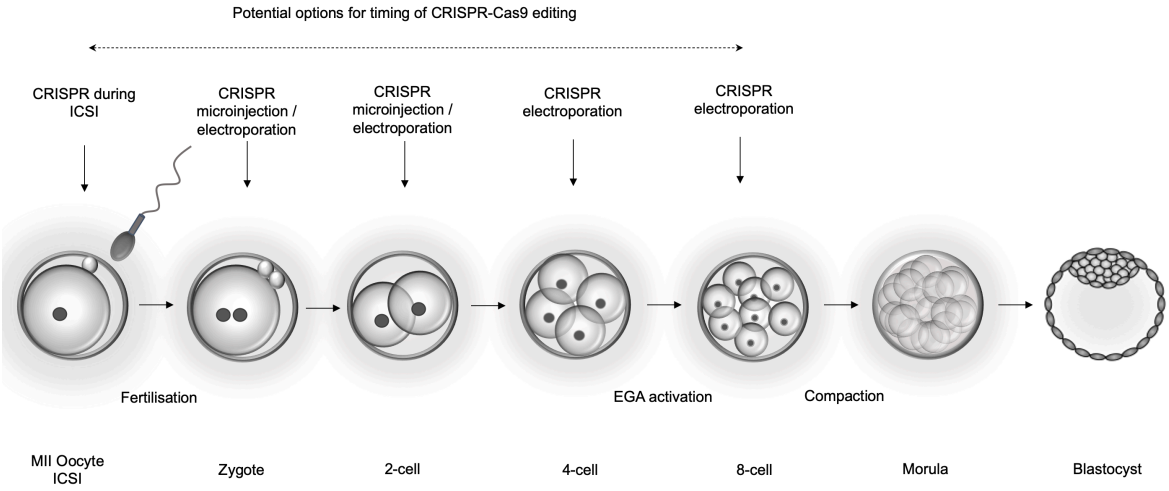


Figure 4: Techniques and timing of CRISPR-Cas9 genome editing. The CRISPR components can be introduced during fertilisation of MII oocytes by ICSI using microinjection technique; microinjection at the zygote stage into the pronuclei or into the cytoplasm; or electroporation at later stages of preimplantation development (4-8-cell stage, morula or blastocyst).

The third experimental chapter of this thesis set out to examine whether CRISPR-Cas9-based genome editing technology introduces DNA damage that is challenging for preimplantation human embryos to repair. It is hypothesised that human embryos prior to EGA are highly susceptible to persistent genotoxic damage because their ability to successfully repair DNA damage is compromised. If DNA repair is indeed deficient at this stage of development, unresolved DSBs are likely to induce a state of genomic instability that could lead to mitotic arrest, and ultimately be deleterious to embryo survival. How (if) human preimplantation embryos resolve DNA damage may have significant implications for those considering the use of GE technologies for the treatment of inherited disorders as well assisted reproductive technologies.

Chapter 1: Development and clinical application of a novel method for PGT of β-globin mutations

1.1 INTRODUCTION

One of the most important reasons for referral for PGT-M worldwide is abnormality of hemoglobin synthesis caused by recessive mutations in globin genes, of which mutations in the hemoglobin beta (*HBB*) gene are the most common. *HBB* is a member of the β-globin gene cluster located on the short arm of chromosome 11 (Figure 1.1).

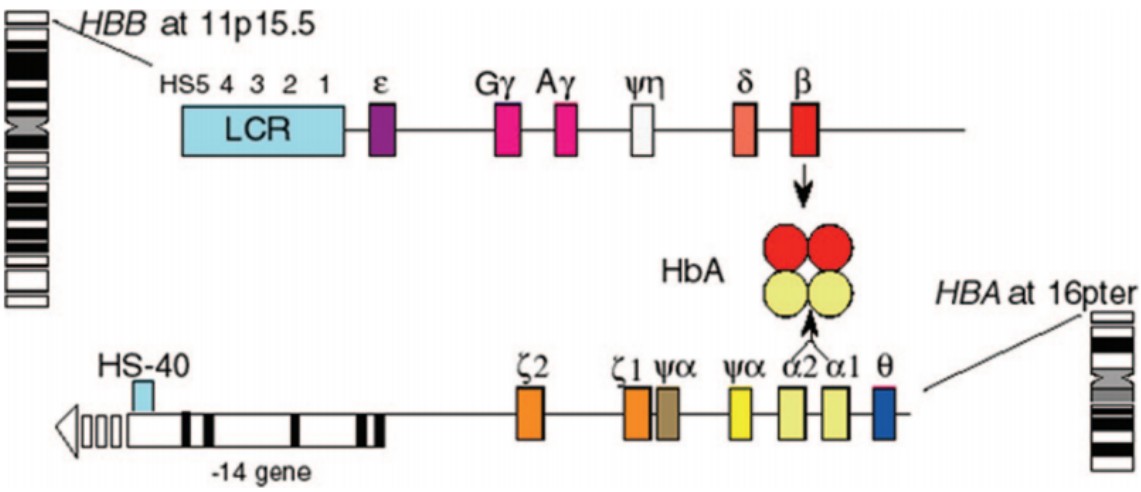


Figure 1.1: B- and α-globin gene clusters - chromosome localisation and structure. Two subunits of β-globin assemble with two subunits of α-globin into hemoglobin A (HbA) (Cao and Galanello 2010).

Sickle cell anaemia (SCA) is the most common inherited blood disorder in the United States and affects approximately 1 in 500 African Americans. Caused by the point mutation Glu6Val in the *HBB* gene, SCA is characterised by abnormal production of hemoglobin S (HbS) (Rees

et al., 2010). Currently, approximately 8% of all African Americans are estimated to carry the Glu6Val mutation. Higher frequencies of the pathogenic allele in regions of the world affected by malaria appear to be the result of selective pressure due to the heterozygous state conferring protection against this parasitic disease (Ashley-Koch et al., 2000). HbS frequencies in population affected by malaria are presented in Figure 1.2.

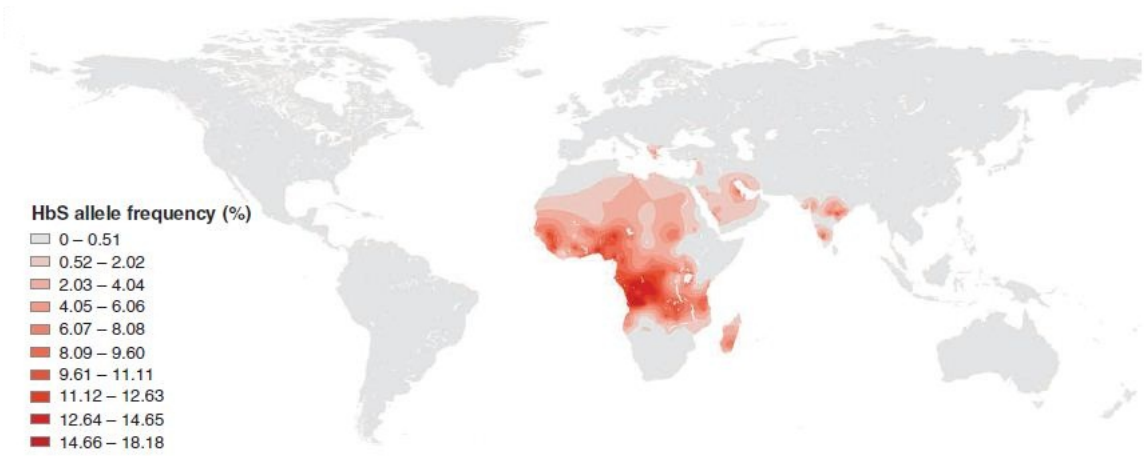


Figure 1.2: Frequencies of HbS allele in regions affected by malaria (Piel et al., 2010).

B-thalassemia belongs to one of the most common autosomal recessive disorders worldwide and is caused by the reduction or absence of β-globin synthesis. Imbalance in the ratio of α- and β-chains leads to abnormal erythropoiesis and the shortage of mature red blood cells (Ottolenghi et al., 1975). The diversity of phenotypes reflects the heterogeneity of mutations in the *HBB* gene. B-thalassemia is highly prevalent in the Mediterranean, Middle East, Far East, Central Asia and India, with the highest frequencies reported in Cyprus (14%), Sardinia (12%), and Southeast Asia (Cao and Galanello 2010). A complete list of mutations associated with β-thalassemia can be found in *HbVar* database, according to which 336 different *HBB* gene variants have been identified to date, mostly as a result of single nucleotide substitutions,

small deletions, and insertions (Cao and Galanello 2010). Genotypic heterogeneity varies greatly depending on the geographical location, with relatively few haplotypes associated with the disease comprising the majority of all cases in each particular region (Figure 1.3, Cao and Galanello 2010). The diversity of disease-causing mutations in the β-globin gene presents a problem for the conventional PGT-M based upon multiplex PCR, requiring a customised design for each combination of mutations. There are tens of thousands of possible combinations of mutations that could cause β-thalassaemia, meaning that a similar number of unique PGT-M protocols would potentially need to be developed and optimized, adding to costs and delays to patient treatment.

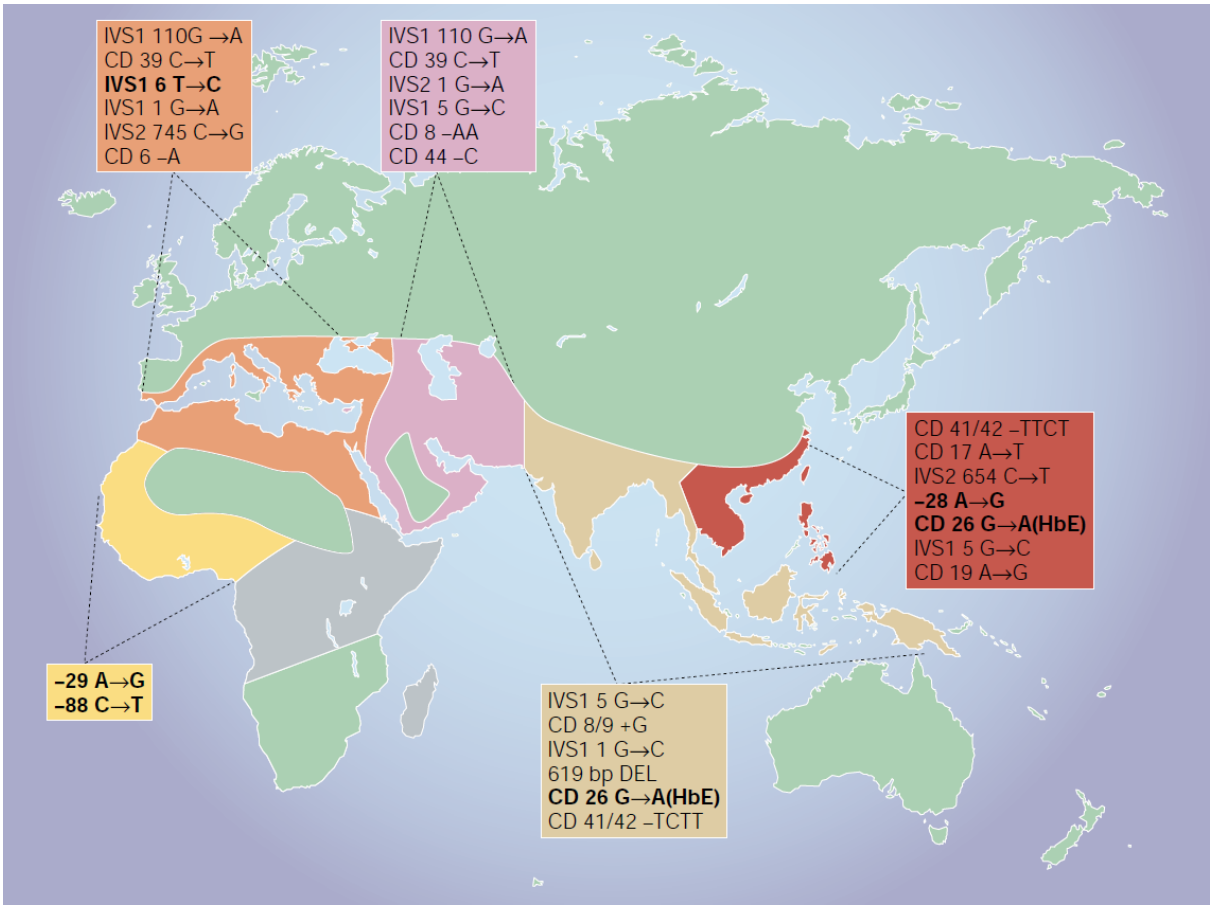


Figure 1.3: Global distribution of β-thalassemia mutation. The most common mutations are shown in bold, the colour is coded based on the region and demonstrate the heterogeneity of β-thalassemia mutations depending on the geographical locations. For areas in grey, little is known about the disease causing mutations. (Weatherall, 2001).

High frequencies of mutations in the *HBB* gene among certain populations have prompted an increase in the availability of genetic counselling and population screening in high-risk areas (Cao and Galanello 2010). As a consequence, the number of individuals affected by SCA and β-thalassemia in industrialised countries continues to decrease. Nonetheless, carrier states remain abundant and continue to render the need for preconception, prenatal, and preimplantation diagnosis in high-risk couples, pregnancies, and embryos, respectively. NGS technology shows a great potential for PGT-M. However, to maximise their value in clinical practice, NGS protocols for monogenic disorders should be designed carefully and in such a way that they can be applied to multiple families, thus ensuring time- and cost-effectiveness. As discussed above, there already exist techniques such as Karyomapping which are applicable to most couples without the need for extensive customisation. However, such tests are relatively expensive and require additional family member(s) to establish the pattern of inheritance (i.e phase). The aims of this chapter were:

- to design a novel PGT-M method, harnessing powerful NGS technology to produce a low-cost and comprehensive method for the diagnosis of β-globin (*HBB*) mutations using a newly developed strategy combining multiplex PCR followed by sequencing

- to deliver a test capable of detecting all mutations found within the HBB gene, applicable to any couple undergoing PGT-M for β-thalassemia and sickle cell anemia

The β-globin gene was chosen because of the difficulties experienced by routine PGT-M methods in approaching the wide array of mutations present in the gene, and because mutations within *HBB* represent the most common cause of inherited disease in the world.

The project aimed to deliver a single protocol capable of sequencing all known mutation sites within the *HBB* gene, in addition to multiple linked SNPs flanking the gene. Development of a comprehensive protocol such as this could potentially be applied to any patient carrying an *HBB* mutation, provided that informativity of SNPs is established prior to the analysis. This chapter describes the design and development of this protocol, as well as its extensive optimization and eventual clinical implementation.

1.2 MATERIALS AND METHODS

1.2.1 Targeted multiplex PCR design of the HBB panel

Selection of suitable SNPs

Polymorphic SNPs in close proximity of the *HBB* locus were selected using the NCBI dbVAR database in combination with the ENSEMBL genome browser, based on the distance from the *HBB* gene and based on the minor allele frequency (MAF). The closer the SNP was to the gene and the higher the variation in the general population (represented by the MAF value), the higher score was allocated to it. Using this ranking approach, twenty-three highly polymorphic SNPs were identified, 12 of which were centromeric SNPs and 11 telomeric SNPs (upstream and downstream from the *HBB* gene, respectively) (Figure 1.4). The MAF cut-off value was set to a minimum of 0.24 to distinguish common polymorphisms from rare variants and only those SNPs with MAF values validated by the 1000 Genome Project were included in the analysis. All of the SNPs were selected within the β-globin gene cluster on chromosome 11, positioned between nucleotides 5246696-5248301 (using the genome assembly GRCh37.p13) in order to minimise the possibility of occurrence of any recombination events (Kubikova et al., 2018).

Telomeric SNPs

	SNP ID	Position	MAF	Alleles
13	rs7945118	5236417	0.3299	G/C
14	rs34220818	5236851	0.498	C/T
15	rs10837620	5243559	0.246	G/A
16	rs12364872	5244144	0.3253	G/A
17	rs10837626	5244299	0.3293	A/T
18	rs10837628	5244404	0.2794	G/A
19	rs2187610	5245406	0.391	G/C
20	rs10768682	5245507	0.2937	T/C
21	rs10837630	5246042	0.2921	C/G
22	rs10837631	5246356	0.248	A/T
23	rs7110263	5246512	0.2897	G/T

Centromeric SNPs

	SNP ID	Position	MAF	Alleles
1	rs3813727	5255912	0.4918	A/G
2	rs7948668	5256647	0.4233	C/T
3	rs4910543	5258827	0.4064	G/C
4	rs4910735	5258852	0.406	A/G
5	rs4910544	5258856	0.406	A/T
6	rs4910736	5258989	0.4099	C/A
7	rs2105819	5259727	0.4071	G/C
8	rs968857	5260458	0.4944	G/A
9	rs968856	5260576	0.4744	A/G
10	rs11036364	5249004	0.4038	A/G
11	rs7936823	5250168	0.3776	A/G
12	rs6578588	5252251	0.355	C/T

Figure 1.4: Selection of SNPs for primer design A) Graphical distribution of 23 selected SNPs with reference to the *HBB* and the *HBD* locus. Telomeric SNPs are displayed as blue lines, centromeric SNPs as pink. The image was retrieved from the NCBI dbVAR database using the GRCh37.p13 reference genome assembly after manually importing the SNP IDs and *HBB/HBD* accession numbers into the database. **B)** Centromeric SNPs, located proximal to centromeres of chromosome 11. **C)** Telomeric SNPs, located proximal to telomeres of chromosome 11. **B**, **C:** Minor allele frequency (MAF) is reported as allele frequencies established by the 1000 Genomes Project and refers to the frequency of less common allele in a pool of all alleles in a tested population. 0.24 MAF cut-off value was chosen empirically to filter out only the common variants. Although MAF is not a measure of heterozygosity, MAF was selected as a criterion because heterozygosity values were not reported for all the SNPs. Where reported, high heterozygosity values (>0.3) were observed for all cases of SNPs, whose MAF values were above 0.24.

Primer design

The initial phases of the study focused on the design of primers for the *HBB* gene as well as the selected SNPs in order to combine the selected panel in a single targeted multiplex PCR. Primer sequences were first designed for the *HBB* locus, using the accession number GCF_000001405.25 and the NCBI assembly GRCh37.p13, in FASTA format. A set of online tools, Primer3 and PrimerPlex2, were used to annotate a unique set of specific primers in order to reduce the possibility of cross-primer interactions in the multiplex PCR. Primers were selected such that individual amplicons overlapped to cover the entire 1,806 nucleotide sequence of the *HBB* gene. After the completion of the *HBB* primer design, NCBI SNP reference sequences (using the SNP ID) were used to design primers to amplify the selected SNPs, using the same set of tools. All of the individual primer sequences were screened through a combination of online tools (OligoCalc, SNPCheck3 and NCBI Primer Blast) in order to assess their self-complimentarity, SNP content and specificity, respectively. Primer pairs, which displayed a high degree of self-complimentarity, contained more than two SNPs within the sequence, or had potential for non-specific annealing to multiple genomic regions, were excluded from the pool and individually redesigned. In total, 24 primer pairs were designed and the sizes of individual target amplicons ranged between 183 and 394 base pairs (primer sequences are represented in Table 1.1).

Table 1.1: Details of the designed primer pairs Sequence information for the primers designed for the *HBB* locus, centromeric SNPs and telomeric SNPs and the expected sizes of target amplicons.

HBB gene

Amplicon Primer ID	Forward Primer	Reverse Primer	Amplicon Size
1	AGTCAGGGCAGAGCCATCTA	GTCTCCACATGCCCAGTTTC	229
2	CAAGACAGGTTTAAGGAGACCAA	ACTTAACCATAGAAAAGAAGGGG	391
3	ATGGGACGCTTGATGTTTTC	TGTACTAGGCAGACTGTGTAAAG	323
4	TGTGTATAACAAAAGGAAATATCTCTG	GCCCTGAAAGAAAGAGATTAGG	328
5	CACATATTGACCAAATCAGGGTA	TGCTATTGCCTTAACCCAGAA	200
6	ATGCCTCTTTGCACCATTCT	CCAGCCTTATCCCAACCATA	183
7	TCCAGCTACCATTCTGCTTTT	GGACTTAGGGAACAAAGGAACC	294
8	CTCGCTTTCTTGCTGTCCAA	ATGCACTGACCTCCCACATT	194

Centromeric SNPs

9	TGGCTGTTCTGTCATGTGTG	CAACCTCTCAAAATTCCCTTGG	288
10	ACCCAGGAATGAAGATCCCA	TCCCTTTCTCTTTCTCTTCCCT	268
11	TGAGTCTGAGGTGCCTATA	ATCTCCTACCTGCTCTGAA	363
12	TCCAAGAGTGTGATGAATAC	CCAGCCAAGAATGTGAAT	380
13	GAACAATGCCTAGAGACA	GAATGGTAATTGACAGAAGG	200
14	CTGACTTCTGATACTATGTC	GCTCATTGTATATTCCTACC	348
15	CTGCGTCTCCAGAATATG	TTGACACCACTGATTACC	394
16	TGGTTCTTCTATGGCTATCT	GTGAAACAGGGTCTTGAAA	322
17	TCACTGGGTCTTGATGTACAGA	GCCGAGCACACACAATTACT	308

Telomeric SNPs

18	TCTGTGATGCCTCCTTTG	TTCTCCAGTGGATTCTTG	189
19	GGATCTCAGTCACCAAGGCT	TGGAATCAACAAGCTAGGGGA	277
20	AATTGCTGGGATTACACATGC	CAACCCAAAGTAGAAACTATCAAGG	244
21	TGAAGCCATTTTTAGATAAACCAA	TGCATCTTGATGATTAGAATTGC	361

22	TGGTTCACCTTTCATTTGTTCA	TCAGACCCTTGTCTTACACCA	249
23	GGGACATGATAAGGGAGCCA	ATCTGCAGTGCTAGTCTCCC	244
24	AGACAACAGAGACAACTAAG	ACAGCTAATGCACATTGG	309

1.2.2 Validation of the targeted HBB panel

Gradient PCR

In order to determine the single most appropriate annealing temperature, at which all primers had maximum specificity to their targets, a gradient of six different temperatures were empirically tested in individual singleplex PCR amplifications (Table 1.2). Unless stated otherwise, each singleplex PCR throughout the study, was carried out with 0.5 µl of control genomic DNA (gDNA, Sigma Aldrich) in a total reaction volume of 15µl, which contained 12.49 µl of nuclease-free water (Roche Diagnostics), 1.5 µl of 10 x PCR buffer (containing 100 mmol/l Tris-HCl, pH 8.3, 500 mmol/l KCl, 15 mmol/l $MgCl_2$, 5Prime), 0.06 µl of 100 µM forward and reverse primer (Eurogentec), 0.3 µl of 10mM deoxynucleotide triphosphates (dNTPs, Thermo Scientific), and 0.09 µl of Hot Master Taq polymerase stock solution with Mg^{2+} (concentration at 5U/ µl, 5Prime). A master mix was prepared for every PCR (allowing an additional 10% volume for losses during pipetting) and individually aliquoted into 200 µl microcentrifuge tubes (Molecular Bioproducts) placed on a cold rack, in UV-irradiated laminar flow cabinet (AirClean 600), and stored in a dedicated clean room. The gradient PCRs took place in the Eppendorf MasterCycler (Eppendorf) and consisted of an initial denaturation step at 96°C for 15 min, followed by 35 cycles of denaturation at 94°C for 15 sec, annealing at six different temperatures for 15 sec, extension at 65°C for 45 sec, followed by a final extension step at 65° for 2 min (annotated as "Gradient Hot Master Taq" program). Success of amplification was assessed using the standard conditions for 1% agarose gel electrophoresis and the DNA bands were visualised after adding 10 µl of the GelRed® Nucleic

acid stain (Biotium) in the UV Transilluminator. Primers for targets, which did not successfully amplify were excluded from the subsequent analysis.

Table 1.2: Primer annealing temperatures tested in the gradient PCR.

Temperature	T1	T2	T3	T4	T5	T6
°C	53.0	54.5	56.7	59.0	60.9	61.8

Targeted multiplex amplification of genomic DNA

After determining the annealing temperature by gradient PCR, multiplex PCR was validated on control human genomic DNA (gDNA). Each reaction was carried out with 0.5 µl of gDNA (concentration 30 ng/µl) in a total reaction volume of 50µl, which contained 22 µl of nuclease-free water, 25 µl of 2 x Qiagen master mix (Qiagen), and 2.5 µl of primer mix (set to a concentration 2µM). Replicates of four, along with a negative control were used for amplifications performed with the conditions: initial denaturation step at 95°C for 15 min, followed by 15 cycles of denaturation at 94°C for 30 sec, annealing at 56°C for 90 sec, and extension at 72°C for 1 min, with a final extension step at 60° for 10 min (annotated as "Preamplification PCR" program).

It was important to minimise the number of 'preamplification cycles' since excessive amplification frequently produces biases when performing multiplex PCR, resulting in overamplification of some fragments and insufficient amplification of others. However, the small number of cycles used (15) produces too little DNA for visualisation on an agarose gel. Therefore, in order to assess the amplification success of the multiplex PCR, aliquots were taken after the 15 cycles had been completed and used for further PCR with individual pairs of primers (i.e. singleplex PCR), allowing amplification of each fragment to a level detectable on

an agarose gel. The second reaction set was run on the "Hot Master Taq" program, using the following conditions: initial denaturation step at 96°C for 15 min, followed by 35 cycles of denaturation at 94°C for 15 sec, annealing at 56°C for 15 sec, extension at 65°C for 45 sec, followed by a final extension step at 65° for 2 min.

Targeted multiplex amplification of single cells, clumps of cells and whole genome amplified DNA

The concentration of isolated gDNA used for the initial validation experiments ranged between 10-20 ng/µl. In embryo biopsy, the amount of isolated DNA from single blastomeres usually does not exceed 5-10 pg. To account for this difference and its potential impact on the multiplex PCR, DNA from single buccal cells and clumps of buccal cells (approximately 5) was amplified in the targeted PCR either directly or after being subjected to whole genome amplification using the MDA method (annotated as SC-DNA, 5C-DNA, SC-MDA and 5C-MDA, respectively). The SC-DNA can be considered equivalent to a blastomere biopsy sample in terms of the quantity of DNA, while the 5C-DNA are equivalent to a trophectoderm biopsy. The conditions used for the targeted PCR were identical to those used for gDNA as previously described.

Single-cell and clumps of cells isolation

Buccal cells were collected by rinsing the mouth with distilled nuclease-free water. Approximately 10 µl of sample solution were transferred into 5 µl droplet of washing buffer (prepared by dissolving 0.1 g of polyvinyl alcohol, PVA in 100 ml of phosphate-buffered saline, PBS, both Sigma Aldrich) on a Petri dish. Cells were washed in at least three drops of washing buffer, then transferred into 200 µl microcentrifuge tubes either individually or in groups of five, and stored at -80°C until DNA isolation took place. From the first half of the

samples, the DNA was isolated and amplified by the multiplex PCR. From the second half of the samples, the extracted DNA was first subjected to whole genome amplification by MDA and then further amplified for the selected loci in the multiplex PCR. This approach intended to compare the targeted PCR efficiency between the DNA from SC/5C and the MDA-amplified DNA from SC/5C.

Alkaline cell lysis and DNA extraction from single cells / clumps of cells

To the tubes containing a single cell or group of five cells 1.5 µl of alkaline lysis buffer (containing 50 mM Dithiothreitol, DTT and 200 mM sodium hydroxide, NaOH, both Sigma) was added. Samples, positive and negative controls were briefly vortexed, centrifuged, and then incubated for 10 min at 65°C in the thermocycler with heated lid. Multiplex PCR was carried out with the 2.5 µl of the isolated DNA and lysis buffer mixture in a total reaction volume of 15µl, which contained 2.89 µl of nuclease-free water, 7.5 µl of 2 x Qiagen master mix, 0.11 µl of primer mix, and 1.5 µl of 0.4 M tricine (Sigma Aldrich). Tubes were placed in the thermocycler and the DNA amplified using the "Preamplification PCR" program.

Multiple displacement amplification (MDA) of single cells / clumps of cells

The second half of the samples containing the single cell and the group of five cells were whole genome-amplified using REPLI-g Single Cell Kit (Qiagen) MDA protocol according to the manufacturer's instruction. MDA products were subsequently subjected to the targeted PCR, using the same reaction set-up described for gDNA and the "Preamplification PCR" program. To assess the success of the multiplex PCR design in single cells on an agarose gel, the preamplification products from SC- and 5C-DNA and MDA-amplified DNA from SC- and 5C were further amplified in a new singleplex PCR, using the "Hot Master Taq" program.

CLINICAL PHASE

1.2.3 Patient selection, IVF and embryo biopsy

Before clinical implementation of the protocol on embryo biopsies, the technique was validated on genomic DNA obtained from five family trios (each composed of the mother, father and a child or prenatal sample) and two couples, together carrying 12 different *HBB* mutations. The gDNA was amplified using the "Multiplex PCR" program, set to 35 amplification cycles, using the same conditions as the "Preamplification PCR" program. An attempt was made to identify informative SNPs in each family member in addition to confirming the carrier genotypes. Table 1.3 refers to the list of mutations present in these families (available from the genetic reports obtained from the medical geneticist). After protocol validation, three of the families, all healthy carriers of β-thalassaemia, used the test clinically for the purpose of PGT-M. In all three cases, the patients underwent ovarian stimulation, and oocytes were collected and fertilized using intracytoplasmic sperm injection. The resulting embryos were biopsied and vitrified at cleavage or blastocyst stages. In two cases, no other family members were available for testing, and one couple had a previous affected pregnancy; a sample of amniotic fluid was included in the study along with the DNA samples extracted from the parents. Analysis of this additional sample allowed the phase of linked polymorphisms to be determined, i.e. revealed which alleles were associated with parental mutations. All patients underwent PGT for β-thalassaemia in different IVF centres and gave consent for NGS-based PGT-M to be carried out in parallel with a validated SNP-array technique (i.e. Karyomapping) (Ben-Nagi et al., 2017; Giménez et al., 2015; Konstantinidis et al., 2015; S. a. Natesan et al., 2014). DNA samples obtained from parents and embryo biopsies were tested at Reprogenetics UK. The

diagnosis of β-thalassaemia in embryos has been previously licensed by the Human Fertilisation and Embryology Authority. The study was approved by Aspire IRB on 26 August 2015 (reference number PGSP-2015) (Kubikova et al., 2018).

Table 1.3: Summary of the mutations present in families affected by β-thalassemia. Genomic DNA samples from the seven families contained mutations at different positions of the *HBB* gene. The exact mutation positions were identified after entering the mutation annotations into HbVar: A database of Human Hemoglobin Variants and Thalassemias.

Member of the family		Genomic locus	Mutation	Disease status
Family 1	Male	5248235	HBB:c.17_18delCT	Carrier
	Female	5247987	HBB:c.135delC	Carrier
	Son	N/A	No	Unaffected
Family 2	Male	5248155	HBB:c.92+5G>C	Carrier
	Female	5247993	HBB:c.126_129delCTTT	Carrier
	Son	5248155/5247993	c.92+5G>C/ c.126_129delCTTT	Affected
Family 3	Male	5248050	HBB:c.93-21G>A	Carrier
	Female	5248160	HBB:c.92G>A	Carrier
	Daughter	5248050	c.93-21G>A	Carrier
Family 4	Male	5247062	HBB:c.316-106C>G	Carrier
	Female	5248004	HBB: GLN39TER	Carrier
	Daughter	5247062/5248004	c.316-106C>G/ GLN39TER	Affected
Family 5	Male	5248155	HBB:c.92+5G>C	Carrier
	Female	5248224	HBB:c.27_28insG	Carrier
	Prenatal Sample	5248155/5248224		
Family 6	Male	5248004	HBB:c.118C>T	Carrier
	Female	5248154	HBB:c.92+6T>C	Carrier
Family 7	Male	5247062	HBB:c.316-106C>G	Carrier
	Female	5248050	HBB:c.93-21G>A	Carrier

1.2.4 Parental DNA extraction and whole-genome amplification of blastomere and trophectoderm biopsies

The genomic DNA was extracted from 4 ml parental blood and from the amniotic fluid sample using the QIAamp DNA Blood Mini Kit (Qiagen) using the standard protocols recommended by the manufacturer. Extracted DNA and cell(s) obtained from embryo biopsies were lysed and subjected to whole-genome amplification using a REPLI-g Single Cell kit (Qiagen) according to the manufacturer's instructions prior to multiplex PCR as described in the previous section (Kubikova et al., 2018).

1.2.5 DNA library preparation, next generation sequencing and data analysis

The processing of samples for NGS followed the steps outlined in Figure 1.5.

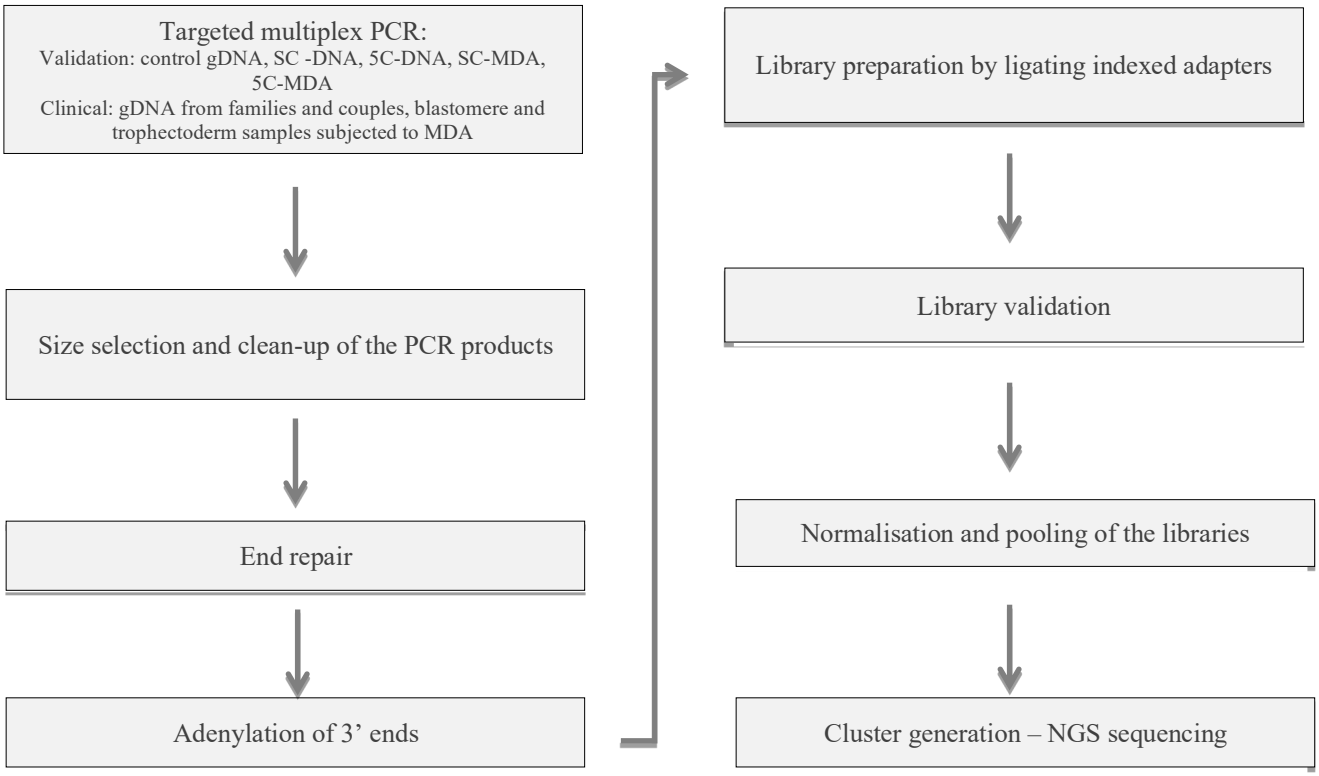


Figure 1.5: NGS work-up. Flow diagram describing the order of steps included in the processing of DNA samples for NGS and the preparation of DNA sequencing libraries using the Illumina TruSeq PCR-Free Library Prep kit (Illumina).

Size selection and clean-up of the PCR products

To remove excess primer-dimers from the multiplex PCR products, the samples were mixed with AmPure XP purification magnetic beads (Beckman Coulter) in a 1:1.6 ratio in a 96 deep-well plate. Samples were pipetted up and down approximately 10 times and incubated for 5 min at room temperature (RT), then transferred onto a magnetic stand. Using this ratio, DNA fragments sized less than 100 bp were eliminated in supernatants that were discarded after the liquid had cleared (clearing of liquid indicated the collection of beads against the wall of the tube in contact with the magnet). The beads were then washed twice with 80% ethanol before adding 52.5 µl of the resuspension buffer (RSB) into each sample well. The plate was removed from the magnetic stand, and the DNA was eluted from the beads into the RSB by gentle pipetting up and down. Following the incubation for 2 min at RT, the plate was placed back on the magnetic stand. When the liquid cleared, 50 µl of the supernatants were transferred into new 200 µl PCR tubes.

End repair

DNA samples were incubated at 30°C for 30 min with 40 µl of End Repair Mix (ERM) in the thermocycler with heated lid. ERM contained 3′–>5′ exonuclease and polymerase enzymes, which generated blunt-ended fragments and added 5′-phosphate groups needed for downstream ligation of the sequencing adapters. Following the incubation, the excess ERM was removed from by performing a second clean-up with the AmPure XP beads using the 1: 1.6 ratio, using the steps described earlier. The DNA was eluted into 15 µl of fresh RSB and transferred from the plate into new 200 µl PCR tubes.

Adenylation of 3′ ends and library preparation

To the samples, 12.5 µl of A-tailing Mix (ATL) and 2.5 µl of RSB were added. After gentle mixing, tubes were placed in the thermocycler with heated lid and incubated at 37°C for 30 min, followed by 70°C for 5 min and 4°C for 5 min. ATL makes fragments compatible with adapters and prevents self-ligation by adding a 3′-A overhang required for the subsequent step. DNA libraries were prepared by adding to each sample tube 2.5 µl of Ligation Mix (LIG) and 2.5 µl of Illumina TruSeq indexed adapters, containing sequences 5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC 3' and 3'GTTCGTCTTCTGCCGTATGCTCTA-index-CACTGACCTCAAGTCTGCACACGAGAAGGCTAG 5'. The last 12 bases of the adapter sequences were complimentary to each other, which allowed them to anneal and form a forked structure after denaturation. While one adapter sequence (known as the universal adapter) always contained the same nucleotides, the second adapter contained a unique index of six nucleotides, which allowed different samples to be distinguished from each other after pooling of multiple samples into a single DNA library and sequencing on a single flow cell (a process known as multiplexing). The TruSeq DNA Sample Prep Kit used in this study provided 24 indexed adapters with distinct index sequences, allowing 24 DNA libraries to be prepared, pooled and sequenced together. The ligation of the adapter to the 3′-A tailed DNA ends was facilitated by DNA ligase, present in the LIG enzyme mix. Excess DNA adapters were removed by performing another two rounds of clean-up with the AmPure XP beads using the 1: 1.6 ratio, as described earlier.

DNA library validation

Indexed DNA libraries were validated for size using the Agilent High Sensitivity DNA kit (Agilent Technologies) according to the manufacturer's instructions on the on the Agilent 2100 Bioanalyzer instrument, employing a High Sensitivity DNA chip. The analysis was carried out using the software provided by the manufacturer. DNA concentration of the prepared libraries was quantified using the Invitrogen Qubit Fluorometer on the Qubit dsDNA assay (both Life Technologies). The average insert library size of 360 bp was then used to convert the obtained ng/µl concentration into molar concentration in nM according to the formula:

$$\frac{\text{concentration in ng/µl} * 10^6}{660 \text{ g/mol} * 360} = \text{nM}$$

Normalisation, pooling and sequencing of the DNA libraries

Each quantified library was normalised with the RSB to the concentration of 4nM. Five µl of each library was then pooled together to result in a 24-sample indexed DNA library. Just before sample loading, 5 µl of the library were denatured with 5 µl of freshly diluted 0.2 M NaOH and further diluted to 8 pM with the HT1 buffer (both Illumina). Finally, 90 µl of 8 pM PhiX control was added to 510 µl of the denatured library to comprise 15% of the total volume as a quality control for cluster generation during the sequencing run. The library was kept on ice until loaded onto the Illumina MiSeq Genome Sequencer and sequenced on a V2 MiSeq Nano flow cell (Illumina).

NGS data acquisition, analysis, and single gene diagnosis

Sequence alignment files for all indexed samples were obtained from the Illumina MiSeq Reporter software in Sequence Alignment/Map format (SAM) and were converted to binary version of SAM (BAM). The BAM files were loaded into the Integrative Genomic Viewer (IGV, Broad Institute) to visualise the sequence alignment using the h19 human genome assembly as a reference for the alignment and the analysis. Regions of interest were inspected after creating tracks by importing all the SNP and mutation exact positions into the software. On the basis of the total number of reads for that position, the percentage of nucleotides that did not correspond to the reference nucleotide was determined. In the case of SC/5C DNA and SC/5C MDA, the alignments were screened for heterozygous loci and compared to reference gDNA, in order to assess ADO and informativity of selected SNPs. For mutations in the β-thalassemia affected families and for PGT-M of embryos, the aligned sequences were screened for a genotype call at every mutation position in each family. The sequences obtained from embryonic DNA were screened for heterozygous loci and compared with the parental DNA to assess allele dropout (ADO) and informativity of selected SNPs. For both categories, the analysis included the determination of sequence coverage at positions of interest and the identification of unexpected variants and polymorphisms within the sequenced DNA, which could potentially serve as additional linkage markers (Kubikova et al., 2018).

1.3 RESULTS

1.3.1 Gradient PCR

Of the 24 primer pairs tested for the suitable annealing temperature in the gradient PCR using the range of six different temperatures, with four primer pairs (10, 14, 22, and 23) no products were detected at any temperature. Those primers were excluded from pooling (Table 1.4, highlighted in dark grey). For the remaining part, specific products were detected, and the majority of amplicons were generated at the temperature range of 54.5-56.7°C. A representative image of the gel showing the size separation of PCR products after the gradient amplifications with primers 7-12 is presented in Figure 1.6. Whenever multiple or no products were detected on the gel at a particular temperature, that temperature was considered unsuitable. Based on this approach, the primer annealing temperature of 56°C was selected as the one where most PCRs generated a specific product.

Table 1.4: Validation of primers by gradient PCR: PCR products were run on agarose gel and specific products detected with each designed primer pair at the corresponding annealing temperature are marked by x.

	ANNEALING TEMPERATURE (°C)					
PRIMER PAIR	53	54.5	56.7	59	60.9	61.8
1			x	x	x	x
2	x	x	x	x		
3		x	x	x	x	x
4	x	x	x			
5	x	x	x	x	x	x
6	x	x	x	x	x	x
7			x	x	x	x
8	x	x	x	x	x	x
9	x	x	x	x	x	x
10						
11	x	x	x	x	x	x
12	x	x	x			
13	x	x	x			
14						
15	x	x	x			
16	x	x	x	x	x	
17	x	x	x	x	x	x
18	x	x	x	x		
19	x	x	x	x	x	x
20		x	x	x	x	x
21	x	x	x	x	x	x
22						
23						
24	x	x	x	x	x	

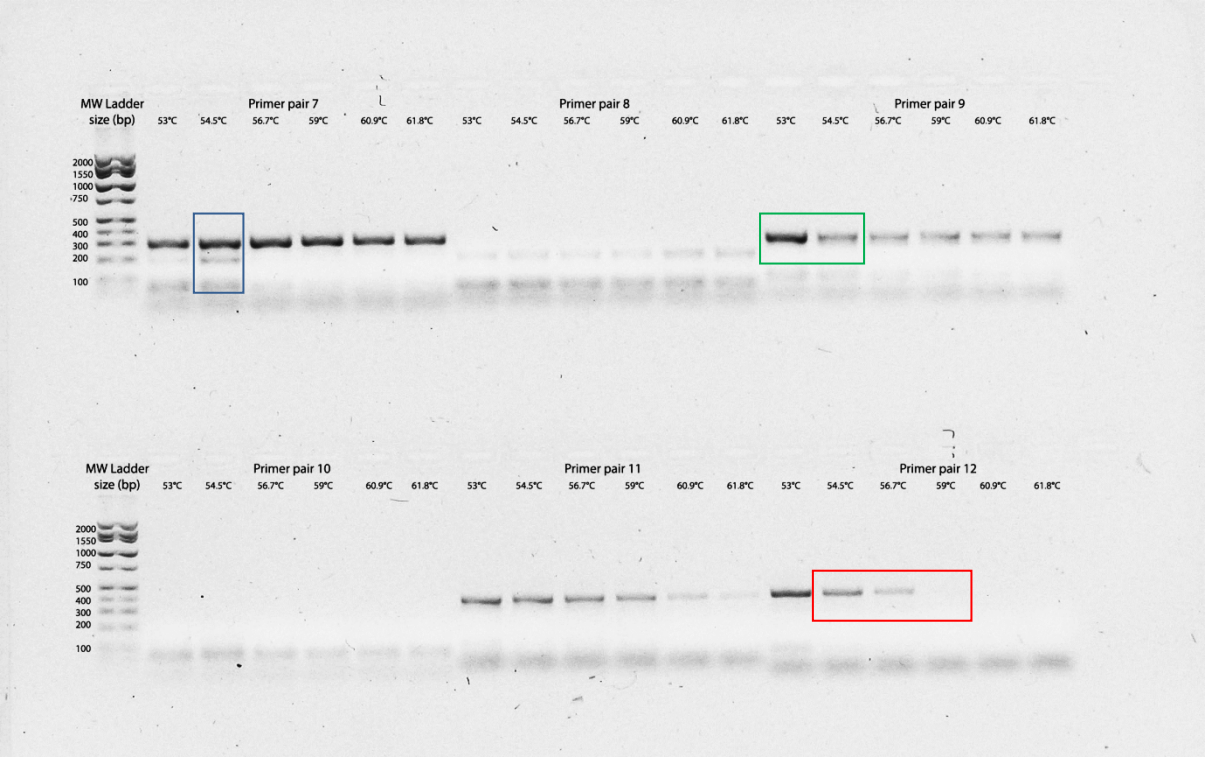


Figure 1.6: Detection of gradient PCR products by agarose gel electrophoresis. A representative image of the gel following amplifications. Five µl of the products were mixed with 6x DNA gel loading dye (Life Technologies) in a 3:1 ratio and loaded on 1% agarose gel, prepared from solid agarose (Invitrogen Life Technologies) and 1 x Tris borate EDTA buffer (Thermo Scientific), containing 10µl of Gel Red nucleic acid stain (Biotium). The analysis was carried out in an electrophoresis apparatus (Biorad) with the following conditions: 35 min at 80V. DNA bands were visualised and compared with 100bp molecular weight (MW) DNA ladder (Life Technologies) in the Benchtop UV Transilluminator (UVP). Highlighted in blue are bands of specific and non-specific products obtained with annealing temperature 54.5°C with primers 7. Highlighted in green are specific products obtained with annealing temperature 53 and 54.5°C with primers 9. Highlighted in red are faint bands obtained with annealing temperature at 54.5 and 56.7 °C and no product amplification at temperature 59°C with primers 12. Numbering of the fragments correspond to the numbering of primer pair that was used in the singleplex PCR.

1.3.2 Multiplex PCR on gDNA

Following a series of validations by a gradient PCR, control gDNA was amplified in the targeted multiplex PCR using the pool of primers designed to cover the sequence of the *HBB* gene and flanking SNPs in replicates of four and one negative control. After 15 cycles of amplification, each fragment was individually further amplified in a new singleplex PCR. The results showed that multiplex PCR was successful in generating all target amplicons (Figure 1.7).

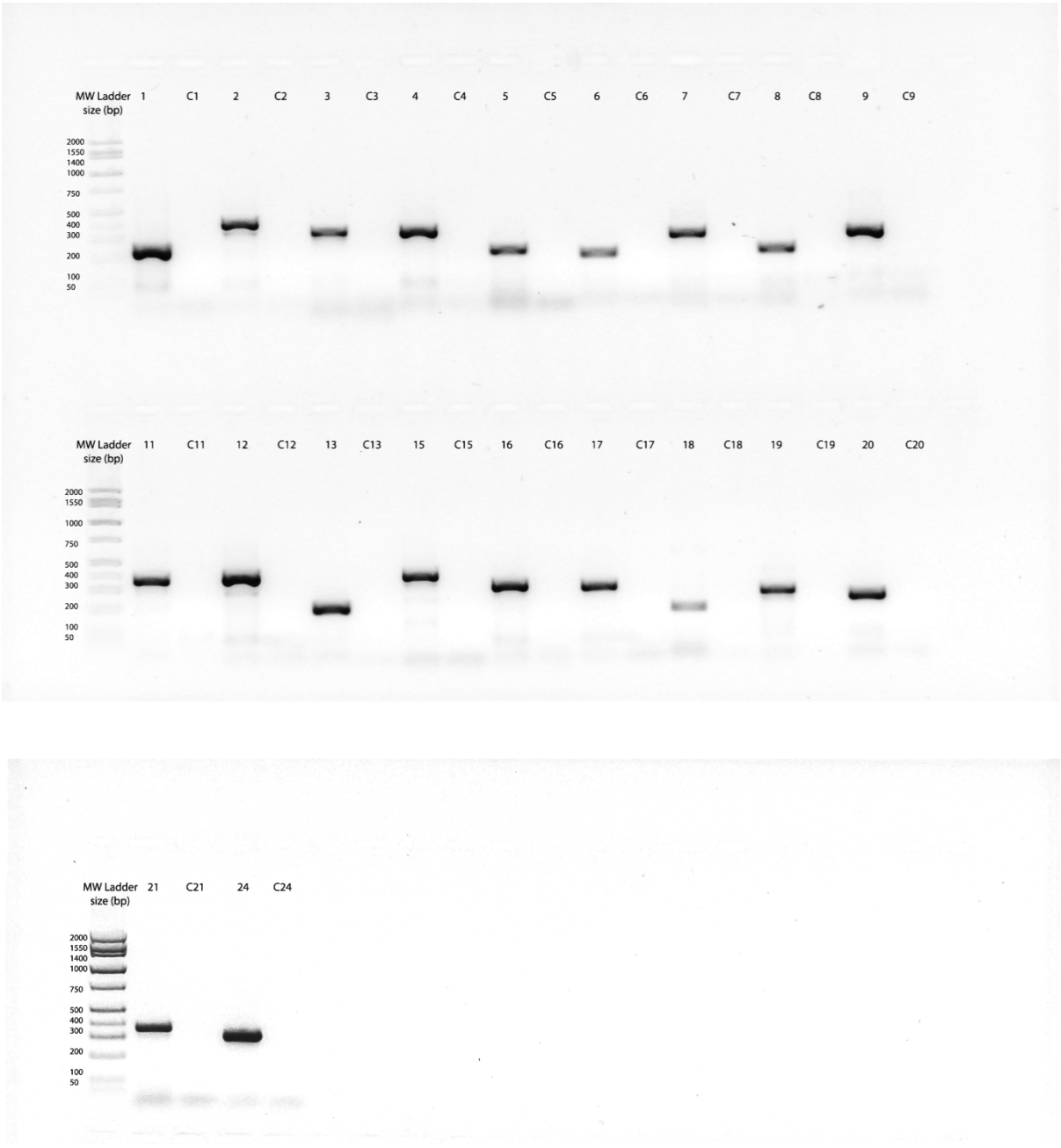


Figure 1.7: Detection of multiplex PCR products from gDNA by agarose gel electrophoresis. A summary image of the gel comparing the sizes of all target amplicons obtained from the multiplex PCR and further amplified in singleplex PCRs along with the negative control. The analysis was carried out as described in the previous figure. Numbering of the fragments correspond to the numbering of primer pair that was used in the singleplex PCR with their respective negative controls (labelled as C#).

1.3.3 Targeted amplification on single cells and clumps of cells with/without MDA

Following successful generation of all target amplicons in the validation reaction using the control gDNA, targeted amplification was attempted using samples composed of single cells and the clumps of cells directly or after MDA. The 15-cycle targeted multiplex PCR was shown to be overall successful on SC- and 5C-DNA samples previously subjected to MDA. Almost all of the 20 fragments were detected on the agarose gel. A representative gel image showing the amplification of fragment 9 after MDA from SC- and 5C-DNA is shown in Figure 1.8 A under columns MDA-SC and MDA-5C. Two exceptions were fragments 5 and 11, where no amplification was detected from SC-MDA (shown in Figure 1.8 B and C, highlighted in blue). The amplification failure of these fragments likely resulted from the insufficient MDA. Although MDA promises to amplify up to 99% of the entire genome, it has been reported that the coverage may be reduced at a single cell level to ~80% (Spits et al., 2006). Sample degradation and DNA fragmentation likely contribute to the lower rates of successful amplification observed in single cell samples compared to the clumps of cells.

The analysis of SC- and 5C-DNA amplified in the targeted panel revealed that the majority of fragments successfully amplified (an example is shown in Figure 1.8 A, under columns SC and 5C). The exceptions were amplicons 15, 12 and 21 (Figure 1.8 D, E and F respectively, shown in red rectangles). Furthermore, few amplicons generated bands of lower intensity, while others produced bands of higher intensity (for comparison of band intensity pattern, see Figure 1.8 A-F, columns SC and 5C). Although even amplicon amplification is ideal, the observed results were expected as it is nearly impossible for all individual reactions to occur at equal

efficiency. Preferential amplification occurs frequently in multiplex PCR set-ups and is likely more common as the number of amplification cycles increases. This validation methodology revealed that although a few fragments were affected by uneven amplification and some had suffered from amplification failure, the majority did amplify to a sufficient level to be detected on the agarose gel.

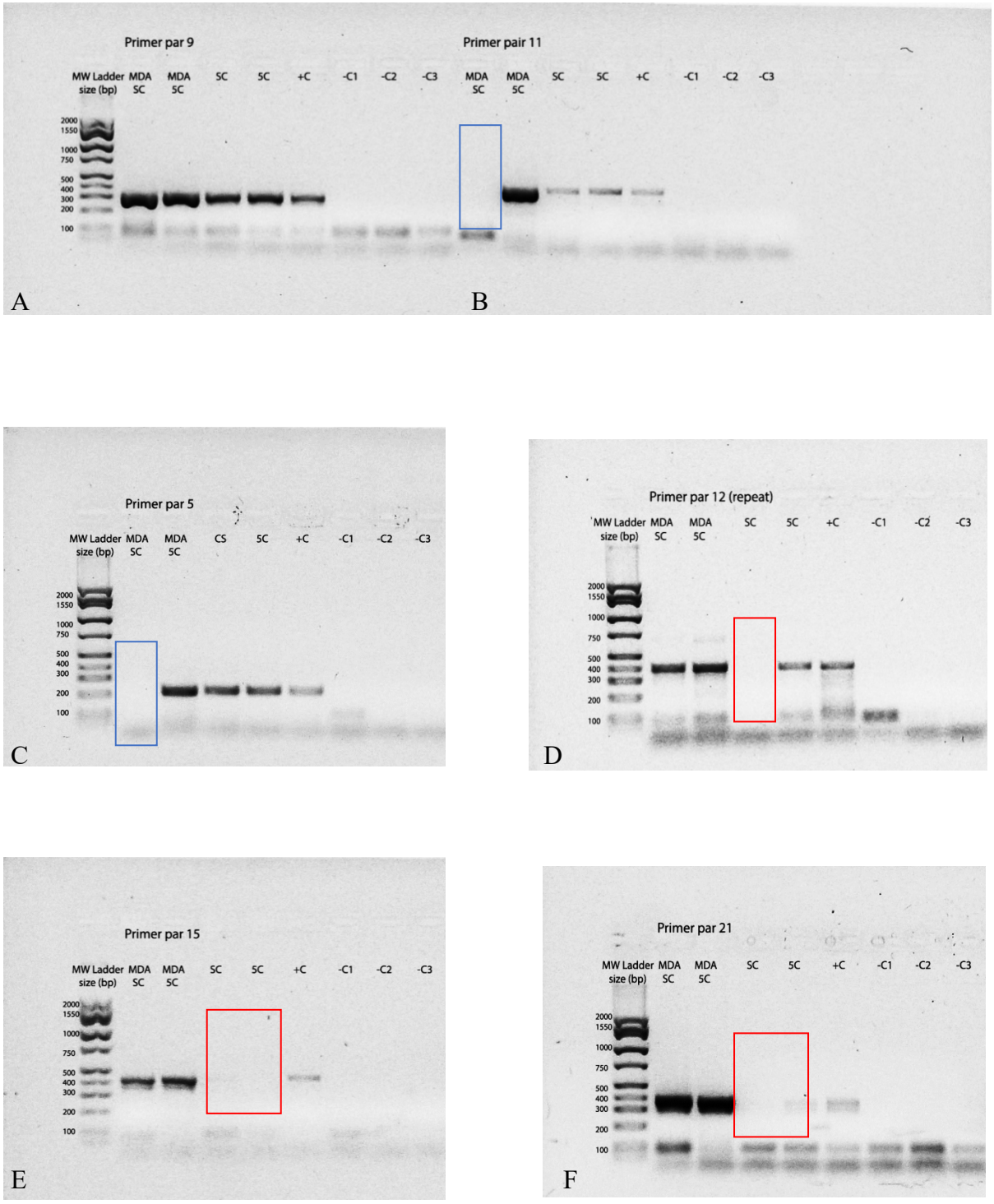


Figure 1.8: Detection of multiplex PCR products from SC- and 5C-DNA and SC- and 5C MDA by agarose gel electrophoresis. A gel image showing the size separation of target amplicons obtained from the multiplex PCR further amplified in individual singleplex PCRs along with positive (+C) and negative controls (-C3). The analysis was carried out as described previously. –C1 was a negative control of SC/5C MDA multiplex PCR mastermix, -C2 was a negative control of SC/5C multiplex PCR mastermix. **A-F:** Amplification of fragments 9, 11, 5, and 21 in SC-MDA, 5C-MDA, SC-DNA and 5C-DNA, respectively. Highlighted in blue is failed amplification of fragments 11 and 5 **(B-C)** in MDA. Highlighted in red is failed amplification of fragments 12, 15 and 21 **(D-F)** in SC and 5C DNA. Analysis for fragment 12 had to be repeated because of contamination.

1.3.4 Validation of the indexed DNA libraries

Indexed DNA libraries were validated for fragment size and success of adapter ligation using the High Sensitivity DNA assay (Agilent). Fragment size analysis revealed the size of the majority of amplicons ranged between 200 and 1200 bp (represented as bands on the Figure 1.9). The target amplicons appeared to be present in the bands within the lower range (between 200-500 bp), assuming the amplicon size range of 183 to 394 bp prior to adapter ligation. After the ligation steps (adding a 65bp adapter), the average amplicon size reached approximately 360 bp, with smaller fragments reaching approximately 250-300 bp and larger fragments reaching approximately 450 bp. A comparison of peaks representing the different amplicons between SC-DNA and gDNA is shown in Figure 1.9 B and C. The absence of fragments smaller than 100 bp suggests the initial clean up with AmPure XP beads was successful. However, the presence of peaks of higher molecular weight than expected indicates the presence of additional fragments, possibly a residual WGA template that was carried over from WGA into the targeted PCR. The fragmented DNA as the larger peaks were more apparent in gDNA sample compared to SC-DNA (Figure 1.9 circled in red).

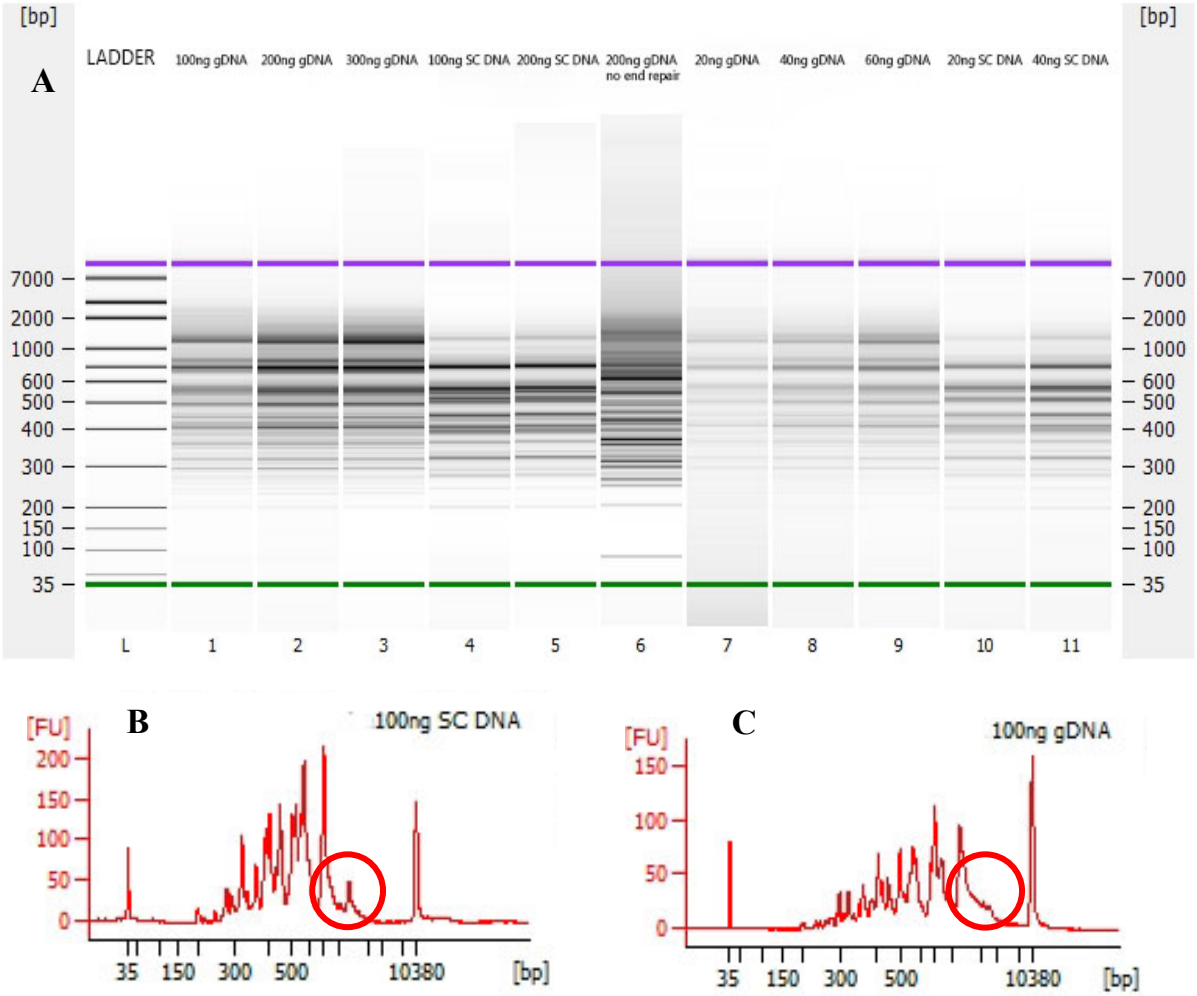


Figure 1.9: Size separation of indexed DNA libraries. A) Size separation of fragments present in the indexed DNA libraries from gDNA and SC DNA loaded at 20, 40, 60, 100, 200 and 300 ng/µl concentration using the High Sensitivity DNA assay (Agilent). **B-C:** Electrograms showing the ligated fragments present in 100ng of SC DNA library **(B)** and 100ng of gDNA library **(C)** loaded onto the High Sensitivity DNA chip. Horizontal axis represents fragment size in base pairs (bp) and vertical axis represents fluorescence units (FU). Individual fragments are represented as peaks. The two circles compare the presence of larger peaks/background DNA between SC DNA and gDNA. First and the last peaks represent MW marker of 35bp and 10380bp, respectively.

The second library validation method was intended for the evaluation of how successful the end repair step was in generating blunt-ended fragments required for downstream adapters ligation. A gDNA library sample subjected to the end repair was run on the High Sensitivity DNA assay simultaneously with the gDNA library without repaired ends. The analysis revealed the excess of adapters in the non-end-repaired DNA, as suggested by the peak at approximately 70 bp, not observed in the end-repaired DNA (Figure 1.10 A and B). The adapters failed to ligate to the amplicons presumably as a consequence of unrepaired overhanging ends and absent terminal 5′-phosphate groups needed for the downstream ligation. Furthermore, a significant shift in the amplicon sizes as well as the absence of significant background at higher molecular weight range, observed in the end-repaired library, is suggestive of failed adapter ligation in the non-end-repaired DNA.

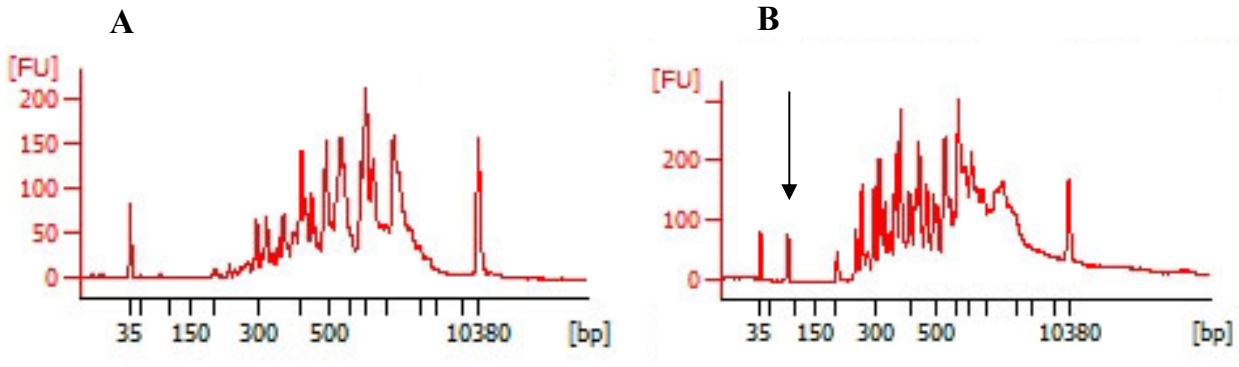


Figure 1.10: Validation of indexed DNA libraries for success of ligation. Electrograms showing the ligated fragments present in 200ng of gDNA library **(A)** and 200ng of gDNA library without repaired ends **(B)** loaded onto the High Sensitivity DNA chip. Horizontal axis represents fragment size in base pairs (bp) and vertical axis represents fluorescence units (FU). Individual fragments are represented as peaks. The arrow highlights the presence of excess unligated adapters in gDNA library without repaired ends. First and the last peaks represent MW marker of 35bp and 10380bp, respectively.

1.3.5 Validation of the protocol on DNA samples from single cells and clumps of cells after targeted PCR with/without MDA – sequencing and data analysis

Sequence coverage analysis

Of the 20 target amplicons, covering the entire *HBB* gene sequence and 18 closely linked extragenic SNPs, complete sequence information was obtained for all samples with exception of one SC-DNA after NGS. Mapping of the *HBB* locus occurred in eight overlapping amplicons (1-8) and was accomplished for all samples. The lowest read depth for a single amplicon (only 5x) was encountered with single cell DNA samples in amplicon 4, while the highest coverage was achieved in amplicons 1-3, reaching 1700 reads in 5C DNA, suggesting that some fragments amplified more efficiently and preferentially over others. This was in line with expectations and did not present a significant concern as long as full gene sequence had been obtained at sufficient coverage in most samples. The NGS protocol generated a mean coverage of β-globin sequences of 888x (1061x for single cells samples and 716x for clumps of cells), which was equivalent to the coverage obtained from the single cells and clumps of cells subjected to MDA before targeted PCR (998x for single cells and 1013x for clumps of cells). For clinical diagnosis, a 30x coverage threshold is usually recommended to ensure adequate representation of germline mutations as well as the effective genome coverage and sequence reproducibility in whole genome sequence data (Telenti et al., 2016). Although there was an incidence of one fragment sequenced with only 4 reads, this only occurred in that one fragment and only in one SC-DNA sample, perhaps as a result of poor sample quality and poor subsequent amplification, a notion supported by the observation that certain SNP loci also had fewer reads in this sample.

For SNPs mapped outside of the β-globin gene sequence, the detection of single nucleotide base variants was obtained for all samples except for the amplicon 14. Calculating the average read depth at variant positions for each sample, the highest coverage (112 reads/base variant) was observed in MDA-DNA from SC, followed by 5C-DNA, SC-DNA and 5C-MDA-DNA with the corresponding values of 83, 84, and 70 reads/base variant, respectively (Table 1.5). SC-DNA 1 was an outlier with only 35 read depth on average (and only 4 sequences in one of the *HBB* fragments) and total amplification failure observed in all fragments except for fragment 9 and 11. This was suspected to be due to insufficient amplification, and as such the SC-DNA 1 was excluded from further analysis. The average read depth at SNP positions outside the β-globin gene sequence was considerably lower compared to that obtained from the β-globin sequence, likely due to some amplicons that were highly overrepresented in the gene sequence compared to others (e.g. amplicons 1-3 have been consistently over-amplified and over-sequenced).

Table 1.5: Mean x coverage at SNP positions.

Sample	Average read depth	Sample	Average read depth
control gDNA 1	70	5C DNA 2	80
control gDNA 2	64	5C DNA 3	124
SC DNA 1	35	SC MDA 1	112
SC DNA 2	121	SC MDA 2	113
SC DNA 3	95	5C MDA 1	70
5C DNA 1	75	5C MDA 2	66

Overrepresentation of amplicon sequences

Within the *HBB* locus, the sequences at each end of some amplicons had higher coverage than central portions, presumably because they were in overlapping regions of adjacent PCR fragments, however this difference was not very marked. However, there was a substantial difference observed in sequence coverage between individual fragments. As the majority of fragments had similar read depth values between different samples, this bias may have resulted from overrepresentation of some amplicons after PCR amplification (amplicon 4). Few amplicons (15, 16, 18 and 21) were poorly represented in all samples with average depth ~10 reads (Table 12). These fragments were most likely less abundant in the PCR products prior to sequencing, which suggests the multiplex PCR system should be optimised further so that all fragments are amplified at comparable efficiency.

Allele drop-out and NGS sensitivity of sequence variant calls

Sequencing multiplex PCR products from SC-DNA, MDA-DNA and reference gDNA was intended to provide an evaluation of how accurate and sensitive the NGS approach may be in generating correct genotypic calls from biopsied embryos. As such, the inspection of those genotypes had to be confined to heterozygous sites as homozygous loci are not informative for problems such as allele dropout. Variants/polymorphisms detected as heterozygous in reference gDNA were analysed for concordance with genotypic calls obtained from SC/5C DNA samples and SC/5C MDA-DNA samples. The results of the analysis fall into two different categories, depending on sequence coverage (Table 12). In total, sequence information was obtained for 24 SNPs (excluding the rs968856 amplified in fragment 14) in all samples except for SC-DNA 1, in which the rs3813727, rs4910543, rs4910735 and rs4910544 were the only variants sequenced (the sample that had amplified insufficiently as described above). Of the 24 SNPs, 18 were known extragenic single base variants flanking the *HBB* locus and six were additionally identified intragenic SNPs. Fifty per cent of the extragenic SNPs had coverage of over 30x after incorporating values obtained from all of the samples and 50% had coverage of less than 30x (the diagnostic threshold). All of the intragenic SNPs were sufficiently covered, with the minimum value of 42x (Table 1.6).

Table 1.6: High and low coverage of SNPs. SNPs in amplicons 9-24 were extragenic and the target variants, for which the primers were initially designed. SNPs 1-6 were intragenic and additionally identified after examining sequences within the *HBB* locus. Read depth values represent an average of all obtained values incorporating the data from every sample category tested (gDNA, SC DNA, 5C DNA, SC MDA-DNA, 5C MDA-DNA). N/A: SNP ID not available as SNP was previously unreported.

High coverage			Low coverage		
Amplicon	Extragenic SNPs	Average read depth	Amplicon	Extragenic SNPs	Average read depth
9	rs3813727	98	14	rs968856	X
11	rs4910543	73	15	rs11036364	16
11	rs4910735	81	16	rs7936823	9
11	rs4910544	81	17	rs6578588	19
12	rs4910736	110	18	rs7945118	12
13	rs2105819	295	21	rs10837628	9
19	rs34220818	122	21	rs12364872	10
20	rs10837620	187	21	rs10837626	16
24	rs10837631	100	24	rs7110263	25
	Intragenic SNPs	Average read depth			
1	rs1609812	246			
2	N/A	243			
3	rs7480526	47			
4	N/A	42			
5	rs10768683	145			
6	rs713040	76			

Genotype analysis in amplicons that reached the diagnostic threshold 30x

Since the accurate analysis of genotypes requires a minimum of 30x coverage, the analysis of heterozygous loci was limited to those that reached this threshold. Of the 15 SNPs located within the *HBB* gene, heterozygosity was detected in three, at positions 5247206, 5247752 and

5247791. These displayed 100% concordance between the variant calls of all samples tested (Table 1.7). A genotypic imbalance was noticed with some alleles being sequenced more often than others, particularly in SC-DNA samples. This is likely due to the very low amount of DNA retrieved from a single cell coupled with preferential amplification. Heterozygous genotypes were further detected in two extragenic SNPs (rs3813727 and rs34220818, located at positions 5255912 and 5236851, respectively) where significantly fewer alleles were affected by genotypic imbalance on average except for samples SC-MDA 1, SC-MDA 2 and 5C-MDA 1 where one allele was sequenced in >90% of all reads (Table 1.9, medium and darker blue). This suggests an extreme case of preferential amplification in the case of 96/3 and 91/9 per cent ratios, and an ADO in the case of the incorrect genotype call at variant rs3813727 in SC-MDA 2, where heterozygous site was called as homozygous one. One case of heterozygosity was also detected in the SC-DNA 1 but the sample was removed from the analysis due to total amplification failure (TAF) in the rest of the of the tested fragments.

Table 1.7: Distribution of genotypes detected in well-covered SNPs where heterozygosity was detected. Darker blue are cases of extreme overrepresentation of one allele (>90%). Darkest blue is the indication of ADO. TAF: total amplification failure. N/A: SNP ID not available as SNP was previously unreported. Read depth = sequence coverage.

SNP ID	N/A	N/A	rs10768683	rs3813727	rs34220818
Position	5247206	5247752	5247791	5255912	5236851
Sample	HBB locus	HBB locus	HBB locus	centromeric	telomeric
control gDNA 1	A=39%	A=33%	C=87%	A=54%	C=46%
	G-61%	T=44%	G=10%	G=45%	T=54%
read depth	154	27	86	85	114
control gDNA 2	A=45%	A=44%	C=90%	A=61%	C=52%
	G=55%	T=53%	G=10%	G=39%	T=48%
read depth	166	32	88	71	94
SC DNA 1				A=53%	
				G=47%	
read depth	TAF	TAF	TAF	98	TAF
SC DNA 2	A=40%	A=90%	C=82%	A=41%	C=28%
	G=6-%	G=10%	G=17%	G=59%	T=72%
read depth	354	31	175	136	184
SC DNA 3	A=34%	A=90%	C=77%	A=84%	C=49%
	G-65%	T=10%	G=23%	G=18%	T=51%
read depth	338	30	164	111	141
5C DNA 1	A=20%	A=94%	C=85%	A=44%	C=53%
	G=78%	T=3%	G=15%	G=56%	T=47%
read depth	184	36	155	77	99
5C DNA 2	A=34%	A=89%	C=76%	A=69%	C=60%
	G=65%	T=11%	G=24%	G=39%	T=40%
read depth	196	27	127	80	83
5C DNA 3	A=20%	A=97%	C=76%	A=52%	C=36%
	G=80%	T=1%	G=24%	G=48%	T=64%
read depth	398	70	255	126	154
SC MDA 1	A=46%	A=49%	C=84%	A=3%	C=61%
	G=54%	T=44%	G=14%	G=96%	T=39%
read depth	337	82	180	124	82
SC MDA 2	A=44%	A=60%	C=76%	A=0%	C=37%
	G=56%	T=31%	G=24%	G=100%	T=33
read depth	306	55	157	107	184
5C MDA 1	A=44%	A=41%	C=82%	A=91%	C-11%
	G-=56%	T=54%	G=17%	G=9%	T=89%
read depth	133	39	99	85	88
5C MDA 2	A=46%	A=69%	C=77%	A=69%	C=40%
	G=54%	T=31%	G=23%	G=31%	T=60%
read depth	107	32	109	78	122

1.3.6 Validation of the protocol on families with β-thalassemia mutations

Genotyping of the familial β-thalassemia mutations

Twelve different mutations were present in the β-globin gene of the seven families, whose gDNA samples were included in the study. All 14 parents were heterozygous carriers of different mutations between the man and the woman and underwent preconception genetic counselling. The complete sequence of the *HBB* gene was successfully obtained for all samples tested with the majority of fragments read on average ~120x (and all individual fragments had more than 30x coverage). Parental genotypes were confirmed in each family along with the mutation status of their existing child and in the one amniotic fluid sample. Furthermore, the analysis of linked polymorphic markers was performed for each family in order to establish whether any were informative for tracking the inheritance of mutant and normal copies of the β-globin gene. Together, this data assisted in the evaluation of the protocol design in terms of its suitability for clinical application.

Family 1

NGS-based genotyping revealed the presence of c.17_18delCT mutation in the father and c.135delC mutation in the mother, confirming their carrier status with the sequence coverage of over 40x. The counts were ~48% of the reference count for the G base at position 5248325 in the female and ~43% of the reference counts for the G base at position 5247987 for the male. In the son's DNA, 100% of the nucleotides at specified positions were in agreement with the

reference, accurately confirming the unaffected genotype with >62 reads. A graphical representation of the aligned sequences at mutation loci present in family 1 is shown in Figure 1.11.

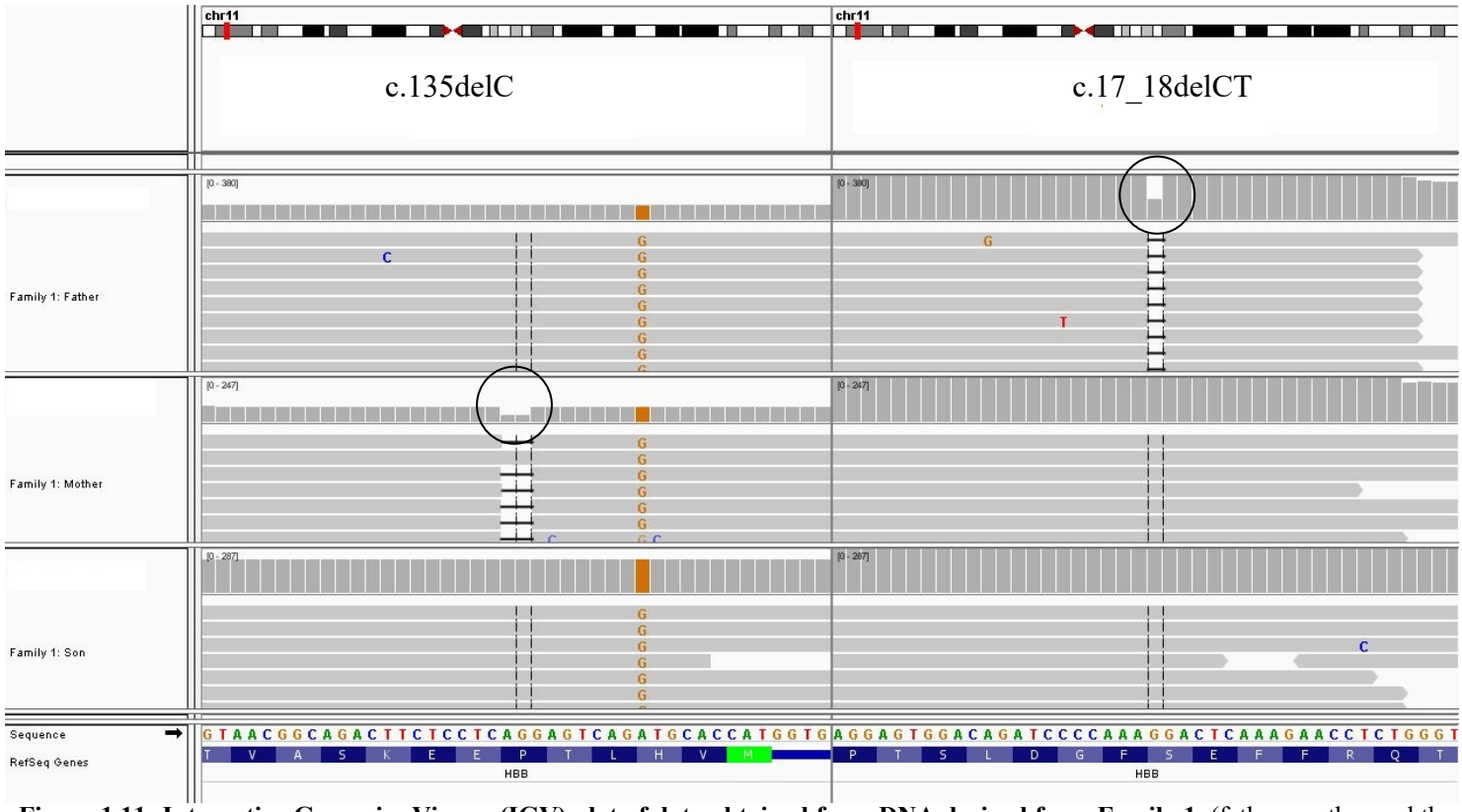


Figure 1.11: Integrative Genomics Viewer (IGV) plot of data obtained from DNA derived from Family 1. (father, mother and the son), representing genotypes found among the tested DNA. Each plot includes a vertical bar graph (columns on top) indicating the depth at each base. Letter codes are indicated at the bottom and represent normal human genome reference sequence. Each plot contains multiple horizontal bars, each representing an individual sequence read, with occasional miscalls indicating a different base was called at that position. Black lines at positions of c.17_18delCT and c.135delC mutations indicate a deletion was detected. Both parents were carries of their respective deletions, as indicated on the vertical bar graph (in circle). Genotype calls that are not in concordance with reference base for that position (in this case G base instead of A) are highlighted in orange and indicate a polymorphism present in all three members of family 1.

Family 2

Analysis of the sequences demonstrated the presence of c.92+5G>C substitution in father (coverage 43x) and c.126_129delCTTT deletion in the mother in the 71 of all reads. The counts were <65% of the reference count for the base C at position 5248155 in male and by ~49% reduced as a consequence of the deletion at position 5247996 in female. For the son,

both mutations were present at specified positions with the coverage of 95x, accurately confirming the affected genotype.

Family 3

NGS-based genotyping revealed the expected carrier genotypes in both the father and the mother with sequence coverage of over 90x in both samples. In the father, c.93-21G>A substitution was confirmed with 46% of the counts for the base C in concordance with the reference at position 5248050. In the mother, c.92G>A substitution was confirmed with 45% of the counts for the base C in concordance with the reference base at position 5248159. The daughter's carrier state for c.92G>A substitution was confirmed by 45% of the counts for the base C in agreement with the reference at position 5248159, with the sequence coverage of 67x.

Family 4

Analysis of the sequences demonstrated the presence of c.316-106C>G substitution in the father and GLN39TER substitution in the mother with sequence coverage of over 75x in both cases. The counts were 40% of the reference count for the base C at position 5247062 in the male and 44% of the reference count for the base A at position 5248004 in the female. Both mutations were called at specified positions with the read depth of over 83x, accurately confirming her affected genotype.

Table 1.8 summarises the genotyping results for familial β-thalassemia mutations present in the samples from four affected families in the pre-clinical phase in addition to the three couples who proceeded with the clinical testing (F5, F6 and F7, discussed in a separate Results section).

Table 1.8: Summary of the genotypes for β-thalassemia mutations in tested families. Each pair of mutations for which the parents are carriers is indicated in the top row with their respective annotations. The four families are separated by different shades of red fill. Carrier genotype is indicated in green and the affected genotype in black. Unaffected genotype is not coloured.

Mutation analysis														
Mutation	HBB:c.17_18delCT	HBB:c.135delC	HBB:c.92+5G>C	HBB:c.126_129delCTTT	HBB:c.93-21G>A	HBB:c.92G>A	HBB:c.316-106C>G	HBB:c.118C>T	HBB:c.92+5G>C	HBB:c.27_28insG	HBB:c.118C>T	HBB:c.92+6T>C	HBB:c.316-106C>G	HBB:c.93-21G>A
Position	5248235	5247987	5248155	5247996	5248050	5248159	5247062	5248004	5248155	5248224	5248004	5248154	5247062	5248050
Sample														
F1 Male	G/-													
F1 Female		G/-												
F1 Son	G/G	G/G												
F2 Male			C/G											
F2 Female				A/-										
F2 Son			C/G	A/-										
F3 Male					C/T									
F3 Female						C/T								
F3 Daughter					C/C	C/T								
F4 Male							C/G							
F4 Female								A/G						
F4 Daughter							C/G	A/G						
F5 Male									C/G					
F5 Female										A+C				
F5 Prenatal Sample									C/G	A+C				
F6 Male											G/A			
F6 Female												G/A		
F7 Male													G/C	
F7 Female														T/C

Analysis of linked informative SNPs

Differential parental haplotypes derived from a set of linked polymorphisms can allow the pattern of chromosomal inheritance in the vicinity of the locus of interest to be deduced, and thus prevents misdiagnosis as a consequence of ADO at individual mutation sites. Of the 18 SNPs targeted outside the *HBB* gene sequence (either upstream or downstream), the genotypes were successfully obtained in 17 SNPs. The one exception was the rs968856, amplified in fragment 14, which suffered a total amplification failure in all DNA samples that have been analysed in the study. In the remaining SNPs, the majority (13/17) were sequenced at sufficiently high coverage, with a minimum of 50x and a maximum of 200x. The four exceptions were rs11036364, rs7936823, rs12364872, and rs10837628, where the sequence depth only reached between 15-18x. A complete list of genotypes for all extragenic informative SNPs analysed for families 1-4 that took part in this pre-clinical study is included in Appendix 1. In addition to extragenic SNPs, five SNPs were identified within the *HBB* gene sequence of the four families carrying β-thalassemia mutations. Although the purpose of complete *HBB* sequencing was to primarily to permit analysis of a wide spectrum of β-thalassemia mutations, this strategy also proved highly successful in identifying additional (and in some cases novel) intragenic SNPs. The additional variants identified within the *HBB* locus of these four families with their respective SNP IDs are presented in Table 1.9. The rs ID for these additional SNPs was determined by browsing through the dbVAR database at the particular genomic positions where the variation had been observed.

Table 1.9: Additional intragenic SNPs identified within the *HBB* locus of four tested families.

	Additional SNPs identified within the HBB gene				
SNP ID	rs1609812	rs7480526	rs713040	rs63750628	rs12574989
Position	5247141	5247733	5248243	5248282	5246514

Overall, of the 22 SNPs sequenced (17 extragenic and 5 intragenic), each was informative for at least one family. On average, families had 8 informative SNPs, although the numbers significantly differed between the families. While family 1 only had one informative SNP (rs10837626), family 2 had 19 (16 extragenic and 3 intragenic), summarised in Table 1.10.

Table 1.10: Informative SNPs identified in tested families.

Linkage Analysis	Family 1	Family 2	Family 3	Family 4
Extragenic informative SNPs	1	16	9	8
Additionally identified SNPs	0	3	1	1

Of the 22 SNPs analysed in this preliminary study, 12 were informative for at least two families, while rs11036364 and rs7936823 were shown to be informative for families 2, 3, and 4. A summary of informativity for all SNPs analysed is presented in Table 1.11.

Table 1.11: Summary of SNPs based on informativity in tested families.

Summary of SNPs based on informativity	
Sequence covered for	22/23 SNPs
Informative for at least one family	22/23 SNPs
Informative for >2 families	12/23 SNPs
Informative for 3 families	2/23 SNPs

An example IGV plot representing the variety of genotypes detected in the four families for rs7936823 is presented in Figure 1.12. Considering that conventional PGT-M methods, based upon PCR and long-established methods for the analysis of DNA, typically rely upon two to four linkage markers to produce a diagnosis (Kuliev et al., 1998), the results demonstrated that the NGS approach may prove highly successful in providing additional linkage markers, making diagnoses even more accurate and robust. A graphical representation of how the

information obtained from IGV data plots can be used to interpret the genetic status of a child using parental haplotypes based on linkage markers is shown in Figure 1.13.

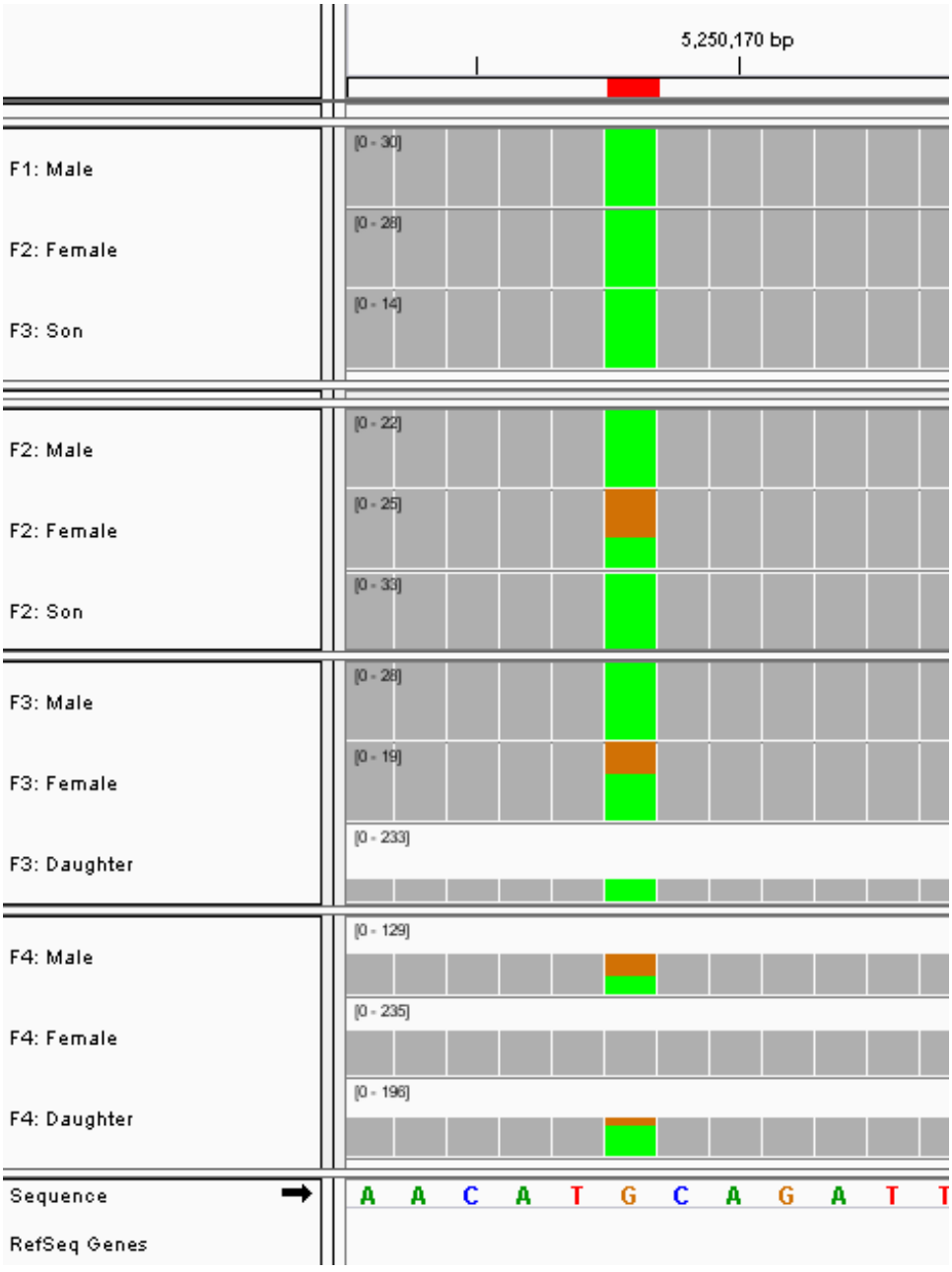


Figure 1.12: The variety of genotypes present in the four tested families detected for rs7936823 SNP, represented as IGV plot of data. Each plot includes a vertical bar graph (columns on top) indicating the depth of reads at each base. Letter codes are indicated at the bottom and represent normal human genome reference sequence. Genotypes in agreement with the reference genome (which is homozygous for G at this site) are marked in grey. For rs7936823 genotypes, A base is marked in green, G base is marked in orange. Combination of both in one DNA sample indicates heterozygosity at that position (A/G).

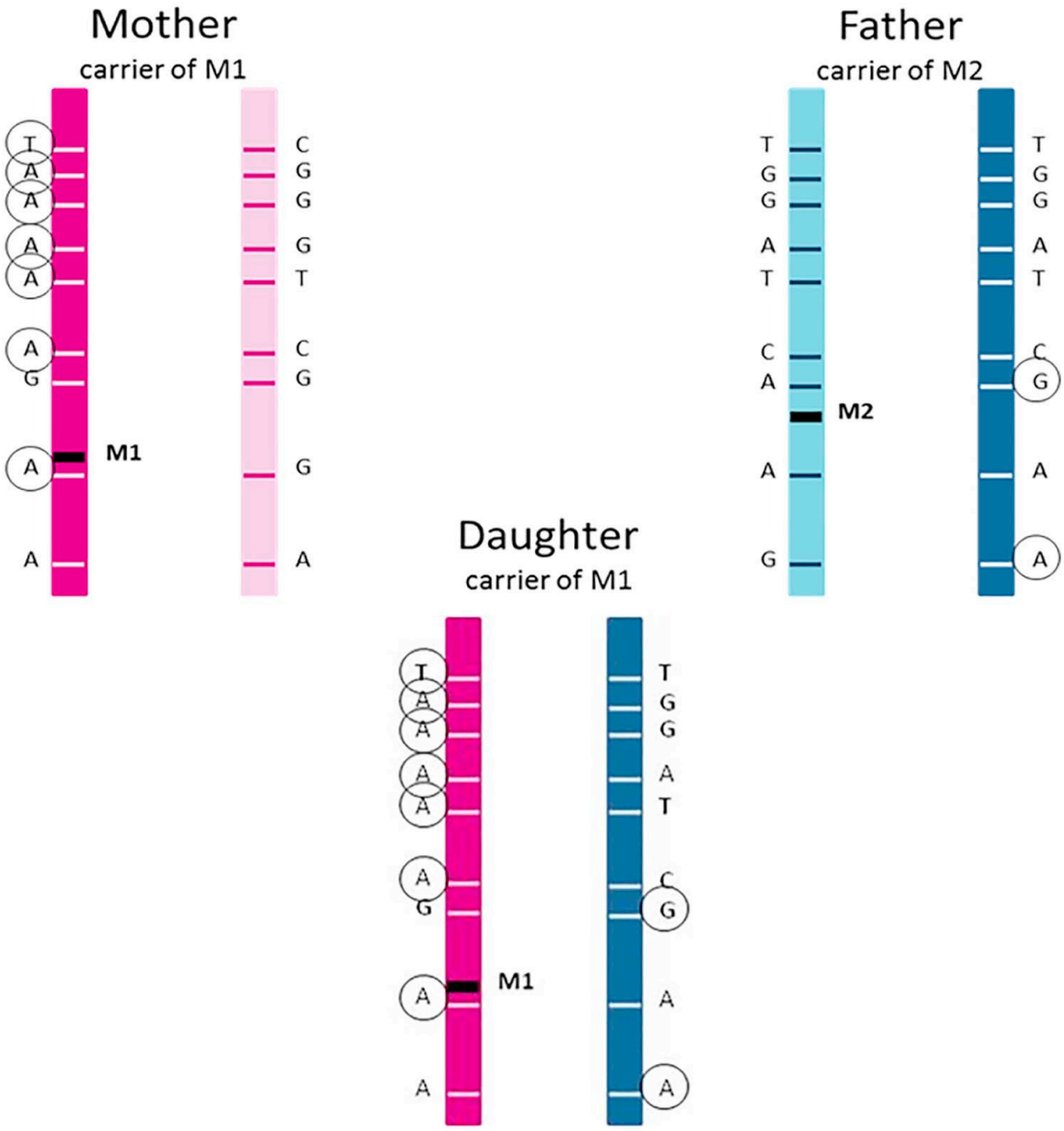


Figure 1.13: Demonstration of the principle of linkage analysis and targeted mutation detection on Family 3. For both parents, carrier states were confirmed by direct mutation detection. Linkage analysis identified nine informative single nucleotide polymorphism (SNP) near the *HBB* gene in this family. The father is heterozygous for two SNP for which the mother is homozygous (blue in circle) whereas the mother is heterozygous for seven SNP for which the father is homozygous (pink in circle). The inheritance of parental chromosomes in the daughter is thus inferred from nine SNPs as well as direct mutation detection. The daughter has inherited the affected maternal haplotype, associated with the M1 mutation (c.92G>A), and unaffected paternal haplotype, and is therefore a carrier of M1. This was further confirmed by the detection of the maternal c.92G>A mutation in the daughter (Kubikova et al., 2018). Of note, none of markers were separated by a distance larger than 25kb.

1.3.7 HBB gene mutation detection in clinical samples

After the validation phase, three couples (F5, F6 and F7) proceeded to clinical implementation of the newly designed protocol. In one of these couples, a previous affected pregnancy and a sample of amniotic fluid was analysed and confirmed as compound heterozygote for the two parental mutations tested. In the first case, a single blastomere was biopsied from cleavage stage embryos whereas in the other two cases, approximately five cells from the trophectoderm were biopsied from blastocyst stage embryos. All samples were whole-genome amplified using the MDA method (Repli-G Single Cell Kit, Qiagen) and subjected to simultaneous testing by Karyomapping (Karyomapping was performed by a member of clinical staff as part of the clinical service at Reprogenetics UK). All 21 embryo samples amplified after MDA and NGS sequencing analysis was successful for each of these. The sequence coverage for each of the amplicons was >1000x in all cases except for embryo 9 where no reads were obtained and Karyomapping analysis confirmed nullisomy of chromosome 11. The obtained genetic status at the mutation sites were concordant with the results acquired from Karyomapping and the genetic reports obtained from the medical geneticist prior to the case investigation. Of note, application of the targeted amplification protocol to single buccal cells and clumps, followed by NGS, generated a mean coverage of β-globin sequences of 888x (1061x for single cells samples and 716x for clumps of cells), which was equivalent to the coverage obtained from the clinical samples subjected to WGA before targeted PCR. These results suggest that it is possible to use this protocol on embryo biopsies directly, without compromising the read depth, eliminating WGA and thus further reducing costs (Kubikova et al., 2018).

1.3.8 Linkage analysis in clinical samples

In case 1, 10 informative SNPs were identified and used to assist determination of embryo mutation status: five associated with the maternal mutation and five associated with the paternal mutation (Table 1.12 A). All SNP assessed by the protocol described here were located outside the *HBB* gene but within 14 kb of upstream and downstream distance. It is worth noting that, for case 1, although the diagnoses of individual embryos obtained from targeted NGS were concordant with those obtained from the standard PGT procedure, the results from Karyomapping were considered suboptimal, associated with low call rates for individual SNPs (60-75% for half of the embryos) and unusually high ADO rates (>40%, 3-4 times higher than typically expected) in all embryo samples tested, likely a consequence of suboptimal WGA after blastomere biopsy, degradation of the biopsy specimens, or both. In embryo 9, no diagnosis was possible using the targeted sequencing method owing to complete absence of chromosome 11 and in embryo 13 only paternal alleles linked to *HBB* were detected. Karyomapping results were consistent with these findings, confirming the presence of chromosome 11 monosomy in embryo 13 and nullisomy in embryo 9. Despite using the same low-quality WGA template, the targeted NGS showed excellent performance in all 13 embryo samples that had provided a Karyomapping and NGS result, with no incidence of ADO observed in any of the informative SNP and mutation loci and >1000x sequencing coverage for the entire *HBB* gene and flanking polymorphisms. In case 2, 13 informative SNPs were identified, out of which two were located within the *HBB* locus. Two of the 13 SNPs were informative for the paternal mutation, whereas the rest were informative for the maternal mutation. For one embryo (number 6) only paternal alleles for SNPs linked to *HBB* were detected. Karyomapping analysis confirmed another instance of chromosome 11 monosomy (Table 1.12 B). In case 3, 10 informative SNPs were analysed, two located within the *HBB*

gene sequence itself (Table 1.12 C). Six of these were informative for the paternal mutation and four for the maternal mutation. Interestingly, three of the maternal SNPs were not among those deliberately targeted by the protocol and appear to be rare variants, which are not present within the dbVAR database. Surprisingly, of 141 heterozygous sites sequenced no instances of ADO occurred in any of the clinical samples tested using the NGS protocol, including the poor-quality WGA products from case 1. This contrasts with an ADO rate of around 12.5% for SNP genotyped using Karyomapping in case 3 and exceeding 40% in case 1, and suggests that the targeted NGS approach is highly sensitive and of excellent diagnostic accuracy (Kubikova et al., 2018). The complete list of genotypes at SNP and mutation positions detected and the corresponding coverage for each allele in all embryo samples is included in Appendix 2.

Table 1.12: Complete embryo results - mutation and linkage analysis: (a) case 1; (b) case 2; (c) case 3. Single nucleotide polymorphisms where heterozygosity was detected in one parent but not the other were selected for identification of disease-associated alleles in the embryo whole-genome amplification products subjected to targeted multiplex amplification in three clinical cases. In green are the alleles associated with the paternal *HBB* mutation and in red are the alleles inherited together with the maternal *HBB* mutation. Black alleles represent the disease-free genotype. A, affected embryo; C, carrier; CA, chromosomally abnormal; E, embryo, MC, carrier of maternal mutation; N/A, not applicable, the single nucleotide polymorphism identifiers do not exist for these variants as they were not present within the dbVAR database; PC, carrier of paternal mutation; SNP, single nucleotide polymorphisms; U, unaffected embryo. (Kubikova et al., 2018).

(A) Case 1	SNP ID	Position	Mother	Father	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
Linked SNP	rs3813727	5255912	G/G	A/G	G/G	A/G	A/G	A/G	A/G	A/G	A/G	G/G	no reads / copy of chromosome 11 detected	G/G	G/G	G/G	G	A/G
	rs4910736	5258989	A/A	A/C	A/A	A/C	A/C	A/C	A/C	A/C	A/C	A/A		A/A	A/A	A/A	C	A/C
	rs2105819	5259727	C/C	C/G	C/C	C/G	C/G	C/G	C/G	C/G	C/G	C/C		C/C	C/C	C/C	G	C/G
	rs7936823	5250168	A/A	A/G	A/A	A/G	A/G	A/G	A/G	A/G	A/G	A/A		A/A	A/A	A/A	G	A/G
	rs6578588	5252251	C/C	C/T	C/C	C/T	C/T	C/T	C/T	C/T	C/T	C/C		C/C	C/C	C/C	T	C/T
	rs7945118	5236417	C/G	G/G	C/G	C/G	C/G	G/G	C/G	C/G	C/G	C/G		C/G	G/G	G/G	G	C/G
	rs34220818	5236851	C/T	T/T	C/T	C/T	C/T	T/T	C/T	C/T	C/T	C/T		C/T	T/T	T/T	T	C/T
	rs12364872	5244144	A/G	G/G	A/G	A/G	A/G	G/G	A/G	A/G	A/G	A/G		A/G	G/G	G/G	G	A/G
	rs7110263	5246512	T/G	G/G	T/G	T/G	T/G	G/G	T/G	T/G	T/G	T/G		T/G	G/G	G/G	G	T/G
	rs10837626	5244299	T/A	A/A	T/A	T/A	T/A	A/A	T/A	T/A	T/A	T/A		T/A	A/A	A/A	A	T/A
Female mutation	c.92+6T>C	5248154	G/A	A/A	G/A	G/A	G/A	A/A	G/A	G/A	G/A	G/A		G/A	A/A	A/A	A	G/A
Male mutation	c.118C>T	5248004	G/G	G/A	G/G	G/A	G/A	G/A	G/A	G/A	G/A	G/G		G/G	G/G	G/G	A	G/A
Diagnosis			C	C	MC	A	A	PC	A	A	A	MC	CA	MC	U	U	CA	A

(B) Case 2	SNP ID	Position	Mother	Father	E4	E6	E7
Linked SNP	rs4910543	5258827	C/G	C/C	C/G	C	C/C
	rs4910735	5258852	A/G	A/A	A/G	A	A/A
	rs4910544	5258856	A/T	A/A	A/T	A	A/A
	rs4910736	5258989	A/C	A/A	A/C	A	A/A
	rs2105819	5259727	C/G	C/C	C/G	C	C/C
	rs11036364	5249004	A/G	A/A	A/G	A	A/A
	rs6578588	5252251	C/T	C/C	C/T	C	C/C
	rs10837620	5243559	G/A	G/G	A/G	G	G/G
	rs12364872	5244144	G/A	A/A	A/A	A	G/A
	rs10837628	5244404	A/G	A/A	A/G	A	A/A
	rs10837631	5246356	T/A	T/T	A/T	T	T/T
	rs7480526	5247733	C/A	A/A	A/A	A	C/A
	rs63750628	5248281	G/G	G/A	G/A	A	G/G
Female mutation	c.93-21G>A	5248050	T/C	C/C	C/C	C	T/C
male mMutation	c.316-106C>G	5247062	G/G	G/C	G/C	C	G/G
Diagnosis			C	C	PC	CA	MC

(c) Case 3	SNP ID	Position	Mother	Father	TE1	TE2	TE4	TE5
Linked SNP	rs713040	5248243	G/G	G/A	G/G	G/A	G/A	G/A
	rs1609812	5247141	A/A	A/G	A/A	A/G	A/G	A/G
	rs10837631	5246356	A/T	A/A	A/A	A/A	A/A	T/A
	rs7110263	5246512	G/G	G/T	G/G	G/T	G/T	G/T
	rs12574989	5246514	C/C	C/T	C/C	C/T	C/T	C/T
	rs10837626	5244299	A/A	A/T	A/A	A/T	A/T	A/T
	N/A	5243559	G/A	G/G	G/G	G/G	G/G	G/A
	N/A	5243613	C/T	C/C	C/C	C/C	C/C	C/T
	N/A	5236740	G/A	A/A	G/A	G/A	G/A	A/A
	rs7945118	5236417	G/G	G/C	G/G	G/C	G/C	G/C
Female mutation	c.27_28insG	5248224	A+C	A	A+C	A+C	A+C	A
Male mutation	c.92+5G>C	5248155	C/C	C/G	C/C	C/G	C/G	C/G
Diagnosis			C	C	MC	A	A	PC

1.4 DISCUSSION

Next-generation sequencing technology has developed rapidly in recent years and has found numerous research and clinical applications. In the context of PGT, NGS has principally been used for detecting chromosome abnormality (Fiorentino et al., 2014a, 2014b; Wells et al., 2014; Zheng et al., 2015), although interest in its application for the diagnosis of inherited single gene disorders has been growing (Chen et al., 2016; Ji et al., 2019; Ren et al., 2016; Treff et al., 2013c; Yan et al., 2015). In this study, an NGS protocol for PGT of β-thalassaemia and sickle-cell anaemia that is applicable to most couples at risk of having children affected by these disorders was designed, optimized and implemented. Preimplantation genetic testing strategies involving WGA of embryo biopsy specimens, followed by targeted re-amplification and NGS, have recently been reported, allowing successful diagnosis of monogenic disorders, discussed in more detail in the following sections.

Treff et al. (2013) conducted a first proof-of-concept study where a sample of trophectoderm was used to extract DNA and was screened for a familial *CFTR* point mutation using NGS. Similarly to this study, the authors showed the correct genotype could be detected with 100% concordance with respect to results obtained from conventional PCR-based mutation screening and linkage analysis (Treff et al., 2013b). While the study provided a useful indication that NGS might be a suitable method for detection of single gene disorders, it was limited to the interrogation of a single *CFTR* mutation site. The lack of data from linked polymorphisms, which can serve as a back-up for direct mutation detection by revealing disease-associated haplotypes, would leave the protocol highly susceptible to misdiagnosis caused by ADO.

Yan et al. (2015) and Ren et al. (2016) developed strategies for the diagnosis of selected inherited disorders, combining direct mutation detection and linkage analysis, and reported live births after the clinical use of their methods. The method used in these studies is known as Mutated Allele Revealed by Sequencing with Aneuploidy and Linkage Analyses (MARSALA) and it relies on sequencing of the whole genome at low coverage (~0.3x) to reveal a spectrum of cytogenetic abnormalities based on the presence/absence of reads corresponding to the genome reference. In addition, disease-specific primers were used to accomplish a several thousand-fold enrichment of sequence reads covering the mutation site as well as informative SNPs linked to the disease allele. Although NGS-based approaches are expected to eventually replace all conventional PGT methods, the current strategy of mutation site enrichment still requires extensive test customisation in most cases. Indeed, the effort required to create a custom-NGS protocol, involving targeted mutation detection and simultaneous analysis of linked polymorphisms may necessitate as much, or more, work than required for a conventional PGT protocol. This makes such approaches impractical for the diagnosis of rare disorders or conditions caused by a wide spectrum of mutations (e.g. β-thalassemia and cystic fibrosis), as in such cases the protocols created after extensive laboratory work may only applicable to a single family. Chen et al. (2016) described an NGS-based PGT approach requiring less customization, which was achieved by using a specially designed sequence capture array followed by NGS to provide data on the genotype of over 24,000 SNPs. The information gathered was subsequently used to assemble haplotypes, allowing diagnosis of embryos that inherited mutant copies of the *PKD2* gene encoding polycystin 2 from their parents based upon linkage analysis. This sort of NGS strategy is particularly useful because it delivers high fidelity, sensitivity, and throughput and, as it focuses on the inheritance of common polymorphisms rather than family specific mutations, requires little workup for each case. One of the limitations of this approach, however, is the cost associated with obtaining an adequate

depth of sequencing when assessing a large number of genomic regions. In the present study, although the entire sequence of the *HBB* gene was sequenced at high depth, as well as additional multiple closely linked polymorphic sites, the total proportion of the genome investigated remained small, equivalent to less than 10 kb of DNA sequence. This allows simultaneous analysis of large numbers of samples in a single sequencing run, potentially reducing costs. In ideal circumstances, each new test should be compatible with aneuploidy screening to assist in distinguishing chromosomally abnormal embryos from their euploid (and likely more viable) siblings. For this purpose, a straight-forward approach that uses gene panels to enrich for the mutation alleles and linked markers (similar to the one designed here) combined with whole genome sequencing (WGS) at ultra-low coverage to reveal cytogenetic abnormalities, as described by Wells et al. (2014), could be most optimal. A novel method known as PGD-SEQ combining these two approaches has recently been reported in over 200 cases, applied to various different disorders using gene panels that have been enriched by approximately 200 linked markers of high minor allele frequency values (Alcaraz Mas et al, 2019).

1.4.1 Clinical considerations: Mutation screening, linkage analysis, family proband and consanguinity

Most recently, there have been studies reporting PGT methods that utilise linkage analysis alone, focusing on common SNPs and abandoning customisation altogether. This can be achieved using the same principle of genome-wide SNP-haplotyping as used in Karyomapping. Ji et al. (2019) reported clinical application and live births after NGS-based SNP-haplotyping for the PGT-M of primary open angle glaucoma. A similar approach was used by Backenroth et al. (2019) who applied the new technique coupled with an analytical pipeline to four couples

undergoing PGT-M for different disorders. Although PGT methods based purely upon linkage analysis have the advantage of providing a more generic approach with no patient-specific work-up, a challenge for such strategies is that they cannot be used in cases in which the phase of the SNP alleles cannot be determined, i.e. where it is unclear which alleles are located on the same chromosome as the mutant gene. Deduction of phase requires DNA samples from the patients requesting PGT, and also from additional family members who have been previously tested and are of known mutation status (close relatives such as the parents or children of the couple are ideal). Testing of polymorphisms in these extra samples allows the inheritance of specific alleles to be traced through the family, revealing those consistently associated with normal and mutant gene copies. A lack of DNA samples from close relatives is a common occurrence in PGT; this is a consideration in about a one-quarter of all referred cases (data from Reprogenetics, UK). Sometimes, the couple are reluctant to discuss the fact that they are undergoing PGT with other members of the family, other times key relatives may be deceased or unavailable for other reasons, and on some occasions a patient may carry a *de novo* mutation, not present in any relative. Recently, an elegant method of overcoming the requirement of a family proband has been reported and uses MARSALA at ~3x of genome coverage to genotype single sperm cells for acquisition of linkage information that is later used to diagnose the embryos (Wu et al., 2018).

Another approach that obviates the need for a family relative of known disease status would be to use long-read sequencing (such as the one developed by Oxford Nanopore Technologies) on the parents to determine which alleles are inherited together with the mutation prior to testing of the embryos. In this case, the aim is to sequence through the mutation site along with potentially hundreds of SNPs in a single read (approximately 10kb), allowing easy phasing of the parental haplotypes.

Another issue for PGT strategies based entirely on linkage analysis is the possibility that informative alleles, permitting the two chromosomal copies to be distinguished, may be difficult to find. This is a particular problem when offering PGT to consanguineous families as many parts on the genome may be identical owing to shared ancestry. The data presented here also supported this view. While in one pre-clinical case 19 informative SNPs at which the two parents had different genotypes were detected, only one informative SNP was identified in one other case. With no access to the information related to genetic background of these families, the analysis suggested that family 1 had similar genetic background, sharing the same alleles at the majority of variant positions, potentially as a consequence of consanguineous marriage. In instances such as these, PGT cannot be carried out without direct detection of the causative mutation(s) in the embryos produced. Here, rather than targeting specific mutations, of which there are a large number in *HBB*, a method was created that provides information on the entirety of the coding region and splice junctions of the gene, as well as selected flanking sequences containing sites of common polymorphism. This permits direct detection of virtually all *HBB* gene mutations, effectively eliminating the requirement for patient-tailored test design. In non-consanguineous families, the ability to trace the inheritance of defective *HBB* genes by using multiple linked polymorphisms, as well as direct mutation detection, results in a highly redundant test, greatly increasing diagnostic accuracy. In the context of a PGT case, the mutation sites and polymorphisms can each provide their own independent diagnosis. Linkage analysis alone can potentially provide a reliable diagnosis in circumstances where parental mutations have not been identified before PGT being undertaken, in cases in which mutations are refractory to detection due to technical limitations, or in instances where one of the mutation sites fails to amplify appropriately as a result of ADO or other technical problems affecting the embryo biopsy sample.

1.4.2 Technical considerations: sequence coverage, detection of base variants and allele drop-out

The uniformity of sequence coverage could be improved by optimisation of the multiplex amplification to reduce the relative over- and under-representation of some amplicons. This could be done by adjusting the individual primer concentrations in order to balance amplification efficiency. Various manufactures advise optimisation of primer concentration as primer performance is influenced by internal stability, melting temperature, secondary structure, and interference with each other (Sint et al., 2012). Alternatively, poorly represented amplicons and primers for fragments which suffered from TAF could be re-designed, with emphasis on avoiding the presence of SNPs in the key sites where primers anneal. Avoiding polymorphisms and mutations in the primers proved to be particularly problematic for the design of the protocol described in this thesis as more than 300 different mutations and multiple polymorphisms have been described within the *HBB* gene (Cao and Galanello, 2010). Sequence variation is even more frequent in regions flanking the gene, where SNPs linked with a particular genotype can interfere with primer annealing and cause preferential amplification of some alleles over others (Chapuis and Estoup, 2007). Low GC content in some portions of the β-globin gene cluster also restricts primer design as primers with GC content below 30% may be less specific (Sint et al., 2012). Primer specificity presents an additional challenge due to *HBB* displaying high sequence homology with other genes in the *β*-globin cluster, particularly δ-globin (*HBD*), with which the human *HBB* gene shows 93% homology, as reflected in the similarity of their encoded proteins that only differ in 10 amino acids (Steinberg and Adams, 1991; Moleirinho et al. 2013). For this reason, all primers candidates should be screened for homology using the NCBI Blast software, as was initially done in this experiment.

The results from this study suggested that at higher sequencing coverage, the technology employed (MiSeq, Illumina) has excellent variant calling accuracy in PCR amplified single cell and MDA DNA. The accuracy of sequence variant calls, based on the SNPs with high coverage at heterozygous variant was estimated to be 98.21% with just one instance of ADO in the case of a single cell amplified using MDA and this sample appeared to be affected by suboptimal amplification, suffering from total amplification failure at multiple sites. This was not surprising as previous studies have indicated that MDA amplifies ~80% of the genome in single cells and ADO rates can be as high as 50% (Spits et al., 2006). The fact that no other ADO was detected in the samples derived and directly amplified from single cells (equivalent to a cleavage stage biopsy) and 5 cells (equivalent to a sample of trophectoderm) suggests the NGS is highly-sensitive, although proper sensitivity measures involving a high number of replicates for each sample category would need to be carried out to confirm this observation. The results also suggest that the methodology may be equally suitable for PGT-M performed at different embryonic stages with or without the intermediate WGA.

The extreme proximity to the *HBB* gene of the SNPx tested (≤25 kilobases in all cases) makes it highly unlikely that meiotic recombination would ever occur between the polymorphic sites and the sites of mutations (indeed this would be impossible for the intragenic SNPs analysed). The diagnoses should be highly resistant to errors owing to problems caused by failure of individual loci to amplify, preferential amplification and ADO. During this investigation, however, considering the embryo biopsy samples with a normal number of chromosome 11 copies, no ADO was detected at any of the 141 heterozygous sites sequenced (0%). Rates of ADO after MDA are influenced by the number of cells within the biopsy specimen and other technical factors. Hou et al. (2015) used the same type of MDA protocol as used in the present study and detected a 12.5% ADO rate on single cells re-sequenced at 30x sequence depth. In the present study, the incidence of ADO observed after Karyomapping was generally of a

similar level, although more than 50% of SNP loci were affected in some samples. The fact that ADO was so low in the present study could be attributable to the high sequencing depth used (>1000x). In theory, this should increase the sensitivity for the detection of alleles, which are substantially under-represented owing to extreme preferential amplification of the alternate allele in a heterozygous sample. Indeed, this was shown to be the case in some of the cleavage stage biopsied embryos analysed in case 1 (embryos 1, 2, 3 and 5), where preferential amplification led to over 97% of total reads being attributable to one of the two alleles at several sites (Table S3). Some investigations into the origins of ADO have suggested that this phenomenon occurs as a result of DNA strand breakage that consequently prevents PCR amplification by interrupting the synthesis of the growing strand (Piyamongkol et al., 2003). In the data from the current study, however, it appears that preferential amplification bias is significantly more common than true ADO. Therefore, it could be hypothesised that the ADO phenomenon arises as a result of the uneven amplification that appears to involve complete loss of one allele due to the limited sensitivity of the detection methods employed. Increasing the overall sequencing depth should then prove to be useful, particularly for the detection of severely under-amplified alleles.

In addition to the 17 well-characterized polymorphisms specifically targeted by the PGT protocol described here, an extra three previously uncharacterized intragenic sequence variants/polymorphisms were detected outside the *HBB* gene sequence (at positions chr11:5243559, chr11:5243613 and chr11:105236740). These provided a useful additional source of linkage data, further supplementing the diagnosis. The ability of NGS to detect novel polymorphisms and variants unique to individual couples reduces the risk of encountering low informativity when using the method. The use of WGA before targeted PCR and NGS, also meant that a resource of material was available for further testing if desired, e.g. repeat of the original PGT analysis. Furthermore, if desired, low-pass next generation sequencing of the

WGA templates generated using multiple displacement amplification can be used to establish the cytogenetic status of the embryo (Wells et al., 2014).

1.4.3 Future perspectives in PGT-M

Towards low-cost PGT

The combination of minimal work-up and high throughput provided by this protocol resulted in an economical test. The issue of cost is of great relevance in this particular case, given the fact that many regions of the world where *HBB* mutations are of high prevalence are relatively resource poor. The experience from the present study confirms that NGS can provide a rapid, streamlined and potentially cost-effective solution for couples seeking to use PGT to avoid genetic disease transmission. It is expected that, in the future, additional NGS protocols will be developed for the testing of other single gene disorders where mutation heterogeneity leads to problems for conventional PGT methods. In general, these tests have the potential to become even cheaper when performed using technology that does not require a significant capital investment. The use of devices such as the MinION, pioneered by ONT, could reduce costs per-sample by avoiding the need for a highly equipped laboratory, while the introduction of novel molecular methods may mean that highly skilled staff become less of a necessity. For example, the DNA could be amplified isothermally, using a recombinase polymerase, without the need of a PCR cycler and processed on an easy to use lateral flow assay. This approach would utilise very little laboratory equipment and could be performed locally, possibly even by a non-specialised nurse who is looking after the patients (Kubikova and Wells, 2020).

PGT for polygenic disease

The easy access to genomic technologies and the continuous decrease in their cost is accelerating the development of more comprehensive testing platforms. Recently, a single universal platform was developed and validated to test for aneuploidies, structural rearrangements, monogenic disease as well as traits controlled by polygenic inheritance (Treff et al., 2019). The aim of this methodology is to rank embryos, prioritizing those with low predictive values for polygenic diseases such as hypothyroidism and type 1 diabetes. It can be argued that polygenic disease risk prediction in preimplantation embryos is desirable and should be considered a valuable addition to the important health-related data revealed during the course of PGT. However, expansion of embryo testing to cover multiple genetic variants that, together confer a degree of risk, sometimes interacting with each other and with the environment in ways that are not fully understood, represents a significant deviation from the traditional application of PGT (Kubikova and Wells, 2020).

For the most part, the last thirty years has seen PGT utilised for the same range of inherited conditions for which prenatal testing has generally been considered appropriate - monogenic disorders that have a clear association between mutation and phenotype. Where PGT has shown a degree of divergence from prenatal diagnosis has been in its wider application to late onset disorders (e.g. Huntington Disease) and conditions associated with incomplete penetrance (e.g. BRCA mutations that predispose to breast cancer) and, most strikingly, the use of preimplantation testing to identify embryos that are HLA compatible with an affected sibling, who is in need of a stem cell transplantation (Verlinsky et al., 2001). The introduction of PGT for polygenic disease raises technical, logistic and ethical questions. It will inevitably require an expansion of genetic counselling efforts to support its clinical utilization, and some might argue that it is ethically questionable to test for such conditions. As new predictors become available, it likely that all of the embryos tested will have an increased risk of something, be it

cardiovascular disease, diabetes, cancer or other medically important problems. How will an embryo with an increased risk for one condition be weighed against a sibling embryo with an elevated risk for a different disease. There are fears that prioritizing embryos that are "genetically superior" based on their disease risk scores might create a hostile environment, potentially stigmatizing those who develop disease or those who suffer from disabilities. On the other hand, if predictions related to the development of serious health issues are demonstrated to be accurate, some might argue in favour of protecting the right for reproductive autonomy and parental choice. Most controversial of all, at least one company promoting the use of PGT to manage polygenic disease risk, has proposed testing to predict intellectual capacity, with a view to avoiding transfer of embryos at risk of producing individuals with a low IQ. As technologies continue to evolve, this area will clearly remain an active area of ethical debate (Kubikova and Wells, 2020).

Whole genome sequencing (WGS) of the preimplantation embryo

Considering the decreasing cost of sequencing-based methodologies, one might envisage a day when the most economical and straightforward strategy for PGT is simply to sequence the entire genome of each embryo. This would eliminate the need to develop patient-specific or disease-specific tests (Spath and Wells, 2015). Any chromosomal abnormalities would be revealed as well as any gene mutations (essentially delivering PGT-M and PGT-A in a single test) and, if desired, polygenic risk scores and information about long-term health could be acquired from the same data. It is likely that mutations that impair embryonic development would also be detected in some embryos, assisting in the prioritisation of viable embryos for transfer to the uterus during IVF cycles. The frequency of such mutations is currently unknown, but their presence may provide a partial explanation for the fact that at least one-quarter of

chromosomally normal and embryos of high morphological grade fail to produce a viable pregnancy after transfer to the uterus. In terms of the cost of the technology, there are already commercially available sequencers that do not require a large capital investment and it is only a matter of time before the cost of sequencing an individual genome falls to a level where it is conceivable that all embryos suitable for biopsy could be examined (Kubikova and Wells, 2020).

In the past, deficiencies in methods of genome amplification and DNA sequencing precluded the application of whole genome sequencing to human preimplantation embryos. However, these technological hurdles have been largely overcome. One of the few challenges still remaining is the requirement for data storage. This may well become the single most important factor on the clinical scale, especially when considering that some regulatory bodies currently require the data obtained from genetic testing to be stored for decades or even indefinitely. The human genome consists of more than 3 Gigabases, which means that genetic laboratories will face significant and ongoing capital investments into IT infrastructure and data storage. Given the highly sensitive nature of genetic information, back-up systems and robust data privacy protection mechanisms will all need to be in place (Kubikova and Wells, 2020).

Higher resolution and increased genomic coverage would, without doubt, lead to an even more comprehensive analysis of the embryo's genetic status. However, considering that in any given embryo, only a fraction of loci will be directly relevant to disease and that many (perhaps the majority) of the variants discovered would have an uncertain impact on health, the balance between the additional value provided by WGS must be weighed against the great increase in the bioinformatic and patient counselling burden (Spath and Wells, 2015). One might learn about clinical predispositions of a late onset nature that the future individual might not wish to know. Furthermore, WGS might also detect mutations in the embryo that provide an unintended diagnosis of one of the parents e.g. a mutation predisposing to cancer (such as

BRCA1/2) or neurological degeneration (like Huntington disease). Nonetheless, as understanding of the genome and how it functions improves, the depth of information provided by whole genome sequencing will have growing clinical utility and it seems highly likely that it will lead to eventual improvements in embryo diagnoses and IVF success rates. For the most part, the argument that remains to be debated, as with polygenic disease, is an ethical one (Kubikova and Wells, 2020).

Chapter 2: Germline genome editing: tools development and technical considerations for clinical application

2.1 INTRODUCTION

Genome editing is defined as technology that utilises programmable nucleases in order to cut and paste genetic information in a specific manner (Kim, 2016). The ability to genetically modify living cells is fundamental to the idea of eliminating genetic disease from the germline. However, achieving specific and targeted changes has been, and still remains, a considerable challenge. The first historical attempts to introduce specific modifications into eukaryotic cells relied on providing a target DNA template to employ homologous recombination (HR) (Capecchi, 2005). HR functions to repair double-stranded DNA breaks (DSBs) in all eukaryotic organisms and the idea behind this approach is that a homologous DNA molecule containing the desired alternative DNA sequence can be recombined into the cells at the targeted site by harnessing the endogenous HR machinery of the cell (Kim, 2016). Unfortunately, HR is extremely rare in the majority of living cells and therefore, the efficiency of incorporation of the donor DNA template is very low, preventing the routine use of gene targeting in this form (Capecchi, 2005).

An improvement came in the early 1990s with the engineering of targets specificity using zinc finger nucleases (ZFNs) and transcription-activator-like effector nucleases (TALENs). ZFNs utilise zinc finger protein domains that bind to a 3-bp motif in a modular manner, making them ideal as building units for engineering sequence-specific DNA binding nucleases (Kim, 2016). TALENs, on the other hand, recognise a single base in each repeat domain, allowing up to four different domains to be mixed and matched to generate a novel DNA binding protein. However, implementation of both of these programmable nucleases generates considerable off-target effects, defined as non-specific cleavage that gives rise to undesirable DNA modifications at

sites other than the intended target site, with the potential to result in cytotoxicity (Hye et al., 2009; Kim, 2016). Furthermore, since the target specificity is determined by modification of the DNA binding domain, the application of these nucleases is limited to cases where such enzymes can be successfully engineered, at significant cost of time and resources.

The field of GE was democratised and expanded dramatically in 2012 with the development of CRISPR-Cas9 technology that utilises RNA-guided programmable endonuclease (RGEN) (Cho et al., 2013; Mali et al., 2013). Here, the target specificity is determined by a custom single stranded guide RNA (sgRNA) that forms a ribonucleoprotein (RNP) complex with the bacterial endonuclease Cas9. In bacteria, the CRISPR-Cas9 system has evolved to facilitate adaptive immune responses to protect the cells from bacteriophage infection and horizontally acquired gene elements (Kennedy and Cullen, 2015). The majority of efforts to edit DNA using this system both *in vitro* and *in vivo* utilised a type II Cas9 derived from *Streptococcus pyogenes* (Spy) (Kennedy and Cullen, 2015). In this system, the guide RNA contains a 23bp sequence complimentary to the target DNA sequence that is located adjacent to a protospacer motive known as PAM (defined as any nucleotide followed by GG – e.g. NGG). The RNP can be readily modified by replacing the guide to target virtually any site in the genome quickly and cheaply, a process analogous to PCR primer design, eliminating the need for elaborate enzyme engineering and assembly. In the context of GE for the purpose of eliminating an unwanted mutation from the germline, the most straightforward approach is to microinject the RNP components into an early zygote or, even earlier, into an MII oocyte during intracytoplasmic sperm injection (ICSI) (Figure 2.1A). The intention is that the desired editing occurs prior to the first cell division in order to maximise the likelihood of delivering the modification to all cells of the future organism. Upon successful recognition of target sequence and PAM, Cas9 flanks the targeted region and cleaves the DNA strand, generating a DSB. Most living cells resolve the induced DSBs predominantly by two conserved mechanisms,

NHEJ and, less frequently, HDR (Figure 2.1B). During the process of reconnecting the two ends of the cleaved DNA strand, NHEJ introduces insertions and deletions (indels), which typically results in disruption of the targeted gene and a consequent loss of function (Figure 2.1A). This type of repair mechanism is not ideal for use in germline genome editing, unless the intention is to disrupt the gene function and not the correction of the mutant copy of the gene. In the latter scenario, HDR is a preferred mechanism since it rebuilds the site of breakage using a homologous DNA molecule as a template, potentially restoring the wildtype genotype. In cases where an embryo is heterozygous for a mutation, the second copy of the gene would remain uncut (as it does not contain the targeted mutation). In nature, repair of DNA damage via HDR usually involves the use of the undamaged copy of the gene/sequence as a template for replacement and correction. Repair using the undamaged/unedited copy of the gene may also occur in the experimental context (i.e. *in vitro*) (Figure 2.1A). However, it is also possible, and potentially more efficient, to supply an exogenous (synthetic) homologous DNA fragment for the cell to use as a template. In cases where the intention is to correct a mutation, the exogenous DNA would be similar or identical to the wild-type sequence. Needless to say, the addition of a synthetic DNA template is indispensable in cases where neither cellular copy of the gene has a wild-type sequence (e.g. homozygous recessive mutation).

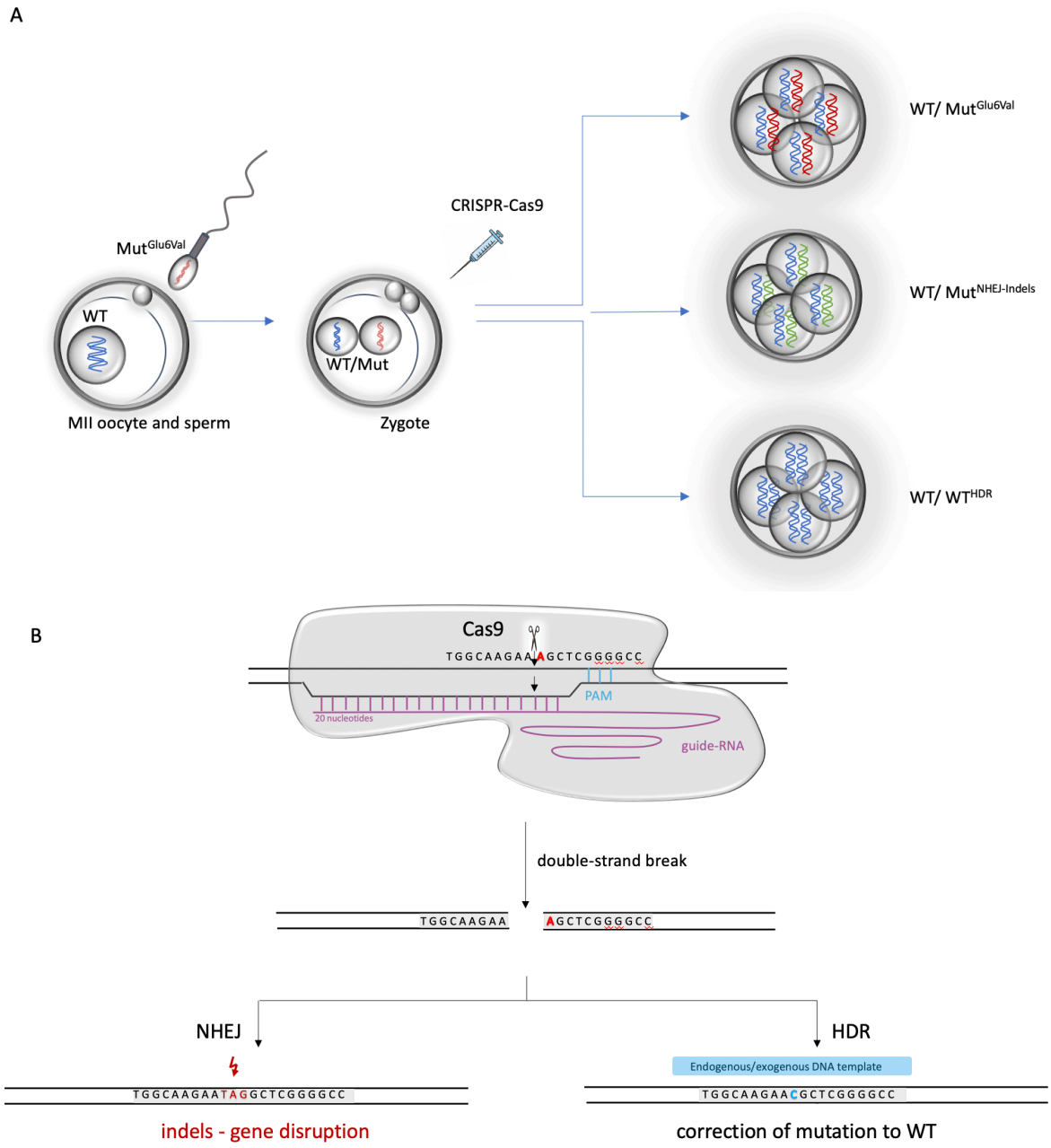


Figure 2.1: Schematic representation of CRISPR-Cas9 germline genome editing. A) Microinjection of CRISPR-Cas9 components after fertilisation at zygote stage using an example of a male carrier of Glu6Val mutation found in the *HBB* gene (the sickle cell anemia mutation). Targeting of only one of the parental alleles (the paternal Glu6Val mutation) can result in three possible outcomes observed during cleavage stage: 1) the editing is unsuccessful and does not alter the genotypes of the blastomeres that continue to carry one maternal wild-type copy of the *HBB* gene and one paternal copy containing the Glu6Val mutation (heterozygous recessive). 2) Induced DSB is repaired via NHEJ, resulting in cleavage stage blastomeres that carry a copy of maternal wild-type allele and the second paternal allele containing insertions and deletions (indels) induced by NHEJ. 3) Induced DSB is repaired via HDR where maternal copy of the gene (wild-type) serves as a template for strand replacement and correction. The homozygous wild-type genotype is restored and disease allele eliminated. **B)** Detailed representation of the RNP complex binding to target DNA upon the recognition of the complimentary DNA sequence and PAM motif and the resulting genotypes upon DNA cleavage and repair via NHEJ and HDR. NHEJ: Non-homologous end joining, WT: wild type, HDR: homology directed repair.

If consensus regarding germline GE can be reached and the controversy surrounding germline GE is resolved, it still remains to be determined whether GE could ever be considered for clinical application to correct germline mutations from the point of technical feasibility. In the second experimental chapter of this thesis, technical considerations for genome editing technology in terms of its potential clinical application in human preimplantation embryos are addressed. The work described here was conducted as a collaboration with the laboratory of Dr. Kathy Niakan (The Francis Crick Institute, London). It used human zygotes donated for research that were targeted using a CRISPR-Cas9 system for the first time under the HEFA approval and direct oversight. The gene target for this study was the well-known developmental regulator *POU5F1* (the *Homo sapiens* version of *OCT4*), a pluripotency and a transcription factor. The zygotes were microinjected with the mixture of Cas9 protein and sgRNA specifically targeting the *POU5F1* locus, with the intention of disrupting the Exon 2 reading frame and subsequent loss of OCT4 expression. The embryos were cultured in a time-lapse incubator where their development could be continuously monitored up to blastocyst stage. Following mitotic arrest or reaching of the blastocyst stage, the embryos were disaggregated into clumps of cells or individual blastomeres, and these were subjected to genetic analyses to investigate the resulting genotypes using a number of novel molecular methods and bioinformatics tools. The main aims of this study were:

- to assess on-target efficiency of the CRISPR-Cas9 system in inducing modifications to disrupt the coding sequence of the *POU5F1* gene leading to loss of OCT4 expression in human preimplantation embryos using a variety of bespoke and newly developed molecular and bioinformatics methods

- to simultaneously investigate potential off-target consequences of the CRISPR-Cas9 editing

- to reveal which repair mechanism had taken place (if any) by investigating the on-target mutational spectrum characteristic of DNA repair signatures and embryonic mosaicism

Although other cellular and animal models have been studied much more extensively and laid down the foundation for this work, it was hypothesised that significant differences in the underlying biology exist between these models and early human embryos, most likely concerning DNA repair and accessibility, and could, therefore, carry direct implications on the outcomes of the potential treatment. Instances of embryonic mosaicism, where one embryo contained more than one edited genotype, were also investigated, as such instances would be a concern for any future clinical application of GE technology. Unless mentioned otherwise, all the work has been carried out by me in the laboratory of Prof. Wells, University of Oxford. While the second experimental chapter of this thesis focuses on the technical considerations related to the application of germline GE in preimplantation embryos, the third experimental chapter looks in more detail at the fundamental biological processes related to DNA repair and genomic instability of early human embryos by exploring the same samples using a different array of molecular and bioinformatic methods.

2.2 MATERIALS AND METHODS

In this study, the *POU5F1* gene, the human homolog of *OCT4,* was targeted. The principal reason for choosing this particular gene was to examine its role in early human development. The concept behind targeting of the exonic sequence of the gene was that in the absence of an HDR template, the NHEJ would introduce indels that would disrupt the gene structure, and this would subsequently lead to loss of OCT4 expression. The initial phase of this study, described in the next paragraphs - sgRNA design and selection, human embryonic stem cells (hESC) transfection experiments as well as the microinjection and targeting of human zygotes - were carried out by Dr. Norah Fogarty and Prof. Kathy Niakan at the Francis Crick Institute, whose principal interest was to study the consequence of OCT4 loss in the edited embryos in the context of human embryogenesis. Nonetheless, it is essential to include these sections in the thesis, as the subsequent experiments relied heavily on the success of this important preliminary and validation work. In our laboratory, the main focus was on the genetic analysis of these edited embryos from the technical viewpoint, in order to examine the efficiency, specificity and the potential of genome editing technology for clinical application as treatment of embryos affected by inherited disorders. As such, our emphasis remained on the elucidation of the potential of CRISPR-Cas9 technology from purely clinical perspective. Nonetheless, the methods undertaken and the main conclusions relating to the OCT4 targeting will be described as these carry biological and clinical importance (discussed in Chapter 3).

2.2.1 Ethics

The Francis Crick Institute obtained ethical approval for this study from the Human Fertilisation and Embryology Authority (HFEA) under the licence number 0162 and from the Health Research Authority's Research Ethics Committee (Cambridge Central reference number 12E/EE/0067). The study was subjected to a review by the HFEA Licence Committee and all the experiments carried out were in compliance with the HFEA Code of Practice with regular inspections by the HFEA. The embryo donors were recruited from the Bourn Hall clinic. All donated embryos were cryopreserved at the pronuclear stage as surplus to IVF treatment. Informed consent was obtained from all donating couples prior to the study initiation. The donors had all the necessary information regarding the treatment of the embryos and that some of them would be genetically modified. Counselling was made available to explain all the procedures, including those relating to genome editing techniques. The donors were informed that the development of all of their embryos would be stopped at a maximum of 14 days post-fertilisation and subsequent biochemical and genetic studies would be performed on the material obtained from the embryos. They were further informed that the acquired data would be used to disseminate knowledge in scientific journals and at scientific conferences. No genetic tests would be carried out on the donors. The donors were not financially reimbursed for their participation in the study.

2.2.2 sgRNA design and selection

Candidates for sgRNA targeting exons of *POU5F1* were selected using an online tool (CHOPCHOP) using the default parameters. The candidates were screened to exclude guides containing SNPs with minor allele frequencies above 0.1 (validated by 1000 Genomes project).

The exception was sgRNA targeting exon 4 which has a single SNP with a MAF 0.32. The guide was kept as it had the highest *in silico* cutting score predicted by CHOPCHOP. Overall, 4 candidates fitted the criteria and were selected for further validation. These included transfection of hESC line to establish which guide has highest cutting efficiency. The genomic location and sequences of these candidates are represented in Figure 2.2 (Fogarty et al., 2017).

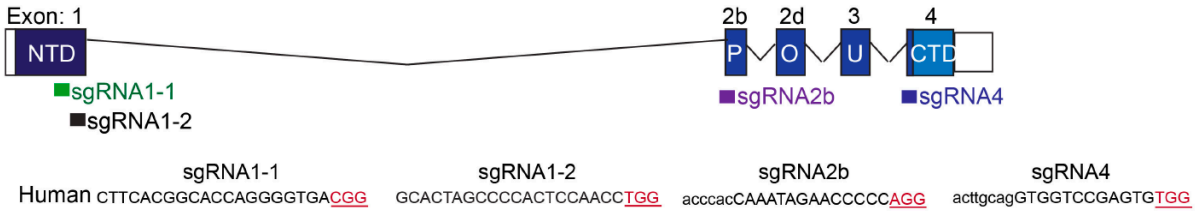


Figure 2.2: Targeting of the POU5F1 locus by 4 different sgRNA candidates and their respective genomic positions within the POU5F1 gene (Fogarty et al., 2017). PAM sequences in red.

2.2.3 Transfection and targeting of hESC line

To screen candidate sgRNAs, human embryonic stem cells (clone H9) were utilised as an unlimited resource that, to some extent, reflects the cellular context of the human preimplantation embryo. Isogenic hESCs constitutively expressing the Cas9 gene were engineered, together with a tetracycline inducible sgRNA (one of the four candidates), thereby allowing comparative assessment as schematically represented and described in Figure 2.3.

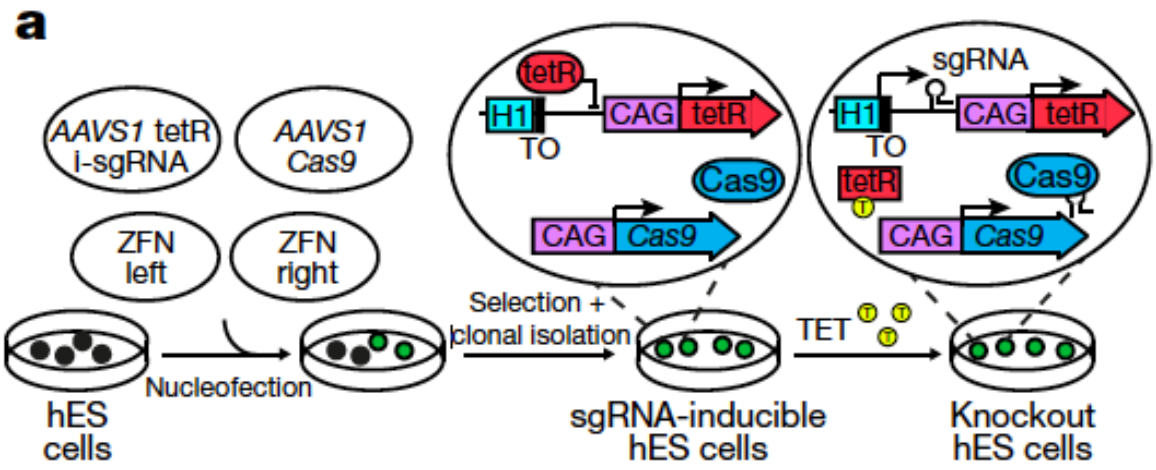


Figure 2.3: Generation of inducible knock-out hESC lines: The selected candidate sgRNA guide sequences were cloned into the pAAV-Puro_siKO-TO vector. First, the complimentary strands were annealed and ligated into AarI-digested plasmids between the H1-TO inducible promoter and the sgRNA sequence. The vectors carrying the sgRNA and *Cas9* were subsequently inserted into one of the two alleles of the *AAVS1* locus by homologous recombination facilitated by ZNFs. Cells were cultured in the presence of antibiotic and 10 μ M of ROCK inhibitor Y-27632 (Sigma-Aldrich) for 24 hours prior to nucleofection, then washed in PBS (Life Technologies) and dissociated with Accutase (Life Technologies) for 5 min at 37 °C. Colonies were mechanically separated into clumps of 2–3 cells and counted. 2 × 106 cells were nucleofected in 100 μ l with a total of 12 μg DNA (4 μg each for the two ZFN plasmids, and 2 μg each for the two targeting vectors) using the Lonza P3 Primary Cell 4D-Nucleofector X Kit and the cycle CA-137 on a Lonza 4D-Nucleofector System. Cells were incubated for 5 min at room temperature, after which antibiotic-free KSR containing 10 μM ROCK inhibitor was added. After another 5 min the cell suspension was distributed on pre-plated DR4 (Applied Stem Cell) drug resistant MEF feeders in antibiotic-free KSR medium. Four days after nucleofection, cells underwent double antibiotic selection with 0.5 μg ml^{-1} Puromycin (Sigma-Aldrich) and 25 μg ml–1 Geneticin (G418 Sulphate (Gibco)) for 7 days. Targeted colonies appeared after 4–8 d and were mechanically picked and clonally expanded at 10–14 days after transfection. Extensive genotyping was carried out on the targeted clones to check for correct AAVS1 gene targeting and to exclude the presence of randomly integrated plasmids. Briefly, genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega). Site-specific integration was checked for both 5′ and 3′ ends of each of the two targeting vectors (Cas9 and inducible sgRNA). Clones were also screened for the absence of the wild-type locus (indicating homozygous targeting) and for the absence of amplicons for both the 5′ and 3′ ends of the targeting vector backbones (to ensure there was no random integration of the plasmid (Fogarty et al., 2017).

2.2.4 Human ESC line culture conditions

Clonal H9 human ES cells (both untreated and engineered inducible knock-outs) were cultured in feeder- and serum-free conditions in mTeSR1 (Stem Cell Technologies) on growth factor-reduced Matrigel-coated dishes (BD Biosciences). Tetracycline hydrochloride (Sigma-

Aldrich) was used at 1μg ml^{-1} to induce guide expression. Human ES cells underwent routine mycoplasma screening and karyotyping.

2.2.5 Genomic DNA extraction and genotyping of hESC

Human ESC cells (both untreated H9 controls and sgRNA[1-1/1-2/2b/44]-Cas9 targeted) were lysed using proteinase K 10 μg/ml in lysis buffer (100 mM Tris buffer pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl)) overnight at 37 °C. Following the treatment, genomic DNA was extracted using a phenol-chloroform protocol combined with ethanol precipitation. The cells were collected in bulk every 24h for up to 4 days, in order to facilitate a time-course analysis of genotypes. The extracted DNA was amplified to enrich each of the *POU5F1* exons using the primers pairs listed in Table 2.1 on the "Hot Master Taq" program (initial denaturation step at 96°C for 15 min, followed by 35 cycles of denaturation at 94°C for 15 sec, annealing at 56°C for 15 sec, extension at 65°C for 45 sec, followed by a final extension step at 65° for 2 min) then multiplexed and sequenced on the MiSeq (Illumina) (described later in Materials and Methods).

Table 2.1: PCR primers used to amplify exonic regions of *POU5F1*.

Site	Forward oligo	Reverse oligo
sgRNA1-1 on-target site	CTGTGGGCCCCAGGTT	ATCAGGCTGCCCTGTCAT
sgRNA1-2 on-target site	TGGAGGTGATGGGCCAG	ACCAGGGGTGACGGTG
sgRNA2b on-target site (used primarily)	AGGGGAGATTGATAACTGGTGT	ACTAGGTTCAGGGATACTCCTTAG
sgRNA4 on-target site	TGTCCTCCTCTAACTGCTCT	CAGAGGAAAGGACACTGGTC

2.2.6 Single guide RNA2b generation and ribonucleoprotein (RNP) preparation

Following the validation of four sgRNA candidates in hESC lines, sgRNA2b was selected as the most suitable guide that cuts with highest efficiency and was used in all downstream

experiments involving human embryos. The guide RNA (from now on termed as sgRNA2b) was prepared by *in vitro* transcription. First, the sgRNA was cloned into the px330 vector (Addgene) using the Bbs1 restriction site and the sequence was amplified to check for correct incorporation (primer sequences detailed in Table 2.2 using high fidelity polymerase (New England Biolabs). The PCR amplicon was *in vitro* transcribed using the MEGAShortscript T7 kit (Thermo Fisher), purified using the Zymo RNA kit (Zymo Research) and stored at -80 °C until further use. Prior to zygote microinjection, the sgRNA mixed with recombinant Cas9 protein (Toolgen) at 37 °C for 15 min according to the manufacturer's instructions.

Table 2.2: PCR primers used to amplify the target *POU5F1* in hESC cells and a T7 oligo designed for *in vitro* transcription.

***In vitro* transcription of human sgRNAs**

sgRNA	Forward	Reverse	T7 Guide Primer
sgRNA2b	CACCGACCCACCAAATAGAACCCCC	AAACGGGGGTTCTATTTGGTGGGTC	TTAATACGACTCACTATAGGACCCACCAAATAGAACCCCC

2.2.7 Microinjection of sgRNA2b-Cas9 RNP into human zygotes and embryo culture

In total, 22 human zygotes were *POU5F1* targeted using the validated sgRNA2b-Cas9 RNP and 20 zygotes were microinjected with Cas9 protein alone to control for the microinjection procedure. Microinjections were carried out in Global Total medium supplemented with HEPES buffer under mineral oil on a heated microscope stage using a holding pipette (Research Instruments) and a Femtojet micromanipulator (Eppendorf) set to 20 constant and 40 injection pressure. The RNP components were microinjected into both zygotic pronuclei by Dr. Kathy Niakan. Embryos were cultured in drops of Global medium (LifeGlobal) under a

layer of mineral oil (Origio) and supplemented with 5 mg/mL of protein supplement (Life Global, LGPS-605) in a time-lapse incubator (EmbryoScope+, Vitrolife) for up to 6 days.

2.2.8 Blastomere disaggregation, embryo biopsy and whole genome amplification

Single guide RNA2b-Cas9- and Cas9-microinjected embryos were micro-dissected into individual blastomeres or a clump of approximately 5 cells of trophectoderm, depending on the development stage they reached (two-cell embryo, cleavage stage or blastocyst). Samples were sent to the University of Oxford and their genomic DNA was whole genome amplified using SurePlex amplification system (Illumina). Briefly, the 22 sgRNA2b and the 20 control samples were first suspended in 2.5 µl of 1xPBS in a 0.2 ml PCR tubes, then subjected to cell lysis and DNA extraction reaction using 2.5 µl of cell extraction buffer, 4.8 µl of extraction enzyme dilution buffer and 0.2 µl of cell extraction enzyme per reaction. The reactions were spun down and placed in a thermal cycler and ran on the following program:

1. 75 C for 10 min x1 cycle
2. 95 C for 4 min x1 cycle
3. Hold at 4 °C

Next, pre-amplification was performed by adding 4.8 µl of SurePlex pre-amp buffer and 0.2 µl of SurePlex pre-amp enzyme to each sample tube. The tubes containing the pre-amp cocktail were flicked and spun down, then placed in a thermal cycler and ran on a following program:

1. 95 °C for 2 min x1 cycle

2. 95°C for 15 sec
   15°C for 50 sec
   25°C for 40 sec
   35°C for 30 sec
   65°C for 40 sec
   75°C for 40 sec
   } 12x cycles
3. Hold at 4 °C

The final amplification step was performed by adding the following reagents to each sample tube: 25 µl of SurePlex amplification buffer, 0.8 µl of SurePlex amplification enzyme and 34.2 µl of nuclease free water. The tubes containing the amplification cocktail were flicked and spun down, then placed in a thermal cycler and ran on a following program:

1. 95°C for 2 min 1x cycle
2. 95 °C for 15 sec
   65 °C for 1 min
   75 °C for 1 min
   } 14x cycles
3. Hold at 25 °C

To determine the success of the amplification, 5 µl of each amplified sample was combined with 2 µl of gel loading buffer (2X) and ran on a 1.5% agarose 1x TBE gel until resolved.

2.2.9 Targeted amplification of the POU5F1 locus in WGA embryonic DNA

Out of the 22 RNA2b-Cas9- and 7 Cas9-microinjected embryo samples, 16 samples amplified successfully using this protocol, three failed WGA and were excluded from further analysis and three showed suboptimal amplification (faint smears on the agarose gel). The RNA2b-Cas9 samples that showed evidence of amplification were processed along with three Cas-9 microinjected control samples for genotype analysis. This was carried out by a PCR-based enrichment of the *POU5F1* locus in the WGA products. The primer sequences used to amplify the 244 bp of the exon 2b were selected according to the criteria described in page 49 and are listed in Figure 2.4. The PCRs were performed using the same reaction set-up and ran on the "Hot Master Taq" program. The resulting PCR amplicons were quantified using high sensitivity dsDNA Qubit assay (Thermo Fisher Scientific) to establish whether concentrations were in an acceptable range for further sample processing (approximately 3–5 ng/µl). Gel electrophoresis was then performed to confirm the size of the PCR product corresponded to the expected amplicon size. Of the 19 WGA products tested from the RNA2b-Cas9 embryos, six failed the targeted PCR amplification and were excluded from further analysis. The rest were processed for genotype analysis using targeted deep sequencing on Miseq.

2.2.10 Targeted sequencing of POU5F1-enriched WGA human embryo DNA and POU5F1-enriched hESC DNA and data analysis

Libraries for sequencing the *POU5F1* enriched WGA products as well as the *POU5F1* enriched hESC DNA were prepared using the TruSeq DNA PCR-free library preparation kit (Illumina). Amplicons were first cleaned with the AMPure XP (Beckman Coulter) beads using a 1.6x sample to beads ratio and followed by a double wash in 80% ethanol while the samples

were kept on a magnetic stand. The DNA was eluted by resuspending the beads/DNA mix in 20 µl of resuspension buffer (RSB), then placed on a magnetic stand until the liquid cleared and the samples were transferred into new 0.2 ml microcentrifuge tube. Library preparation, multiplexing and sequencing were performed as described before on pages 58-62 using the MiSeq Reagent Kit v2 (Illumina) along with 10% PhiX genomic control as outlined in Figure 2.4.

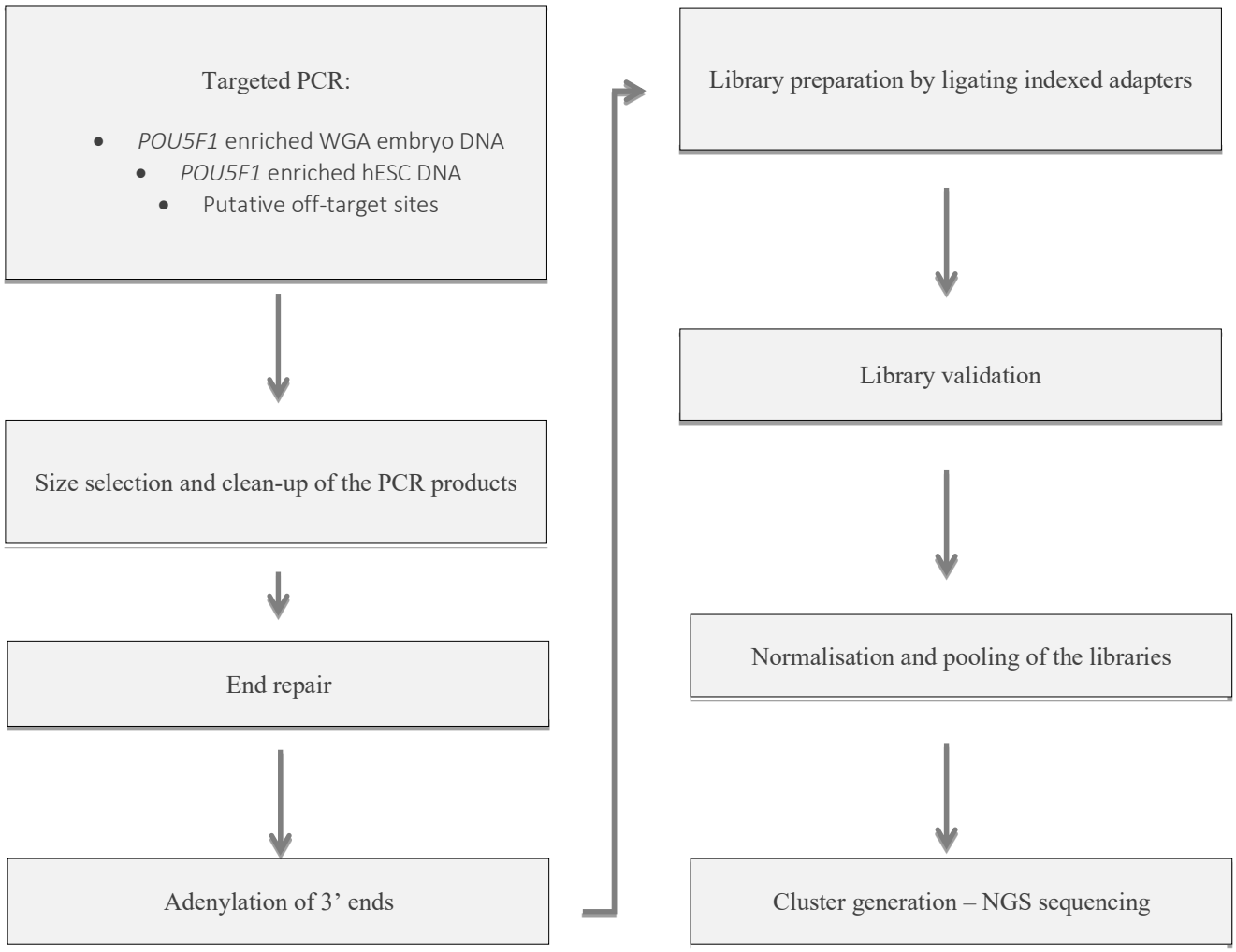


Figure 2.4: NGS work-up. Flow diagram describing the order of steps included in the processing of DNA samples for NGS and the preparation of DNA sequencing libraries using the Illumina TruSeq PCR-Free Library Prep kit (Illumina, UK).

The run generated paired-end (2 x 150 bp) dual indexed reads. The sequences were analysed using the online CRISPR Cas-Analyzer tool (Park et al., 2017) by uploading the FASTQ files and defining the target DNA sequence and unedited sequence as a reference. The tools were used to align the reads and to determine the percentage of non-wild type reads resulting from editing, as well as assessing the position and size of each indel for all of the PCR amplicons evaluated. The genotypes were additionally confirmed using the IGV software.

2.2.11 Evaluation of the putative off-target sites

In order to evaluate the possible off-target effects of non-specific cutting, MIT CRISPR design tool was used to predict putative off-target sites *in silico*. The top off-target indicated sites were selected for further analysis if they contained fewer than three mismatches compared to the original sgRNA2b sequence (Table 2.3). In addition, a global alignment of the original sgRNA2b sequence was performed using NCBI Nucleotide Blast against the hg19 genomic reference in order to identify if the first 12 bp of the sgRNA2b (defined as the 'seed' sequence) aligned elsewhere in the genome and if so whether it contained a PAM motif adjacent to the seed sequence. For the shortlisted candidates, a new set of primers were designed to encompass approximately 250bp centred around the putative off-target cut site in order to maximize the detection of a variety of mutations and to ensure that each amplicon was sequenced continuously from the forward and reverse barcode. The primer sequences for these regions (located in chromosomes 1, 8, 10 and 12) were designed according to the same criteria as described earlier and the sequences are listed in Table 2.4. For each primer set a gradient PCR was performed to establish the most suitable primer annealing temperature followed by agarose gel electrophoresis to visualise the DNA bands (as described before). Once the correct annealing temperature was determined, the primers were pooled into a multiplex primer pool and multiplex amplifications were performed on the WGA products derived from the human

sgRNA2b-Cas9 and Cas9 microinjected embryos. Each reaction was carried out with 0.5 µl of the DNA sample in a total reaction volume of 50µl, which contained 22 µl of nuclease-free water, 25 µl of 2 x Qiagen master mix (Qiagen), and 2.5 µl of primer mix (set to a concentration 2µM, prepared by adding 2 µl of each stock primer solution of 100 µM concentration up to a total volume of 100 µl with nuclease-free water). Replicates of four, along with a negative control were used for amplifications performed with the conditions: initial denaturation step at 95°C for 15 min, followed by 15 cycles of denaturation at 94°C for 30 sec, annealing at 56°C for 90 sec, and extension at 72°C for 1 min, with a final extension step at 60° for 10 min. The amplicons were sequenced along with the rest of the samples using the same library preparation kit and sequencing method as described earlier and the obtained reads as well as the VCF files (variant calling files) were subjected to the same investigations using the CRISPR Cas-Analyzer and IGV.

Table 2.3: Putative off-target sequences tested.

Exon2b sgRNA: ACCCACCAAATAGAACCCCCAGG

Putative off-target sequence	Score*	Mismatches	UCSC gene	Gene ID	Locus	PAM
ACCCATCAAATACAACCCCCCAG	1.7	2MMs [6:13]			chr19:-9072444	No NGG PAM
AGCCACCAGGTAGAACCCCCAAG	1.5	3MMs [2:9:10]			chr3:+101807899	No NGG PAM
CCTTCCCAAATAGAACCCCCAGG	1.3	4MMs [1:3:4:5]	NR_036440	*POU5F1P3*	chr12:+8286916	12p13.21
CCTTCCCAAATAGAACCCCCCAG	1.3	4MMs [1:3:4:5]			chr3:+128394390	No NGG PAM
CCTTCCCAAATAGAACCCCCAGG	1.3	4MMs [1:3:4:5]	NM_001159542	*POU5F1B*	ch8:-128*428616*	8q24.21
CCTTCCCAAATAGAACCCCCAGG	1.3	4MMs [1:3:4:5]	NR_034180	*POU5F1P4*	chr1:-155403476	1q22
TATTCCCAAATAGAACCCCCAGG	N/A	5MM [1:2:3:4:5]	NR_131184	*POU5F1P5*	chr10:68010749	10q21.3

*Score based on the MIT CRISPR Design tool when available. Three sequences predicted from the *in silico* tool lacked an NGG PAM site required for the binding of the Cas9 nuclease and were subsequently excluded from further analysis. Mismatches are in red.

Table 2.4: PCR primers used to amplify the regions of putative off-target cut sites of sgRNA2b-Cas9.

NR_036440 putative off-target site	CCTGCACGAGGGTTTCTG	AAGGAGTCCCAGGACATCAA
NM_001159542 putative off-target site	AACCCGGAGAAGTCCCAG	TGTTGTCAGCTTCCTCCAC
NR_034180 putative off-target site	GCAGGAGTCCCAGAACATC	GGGTTTCTGCTTTGCATGTC
NR_131184 putative off-target site	CCAGTCCCAGGACATCTCAA	ACTTCTGCAGCAAGGGC

Amplification of the sgRNA2b on-target site was initially performed using a pair of primers generating a 244 bp fragment in all WGA embryonic DNA samples. Of the 23 WGA products examined from the *POU5F1* targeted embryos, six failed the targeted PCR. Any sample that failed targeted amplification three times was subjected to an alternative amplification strategy using a different pair of primers designed to encompass larger fragments (800 bp and 1000 bp). This was carried out in order to minimise the possibility of failed detection as a result of an unexpected SNP in one of the primer annealing sites, or due to the presence of a deletion, induced by CRISPR-Cas9, extending far enough from the targeted site to encompass one of the PCR primer sites in one or both alleles. Additionally, we designed a set of primers adjacent of the cut site for the same reason. The sequences of these alternative primers are listed in Table 2.5.

Table 2.5: Additional PCR primers used to amplify regions of the 2b exon of *POU5F1* in samples that failed amplification using the primary pair.

sgRNA2b on-target site (used primarily)	AGGGGAGATTGATAACTGGTGT	ACTAGGTTCAGGGATACTCCTTAG
sgRNA2b on-target site 1000bp	CGCCCAGCAAAGAACTTCTA	GAGAACCACTGCACCAAAGA
sgRNA2b on-target site 800bp	TGCATGAGTCAGTGAACAGG	GAGAACCACTGCACCAAAGA
sgRNA2b on-target site 140bp	CATGGGTGAGGGTAGTCTGC	TGGGATATACACAGGCCGAT
sgRNA2b adjacent to on-target site1	GCCTGACTGCTTGGACATTC	GGCTCGAGAAGGATGTGAGT
sgRNA2b adjacent to on-target site2	CAGGTGGTGGTGTGAAAAGG	TCGTAGCTCTCCGTCTTTGG
sgRNA2b adjacent to on-target site3	CAGATGGTCGTTTGGCTGAA	TCTGGGAAGAGGTGGTAAGC
sgRNA2b adjacent to on-target site4	CTTCAGGAGCTTGGCAAATTG	AGGGGAGATTGATAACTGGTGT
sgRNA2b adjacent to on-target site5	ACCCATTCCCTGTTCACTGA	GCCAGGGTCTCTCTTTCTGT
sgRNA2b adjacent to on-target site6	TCTCTCACTCAAGTATCACCCC	AAAGCAAGCTGGGGAGAGTA

2.2.12 Development and validation of PCR-free Cas9-mediated protocol for enrichment and long-read sequencing of the POU5F1 locus

The last experimental section of this chapter focused on the development of novel strategy for sequencing of the *POU5F1* locus using the long-read Nanopore technology. Traditionally, long-read targeted sequencing of DNA derived from embryo biopsy specimens has been considered a challenge, principally due to the limited amount of DNA material obtained from the biopsy. One strategy is to employ WGA systems such as MDA that generate fragments of up to 10 kilobases, and subject the sample to whole genome sequencing (WGS). The major limitation of this approach is that it is very unlikely to obtain sequencing coverage in the regions of interest from WGS unless a considerable depth of coverage is obtained (approximately 30x), and this can be cost-prohibitive. Furthermore, introducing long range PCRs to amplify regions of interest is problematic due to additional bias introduced from the PCR and the increasing likelihood of allele drop-out (ADO), which can affect up to 50% of all blastomere samples derived from cleavage stage embryos (Spits et al., 2006). The novel protocol described in the next sections involves the use of CRISPR-Cas9 technology to cleave DNA sample in the upstream and downstream of region of interest (in this case *POU5F1*), in order to facilitate PCR-free enrichment of the targeted locus, followed by Nanopore sequencing. This methodology would enable the detection of larger indels, otherwise impossible to detect with traditional PCR approach where a primer annealing site might be lost due to an unintended modification present in the sequence. Additionally, sequencing of larger fragments would increase the likelihood of detection of additional heterozygous loci, and subsequently allow for detection of instances of ADO by observing loss of heterozygosity in the affected region. Furthermore, this approach has the potential to detect and characterise structural variation, a possible unintended consequence of editing and repair that has, so far,

not been evaluated in the context of human germline genome editing. Finally, the proposed technique would allow the detection of diverse spectrum of mutations, possibly extending kilobases away from sites where the modification was intended, otherwise impossible to detect with traditional PCR approach where only a short fragment of several hundred base pairs is generated and sequenced. In order to validate the proposed approach, a proof of concept study was carried out using hESCs that were subjected to WGA by MDA. The MDA products were subsequently processed using the newly developed protocol and sequenced on a MinION (ONT). The data was analysed using a number of bespoke bioinformatics tools.

Sample preparation – hESC culture, DNA extraction and WGA by MDA

The hESC H9 cells were cultured as described in the previous sections and collected in bulk in a 1.5 ml low bind DNA tube (Qiagen). The cells were spun down and the DNA was extracted using the QIAamp DNA mini kit (Qiagen). Briefly, 180 µl of buffer ATL and 20 µl of Proteinase K were added to the tube containing approximately 10^6 cells and the tube was incubated at 56 °C for 15 min in a heat block. After the incubation, 200 µl of buffer AL was added and the sample vortexed and incubated at 70 °C for 10 min. Next, 200 µl of ethanol (100%) was added into the mix and briefly centrifuged. The tube containing the mix was then transferred into the QIAamp mini spin column where 500 µl of buffer AW1 was added and a new 2 ml collection tube was placed underneath. The column was centrifuged at 60000 x g for 1 min. Next, we discarded the flow-through from the collection tube and placed the spin column into a new 2 ml collection tube with 500 µl of buffer AW2. The column was centrifuged at 20,000 x g for 3 min. The flow-through was discarded from the collection tube and the column was placed in a new 1.5 ml tube. After adding 200 µl of buffer AE to the mix, the column was centrifuged at 6000 x g for 1 min to elute the DNA. The DNA was quantified using Qubit High Sensitivity dsDNA assay with the expected yield of approximately 10-15 µg.

Multiple displacement amplification (MDA)

The DNA samples extracted from the hESCs were subjected to WGA by MDA using the Repli-g Single Cell kit (Qiagen). First, we prepared the buffer D2 by adding 3 µl of 1 M DTT to 33 µl of reconstituted buffer D2. We then diluted the extracted DNA to approximately 10 pg/µl to mimic the amount that would be obtained from embryo biopsy and we transferred 1 µl of the diluted DNA into a separate 0.2 microcentrifuge tube. Next, 3 µl of buffer D2 was added to the sample and incubated in a thermal cycler at 65 °C for 10 min. Stop solution (3 µl) was then added to terminate the lysis reaction. The amplification was carried out by adding 9 µl of water, 29 µl of the Repli-g reaction buffer and 2 µl of the Repli-g DNA polymerase to the sample tube containing the mix from the previous reaction. The tubes were briefly vortexted, spun-down and placed in a thermal cycler set to 30 °C for 3 hours. The expected DNA yield of approximately 10 µg was quantified by Qubit and the fragment length distribution was checked in the TapeStation (Agilent).

Cas9-mediated in vitro cleavage of hESC MDA DNA

In the experimental design of this protocol, we used the excision approach where sgRNAs are designed to target either side of the region of interest (*POU5F1*). In order to excise the region of interest, we used two previously designed sgRNA guides that were known to cleave DNA with high efficiency, sgRNA1 and sgRNA2d, located in exons 1 and 2d, respectively (sgRNA1 of sequence CTTCACGGCACCAGGGGTGACGG and sgRNA2d of sequence GTTTGGCTGAATACCTTCCCTGG). This method should provide highest coverage of the *POU5F1* locus (approximately 4 kb), containing the previously targeted exon 2b in the middle of the sequence (Figure 2.5). This approach should yield coverage on either side of each of the two cut sites, further away from the region of interest, extending the read coverage beyond 4 kb, which is advantageous for the detection of structural variation. The sgRNA guides along

with the S. Pyogenes HiFi Cas9 and tracrRNA (all supplied by IDT) were assembled into individual RNPs, then pooled together and used in an *in vitro* cleavage reaction with the hESC MDA products.

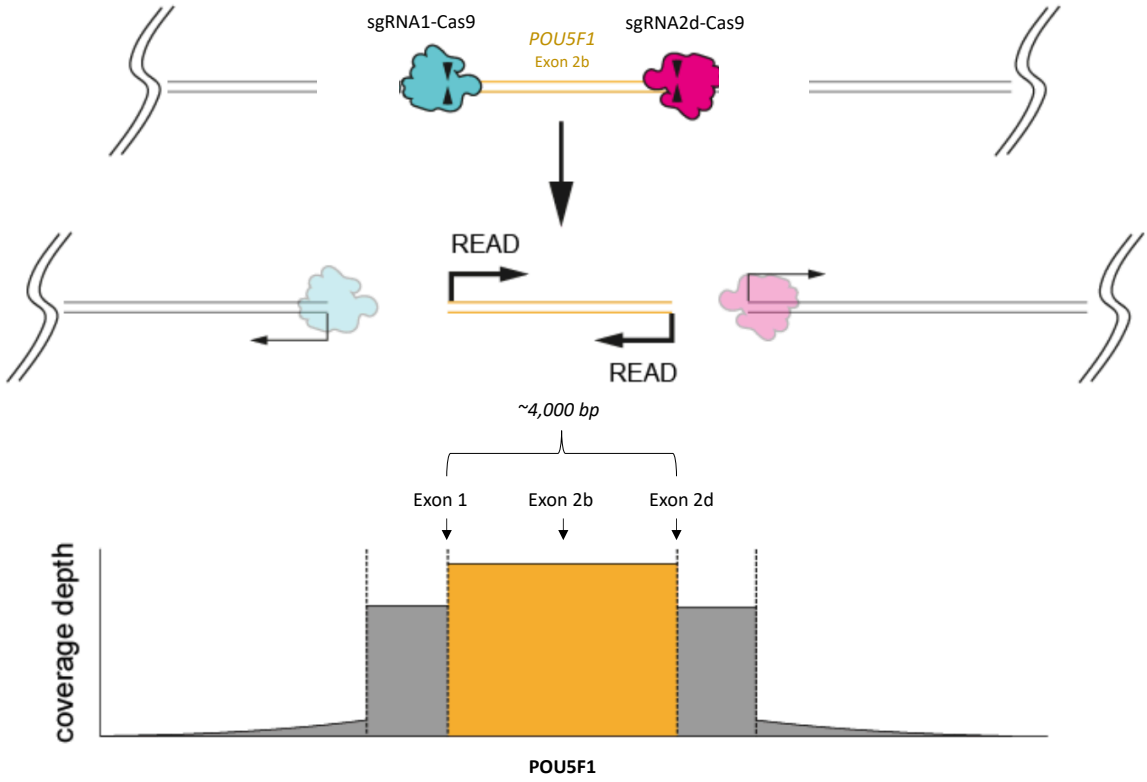


Figure 2.5: Theoretical experimental design and relative positioning of the sgRNAs with respect to the region of interest (*POU5F1* exon 2b) and the expected coverage. In yellow is the target sequence of interest, surrounded by the respective sgRNA-Cas9 RNP binding on either site. The Cas9 enzyme binds directionally, leaving the inner ends exposed for sequencing (producing high read coverage of the region of interest represented in yellow bar, while remaining bound on the outer generated ends, producing 3-fold decrease in coverage of the outer sequences (represented in grey bars).

First, we assembled the sgRNA pool by adding 8 µl of Duplex buffer, 1 µl of each sgRNA and 1 µl of tracrRNA (all IDT) into a reaction tube and mixed by pipetting. The mix was denatured at 95 °C for 5 min and allowed to cool to RT. Next, the RNP complex was formed by adding 10 µl of CutSmart buffer (NEB), 79.2 µl of nuclease-free water and 0.8 µl of HiFi Cas9 into the reaction mix. The RNP was incubated for 30 min at RT. Next, the DNA samples were dephosphorylated with calf intestinal phosphatase (CIP). This step blocks the DNA ends by

removing the terminal 5' phosphate and is essential to avoid sequencing the rest of the DNA present in the sample. During this step, 1 µg of MDA was suspended in 24 µl of nuclease-free water was mixed with 3 µl of CIP (NEB) and incubated at 37 °C for 10 min, then at 80 °C for 2 min to terminate the reaction and then placed on hold at RT. In the next step, 10 µl of the pre-assembled Cas9 RNP, 1 µl of Taq polymerase and 1 µl of 10mM dATP were added to the dephosphorylated DNA. This process cleaves the target and dA-tails all available ends (in this case only the ends that have been cleaved by sgRNA-Cas9) and activates the cut sites for ligation (schematic diagram summarising the steps is presented in Figure 2.6). The contents of the tube were carefully mixed by inversion and flicking, then placed into a thermal cycler to incubate at 37 °C for 60 min, then at 72 °C for 5 min followed by hold at 4 °C. This experimental approach is summarised in Figure 2.7.

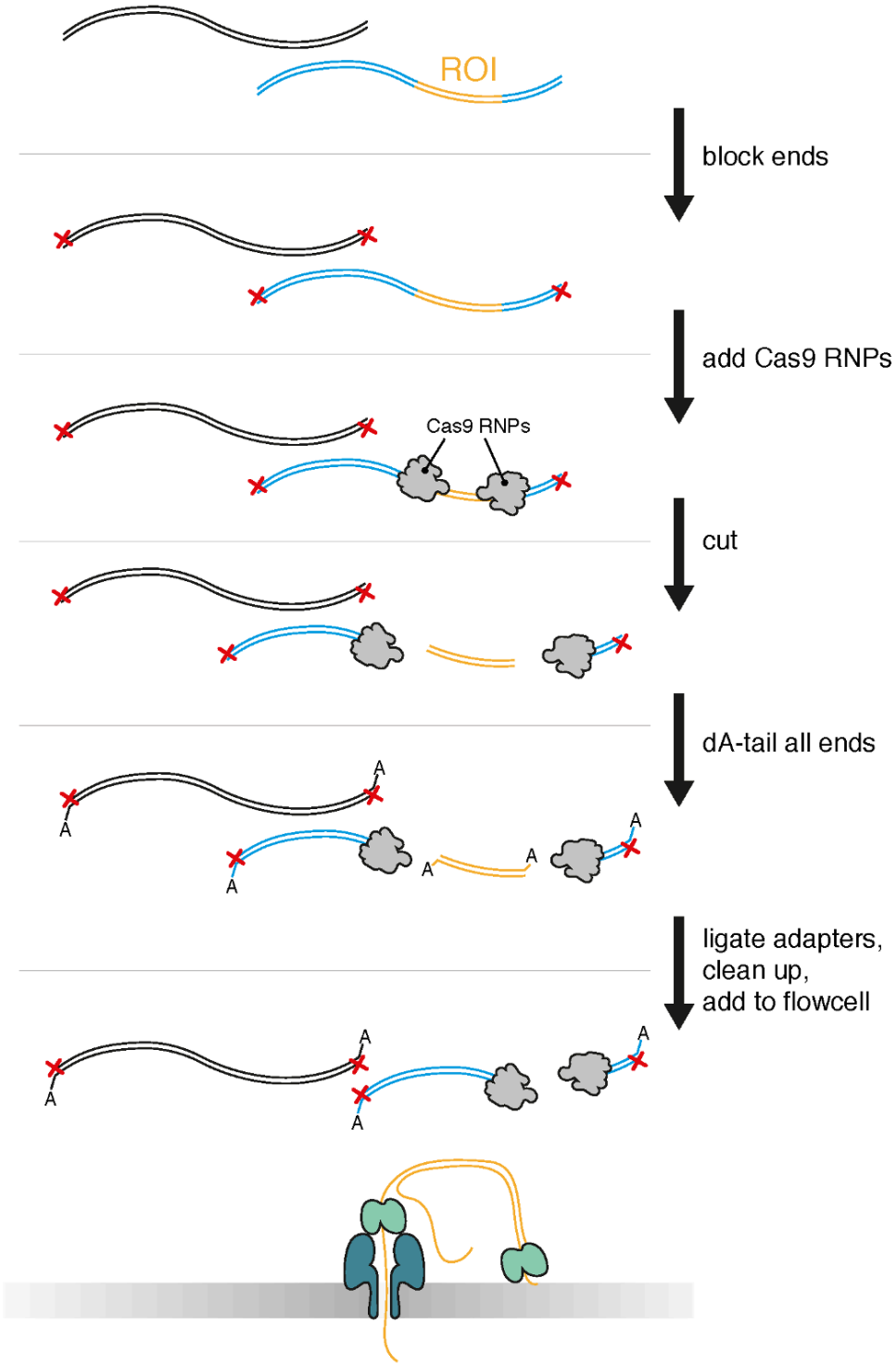


Figure 2.6: **Schematic diagram summarising the steps of the Cas9 enrichment experiment.** The DNA sample is first subjected to dephosphorylation to block all DNA ends from adapter ligation. The target DNA is then cleaved using a specific sgRNA-Cas9 enzyme. This specific cleavage exposes DNA ends for subsequent ligation of sequence adapters. The entire sample is loaded onto a flow cell, however, only the enriched region with ligated sequence adapters is sequenced (Adopted from Oxford Nanopore Technologies online Community resource).

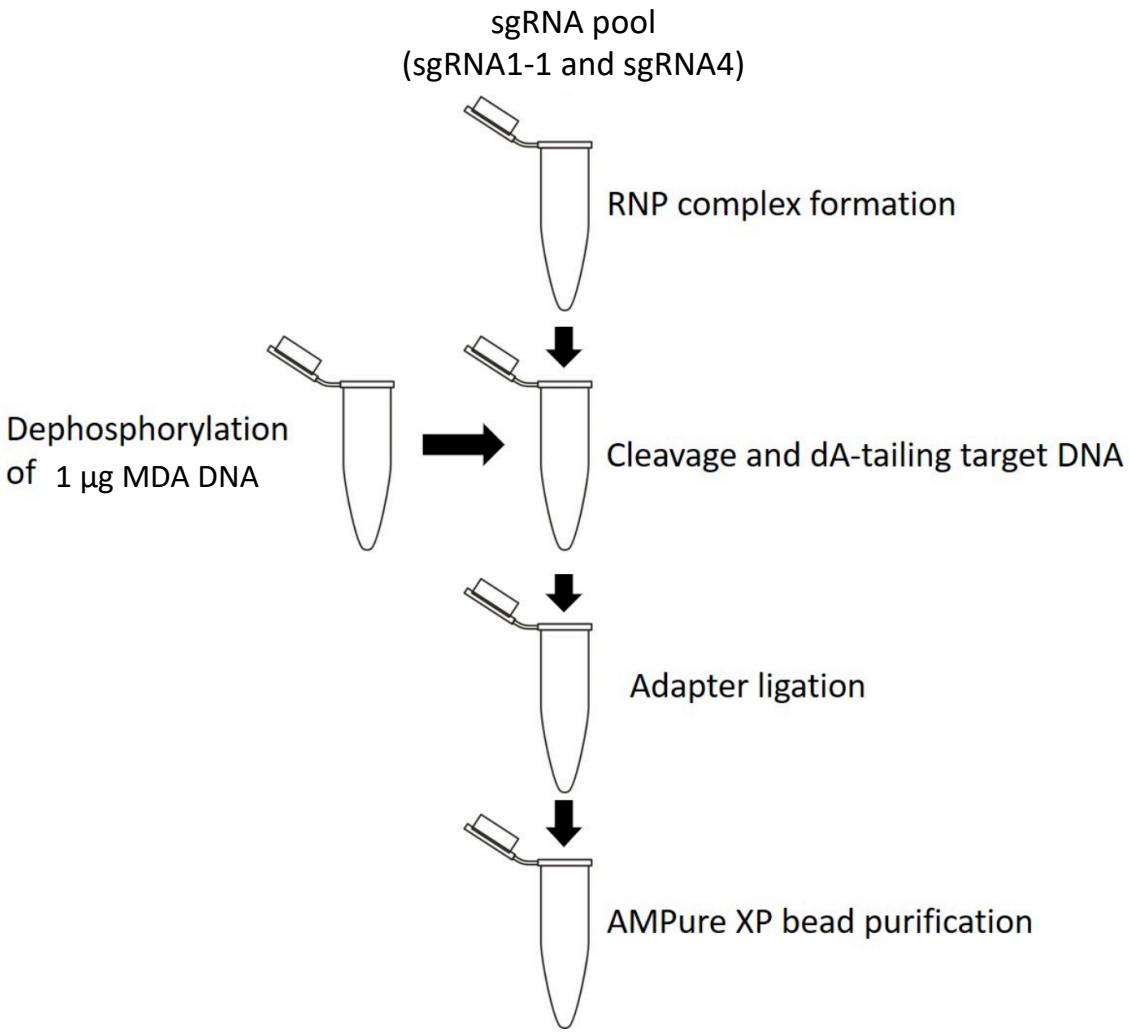


Figure 2.7: Steps of the Cas9 enrichment protocol using the 'excision approach'.

Library preparation

The cleaved and dA-tailed DNA was then used to prepare the library for sequencing using the Ligation Sequencing Kit SQK-LSK109 (ONT). First, the AMX adapters, containing the motor proteins used later anchor the DNA onto the sequencing pore embedded in a semi-synthetic membrane, are ligated to the target DNA. The mix of 20 µl of ligation buffer LNB, 3 µl of nuclease-free water, 5 µl of AMX adapter mix (all ONT) and 10 µl of Quick T4 DNA ligase

(NEB) was prepared in a separate tube and pipetted up and down, then added to the cleaved and dA-tailed sample. The reaction was incubated for 10 min at RT. The final step involved the clean-up of the excess adapters using the AMPure XP bead system (Beckman and Coulter). One volume of TE (pH 8) buffer was added to the ligation mix, then mixed, followed by adding 0.3x volume of the beads. The mixture was gently mixed and incubated at RT for 10 min. Next, the beads were placed on a magnetic stand and allowed to pellet until the liquid cleared. The supernatant was discarded and the beads where washed twice with 250 µl of long fragment buffer while on the magnetic stand. The library was eluted into 13 µl of elution buffer after resuspending the bead mix in the elution buffer and placing back onto the magnetic stand until the liquid was clear.

2.2.11 Long-read Nanopore sequencing, data collection and analysis

The cleaned-up library was next loaded onto the ONT SpotON flow cell using the standard procedures for loading provided by the ONT. The final loading volume was 50 µl, prepared by adding together 25µl of the sequencing buffer, 13 µl of the loading beads and 12 µl of the DNA library. Once the flow cell was loaded, it was placed into the MinION and the sequencing run was initiated and allowed to run for approximately 24 hour under the default criteria. The data was available in real-time but was only collected after the run had finished in the fastq format. The fastq files were processed in a suite containing a set of different command line tools to index and align the reads against the human reference genome (build hg38) in Minimap2. The resulting SAM files were then converted into BAM format, sorted and indexed in Samtools. Next, the reads were processed in R package (using Rsamtools and GenomicAlignments scripts) to prepare summary statistics and to filter out the on-target sequence reads from the

off-target reads (using seqtk script and a custom BED file containing the genomic coordinates for the *POU5F1* region of interest). The reads were visualised in IGV (Broad Institute).

2.3 RESULTS

2.3.1 Selection of sgRNA for targeting of the POU5F1 locus

As previously described in the Material and Methods of this chapter, to target *POU5F1* a standard *in silico* prediction tool was used to design four sgRNAs, whose specificities and cutting efficiencies were validated *in vitro* using hESCs as a model. The cells were engineered to constitutively express the Cas9 gene, together with one of the four tetracycline inducible sgRNAs. To compare the on-target efficiencies and mutation spectrums induced by candidate sgRNAs, a time-course on-target genotypic analysis was performed up to day-4 after the induction, presented in Figure 2.8. Of the four candidates tested by targeted deep sequencing on a MiSeq, sgRNAs 1-1 and 2b induced indels with the highest efficiencies (in over 80% of the cells) on day 4. The analysis further revealed that sgRNA2b was able to induce 20% of the cells after 24 hours and the proportion steadily increased until day-3, after which it stabilised. This appeared superior to the sgRNA1-1 that induced clones where the proportion of the induced cells decreased after 48 hours and increased again after day-3.

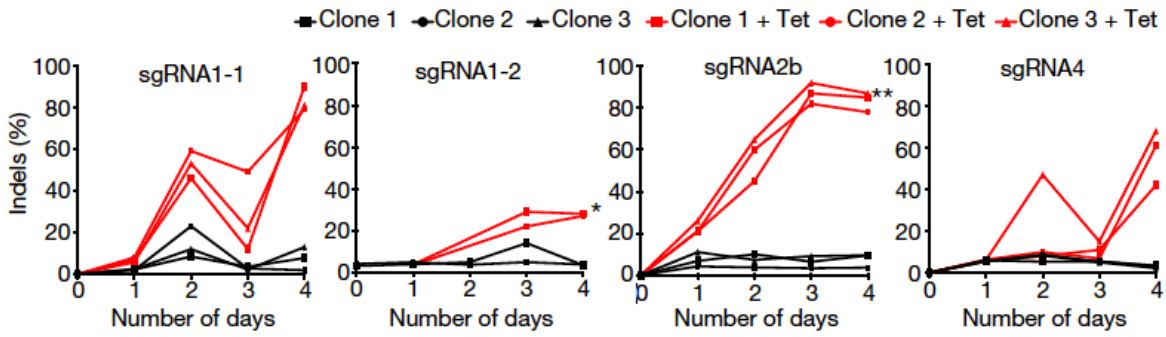


Figure 2.8: Quantification of indel mutations detected at each sgRNA on-target site after 4 days of sgRNA2b induction (+ Tet). n = 2 (sgRNA1-1); n = 3 (sgRNA1-2, sgRNA2b or sgRNA4 clones). One-way ANOVA compared to uninduced human ES cells (Fogarty et al., 2017). Tet: tetracycline.

The candidate sgRNAs were then screened to reveal the on-target mutation spectrums, presented in Figure 2.9. Of the four sgRNAs tested, sgRNA2b exhibited the highest cutting efficiency, with mutations present in the highest number of cells (represented here as number of cells rather than a percentage). In particular, 1 bp deletion appeared by far to be the most common indel, detected in almost 60,000 of the cells, followed by 2bp deletion, 3 bp deletion and 14 bp deletion, detected in almost 10,000 cells, and finally a 1 bp insertion detected in approximately 2250 cells. The mutational profiles of sgRNA2b appeared to have produced smaller and more consistent indels to the rest of the sgRNAs tested. Smaller deletions were preferred over the larger ones, since they would most likely not compromise the success of the targeted PCR by the potential loss of primer annealing site. Taken together, sgRNA2b was selected as the sgRNA to be used in all subsequent experiments and for the targeting of *POU5F1* in human zygotes.

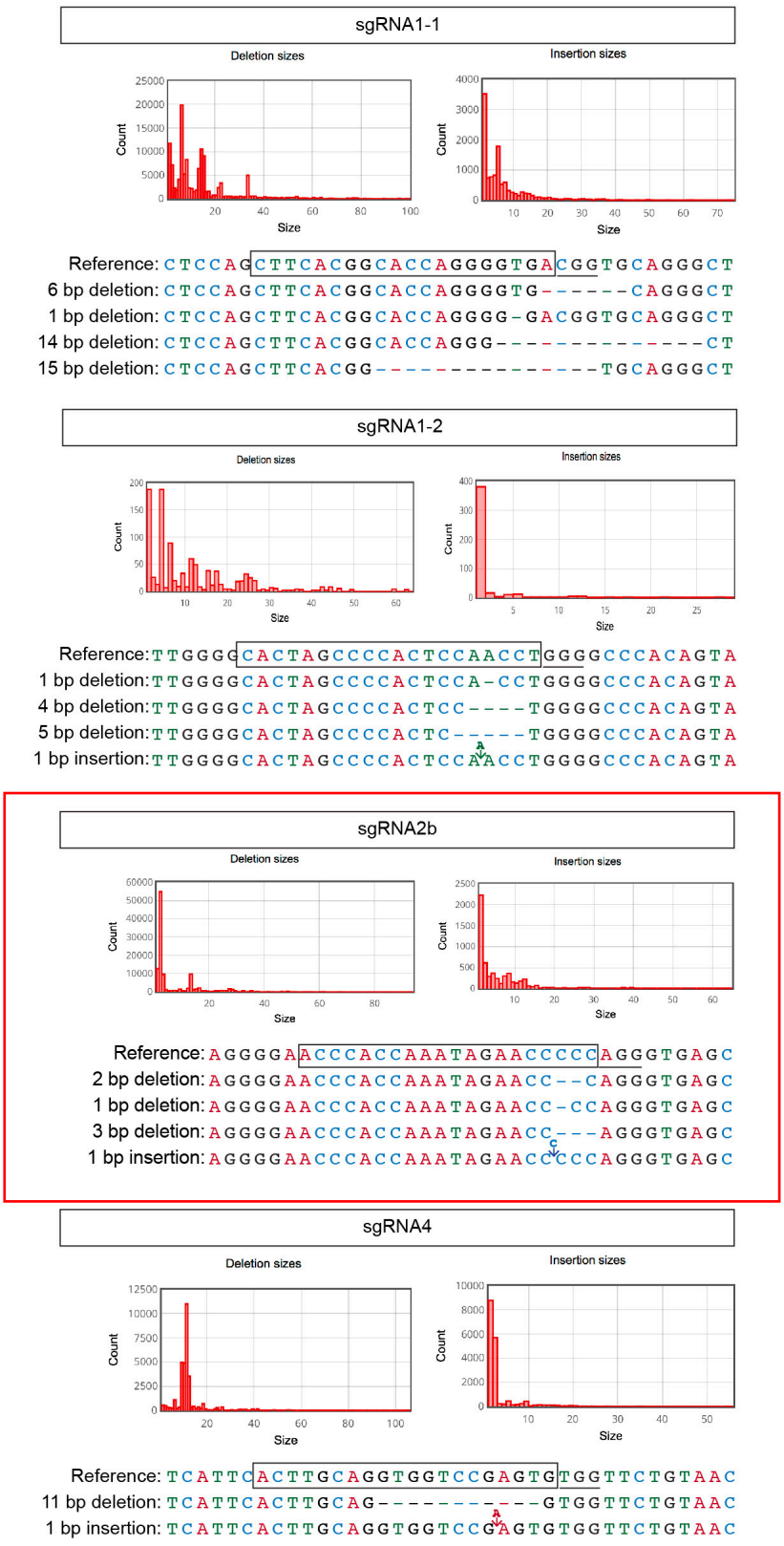


Figure 2.9: On-target mutation analysis in human hESC induced to express sgRNA1-1, sgRNA1-2, sgRNA2b or sgRNA4. CRISPR Cas-analyzer was used to show the frequent types of indel mutations and corresponding sequences observed in human ES cells induced to express sgRNA1-1, sgRNA1-2, sgRNA2b or sgRNA4. The cells were induced to express each sgRNA for 4 days and the data shown are representative of the types of indel mutations observed in other clonal lines (n = 2 sgRNA1-1 clones; n = 3, sgRNA 1-2, 2b or 4 clones) and across time (from 1 to 4 days following induction of each sgRNA). The best preforming sgRNA2b is marked in red rectangle (Fogarty et al., 2017).

2.3.2 Whole genome amplification and targeted PCR validation on clumps of cells and single cells

The next step after the sgRNA selection was to determine whether WGA would be capable of producing suitable DNA templates for PCR amplification of the targeted site and to determine the most suitable technique for WGA. WGA was required as an intermediate step in order to allow a sufficient amount of DNA to be generated for all the subsequent analyses, including cytogenetic testing (covered in Chapter 3). Two WGA methods were evaluated, Sureplex and MDA. The Sureplex system can be successfully combined with comprehensive chromosome screening and was therefore preferred the technique (Wells et al., 2014, 2008). However, since the generated fragments are significantly shorter than those obtained with MDA (on average 650 bp vs. 3 kb, respectively), it remained to be shown whether the portion of exon 2 of *POU5F1* can be successfully amplified in the targeted PCR. To asses this, clumps of buccal cells were isolated subjected to Sureplex or MDA, followed by targeted PCR and resolved via agarose gel electrophoresis. The results showed successful amplification of the *POU5F1* locus with both techniques, as represented by 244 bp bands in Figure 2.10.

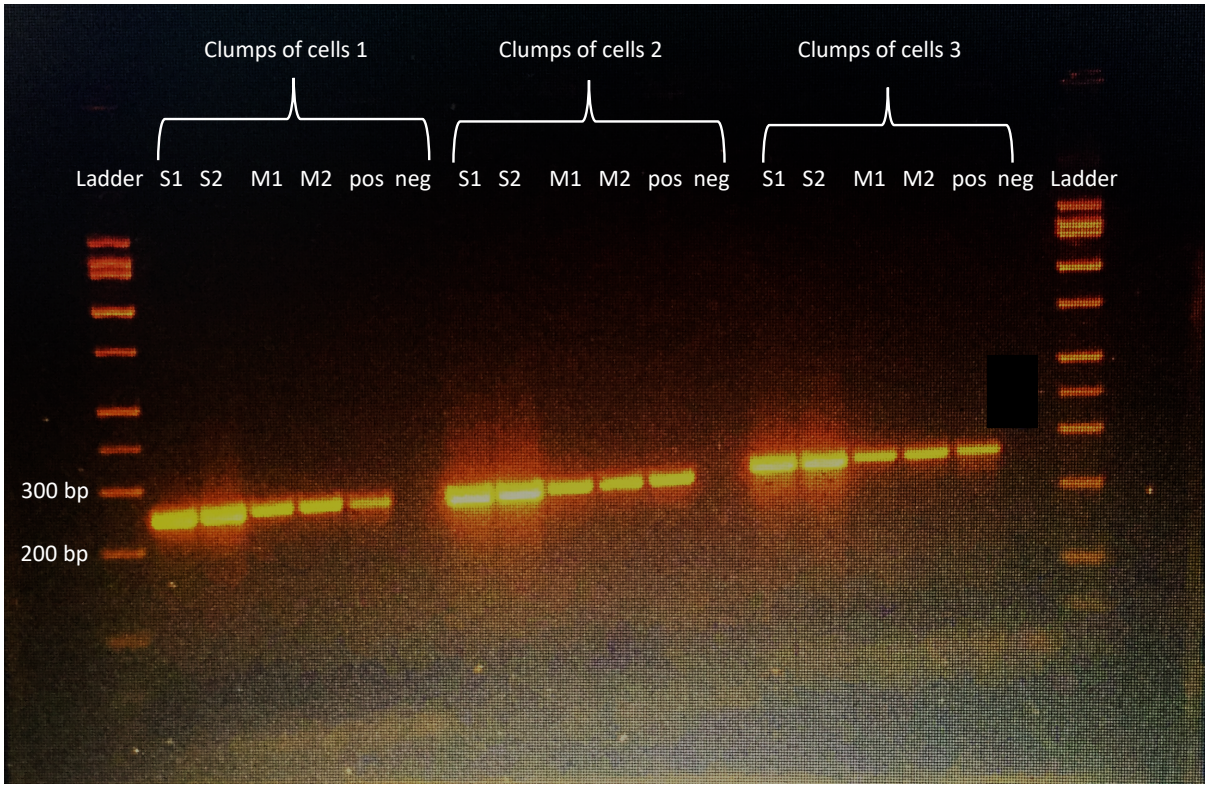


Figure 2.10: The gel image of *POU5F1* amplicons amplified from isolated clumps of cells subjected to Sureplex (S1, S2) or MDA (M1, M2), performed in triplicates. Pos: positive control – genomic DNA. Neg: negative control – PCR master mix. Expected amplicon size: 244 bp.

We next assessed whether Sureplex amplification, followed by targeted PCR of the *POU5F1* locus could be successfully performed in isolated single cells, since it was likely that when it came to the human embryo samples, the embryos would in some cases have to be disaggregated into single blastomeres (e.g. in case of two cell embryos where one blastomere was subjected to on-target analysis and the second blastomere to a transcriptomic analysis). Similarly to the clumps of cells, single cells were subjected to Sureplex amplification, followed by targeted *POU5F1* PCR, and the resulting amplicons resolved on an agarose gel (Figure 2.11). The results were encouraging, with all six single cells showing successful amplification of exon 2b.

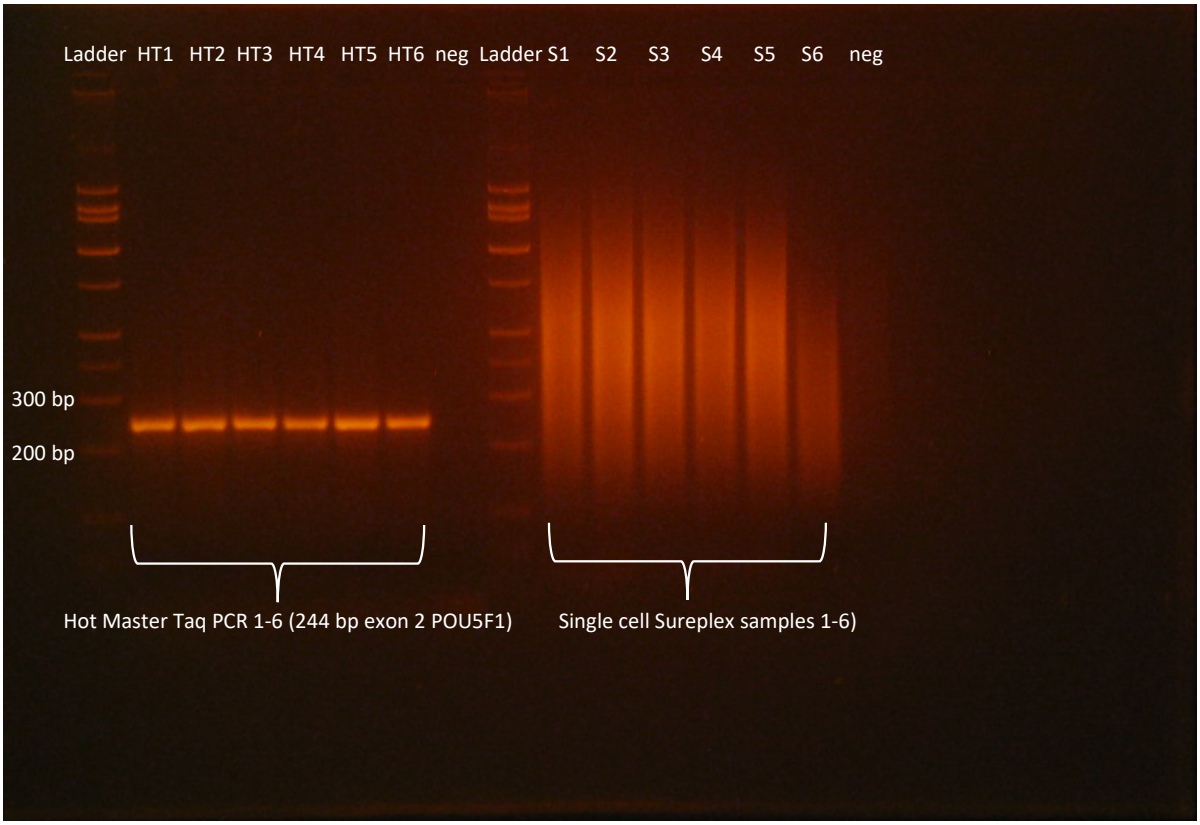


Figure 2.11: The gel image of *POU5F1* amplicons amplified from isolated single cells subjected to Sureplex (S1- S6), followed by targeted PCR (HT1-HT6), performed in replicates of six. Neg: negative control – Sureplex master mix, PCR master mix. Expected amplicon size: 244 bp.

After the sgRNA2b generation and ribonucleoprotein (RNP) preparation, 22 human zygotes were microinjected with sgRNA2b-Cas9 and 7 were microinjected with Cas9 RNP as controls (all injections undertaken by Dr. Kathy Niakan). From the resulting embryos, we obtained a total 33 samples consisting of individual blastomeres, several blastomeres together or a sample of trophectoderm (in some cases multiple biopsies were received from the same embryo). All biopsied samples were subjected to WGA by Sureplex, followed by targeted PCR to amplify the *POU5F1* exon 2b fragment. The success of Sureplex amplification was evaluated by agarose gel electrophoresis, by observing a characteristic smear in the range of 200-1000 bp. A representative image taken after processing a batch of biopsies from the targeted embryos is shown in Figure 2.12.

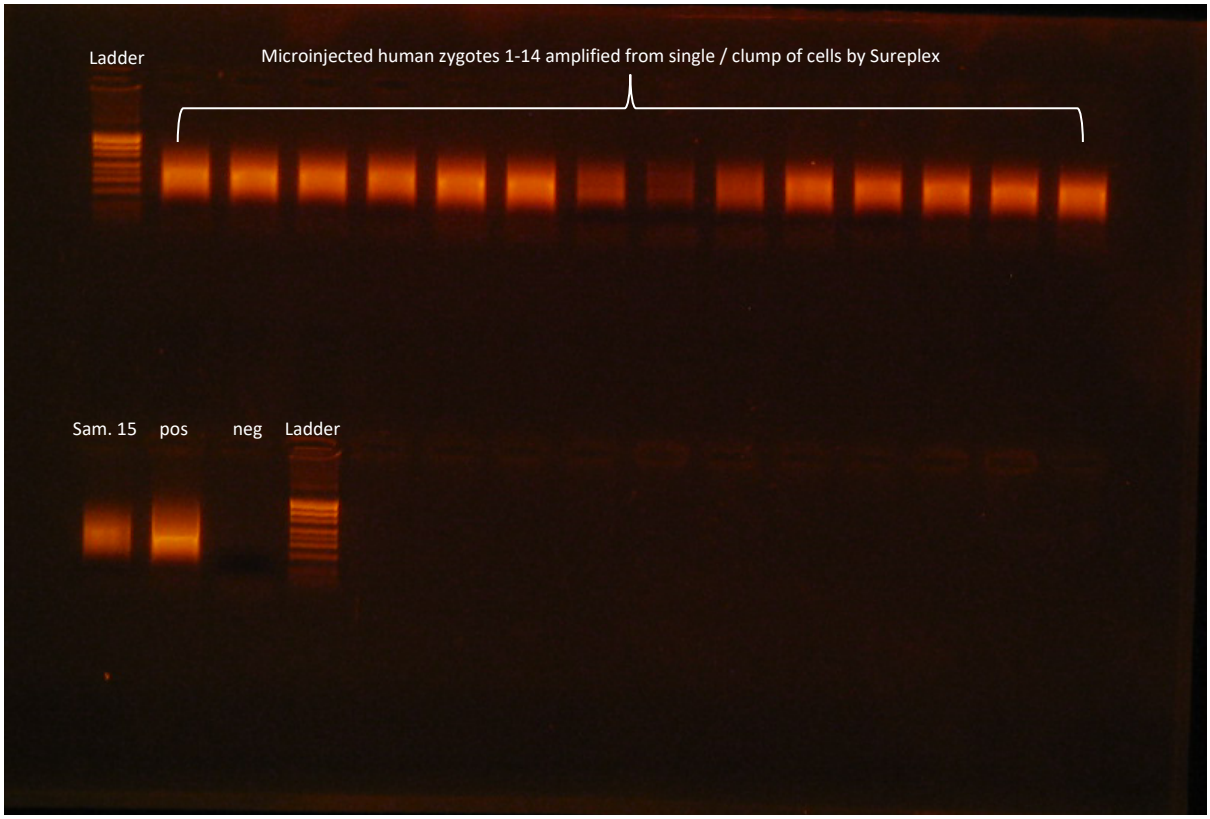


Figure 2.12: The gel image of Sureplex products amplified from *POU5F1*- targeted zygotes (samples 1-15). Pos: positive control – control genomic DNA. Neg: negative control – Sureplex master mix.

Of the samples tested, only three did not amplify at all (samples C5, C6 and C11). Three additional samples showed evidence of poor amplification (weak smears), but these samples were nonetheless processed further and included in the subsequent analyses. Overall, the amplification rate was 91% (Table 2.6).

Table 2.6: Summary of sample IDs, success of amplification embryo biopsy details.

	EMBRYO	BIOPSY ID	SAMPLE TESTED	AMPLIFICATION SUCCESS
CRISPR-CAS9sgRNA EDITED EMBRYOS	C1	C1	Entire 3-cell arrested embryo	yes
	C2	C2	3-5 cells from the TE	yes
	C3	C3	1 cell from 8-cell arrested embryo	yes
	C4	C4	1 cell from 2-cell arrested embryo	yes
	C5	C5	1 cell from 5 cell arrested embryo	no
	C6	C6	1 cell from 2-cell arrested embryo	no
	C8	C8(1)	3-5 cells from the TE	yes
		C8(2)	3-5 cells from the TE	yes
		C8(3)	3-5 cells from the TE	yes
	C9	C9	2-3 cells from the TE	yes
	C10	C10	1 cell from 8 cell arrested embryo	yes
	C11	C11	1 cell from 8 cell arrested embryo	no
	C12	C12(1)	3-5 cells from the TE	yes
		C12(2)	3-5 cells from the TE	yes
	C13	C13	1 cell from 5-cell arrested embryo	yes
	C14	C14	1 cell from 2-cell arrested embryo	yes
	C15	C15	1 cell from 6-cell arrested embryo	yes
	C16	C16(1)	3-5 cells from the TE	yes
		C16(2)	3-5 cells from the TE	yes
	C17	C17	1 cell from 4-cell arrested embryo	intermediate (weak smear)
	C19	C19	1 cell from 8 cell arrested	intermediate (weak smear)
	C20	C20	3-5 cells from TE	intermediate (weak smear)
	C21	C21	2 cells from morula	yes
	C22	C22	1 cell from 4 cell arrested embryo	yes
	C23	C23	1 cell from 6 cell arrested embryo	yes
	C24	C24	2-3 cells from TE	yes
CRISPR-CAS9 EDITED CONTROLS	5K	5K	3-5 cells from the TE	yes
	7K	7K	3-5 cells from the TE	yes
	8K	8K	3-5 cells from the TE	yes
	1K	1K	3-5 cells from the TE	yes
	2K	2K	3-5 cells from the TE	yes
	3K	3K	3-5 cells from the TE	yes
	4K	4K	3-5 cells from the TE	yes

Products from Sureplex WGA subsequently underwent a further targeted PCR to enrich the samples for the 244 bp *POU5F1* exon 2b fragment prior to sequencing, with the aim of revealing the mutational spectra generated in the edited embryos. The results obtained from the sgRNA2b microinjected embryos are presented in Figure 2.13. Of the 23 Sureplex products obtained from the sgRNA2b edited embryos, seven failed the targeted PCR: samples C2, C8(3), C13, C17, C19, C20 and C23. The latter three were the insufficiently amplified samples and, therefore, failure of the targeted PCR was not unexpected. All of the failed samples were then subjected to repeated targeted PCRs, including using a set of alternative primers targeting the same portion of *POU5F1* (exon 2b) but using different annealing sites. Additionally, any sample that failed targeted amplification three times was subjected to targeted amplification using a third pair of primers, in this instance designed to encompass a larger fragment (800 bp and 1000 in two separate amplicons), in order to reduce the possibility of failed amplification as a result of a deletion of a few tens of base pairs, including one of the primer annealing sites.

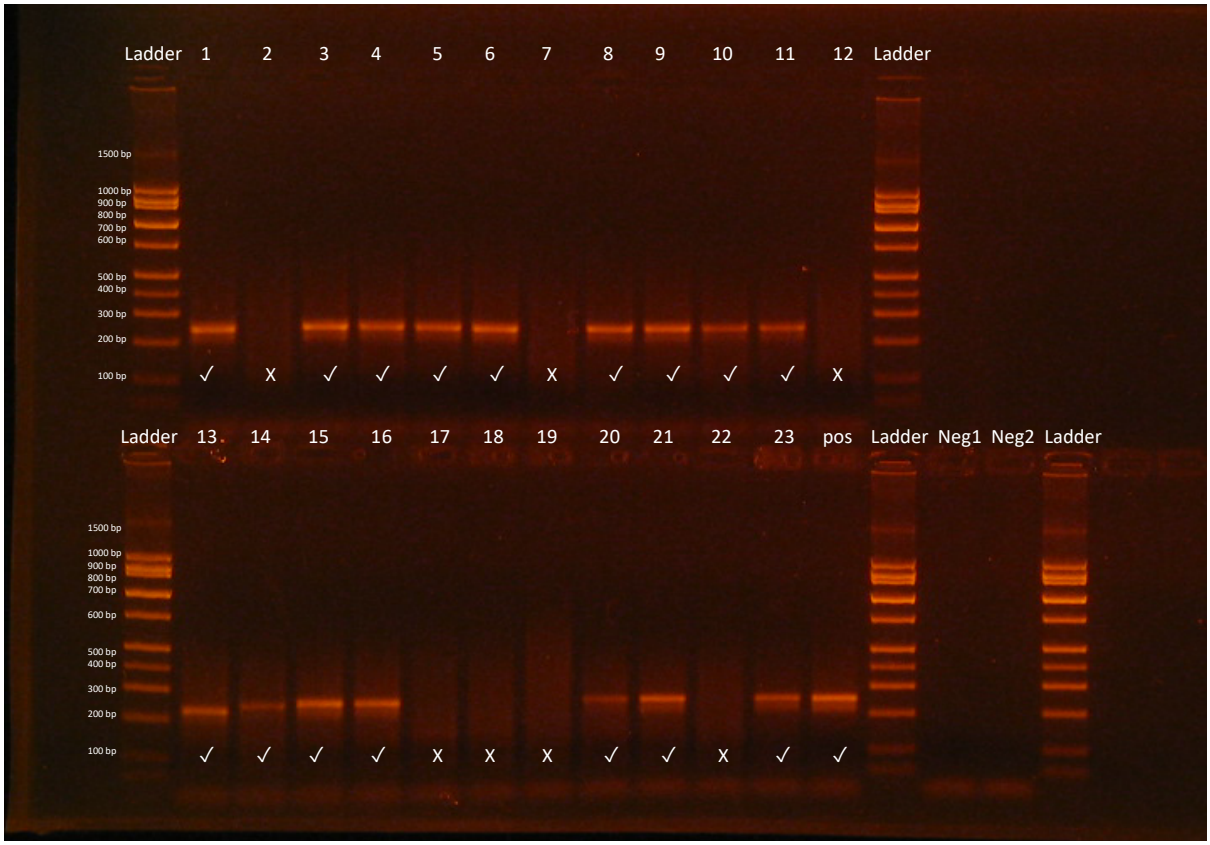


Figure 2.13: The gel image of *POU5F1* amplicons amplified from the *POU5F1* targeted human embryos subjected to Sureplex (samples 1-23). Pos: positive control – genomic DNA. Neg1: negative control – Sureplex master mix. Neg2: negative control - PCR master mix. Expected amplicon size: 244 bp.

Unfortunately, the utilisation of alternative primers did not succeed in generating additional data for any of the failed samples. One possible explanation for this is that CRISPR-Cas9 editing could have caused a large deletion, ranging over hundreds (or even thousands) of base pairs of DNA, eliminating one or both annealing sites for all of the different pairs of primers described above. To explore this possibility, seven additional primer pairs were designed in overlapping amplicons to encompass a larger portion of the *POU5F1* (up to 1500 bp). The idea was to test these individually (each amplicon was ~200 bp in length). If individual amplicons close to the CRISPR-Cas9 targeted site failed to yield PCR products, but were flanked by amplicons that successfully amplified, then a forward primer from an amplified fragment upstream could be paired with a reverse primer from an amplified fragment downstream, generating a PCR fragment (keeping in mind that if the deletion was present, it would most

likely span across only one or two annealing sites). These additional PCRs also failed to generate products in all tested samples. It is possible that in some cases deletions induced by CRISPR-Cas9 span across much larger regions, possibly loosing multiple primer annealing sites and interfering with the amplification of those alleles. Unfortunately, the success of conventional PCR in such scenarios would likely be very limited. An alternative explanation could be that the region spanning the on-target site could not be amplified due to unresolved DSBs induced by CRISPR-Cas9. A long-read sequencing approach combined with a PCR-free *POU5F1* fragment enrichment strategy might be much more suitable for this application. This approach was later developed and tested, and the results will be reported later in this chapter.

2.3.3 On-target genotype analysis in POU5F1 targeted human embryos

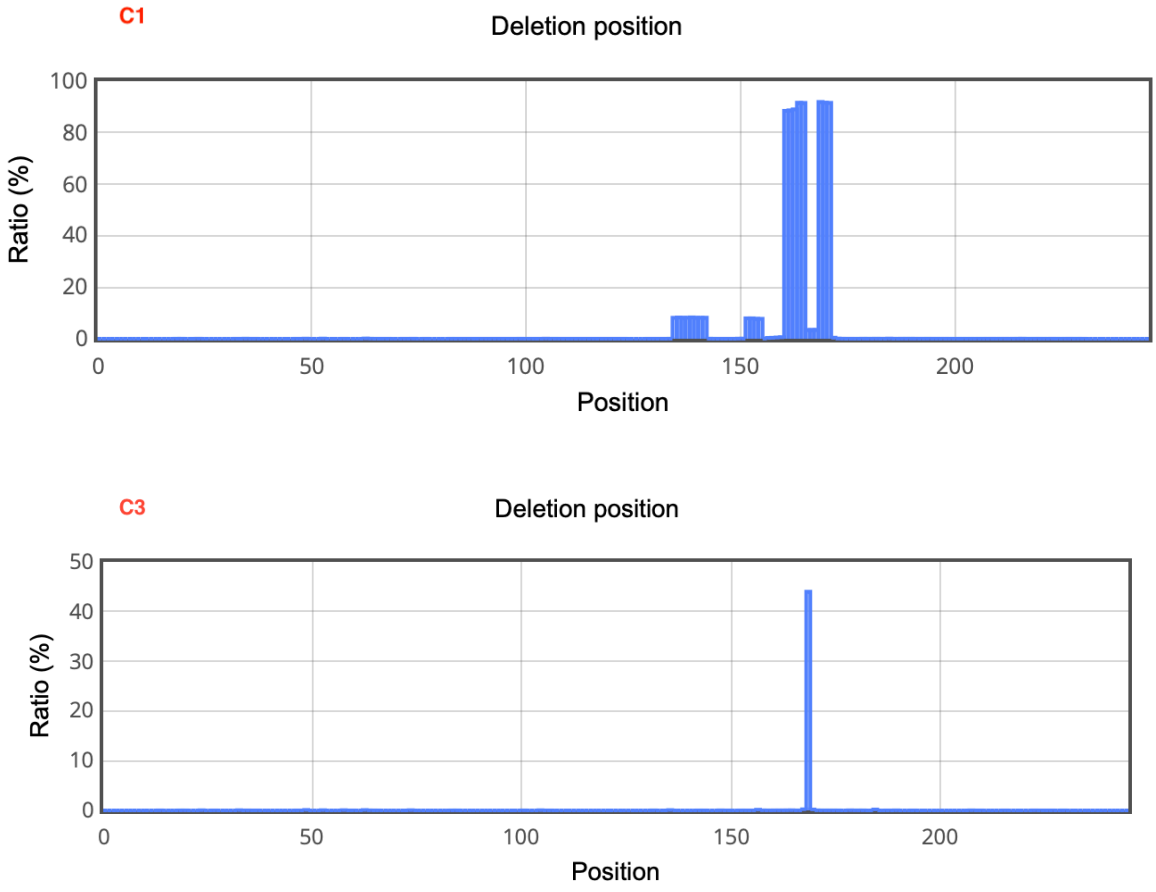
The 16 sgRNA2b-edited embryo samples that successfully amplified by Sureplex and targeted PCR were subjected to deep targeted sequencing on a MiSeq along with three Cas9-microinjected controls. Seventeen of these samples generated a result, allowing interrogation of the 244 bp fragment encompassing the CRISPR-Cas9 targeted site within exon 2b of the *POU5F1* gene. While the control samples and the unsuccessfully edited samples were expected to be associated with unmodified gene sequence, it was anticipated that edited embryos would include indels at the cut site or in the near proximity, corresponding to the mutational spectra observed in the sgRNA2b edited hESCs. The generated fastq files containing the sequenced reads were uploaded onto CRISPR Cas-Analyzer along with the hg19 reference fasta file. The reads were then aligned against the reference in the specified interval to filter out the reads that contained unmatched bases and indels. The summary of the genotypes obtained from this analysis are presented in Table 2.7.

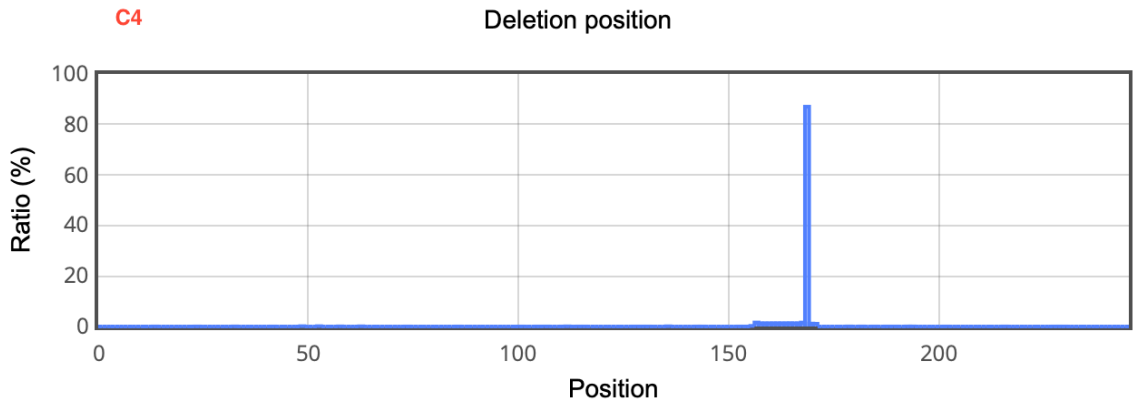
Table 2.7: Summary of the on-target genotypes in sgRNA2b-Cas9 targeted human preimplantation embryos. The indel frequency is defined as the proportion of reads with indels detected.

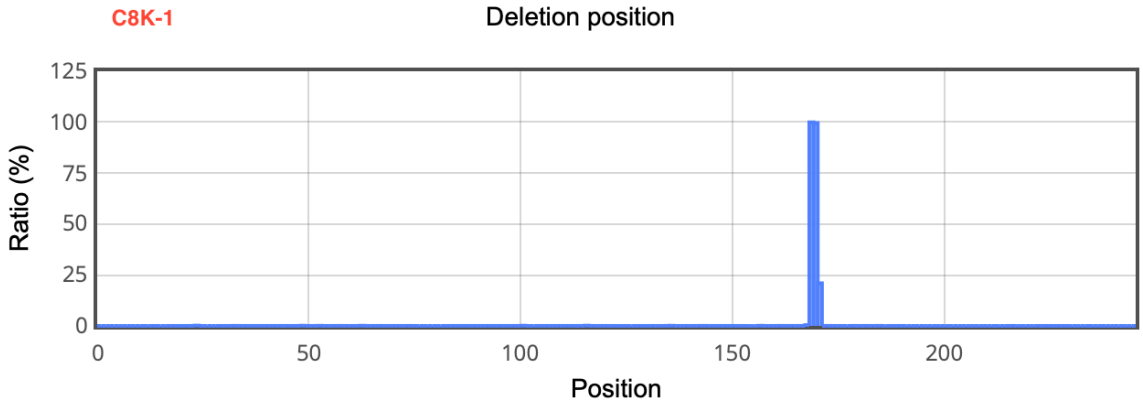
	Embryo	Indel Frequency	Indel genotype	Number of distinct mutations detected by Cas-Analyzer
CRISPR-Cas9 sgRNA Edited Embryos	C1	86%	8 bp dels and one more complex indel	3
	C3	45%	1 bp del	1
	C4	87%	1 bp del	1
	C8	100%	2bp del / 3 bp del	2
		100%	2bp del / 3 bp del	2
	C9	100%	2bp del / 3 bp del	2
	C10	66%	1 bp del	1
	C12	100%	1 bp del	1
		100%	1 bp del	1
	C14	100%	23 bp del	1
	C15	100%	7 bp del	1
	C16	52%	2 bp del	1
		71%	2 bp del	1
	C21	0%	none	0
	C22	0%	none	0
	C24	0%	none	0
Cas9 microinjected Controls	5K	0%	none	0
	7K	0%	none	0
	8K	0%	none	0

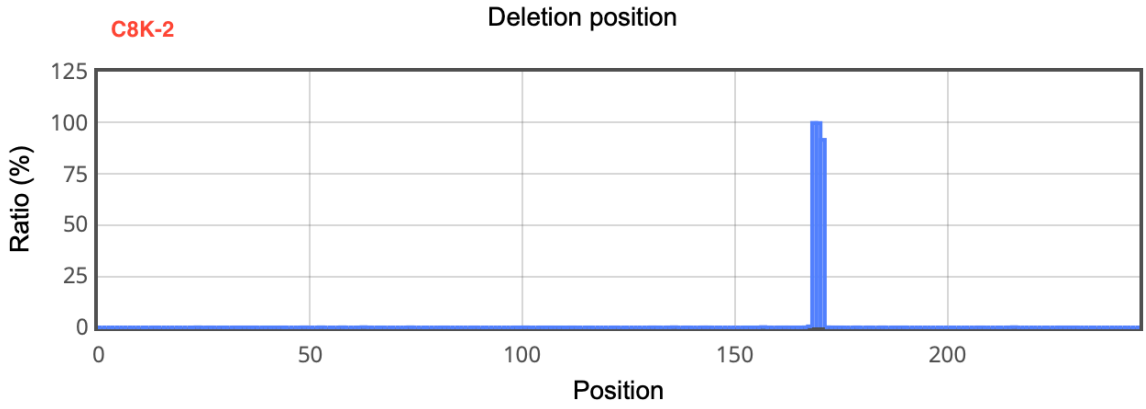
Overall, there were three embryos in which no indels were detected, suggesting that editing had not taken place, or, less likely, that repair had involved HDR using the other gene copy as a template (embryos C21, C22 and C24). These embryos generated results comparable to profiles obtained from unedited Cas9 microinjected controls (5K, 7K, 8K). The rest of the embryos showed evidence of editing, either affecting specific proportion of the reads or affecting 100% of the reads. Given the nature of the indels, the gene was predicted to be knocked out since any indel where the insertion or deletion is not a multiple of 3bp will result in a frameshift mutation and, very likely, the introduction of a stop codon. Whenever 100% of all reads were modified, it was assumed that both maternal and paternal genomes were successfully edited, and this was usually supported by observing two or more distinct mutational genotypes (embryos C8, C9). Embryos C14 and C15, although 100% knock-outs, presented with only one type of indel (23 and 7 bp deletion, respectively). The visual inspection of the reads in IGV later confirmed that both amplicon sequences contained an additional SNP where the samples tested heterozygous, precluding the possibility of ADO. The resulting genotype must have, therefore, been identical in both parental alleles. When only a proportion of the reads were detected to be edited in the range of 40-66%, it was assumed that only one of the two pronuclei was successfully edited and the embryos were heterozygous for editing, and this was usually supported by the presence of a single distinct mutation (embryos C3, C10, C16). In embryos C1 and C4, more complex deletion patterns were obtained, not easily explained from the results of the Cas9-Analyzer. Therefore, the aligned BAM files were uploaded onto IGV and inspected manually. In embryo C1, only one type of mutational pattern was detected, identical to the one present in 86% of the reads detected by Cas9 analyser (8 bp deletion). In embryo C4, two distinct genotypes were detected, one incorporating a 1 bp deletion identical to the one found by the Cas9 analyser, and the second genotype incorporating a 2 bp deletion and a 5 bp deletion. This embryo was therefore classified as a knock-out but

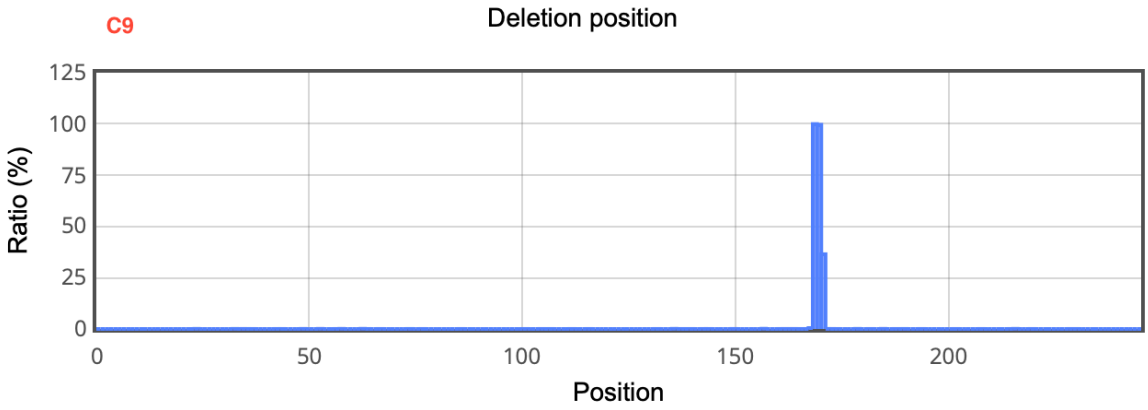
heterozygous for two different types of mutations, most likely a result of differential editing of maternal and paternal genomes. All observed deletions were present in the expected location approximately 3 bp from the PAM site of the respective sgRNA2b. The deletion patterns obtained from individual embryos are presented in Figure 2.14. Overall, the analysis of the generated single-cell amplified genomic DNA samples confirmed on-target genome editing in all but three micro-injected embryos and a stereotypic indel pattern with the majority of samples exhibiting a 1 bp, 2bp or 3bp homozygous deletion, characteristic of NHEJ.

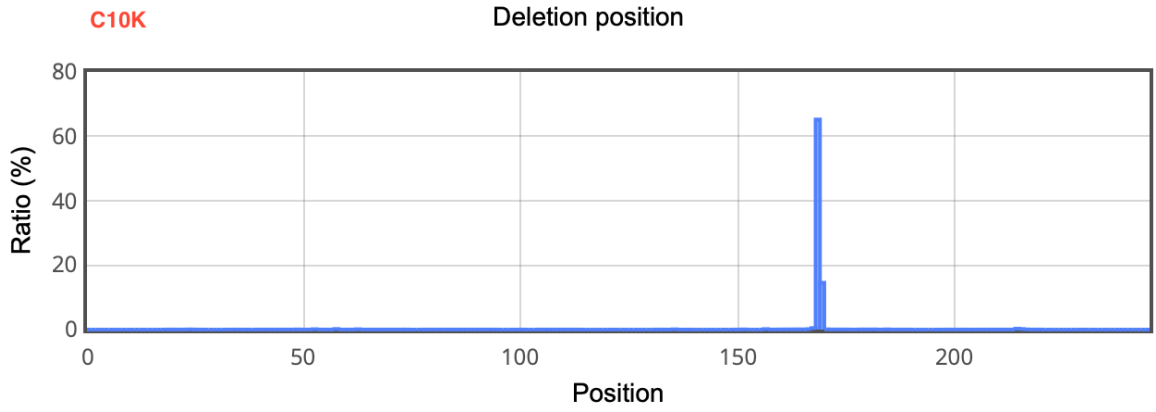


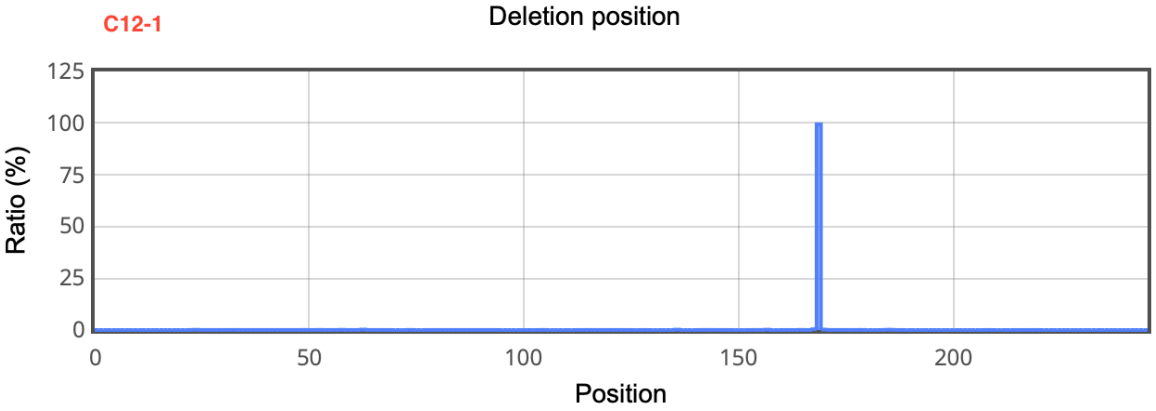

C4
Deletion position
Ratio (%)
100
80
60
40
20
0
0
50
100
150
200
Position

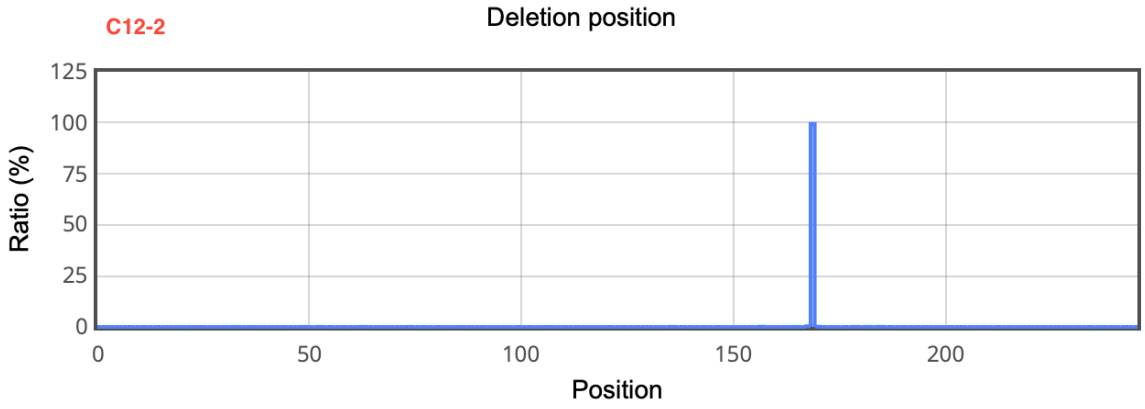

C8K-1
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

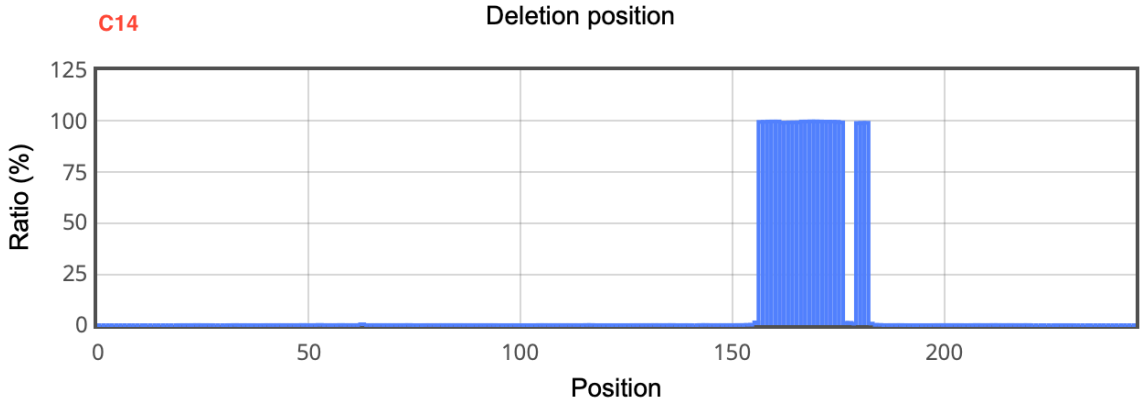

C8K-2
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

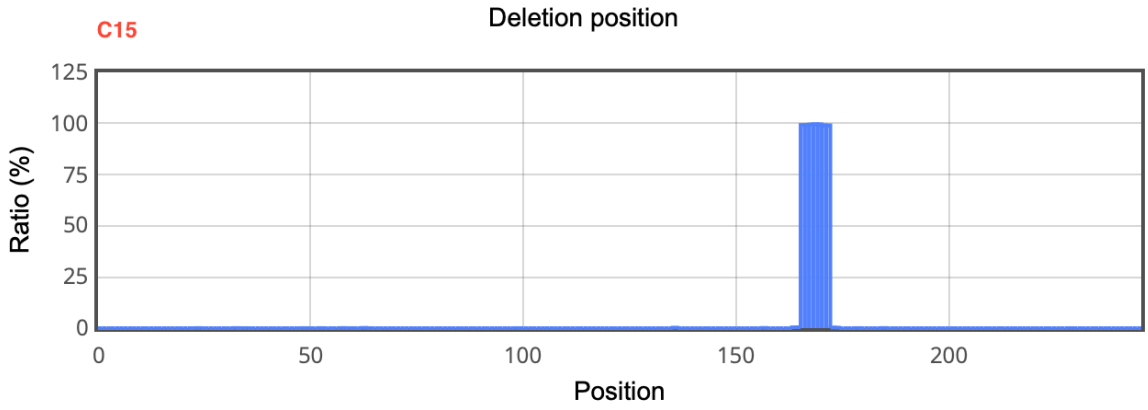

C9
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

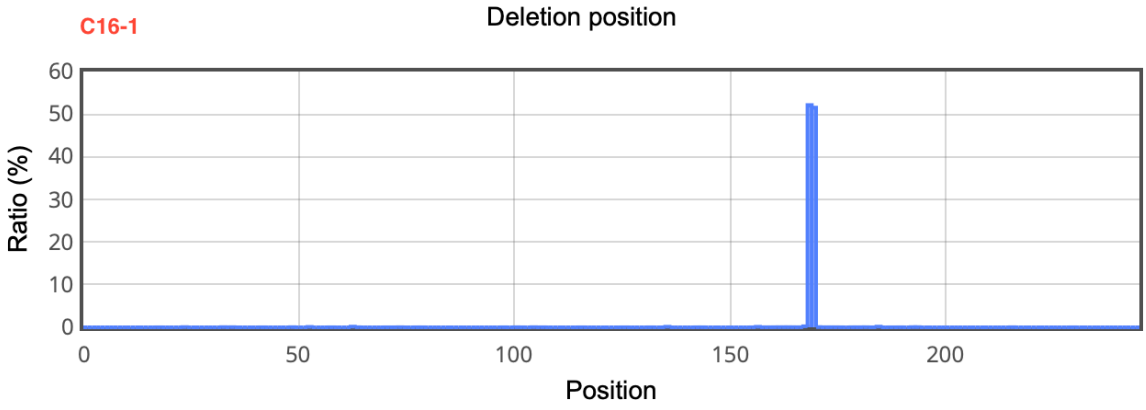

C10K
Deletion position
Ratio (%)
80
60
40
20
0
0
50
100
150
200
Position

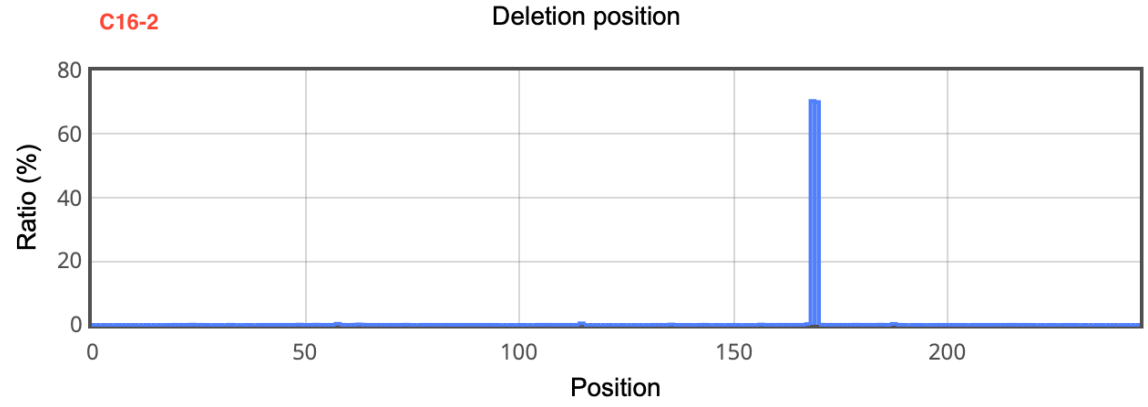

C12-1
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

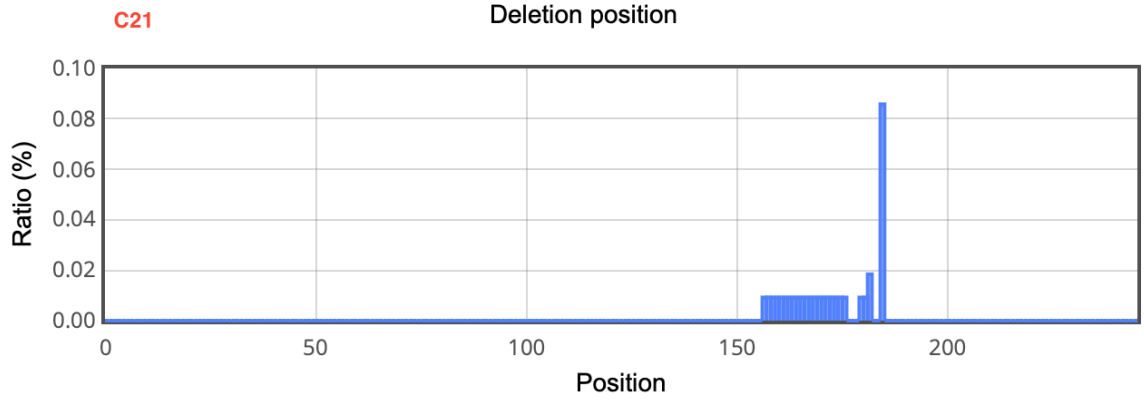

C12-2
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position


C14
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

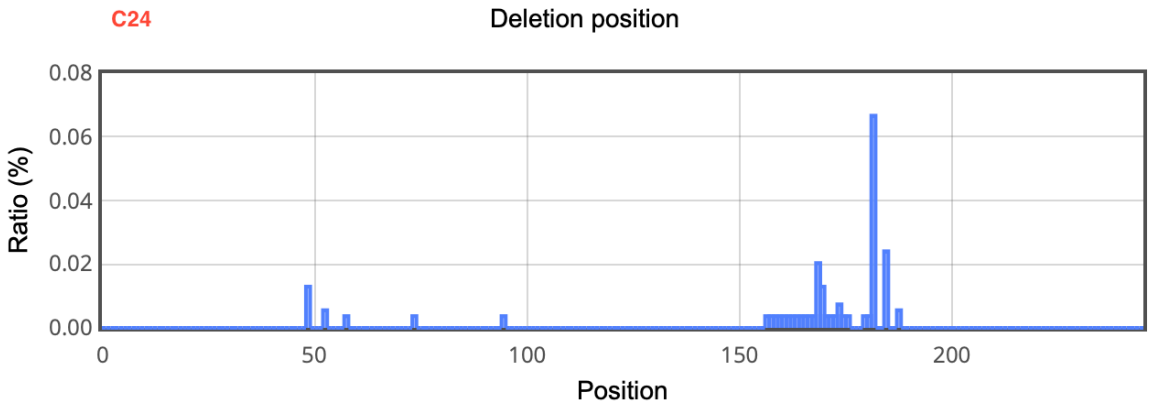

C15
Deletion position
Ratio (%)
125
100
75
50
25
0
0
50
100
150
200
Position

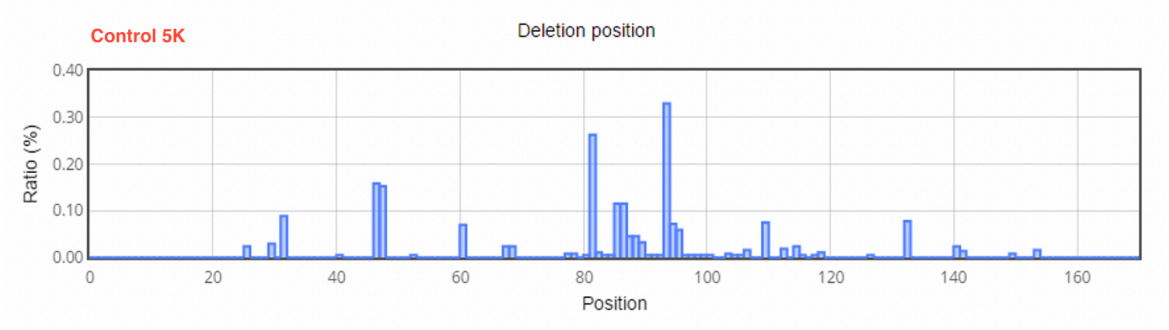

C16-1
Deletion position
Ratio (%)
60
50
40
30
20
10
0
0
50
100
150
200
Position

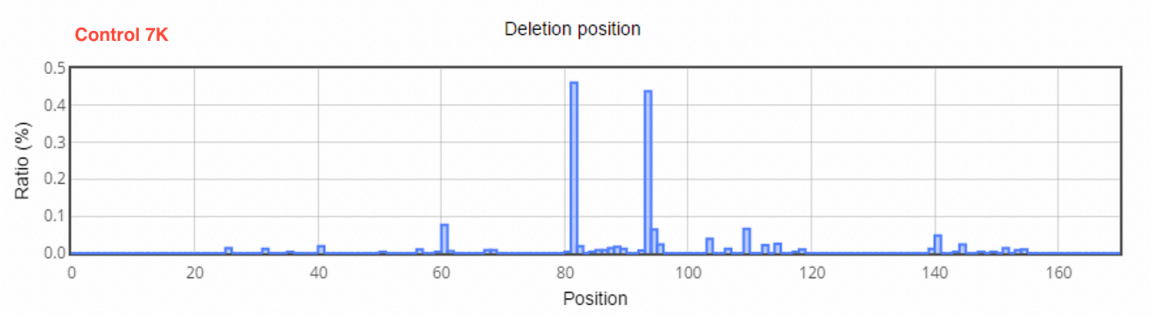

C16-2
Deletion position
Ratio (%)
80
60
40
20
0
0
50
100
150
200
Position


C21
Deletion position
Ratio (%)
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
Position


C22
Deletion position
Ratio (%)
0.040
0.030
0.020
0.010
0.000
0
50
100
150
200
Position


C24
Deletion position
Ratio (%)
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
Position


Control 5K
Deletion position
Ratio (%)
0.40
0.30
0.20
0.10
0.00
0
20
40
60
80
100
120
140
160
Position


Control 7K
Deletion position
Ratio (%)
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
120
140
160
Position

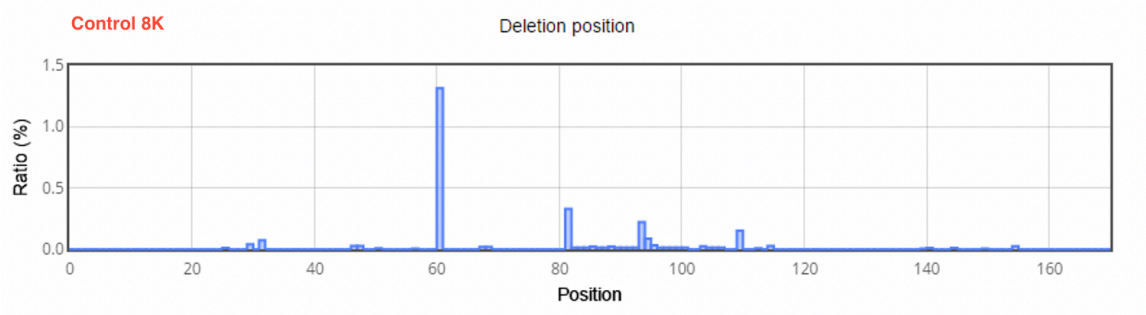


Figure 2.14: Mutational profiles of sgRNA2b-edited embryos and Cas9-microinjected controls. The ratios (%) represent the percentage of all reads at a given position of the sequence, in which a particular deletion was detected. E.g. in embryo C12-1 100% bases contained the designated mutation. In control embryo 8K, the percentage of modified bases is undetectable from the background (<1.5%). The x-axis represents the position in the amplicon (244 bp).

2.3.4 Off-target analysis

The analysis of putative off-target sites initially involved the identification of homologous sequences with substantial identity, obtained from the MIT CRISPR design tool, presented in Figure 2.15.

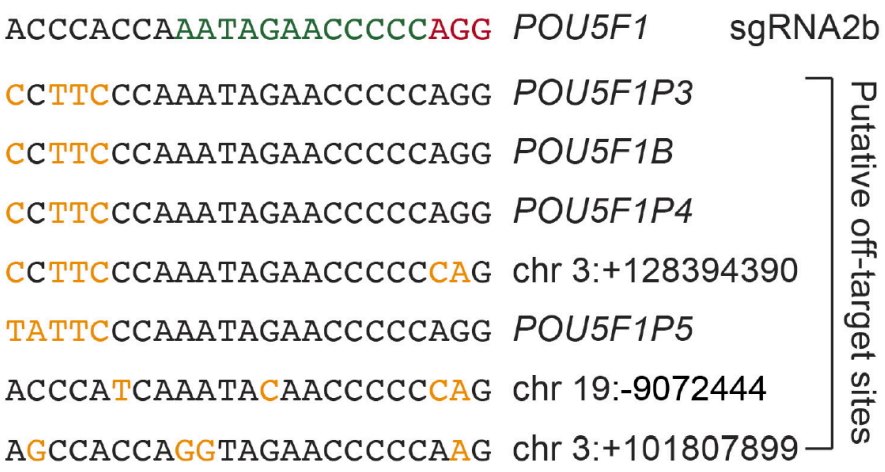


Figure 2.15: Putative off-target sequences and gene annotation. The top sequence represents the original target sequence sgRNA2b. The subsequent sequences represent the putative off-targets of sgRNA2b. The mismatched nucleotides are coloured in orange while the PAM and the seed sequences are coloured in red and green, respectively.

Since three sequences obtained from the MIT CRISPR design tool lacked the essential PAM sites (sequences chr3: +128394390, chr19: 9072444 and chr3: 101807899), we decided to exclude them from testing as the absence of PAM would have prevented the binding of Cas9 to those off-target sequences. The remaining four sequences found in *POU5F1P3, POU5F1B, POU5F1P4* and *POU5F1P5* genes were simultaneously amplified in a targeted multiplex PCR from the Sureplex-amplified sgRNAb2- and Cas9-microinjected embryos and analysed in parallel with the on-target site. The results of the CRISPR Cas9 analyser showed no off-target mutation genotypes present at any of these sites in sgRNA2b-microinjected embryos above the background of PCR errors, also observed in the microinjected controls. This was confirmed by visual inspection of the reads and the variant call format (VCF) files using IGV (Figure 2.16). Targeted deep sequencing revealed that indels had occurred only at the on-target site.

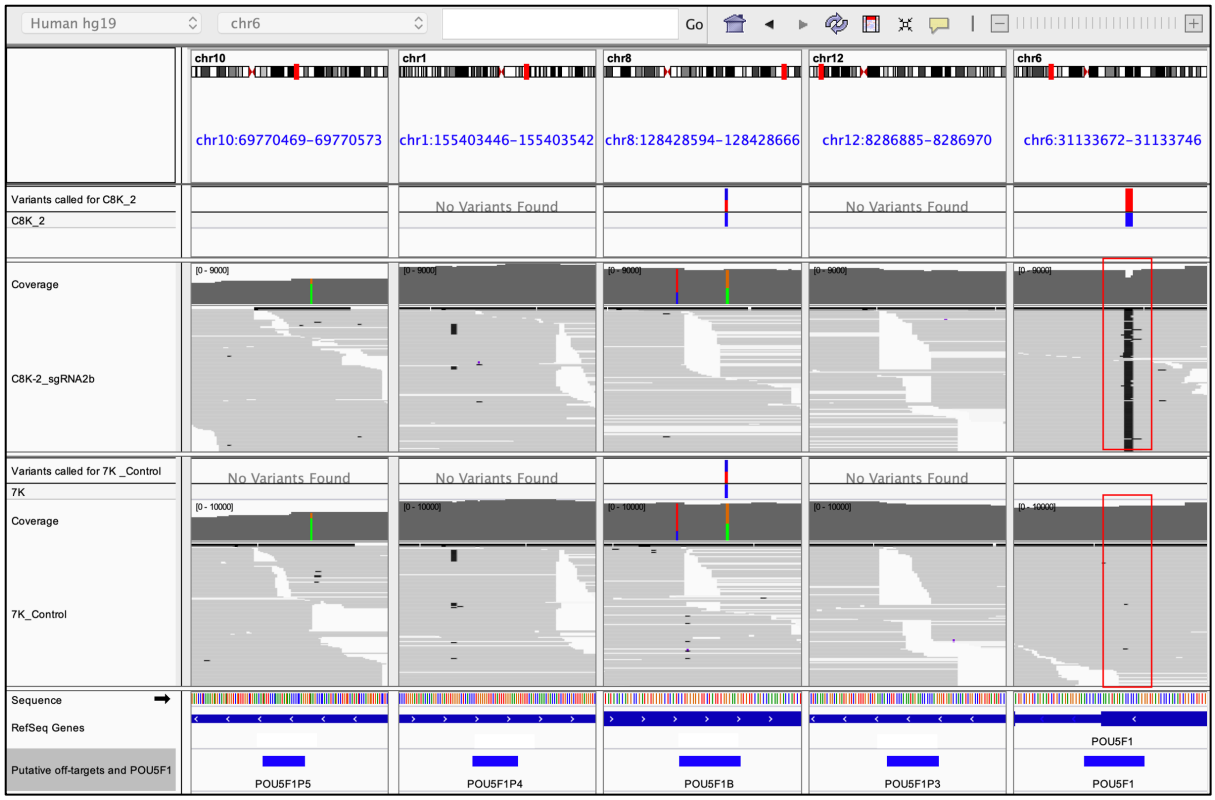


Figure 2.16: IGV plot comparing the genotype results from the putative off-target sites with the *POU5F1* locus in sgRNA2b-Cas9-edited embryo and untargeted control embryo. The top lane designates the chromosomal location with the corresponding gene annotation shown at the bottom in blue. The first four columns represent the variant analysis and read coverage in the off-target gene sequence, while the column on the right represents the actual target site of sgRNA2b. None of the sequences obtained from the targeted and control embryos in the putative off-target sites were edited, as can be seen from the unaffected darker grey and light grey read coverage. The targeted embryo sequence contains edits represented by the black bar highlighted in red. The control Cas-microinjected embryo does not contain any edits in the *POU5F1* target (also highlighted in red).

2.3.5 Development of CRISPR-Cas9 target enrichment and sequencing protocol for the analysis of the POU5F1 locus with long read sequencing by Nanopore – a proof of concept study on hES cells

The bioinformatic analysis workflow for the data obtained from the Nanopore sequencing run is represented in Figure 2.17. This method employs a suite containing multiple tools optimised to produce all the required summary statistics, figures and on-target and off-target sequence filtering.

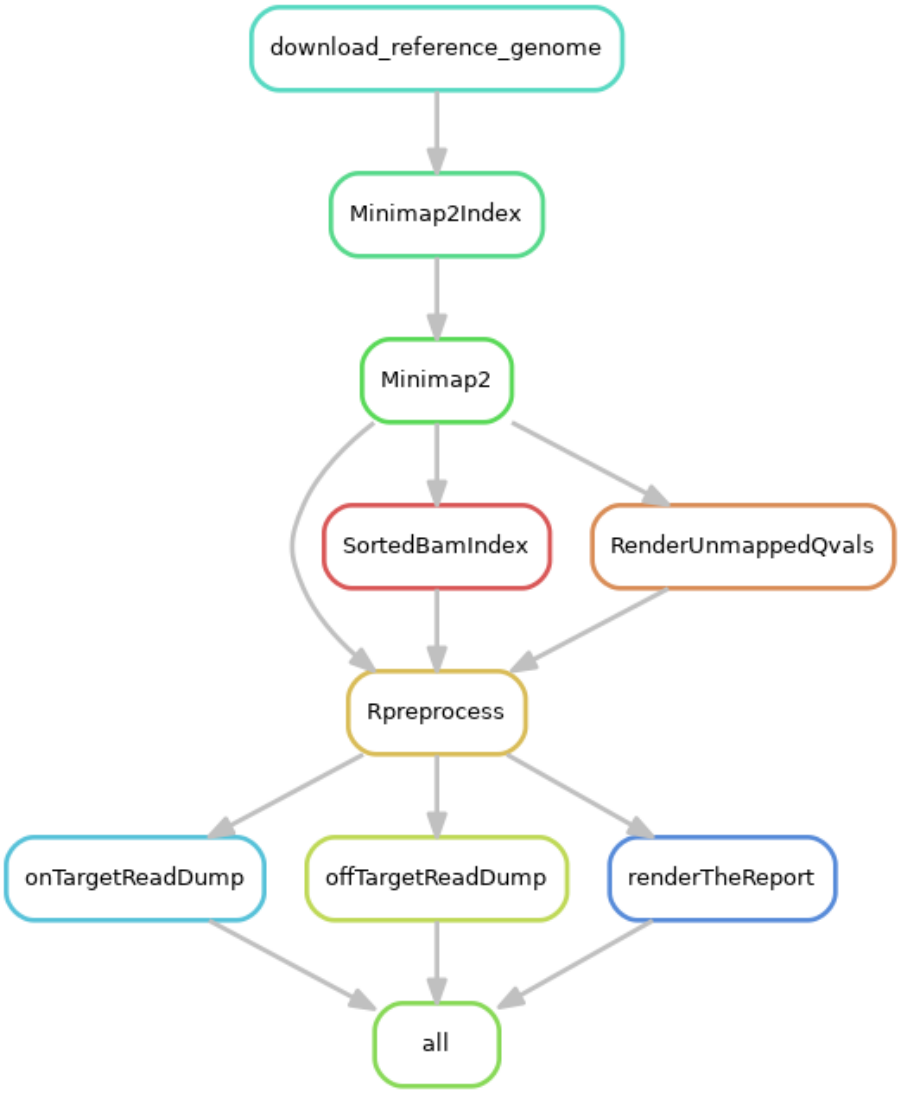


Figure 2.17: Bioinformatic workflow designating the combination of tools used for the Cas9 enrichment analysis.

Workflow:

1. Download the specified reference genome

```
rule download_reference_genome:
    input: ftp.ensembl.org/pub/release-
96/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz
    output: ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz

rule unzip_reference_genome:
    input: ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz
    output: ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa
```

2. Use minimap2 to index the reference genome

```
minimap2 -t 4 -d
ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa.mmi
ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa
```

3. Map DNA sequence reads against the reference genome index using minimap2

```
minimap2 -2 -a -x map-ont --MD -R @RG\tID:Nanopore Cas9 enrichment
tutorial\tSM:Nanopore Cas9 enrichment tutorial -t 4
ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.6.fa.mmi -
```

4. Convert minimap2 output (SAM) into a sorted BAM format using samtools

```
#Sam to bam:

samtools view -S -b input_file.sam > output_file.bam

samtools view output_file.bam | head

#Bam sort:

samtools sort input_file.bam -o output_file.sorted.bam

samtools view output_file.sorted.bam | head

Sorted bam to index:

samtools index output_file.sorted.bam

samtools view -H output_file.sorted.bam
```

5. Prepare summary mapping statistics using Rsamtools and GenomicAlignments using an R script

```
#Coverage plots
singlePlot(names(ontargetUniverse)[1], aggregatedGR)

# to plot the figure for a target called ¨Example¨ we could specify #
singlePlot('Example', aggregatedGR)

strandedPlot(names(ontargetUniverse)[1], aggregatedGR)

# to plot the figure for a target called ¨Example¨ we could specify #
strandedPlot('Example', aggregatedGR)
```

6. Filter for the on-target sequence reads using seqtk (Heng Li (2019))

```
#Executive summary

<img src="Analysis/R/cas9_FAK75845_enrichment_info.png" width="236" />

#Off-target mapping

<table class="table table-striped table-condensed" style="margin-left:
auto; margin-right: auto;">

#For full code see Appendix 3
```

The complete version of R code used to analyse the BAM reads is appreciable in size. In order to keep the thesis concise, the full code can be found in Appendix 3. The appended script was used to generate all of the tables and figures presented in the results section.

2.3.6 Long read sequencing and data analysis – coverage and read length analysis of Nanopore data

Figure 2.18 contains the basic summary of the run. Overall, the run generated 0.28 Gb of data. In terms of the on-target read coverage, only 0.08% of all reads were contained within the targeted region of *POU5F1*. This is considerably lower than the recommended 1-10%, although not unexpected as only a single relatively short region was evaluated. The obtained on-target coverage was 16.6 x (meaning that each base in the target sequence was sequenced on average 16.6 times). The widely accepted diagnostic threshold is set to 30 x, and therefore it would be worth attempting to increase the on-target coverage in the future experiments. The observed on-target coverage is most likely low as a result of low DNA input (~1 μg). The manufacturer recommends 1-10 μg of input DNA and optimisations performed here started using the minimum amount of DNA in order to establish the lower threshold at which satisfactory throughput can be obtained. Improving the coverage should be straightforward to achieve by increasing the input DNA amount. Finally, the non-target depletion refers to the fold decrease in the coverage of the non-targeted bases in comparison to what would be expected if no enrichment took place. A depletion of 3000 x would be ideal in order to better utilise the throughput capacity of a Nanopore flow cell. The observed value can most likely be explained by a) an incomplete dephosphorylation of the non-targeted DNA ends, which subsequently become available for adapter ligation and/or b) excess ligation of sequencing adapters to non-specific DNA ends. Since in a single flow cell there is sufficient capacity to produce several thousand on-target x coverage regardless of the fold depletion, the non-target depletion is not a major concern. In fact, the non-targeted reads might be useful if there was a desire to ascertain chromosome copy number in the same experiment.

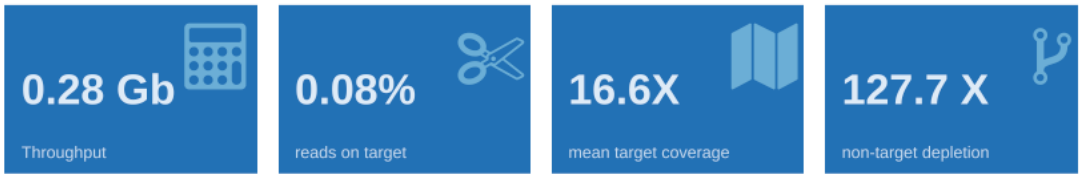


Figure 2.18: Summary of the basic quality control parameters from the Cas9 enrichment Nanopore sequencing run.

Mapping characteristics by genomic segments

Table 2.8 summarises the mapping characteristics by genomic segments. The on-target column is of main interest where in total >50,000 bases were sequenced and mapped within the *POU5F1* selected region. Due to computational constrains, the sequences were only mapped to chromosome 6 (where *POU5F1* is located), and this explains the difference between the overall mapped reads (24,478) and total sequence reads (91,225). With higher computational power it is possible align the reads globally (using the entire genome fasta file as reference as opposed to chr6 fasta). These background reads could then potentially be used to calculate the chromosome copy number by employing an additional bioinformatics pipeline. Of note, a workflow for aneuploidy detection using this sort of 'low-pass' genome sequencing has already been developed and validated on Nanopore reads (Elinati et al., 2020). This could be a useful addition to the analysis of clinical samples where it is important to determine ploidy status as well as the genotypes resulting from genome editing. An example script that could be used to determine the proportion of reads per chromosome is presented in Appendix 4.

Table 2.8: Mapping characteristics by genomic segments. Background represents the total of all sequenced reads, off-targets only the sequences that have been flagged as potential off-target ones and target-flanking the sequences that have been flagged as potential on-target ones, later filtered to actual reported on-targets, highlighted in red. * fastq bases are calculated from the field of the mapped sequences and from the sequence length of unmapped sequences † this table presents only primary sequence mappings ‡ depth of coverage based only on primary mapping reads.

	Background	Off-Target	Target-flanking	On-Target
total sequence reads*	91,225	4,898	35	23
mapped reads (primary)†	24,478	4,898	35	23
bases sequenced	261,931,539	22,427,864	131,780	56,526
bases mapped	118,820,466	22,427,864	131,780	56,526
Fraction of genome (%)	99.293%	0.693%	0.012%	0.002%
Mean coverage (primary)‡	0.15	7.68	1.76	16.63

Evaluation of individual target performance

To gain the best insight on the performance of the Cas9-mediated PCR-free enrichment protocol it is preferable to consider the performance of each discrete target separately. Table 2.9 below highlights the characteristics for the target *POU5F1* region defined within the starting BED file. A mean read quality threshold recommended by ONT is 7, and the value obtained from the work described in this thesis was within the ideal average of 9.79. In terms of mapping the reads to a reference genome, the quality value known as 'MAPQ', corresponds to a probability of an incorrect mapping and is calculated as follows:

```
10^-obtained mapping quality / 10
in this case 10^-0.6
= 0.000001
```

A perfect value of MAPQ 60 indicates that reads are mapping to a single location in the genome (the target location) and the probability of this mapping being wrong is 1 in a million. Low mapping scores (below MAPQ 30 or 40) may indicate either fragmented mapping (blocks of sequence interspersed by regions of no mapping at a single genomic location) or multi-mapping (the sequences can be mapped to multiple locations in the genome) leading to off-target effects. Reads on FWD(%) indicates the percentage of sequence reads that map to the forward strand. If this value is not in the region of 50% then one of the sgRNA probes is not working effectively. The value obtained from our experiment was 43.48, which indicates that one of the sgRNA guides works marginally more effectively, although this difference is negligible.

Table 2.9: Quality control analysis if the on-target reads. The full sequence of the region of interest (4,081 nucleotides) was obtained, represented by the target size. On average, each nucleotide in the target sequence was covered 16.63 times and this constitutes 56,526 nucleotides with corresponding average read quality and mapping quality (mapQuality = MAPQ) scores. Reads on FWD(%) indicates the percentage of sequence reads that map to the forward strand * Reads were counted as all sequence reads where the SAM start location was located within the target interval. This did not correct for sequences on the reverse strand. † Bases were counted as the sum of nucleotides from all reads where the SAM start location was within target region; some of these bases will overlap the flanking region ‡ reads are assessed for strand of mapping; here reads on + strand are summarised as percentage of all.

Target Gene	Target size (nt)	Mean coverage	Read count*	Bases†	Mean readLength	Mean readQuality	Mean mapQuality	Reads on FWD(%)‡
POU5F1	4,081	16.63	23	56,526	2,915	9.79	60	43.48

Graphical review of depth-of-coverage for POU5F1 target gene

The tables presented in the previous two sections have provided a summary of general mapping characteristics and on-target statistics. Plotting depth of coverage across the target regions also allows for an assessment of the performance of the two sgRNA guides used for sequence enrichment. The plot represented in Figure 2.19 reviews the depth of coverage and "leakiness" of sequence coverage beyond the boundaries of the *POU5F1* target region. Overall, the on-target coverage is not uniform, as indicated by the spike rising at chr6:31,177,500 and peaking in the region of chr6:31,180,000. This could be explained by one-sided adapter ligation, most likely due to one guide RNAs being more efficient in *in vitro* cleavage reaction as opposed to adapter ligation on both ends.

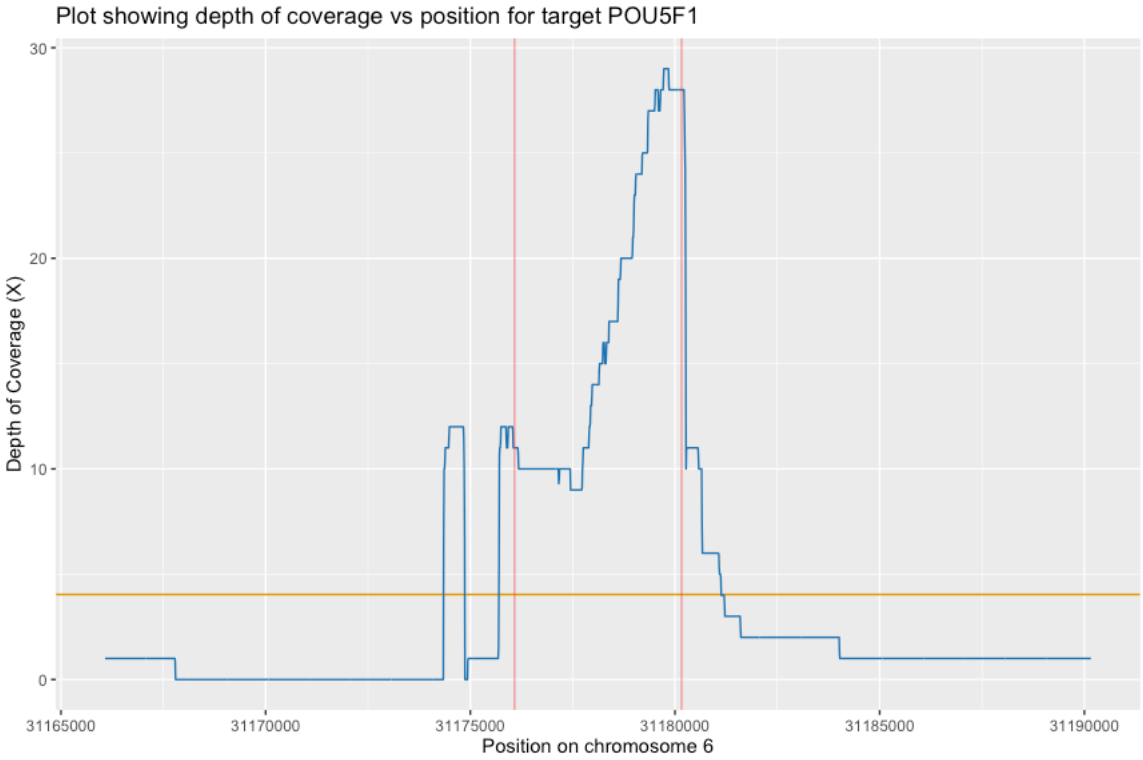


Figure 2.19: Coverage plot 1. The figure above shows the depth-of-coverage around a target region. The on-target region is located within the vertical red-bars and is flanked by the target-proximal regions. The horizontal bar shows the threshold at which an off-target feature would be defined. This plot is for the *POU5F1* target.

Figure 2.20 represents the same coverage plot but instead differentiates between the strand orientation (forward vs. reverse). For optimal results, a 50:50 ratio should be obtained. As can be seen on the plot, the forward strand reads are marginally less abundant in this case, representing 43.48% of all on-target reads. The orientation of the probes is critical to designing a Cas9 cleavage experiment, the orientation being defined by the PAM and target sequence. As Cas9 cuts at the PAM-proximal end of the protospacer 3 bp upstream of the PAM and exposes that strand, the other PAM-distal end remains bound to Cas9, protecting the end from adapter ligation. Since this process is imperfect, sometimes Cas9 can release the DNA, but on average the reads towards the PAM site outnumber the reads away by a factor of at least 3:1 and can exceed 10:1, as was evident in this experiment. For this reason, it is essential to search for guides for the forward strand upstream of the target region (i.e., <Chr6: 31,175,000) and

the reverse strand downstream (i.e., >Chr6: 31,180,000) in keeping with how the guides were designed in this experiment.

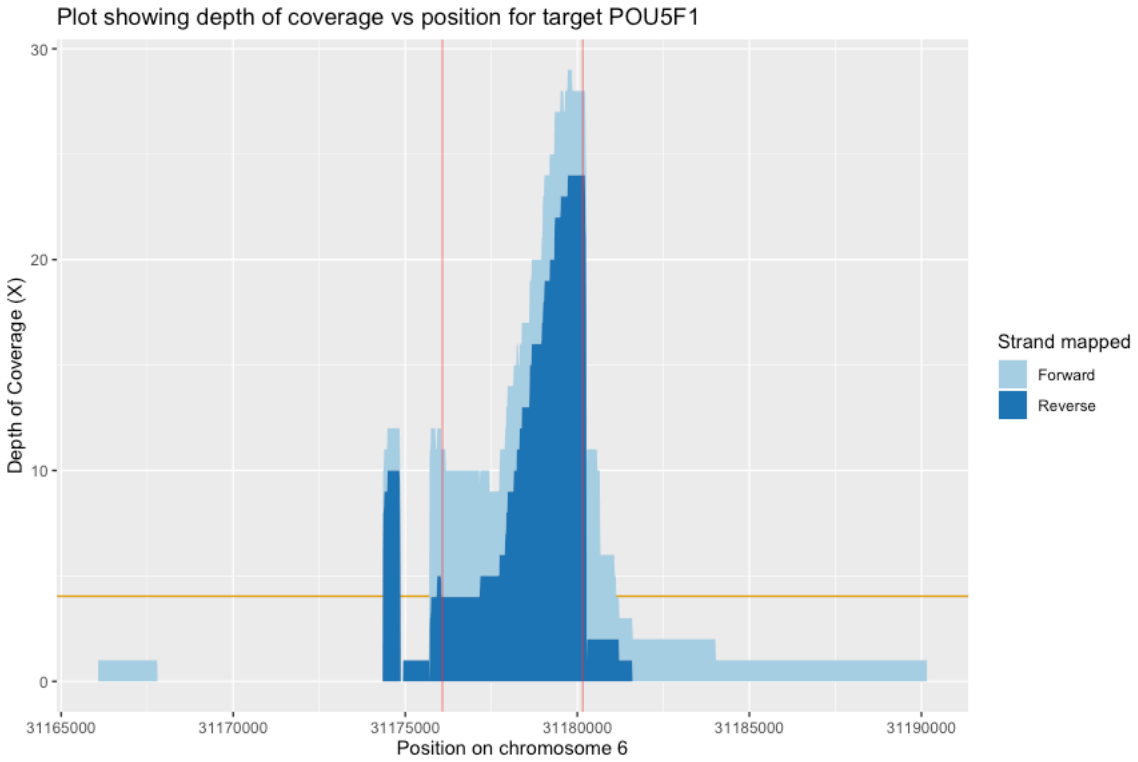


Figure 2.20: Coverage plot 2. The figure above presents the depth of coverage but is shaded by the strand (forward or reverse) to which the reads are mapped. This figure can be used to observe deviations from the expected 50:50 distribution of mapping between the + and - strands. Sequences that extend from the target regions and into the target-proximal regions may indicate suboptimal performance of a sgRNA guide sequence.

Off-target mapping

Table 2.10 represents the top 10 coordinates for potential off-target regions that have been written to an accompanying CSV file and imported into Excel for further analysis. The top 10 regions are ranked by mean depth-of-coverage (ranging between 26 x – 16 x). Although this might appear in a similar range to sequence reads obtained for the on-target *POU5F1* locus, it is necessary to consider the MAPQ values when evaluating the data. In the case of these top putative off-target reads, the MAPQ values range between 1.13 and 3.42. When these MAPQ values are plotted to determine the probability of incorrect mapping (10^-0.113=0.77 and 10^0.342=0.45), the resulting probability ranges between 45%-77%. This can be explained by

imperfect alignment of the reads to the reference sequence, and as such these represent bioinformatic artefacts rather than experimental or biological phenomena. Importantly, all of the on-target reads were of MAPQ 60, while there are virtually no off-target entries with MAPQ above 3.5 in the top candidates and the off-target candidates.

Table 2.10: Summary of the location and characteristics for the off-target regions with the highest depth-of-coverage. Start and end represent chromosomal location. The mean read length is the mean sequence read length for the mapping reads identified; their strandedness is summarised in %FWD reads (the number of sequences that appear on the forward strand) and the read and mapping quality is summarised in MAPQ.

chrId	start	end	width	mean coverage	reads in segment	mean read length	%FWD reads	mean readQ	mean MAPQ
6	104,489,401	104,495,400	6000	26	42	6,087	50.00	9.70	1.13
6	40,236,901	40,241,700	4800	22	55	5,545	53.45	9.51	3.42
6	28,621,201	28,624,300	3100	21	37	5,842	62.16	9.65	2.31
6	78,926,401	78,929,000	2600	21	31	5,235	47.37	10.08	2.65
6	139,210,501	139,216,600	6100	20	43	5,502	48.98	9.54	1.45
6	2,584,401	2,589,500	5100	18	46	4,194	46.00	9.70	3.09
6	106,404,601	106,411,000	6400	18	42	5,917	50.00	10.00	2.15
6	169,371,801	169,372,900	1100	17	27	4,095	48.48	9.64	2.26
6	28,305,801	28,310,900	5100	17	40	6,396	56.82	9.65	3.23
6	110,539,201	110,540,900	1700	16	23	5,485	57.14	9.73	3.30

In order to corroborate this finding, potential off-target entries generated from a separate validation run, set up using the same sample and protocol but instead using genomic DNA derived from hESC rather than MDA product (the results of this run are not presented here). When comparing the off-target entries obtained from the two data sets against each other, there are almost no matched entries in the same genomic coordinates, as suspected and supporting the notion that these are not genuine off-target events caused by non-specific cleavage, but occur at random, a consequence of exposure of 5' phosphates due to DNA strand breakage or

failure to entirely remove phosphate groups prior to CRISPR-Cas9 treatment and ligation. When comparing the sequences from the top entries presented in Table 2.10 with the on-target *POU5F1* sequence using Nucleotide Blast, no identity between the two queried sequences was detected, as expected. One limitation of the analysis pipeline executed in this study is that the sequences were only aligned against chromosome 6, due to computational limitations. As a result, one could only investigate the possibility of off-target effects on this one chromosome, where *POU5F1* is located. *POU5F1* has numerous homologs (e.g. *POU5F2* or *POU5F1B*) and these are located on different chromosomes but it is not possible to investigate these from this data set. In addition, when aligning against one chromosome as opposed to the entire genome, the alignment tool attempts to "force-match" the sequences against the queried chromosome, producing many more erroneous alignments, whose MAPQ values are likely below quality thresholds. In order to enable reaching of high mapping efficiency with long-read data, aligners like Minimap2 perform so called "soft-clipping" of reads. This causes portions of the read that do not match well to the reference genome on either side of the read to be ignored for the alignment as such, but still be reported. When inspecting these in IGV, it is possible to observe that in some areas, particularly in repeated sequences, the reads are frequently soft-clipped and only the matching part of the read is reported. The next logical improvement of the data set studied here would be to filter out the reads whose MAPQ score is below 30 or 40. This will ensure the soft-clipped alignments are eliminated from the data set as are the potential off-targets. An example of genomic segment located in chromosome 6 (*KHDRBS2)* where reads have been soft-clipped and wrongfully aligned is presented in Figure 2.21.

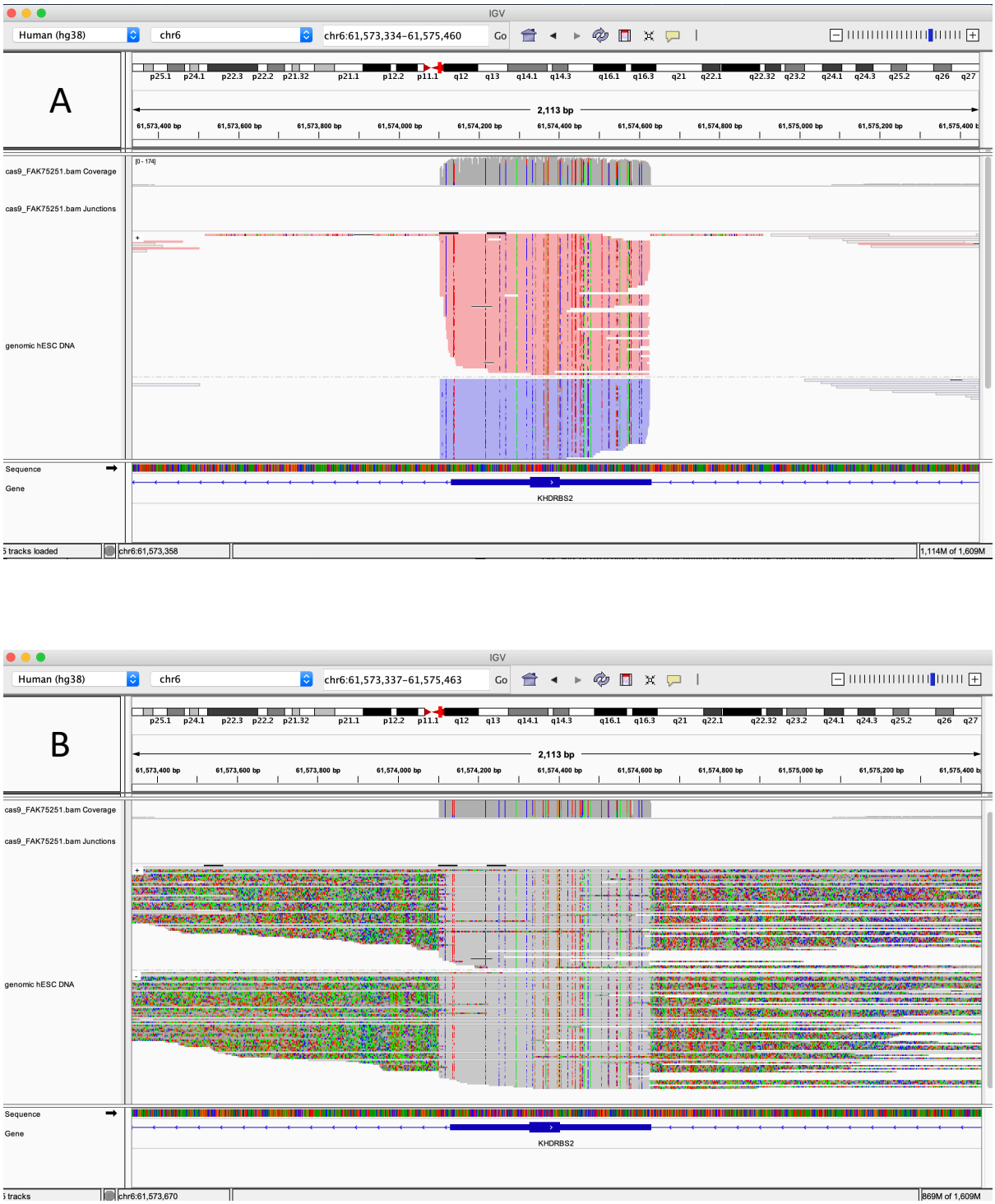


Figure 2.21: Example of soft-clipping in the *KHDRBS2* gene visualised in IGV coverage plots. Soft-clipping of reads may force reads to align somewhere in the genome even if the presented location is not the genuine location of reads, and this is most apparent for repetitive regions of the genome. This particular result was obtained from the sequencing of genomic DNA derived from hESCs (not the MDA DNA) during protocol validation. A) The aligned reads have been soft-clipped to match the genomic context. The reads are colour differentiated according to the read orientation (forward or reverse) with the coverage plot displayed above. B) Modification of IGV settings reveals the ends of the reads that have been soft-clipped. It becomes evident that *KHDRBS2* is not the genuine origin of these reads but rather a bioinformatic artefact. Filtering out reads with MAPQ values below 30 or 40 will effectively eliminate these such wrongfully aligned reads from the data set.

One way of overcoming the current limitation is to increase the computational power of the machine used to analyse the data set and to align the sequences against the whole genome reference (hg38) rather than just one chromosome at the time. When doing this, one can search for the potential off-target effects across the entire genome and investigate carefully whether these are the reads gained from true off-target cleavage or due to imperfect alignments. The latter is easily distinguishable from the real off-targets due to low MAPQ values (usually in the range of 1-3), whereas a true off-target reads are expected to have a MAPQ 60 as well as an appreciable depth-of-coverage. This was of lesser concern during the application of CRISPR-Cas9 for targeted enrichment, since this strategy aims to sequence samples without the need to perform PCR. However, the off-target effects are of considerable concern to those who contemplate the use of CRISPR-Cas9 as therapeutic intervention. In this respect, the workflow presented here proved promising for the investigation of biological off-target consequences of genome editing tools. The widely used *in silico* analyses that yield candidate regions for which a specific PCR must be designed and validated are self-limiting and more laborious. However, the approach described here is simple and truly exploratory, since the off-target cleavage will expose DNA ends that will inevitably be ligated to sequence adapters along with the on-target sequences, allowing all of the potential off-target sites to be sequenced (and thereby revealed) in a single run. This appears to be superior to strategies that examine a pre-defined set of loci and much less expensive in comparison with sequencing of the entire genome at 30 x coverage.

2.4 DISCUSSION

CRISPR-Cas9 based genome editing is an indispensable molecular biology tool with enormous potential to correct disease-causing mutations. Applied to the germline, heritable GE has been proposed as an alternative strategy to reduce the burden of genetic disease, although its clinical application remains a subject of considerable international debate, centred around efficacy, safety and complicated ethics (Human genome editing: Science, ethics, and governance, 2017; Lovell-Badge, 2019). Several groups have conducted initial studies aiming to assess the feasibility of this approach in human preimplantation embryos (Kang et al., 2016; Liang et al., 2015; Tang et al., 2017). Liang et al. (2015) used CRISPR-based editing to cleave the *HBB* gene in human tripronuclear zygotes (3PN) and found that although the achieved editing efficiency is high (52%), the majority of the edited zygotes contain indels and only 14% of the cells utilised the exogenous HDR repair template. Similar low frequencies of HDR were observed in the study of Kang at al. (2016) who used the same system to introduce the naturally occurring CCR5Δ32 allele into 3PN zygotes by providing an exogenous DNA template to facilitate HDR. It became evident that the technique, despite its success in achieving high frequency of editing, faces significant challenges in terms of its technical feasibility and clinical applicability. These relate to the low efficiency of HDR, potential off-target editing in homologous and paralogous sequences to the ones being edited as well as other de novo off-targets, and also the possibility of mosaicism which could complicate prediction of the gene editing outcome following PGT. Perhaps the most obvious challenge, however, lies in the lack of validated tools and methodology to assess these effects comprehensively. The doctoral work

described in this chapter set out to develop such methodology, as well as to add to the growing body of literature in the area of germline genome editing.

2.4.1 Prediction of on-target spectrum of mutations

One possibility to simplify the selection of putative sgRNAs is to use an online algorithm that can predict the mutational spectrum arising from editing and aid in choosing of highly specific sequences. Currently, there are three tools that have been developed and made available for this purpose (Allen et al., 2019; Chakrabarti et al., 2019; Shen et al., 2018). The ability to accurately predict the mutational spectrum largely relies on training cell type specific data (using artificial intelligence) from CRISPR-Cas9 experiments. However, considering that experiments in human embryos are extremely limited, it would be impossible to generate a sufficient amount of data to enable an accurate prediction as each cell type potentially behaves differently, necessitating preliminary testing in the same cell type. As such, studies utilising hESCs might represent the best available alternative, although it is acknowledged that these cells may also differ significantly from those of a preimplantation embryo. In the current study, hESC closely resembled the mutation spectrum observed in *in vivo* zygotes. Indels characteristic of NHEJ were observed in 77% (10/13) of the successfully amplified embryos. The majority of the samples contained a stereotypic 1bp, 2 bp or 3 bp deletion closely resembling the pattern observed in hESCs during earlier work. However, there were three instances in which larger deletions (5 bp, 8 bp and 23 bp) were detected. The tool of Allen et al. (2019) used for prediction of large indels offers a "pluripotent stem cell type" option and the researchers found larger indels as well as microhomology-mediated small deletions to be significantly more prevalent when compared to other cell types. Interestingly, these are favoured at tandem repeats, and one might wonder about potential therapeutic routes for

diseases in which repeated DNA sequences play a role, such as Huntington's disease and Fragile X, one of the most common indications for PGT-M.

2.4.2 Mosaicism

In theory, successful application of CRISPR-Cas9 during fertilisation with sperm should enable delivery of the desired edit to all cells of the future individual. This would be required if the use of the technique was considered for therapeutic interventions, with successful editing having to be confirmed by PGT prior to establishing a pregnancy. Multiple groups have reported genetic mosaicism in the edited embryos, which could impede accurate PGT and this underscores the challenge that clinical application faces (Lea and Niakan, 2019; Kang et al., 2016; Liang et al., 2015). In the context of IVF, chromosomal mosaicism (where a subset of analysed cells is affected by aneuploidy) has been reported to affect between 70% and 90% of all IVF embryos (Taylor et al., 2014). While cytogenetic methods used for PGT have advanced to a point where chromosomal mosaicism is detectable within a trophectoderm biopsy specimen, methods used for assessment of mutations and other variants have tended to be qualitative rather than quantitative and the extent to which they could provide an accurate evaluation of mosaicism in DNA sequence is unclear. If CRISPR-Cas9 editing leads to mosaicism at the targeted site, then any genotyping PGT method would need to have sufficient sensitivity to detect it in order to avoid transfer of embryos harbouring pathogenic mutations. Some believe that even if the methodology achieves highest sensitivity, there is a risk that the trophectoderm biopsy might not be representative of the remaining cells of the embryo. Vilarino et al. (2018) compared the genotyping results after Cas9 editing in sheep embryos and found the concordance between the results obtained from sequencing of the trophectoderm biopsies and the remaining cells of the embryo to be less than 50%, underestimating the

unedited contribution of the embryo. In the dataset presented in this thesis, most of the samples investigated consisted of single cells obtained from embryos arrested prior to/at cleavage stage, precluding the possibility of detecting mosaicism. However, in the two trophectoderm biopsies (samples C8K, C9K) that were processed in this dataset, a small proportion of the unedited sequencing reads were detected, comprising between 13-25% of all reads. Although these samples were classified as complete knock-outs by the Cas9-Analyzer, there is a possibility that these apparently unedited reads represent genetic mosaicism and persistence of one or more unedited copies of the gene. Indeed, there is a possibility that the embryo may be heterozygous in all of its cells (due to successful editing of just one copy of the gene) and that the skew towards edited (knocked out) gene copies could be a consequence of preferential amplification. Future experiments in this area would have to be carried out in order to confirm one scenario or the other. This should ideally involve disaggregation of all of the blastomeres from a significant number of edited embryos, with each analysed separately in order to determine their exact genetic make-up.

2.4.3 Evaluation of off-target consequences

The initial approaches towards the analysis of off-target events following CRISPR-Cas9 editing involve the use of *in silico* tools to predict putative cut sites and then investigate these loci by using targeted amplification and deep sequencing, revealing whether off-target editing took place. The main limitation of this approach is that *in silico* tools cannot reliably and consistently identify all sites of potential off-target activity and, therefore, it would be important for any embryo considered for uterine transfer to be screened for unintended modifications across the entire genome. High-throughput experimental methods, such as Circle-Seq, developed for this purpose, could overcome the limitation of the *in silico* predictor,

but require at least 25 μg of DNA, rendering them incompatible with the minute DNA samples obtained from embryo biopsies. One exception is Digenome sequencing where gDNA is cleaved *in vitro* and then sequenced. The bioanalytical pipeline is then employed to reveal off-target (Kim et al., 2015). The initial efforts to characterise the potential off-target effects in sgRNA2b-Cas9 microinjected embryos relied on *in silico* approaches and direct PCR, followed by deep amplicon sequencing of the putative off-targets (all homologs of *POU5F1*). The investigation did not reveal any off-target activity in these regions, although this approach has limitations, as discussed earlier. In later stages of this work, a CRISPR-Cas9 PCR-free enrichment method was developed, allowing detection of larger deletions and structural variations following GE (discussed later). This technique, although not specifically intended for this purpose, shows a great promise in identifying off-target sites experimentally. It can be used to validate candidate sgRNAs prior to their use in preimplantation embryos in a time- and cost-efficient manner. The technique relies on dephosphorylation and *in vitro* cleavage of gDNA (any human DNA) using the candidate sgRNA-Cas9 complex to expose the cleaved ends for adapter ligation, followed by long-read sequencing. The analytical pipeline for identification of the off-targets is rapid, comprehensive and straight-forward, since only the true target and real off-targets (if any) will be sequenced and mapped to the genomic reference. Quantification of fold-enrichment of the off-target reads gives an indication of the strength of the off-target activity.

2.4.4 Evaluation of potential large indels and structural variations after CRISPR-Cas9-based editing

Ma et al. (2017) set out to apply germline GE to correct a pathogenic 4 bp heterozygous deletion in the *MYBPC3* gene causing cardiomyopathy. In their study, 54 S-phase wild-type

oocytes were fertilised with sperm derived from an affected donor using ICSI in combination with the mixture of Cas9 protein, sgRNA specifically targeting the male mutant locus, and an exogenous oligonucleotide HDR repair template. The HDR oligo was, in this case, harbouring two SNP variants not present in the WT or mutant sequence, in order to allow confirmation of when (if) HDR repair took place. Rather unexpectedly, the group observed a homozygous WT phenotype in 66.7% of the S-phase injected embryos, as opposed to the expected 50/50 ratio. Furthermore, 72.4% of M-phase injected oocytes (n=58) also exhibited this pattern, with only a minority of the embryos harbouring indels characteristic of NHEJ in the paternal allele and HDR repair using the provided exogenous template. The authors concluded the excess of the uniformly homozygous embryos to be the result of the high fidelity HDR mechanism using the maternal WT allele as a template to resolve the lesion caused by the Cas9 enzyme in the paternal strand (Ma et al., 2017). This conclusion received considerable scepticism, as several alternative phenomena could explain the observed result: 1) amplification failure, allele drop-out could and/or extreme preferential amplification of the maternal strand and the lack of detection of the under-amplified allele, particularly if the paternal strand contained a SNP variant in the primer binding sites (not known); 2) CRISPR-Cas9 associated GE generated a larger deletion at the site of editing leading to a loss of one or both primer annealing sites or other PCR-refractory changes (e.g. a large insertion, inversion or translocation), ultimately resulting in a failure to amplify the edited allele and the impression of loss of heterozygosity at the targeted site (as reported by Adikusuma et al., 2018; Cullot et al., 2019; Egli et al., 2017; Kosicki et al., 2018; Owens et al., 2019), and 3) whole chromosome loss. The fact that several groups observed formation of large deletions and complex structural rearrangements at the site of editing could be of concern for several reasons. First, the lack of available methodology to accurately characterise the complexity of the on-target mutations and second, the appreciable frequency with which they have been reported. Adikusama et al. (2018) detected deletions

ranging in size between several hundred bases up to 2.3 kilobases in approximately half of the mouse pre-implantation embryos subjected to CRISPR-based germline editing when using the long-range PCR approach. Similarly, Owens et al. (2019) reported a two kilobase deletion after targeting the *Runx1* locus in mouse embryonic stem cells in 23% of the clones, and these deletions escaped detection by short fragment PCR and were only detected after a long-range PCR had been carried out to amplify the target, consistent with the findings of Adikusama et al. (2018). Kosicki and colleagues reported up to six kilobases deletions and complex rearrangements in 5-20% of the mouse embryonic stem cell clones after targeting the line with CRISPR-Cas9. Finally, the study conducted by Cullot et al. (2019) who used the CRISPR-targeted system to target the *UROS* locus in HEK293T and K562 cells suggested that the editing introduced a megabase scale chromosomal truncation in approximately 10% more cells compared to the untreated controls.

It appears that long-range PCR is sufficient for the detection of most indels that span across several kilobases of DNA when a sufficient quantity of DNA is present in the sample (such as when bulk DNA is isolated from cell lines). However, this approach is problematic when larger indels (megabase scale) are present or when this technique is applied to single-cell DNA derived from preimplantation embryos. Typically, single cells are whole-genome amplified prior to subjecting the DNA to further targeted PCR. When analysing single cell genomes after WGA, it is evident that biased and insufficient amplification can occur with low quality template such as DNA containing breaks, as a result of the sample processing or cell storage (Bäumer et al., 2018). The regions that accumulate DNA damage are known to prevent the amplification process and most of the WGA methods only amplify the genome in relatively short fragments, e.g. on average 630 bp in the case of Sureplex system (Kubikova, unpublished). An exception to this is the MDA system that generates fragments of several kilobases, depending on the quantity and quality of the template material. However, even with

the MDA, the fragment length is a limiting factor for the detection of larger structural variation following the long-range PCR approach.

The novel protocol described in the experimental part of this chapter involved CRISPR-Cas9 technology to cleave the DNA sample in the upstream and downstream of the region of interest (*POU5F1*), in order to facilitate PCR-free enrichment of the targeted locus, followed by Nanopore sequencing. This proof-of-concept intended to allow detection of larger indels, otherwise impossible to detect with a traditional/long-range PCR approach where a primer annealing site might be lost due to an unintended modification present in the sequence or due to insufficient quantity of DNA obtained from embryo biopsy. Additionally, sequencing of larger fragments would increase the likelihood of detection of additional heterozygous loci. Such loci are extremely valuable as they provide information that can be used to distinguish instances of ADO and loss of heterozygosity in the affected region. To our knowledge, this is the first developed tool that allows detection of structural variation, a possible unintended consequence of editing and repair. This has, so far, not been addressed in the context of human germline genome editing, as is apparent from the review of studies that were conducted to characterise mutational spectra following CRISPR-Cas9-based GE in human preimplantation embryos.

In order to validate the proposed approach, a proof of concept study using hESCs that were subjected to WGA by MDA was carried out. The MDA products were subsequently processed using the newly developed protocol and sequenced on a MinION (ONT). The data was analysed using a number of bespoke bioinformatics tools. The study used the excision approach for the *in vitro* cleavage of the MDA-DNA (two different sgRNAs targeting upstream and downstream of the *POU5F1* locus). Therefore, the size of the excised fragment can technically be considered a limiting factor (in this case approximately four kilobases). However, since the long-read Nanopore sequencing following the *in vitro* Cas9 cleavage can occur in both

directions (albeit with less efficiency on the strand that retained the bound Cas9-sgRNA complex), the application allows generation of sequence coverage that exceeds the size of the excised fragment. In order to ensure a uniform coverage, one can use a single cut 'tiling' approach where multiple sgRNAs are designed to enrich for the region of interest in both directions. In fact, pooling of multiple sgRNAs is recommended in order to increase the coverage in the instances where higher coverage is required. Overall, the methodology allowed for more accurate and more comprehensive characterisation of the effects induced by germline GE, including structural variation, large indels and off-target activity, and showed great promise for defining the nature of the DNA repair that has taken place (e.g. NHEJ or HDR). S It is conceivable that such a strategy could become an indispensable tool in elucidating the potential use of germline GE in the treatment of germline genetic disorders at the preimplantation stage in the future.

2.4.5 Future perspectives in the area of germline genome editing

The work described in this thesis, together with the studies that have been published on the use of CRISPR-Cas9-based editing in the human germline, had a pre-clinical focus on establishing whether disease-causing mutations could be disrupted/repaired. They provide a proof-of-principle as well as methodological framework of the technique's utility in preimplantation embryos. Collectively, the body of work suggests high editing efficiency but diverse mutational profiles, and this likely precludes the use of the technique in its current format for correction of inherited disease in the human germline. CRISPR-Cas9 editing could, in the future, be used more precisely to alter the genome and bypass the need to introduce a DSB, thus avoiding the repair by NHEJ and the associated problems of induced mutations at the targeted site. A new generation of GE tools using catalytically-impaired CRISPR-Cas9, known base editors, have

recently been developed and applied in this context (Gaudelli et al., 2017; Komor et al., 2016; Li et al., 2017; Zhou et al., 2017). Base editors could potentially achieve more specific and better-refined editing, given how little is understood about the capacity of preimplantation embryos to carry out DNA repair, particularly prior to activation of the embryonic genome. Instead, base editors induce alterations through the direct irreversible conversion of one base to another (e.g. a single mutant base could be substituted for the wild-type equivalent). If base editors prove to be reliable, they could provide a powerful addition to the growing number of genome engineering and informatics tools possibly leading to clinical application in the context of the human germline.

2.4.6 Ethical considerations for germline genome editing

Technological advances, particularly in the area of molecular biology and genomics, have outpaced the regulatory frameworks set-up to govern them and present a serious ethical concern for many. Until recently, human genome editing was in the realm of science fiction. However, the advent of CRISPR-Cas9 technologies together with reports of the birth of the first genome edited children in China in 2018 have prompted a wide societal debate. Even prior to the application of CRISPR-Cas9-based editing to human embryos, multiple bodies charged with providing guidelines and advice on medical issues, had already launched inquiries into the ethics of the germline genome editing in the context of human reproduction (Kubikova and Wells, 2020).

The application of GE technologies to gametes or preimplantation embryos, in order to correct an inherited disorder, maximises the likelihood that the mutation will be successfully eliminated from all of the cells of the resulting individual. Yet it is this apparent benefit that also makes the use of GE during the preimplantation stage contentious, since any alterations

would be present in the germline and could, in theory, be propagated through future generations. For some this would mean crossing a perceived line. Nevertheless, from a purely clinical perspective, the successful editing of all cells, including the germ cells, would make it possible to permanently eliminate a deleterious mutation from a family, and therefore free future generations from the burden of disease transmission and from the need for further medical interventions. It could be argued that this protects the interest and welfare of the future individual (Kubikova and Wells, 2020).

As treatments improve, more individuals suffering from genetic disease will survive into reproductive age and may want to have genetically related children (Lovell-Badge, 2019). The desire to avoid the transmission of a mutation, without discarding potentially viable embryos following PGT-M, can surely be appreciated in these cases, and has indeed been recognized by the UK Nuffield Council on Bioethics and the US National Institute of Sciences' in their recent reports on the subject. After careful deliberation, the Nuffield Council came to the view that "there are circumstances in which gene editing of human embryos should be permissible". It is likely that to begin with, PGT and GE would have to be carried out in tandem, in order to confirm successful correction and an absence of unintended edits, but it is possible to imagine that if genome editing evolves into a precise and safe tool, PGT might eventually become superfluous (Kubikova and Wells, 2020).

Although GE technologies might not be ready for clinical trials in human embryos at this time, and as the work described here suggests, it may still be some time until the future refinements render the technique efficacious and safe, it would be irresponsible for science not to explore the possibility that they could, in the future, deliver a major advancement in precision medicine, with the capacity to eliminate the majority of inherited diseases. The recently published results obtained from studies using the new generation base editors suggest that these might be significantly safer because of the avoidance of the DSBs (Gaudelli et al., 2017; Komor et al.,

2016). Furthermore, it should be remembered that there are few medical interventions that are entirely without risk and that these must be weighed against the potential benefits of treatment. It is also worth recalling that some procedures that are almost universally embraced today, such as organ transplantation, were once considered controversial (Kubikova and Wells, 2020).

While conversations about the legitimacy of germline GE continue, there is already at least one assisted reproductive treatment that had been successfully implemented in order to permanently correct an inherited genetic defect. Mitochondrial disorders, where a proportion of the mitochondria in each cell harbour a deleterious mutation leading to a dysfunction in adenosine triphosphate (ATP) production and clinical consequences that are frequently severe, have been addressed by removing the meiotic spindle from an affected oocyte (or pronuclei from a fertilized egg) and transferring into the "healthy" cytoplasm of a donor oocyte. This procedure has already been licensed for use in the UK (by HFEA) and, internationally, has led to birth of at least one child who appears to be free of mitochondrial disease (Zhang et al., 2017). The spindle transfer procedure has also been proposed as a method to assist in the treatment of infertile patients with a history of unsuccessful IVF treatment due to poor in vitro embryo development. Data from a mouse model, characterized by high rates of embryo developmental arrest, has shown encouraging results, with rescue of the phenotype after transfer of meiotic spindles to donor oocytes from a different mouse strain (Costa-Borges et al., 2020).

With new and powerful technologies always comes the fear of abuse. These concerns should not be lightly dismissed and need to be the subject of an inclusive public debate. In terms of GE being used as a therapeutic intervention, some argue that there is potential for the technique to applied for non-medical reasons: introducing genetic variants that do not exist in either parent, for the purpose of enhancement rather than cure. Such modifications could be aimed at conferring resistance to pathogens, increasing tolerance to environmental conditions or

enhancement of physical or mental attributes. To address these concerns, it is necessary to build a consensus and decide which traits would be ethically justifiable and acceptable for alteration. It is possible that in well-regulated areas of the world the use of GE would always be confined to treatment of serious medical illness. In line with this notion, it is the view of the Nuffield Council that the use of heritable GE interventions should be "consistent with social justice and solidarity so that it should not be expected to increase disadvantage, discrimination, or division in society". From a technical perspective, it is worth noting that most traits are controlled by polygenic inheritance and as such their enhancement would be technically challenging using the existing approaches, which focus on correction of a single mutation. Public attitudes and prevailing social norms will likely shape how germline GE will be used, positions that might conceivably change over time (Lovell-Badge, 2019). If GE is to be used at all, adequate guidelines and a regulatory framework need to be developed, setting out minimal conditions that would have to be met in respect to clinical and technical issues surrounding the gene to be edited, the need and justification for such modification, and the criteria for efficacy, precision and safety of the tool (Kubikova and Wells, 2020).

Chapter 3: Germline genome editing: considerations for DNA repair in preimplantation human embryos

3.1 INTRODUCTION

With the advent of rapid user-friendly genome editing using CRISPR-Cas9, manipulation of the DNA of human embryos has become a realistic possibility. Such methods can be applied for the study of human preimplantation embryos, providing tremendous opportunities for the enhancement of our understanding of molecular mechanisms operating during the first few days following fertilization and clarifying the function of individual genes. The technique has also been proposed for modification of the human germline, as a potential therapeutic, correcting mutations responsible for inherited disorders. One of the most important questions related to early development, and subsequent embryonic competence, concerns the capacity of embryos to repair DNA damage. The number of studies that have deployed CRISPR-Cas9 to induce DSBs in the DNA is on the rise, however, none of them set out to examine the DNA repair capability in human preimplantation embryos. They nonetheless provided data indicating that human embryos may suffer a deficiency in DNA repair during the first few cell divisions. Additional research, examining the chromosomes of early human embryos, has also indicated that chromosome breakage, a hallmark of compromised DNA repair activity, is particularly common during a 2-3 day window prior to activation of the embryonic genome (Babariya et al., 2017). These findings lead to several important questions: How exactly do early human embryos resolve the double strand breaks in their DNA? Are cellular DSB repair mechanisms functional prior to embryonic genome activation? Are such pathways in early human embryos distinct from other cellular systems? Does DNA damage induced using methods such as CRISPR-Cas9 have the potential to induce genomic instability?

Collectively, evidence from research in unedited human embryos supports the hypothesis that human zygotes are susceptible to DNA damage, as evidenced by: a) spontaneous chromosome breakage resulting in segmental aneuploidy, observed in up to 25% of all cleavage stage embryos (Babariya et al., 2017); b) morphological features associated with suboptimal embryonic development and chromosome fragmentation, such as the presence of micronuclei; and c) extremely low expression of several genes with key roles in cell cycle regulation, DNA repair, checkpoint control and apoptosis (Wells et al., 2005). Based on these observations, the hypothesis in this chapter was that DNA DSB repair is compromised until after the EGA, normally occurring 3 days after fertilisation. CRISPR-Cas9 editing performed in zygotes, prior to EGA, has been seen as an attractive strategy, as delivery of the CRISPR components can be virtually guaranteed, and thus all cells of the resulting embryo will be modified. However, if DNA repair is deficient at this stage of development, unresolved DSBs are likely to lead to cell death or complex structural rearrangements. These effects could induce a state of genomic instability, mitotic arrest, and ultimately be deleterious to embryo viability or (if the embryo survives) to the health of the individual.

How (if) human preimplantation embryos resolve DNA damage may have significant implications for those considering the use of GE technologies for the treatment of inherited disorders. Confirmation of a limited DNA repair capacity would indicate that current GE tools cannot be safely applied at early stages of development. Additionally, findings confirming that human embryos are sensitive to genotoxic damage may also have significant implications for assisted reproductive treatments, potentially helping to guide the formulation new embryo culture systems, with a focus on preserving DNA integrity and supporting DNA repair. The main aims of this chapter were:

- to investigate the incidence of whole chromosome and segmental aneuploidy in human preimplantation embryos targeted with CRISPR-Cas9 technology using an established cytogenetic technique combined with a newly developed bioanalytical workflow

- to assess whether the incidence of sub-chromosomal aberrations detected in targeted human zygotes correlates with the incidence detected in targeted human embryonic stem cells using fluorescence *in situ* hybridisation

3.2 MATERIALS AND METHODS

To address the study objectives of this chapter, the same sample set as the one described in Chapter 2 was used, comprising human zygotes that were targeted with CRISPR-Cas9 technology to introduce DSB in the *POU5F1* locus, encoding the OCT4 pluripotency factor. Following the validation of various sgRNAs targeting different exons of *POU5F1* in hECSs using transfection experiments, the sgRNA2b demonstrated highest editing efficiency and was selected for the downstream application in human zygotes donated for this research. In total, 22 human zygotes were *POU5F1* targeted using the validated sgRNA2b-Cas9 RNP and 20 zygotes were microinjected with Cas9 protein alone to control for the microinjection procedure. The biopsied samples were sent to the Oxford laboratory from the Francis Crick Institute on dry ice where they were processed using next generation sequencing and bioinformatic tools for genotype, off-target, and cytogenetic analysis. The analysis comprised a determination of the whole chromosome copy number and the analysis of segmental aneuploidy, with particular focus on chromosome 6 where the DSBs were introduced by the sgRNA2b-Cas9 RNP complex. The investigation utilised next-generation sequencing followed by analysis with BlueFuse Multi software. The exact steps undertaken in this section are schematically described in Figure 3.1 Segmental aneuploidies affecting chromosome 6 were further investigated using a bespoke bioinformatics workflow, created during the course of this DPhil, that divided the genome into smaller bin sizes. This strategy permits and increased resolution compared to BlueFuse Multi when assessing variation in the copy number of defined chromosomal regions, with the potential for determining the exact location of breakpoints. Finally, three clones of hESCs induced with sgRNAs targeting different exons of *POU5F1* were fixed and fluorescent *in situ* hybridisation (FISH) was carried out in order to quantify the

occurrence of segmental gains and losses on chromosome 6 in hESCs and compare the finding to the results obtained from human targeted zygotes.

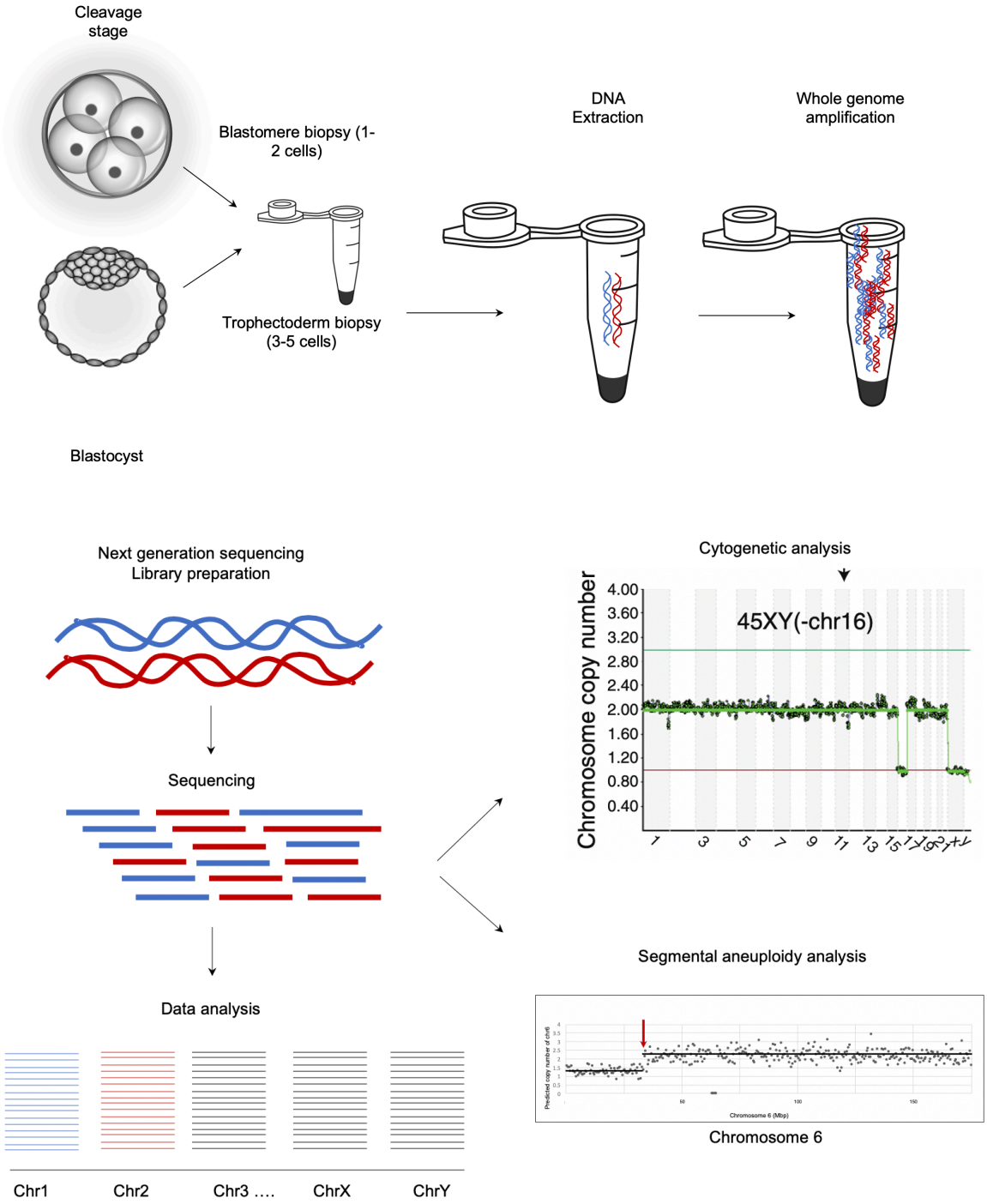


Figure 3.1 Experimental workflow for the cytogenetic and segmental analysis of CRISPR-targeted human embryos. The sgRNA2b-Cas9-microinjected embryos and Cas9-microinjected controls were biopsied at the blastocyst stage or if they failed to reach the blastocyst stage, the samples were biopsied from the arrested embryos (at the 2-cell stage or the cleavage stage). The biopsied samples were whole genome amplified using the Sureplex system and subjected to low-pass whole genome sequencing on the Miseq using the VeriSeq PGS workflow. Data analysis indicate mapping of reads to each chromosome. BlueFuse Multi software was used for cytogenetic analysis to determine the chromosome copy number and segmental aneuploidies on chromosome 6 where the *POU5F1* target was located in addition to a newly-developed bioinformatics workflow.

3.2.1 Low pass whole genome sequencing for cytogenetic analysis of CRISPR-edited human embryos

To determine the whole chromosome copy number and detect any segmental gains and losses, single or multiple blastomeres were biopsied from embryos at the cleavage stage and clumps of approximately three to five cells were micro-dissected from the trophectoderm of blastocysts. The cells were washed through three drops of a wash buffer (PBS/0.1% polyvinyl alcohol), which had previously been tested to confirm absence of contaminating DNA. The cells were transferred to 0.2-ml PCR tubes in a volume of 1.5 µl, lysed and subjected to whole-genome amplification (SurePlex, Rubicon) followed by low-pass next generation sequencing (coverage depth < 0.1×) (VeriSeq PGS kit, Illumina). Libraries were prepared according to the manufacturer's instructions and sequenced using the MiSeq sequencing platform as described in the Materials and Methods of Chapter 1 (pages 58-62). Typically, ~ 1 million reads were generated per sample, of which 60–70% successfully mapped to unique genomic sites. Mapped reads were interpreted using BlueFuse Multi software (Illumina) in order to generate chromosome copy number profiles. This strategy has been extensively validated and is widely used for the detection of whole chromosome losses and gains, as well as segmental aneuploidy, in human embryos undergoing preimplantation genetic diagnosis (Zheng et al., 2015). Analysis of single blastomeres allowed each chromosomal region of at least 5 megabases to be assigned a copy number of 0, 1, 2, 3 or 4 (corresponding to nullisomy, monosomy, disomy, trisomy or tetrasomy) (Norah M. E. Fogarty et al., 2017).

3.2.2 Bioinformatic analysis of segmental abnormalities on chromosome 6

Sequence files from the embryos with detected segmental aneuploidy in chromosome 6 were further interrogated using a bespoke bioinformatic workflow to confirm the suspected breakpoint in the segmental gain/loss. Since the VeriSeq workflow in combination with the analysis in BlueFuse Multi only allows visual inspection of the generated cytogenetic profile, it was impossible to establish with certainty whether the breakpoints associated with segmental losses and gains on the short arm of chromosome were confined to the targeted *POU5F1* locus. In the newly designed protocol, the sequenced BAM files were processed in Bedtools for coverage in chromosome 6 (using the Bedtools coverage sub-command) (Quinlan and Hall, 2010). The reads were split into 350 bins of 500 kilobases in size (10x smaller than those created by the BlueFuse Multi software), with exception five bins on either side of the predicted sgRNA2b-Cas9 cut site, which were approximately 100 kilobases in size (50x smaller than those created by the BlueFuse Multi software). The two immediately adjacent bins were positioned around cut site, so that any gains or losses resulting from unresolved DSB caused by the CRISPR-Cas9 could be clearly detected. The generated tab-separated value files (TSV files) were transferred into Microsoft Excel. The initial steps of the analysis in Excel involved building a reference file. The exact proportion of reads mapping within each bin was determined from the total number of reads that aligned to chromosome 6 and compared to a reference set comprising data compiled from multiple karyotypically normal samples. The resulting values were doubled prior to plotting them to generate a predicted copy number profile of chromosome 6 and to reveal the corresponding segmental gains and losses. The results were then compared with the those obtained from VeriSeq analysis for concordance. The following script was written to generate read counts per bins:

```
bedtools makewindows -g chromosome_6.txt -w 100000 >
chromosome_6.bins.txt

#this is to separate the genome file into bins of 100
kilobases

#then retrive the POU5F1 location from the generated bin
file and adjust the bins according to the cut site

#then calculate coverage per bin

bedtools coverage -a chromosome_6.bins.bed -b
/Users/INPUT.bam > OUTPUT.tsv
```

3.2.3 Fluorescent in situ hybridisation of chromosome 6 p-arm telomeres and centromere in CRISPR-edited human embryonic stem cells

Human ESCs (H9) were transfected with three candidate sgRNAs spanning through Exons 1, Exons 2 and Exon 4, in order to determine the sgRNA-Cas9 complex with the highest efficiency in editing to be used in the subsequent experiments with human zygotes, described in Chapter 2. Following the experiments of low-pass WGS and cytogenetic analysis, multiple edited embryos exhibited segmental abnormalities on chromosome 6, with breakpoints confined to the intended cleavage site of sgRNA2b. This raised an interesting question of whether chromosomal breakage, due to a failure to repair the DNA break induced by CRISPR, is a phenomenon unique to human embryos or whether this is a feature of CRISPR-based genome editing in general. In order to answer this question, the incidence of segmental chromosome abnormality resulting from the cleavage of the *POU5F1* locus was assessed in targeted hES cells - a cell type often used as an experimental model in genome editing work that, to some extent, reflects the cellular context of the human embryo. Interphase FISH was

carried out on cells from three clones of the induced hESC and untreated control cells (clone H9) to examine the number of centromeric and telomeric signals associated with chromosome 6p (where the *POU5F1* is located). For this purpose two differently labelled FISH probes were used (as described in detail below). The hESCs were induced with sgRNA-Cas9 targeting the exon 2 (equivalent to the targeting carried out in human zygotes) and cultured for several days as described before prior to collecting approximately 10 million cells (on Day-2 or -4 post-induction) into a 1.5 mL tube and pelleting the cells at 300 x g force. The supernatant was aspirated and a drop of Carnoy's fixative (methanol:glacial acetic acid in 3:1 ratio, both Thermo Fisher Scientific) was added to the cells collected at the bottom of the tube. The mixture of cells and the fixative were gently agitated before adding each additional drop of the fixative until reaching of approximately 1 mL. The cells were then briefly checked under the brightfield microscope for density and if necessary, additional fresh fixative was added to adjust the concentration prior to storage at -20 °C.

Preparation of FISH slides

Prior to the preparation of the FISH slides, the cells were first pelleted and re-fixed with fresh fixative. The slide warmer was set to 45 °C and glass slides were immersed in a Coplin jar filled with some fresh fixative (or placed at 4 °C until used). The cells were then resuspended using a Pasteur pipette and approximately three drops of the cell suspension were dropped vertically from the distance of approximately 25 cm onto a paper-dried glass slide. The slide was then rotated gently to spread the cells around and placed directly onto slide warmer until completely dry. The slides were assessed under a phase contrast microscope for concentration as to allow for optimal cell density. The slides were dehydrated in ethanol (70%, 90%, 100% for 5 minutes each) and allowed to dry completely for approximately 24 hours prior to pre-treatment with proteinase K and probe hybridisation.

FISH Slide proteinase K pre-treatment

The dried slides were placed in a Coplin jar containing 2x saline sodium citrate (SSC) (Thermo Fisher Scientific) buffer in water bath set to 73 °C for 2 min. The slides were then placed into a Coplin jar containing the proteinase K (Qiagen) and 1x phosphate buffered saline (PBS) (Thermo Fisher Scientific), set to a concentration of 0.1 mg/mL at the temperature of 37 °C for 1 min. The slides were then taken out and washed in 1x PBS for 5 min at room temperature (RT). The slides were then fixed in a solution of 1% paraformaldehyde (Sigma Aldrich) for 5 min at RT, then again washed in 1x PBS for 5 min at RT. Finally, the slides were dehydrated in a sequential immersion into 70%, 90% and 100% ethanol, each for 1 min.

Probe hybridisation

To prepare the probe mixture, 7 µl, of the supplied hybridisation buffer, 1 µl, of each probe (TelVysion 6p SpectrumGreen 5 µL and Vysis CEP 6 D6Z1 SpectrumOrange Probe, 20 µL, both Vysis Abbott Molecular) and 1 µl of purified water was added into a 0.2 mL microcentrifuge tube. The mixture was vortexed and spun down, then placed into 73 °C water bath for 5 min. The tube was then placed into the slide warmer set to 45 °C until ready to apply to target DNA. Next, 10 µl of the probe mixture was applied to each target area on the fixed and dehydrated slides, and immediately covered with the cover slip. The slides were placed in a humidifying chamber for 30 min, then washed in a Coplin jar containing a solution of 70 mL of 0.4X SSC/0.3% NP-40 detergent (Sigma Aldrich), previously equilibrated to 74 °C in water bath. Once the cover slips loosened up, they were removed from the slide and the slides remained in the wash solution. After 2 min, the slides were transferred into a second wash solution of 2X SSC/0.1% NP-40 for 1 min. The washed slides were air dried in darkness and when completely dry, 10 µL of DAPI II counterstain (Vysis Abbott Molecular) was added to the target area and covered with a cover slip. The 6p telomeric regions and chromosome 6

centromeric regions in the fixed cells were then visualised under fluorescent microscope (Zeiss). The centromeric and telomeric signals were then individually counted in each of a minimum of 500 cells. In order to minimise the potential impact artefacts related to overlapping or split signals, two closely situated signals were only considered to represent two distinct loci when they were separated by a distance larger than the diameter of a single signal.

3.3 RESULTS

3.3.1 Whole chromosome and segmental aneuploidy detection by cytogenetic analysis

The CRISPR-edited and Cas9 control zygotes were processed using whole genome sequencing at low-pass to determine whether CRISPR-Cas9 editing can lead to unintentional whole chromosome and/or sub-chromosomal damage. In this analysis, 23 samples were examined (16 CRISPR-edited and 7 controls) and generated copy number profiles using the BlueFuse Multi software. The profiles were used to examine the presence of whole chromosome and segmental aneuploidy, with particular interest in chromosome 6, where the on-target site of the sgRNA2b-Cas9 is located (*POU5F1* locus, p-arm band 21.3). The complete set of results obtained from this analysis is summarised in Table 3.1 while the complete set of copy number profiles can be found in Appendix 5.

Eleven out of 16 (69%) of the samples examined from the targeted group displayed a whole chromosome gain or loss and in three out of seven (43%) in the Cas-9 microinjected control group. The number of whole chromosome gains and losses did not significantly differ between the CRISPR-targeted and control group (p=0.363, two-sided Fisher's exact test) (Figure 3.2). This was expected as CRISPR-Cas9 editing is not thought to affect the rate of whole chromosome abnormalities, predominantly occurring as a result of errors in meiosis (Chiang et al., 2012). In total, whole chromosome abnormality was predicted in 14 out of 23 samples, or 60.9%. In the collected blastomeres from sgRNA2b–Cas9-microinjected embryos arrested up to the eight-cell stage, chromosomal loss or gain was detected in 78% (seven out of nine) of these embryos, which is consistent with rates reported by preimplantation genetic screening

(Maurer et al., 2015; Wells et al., 2014). Trophectoderm biopsies of a subset of blastocysts that developed following sgRNA2b–Cas9 microinjection showed that 43% (three out of seven) were euploid (Table 3.1).

Table 3.1: Summary of the results obtained from the cytogenetic analysis of sgRNA2b-Cas9 microinjected embryos and Cas9 controls using the VeriSeq protocol and BlueFuse Multi software.

	Embryo	Stage	Detected whole chromosome abnormalities	Abnormalities detected on chromosome 6
sgRNA-Cas9 targeted embryos	C1	Cleavage	-15	none
	C2	Blastocyst	none	none
	C3	Cleavage	-X	none
	C4	Cleavage	none	none
	C8 (3 TE samples)	Blastocyst	none none none	none none none
	C9	Blastocyst	+8q.21.3, -9q, +11, -12, +13, -14, +15, +16, -18, +21, +22, +X	+6
	C10	Cleavage	+1p, -2, +3, -7, +9p, +9q, -12, +14	+6p21.3
	C12 (2 TE samples)	Blastocyst	none none	-6p21.3 +6p21.3
	C14	Cleavage	+16	+6p21.3
	C15	Cleavage	+1, +2, -3	+6p21.3 +6q
	C16	Blastocyst	-16 -16	none none
	C19	Cleavage	-5, -21	-6p21.3
	C20	Blastocyst	-9	-6p21.3
	C22	Cleavage	none	none
	C23	Cleavage	-7pq, -8, +9, -17, +22, -X	none
	C24	Blastocyst	-21	none
Cas9-microinjected controls	1K	Blastocyst	none	none
	2K	Blastocyst	none	none
	3K	Blastocyst	none	+6q22.3
	4K	Blastocyst	-21	none
	5K	Blastocyst	none	none
	7K	Blastocyst	-14	none
	8K	Blastocyst	-14	none

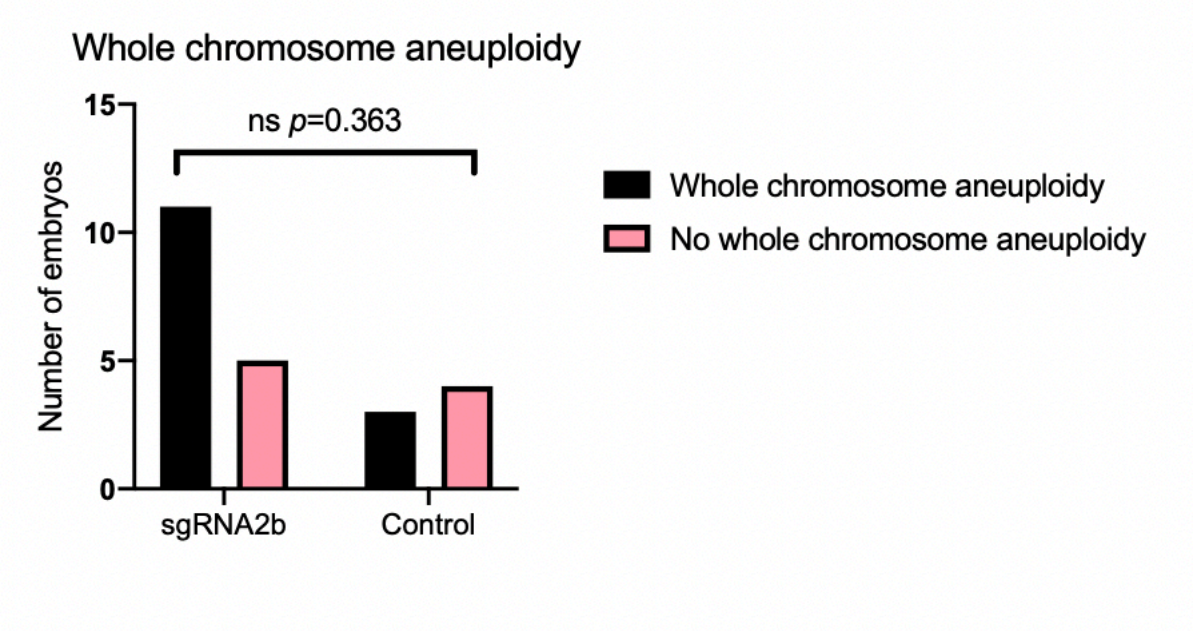


Figure 3.2: The number of control and targeted samples with whole chromosome aneuploidy according to their copy number profiles generated by low-pass whole genome sequencing. The predicted *p*-value is a result of a two-sided Fisher's exact test. The statistical test did not predict significant differences in the rate of whole chromosome aneuploidy between the two groups (p=0.363). ns: non-significant. Figure was generated by GraphPad Prism.

3.3.2 Segmental gains and losses of chromosome 6p arm are prevalent in the sgRNA2b-Cas9-targteted embryo samples

Out of the 23 profiles examined, 69.6% (16 out of 23) samples exhibited two copies of intact chromosome 6 and showed no evidence of segmental aneuploidy affecting any of the chromosomal arms. However, unlike with the rate of whole chromosome aneuploidy, there was a marked increase in segmental abnormalities detected on chromosome 6 p-arm (where editing occurred) between the targeted group and controls. In 37.5% of the targeted embryos (six out of 16) a gain or loss with a breakpoint at 6p21.3 was detected, compared to none in the control group, although this difference was not statistically significant, presumably due to small sample size (p=0.124, two-sided Fisher's exact test, Figure 3.3, Table 3.1). A representative copy number profile of the targeted embryo showing a 6p21.3 gain and control embryo profile is presented in Figure 3.4. The sole control embryo with a segmental aneuploidy (which

represents a frequency of 14.3%) had a gain extending from 6q22.3 to the end of the long arm of the chromosome. This does not appear to be related to the targeted *POU5F1* locus which is on the other arm of the chromosome. Strikingly, multiple segmental anomalies involving the targeted 6p21.3 region were identified in CRISPR embryos. This suggests CRISPR-induced breakage of the DNA at the *POU5F1* locus (although it was only possible to obtain a rough estimate of the location of the breakpoint using the BlueFuse Multi analysis software). Nonetheless, the breakpoints appeared to involve the same area in all targeted samples. The segmental gain/loss of 6p21.3 affected twice as many cleavage stage embryos compared to the blastocysts (four and two, respectively). The distribution of gains and losses was spread evenly, with three embryos exhibiting a gain of 6p21.3, two embryos a loss of 6p21.3 and one embryo (embryo C12) that had two trophectoderm samples tested exhibited a gain extending from 6p21.3 to the end of 6p in one sample and a reciprocal loss in the second trophectoderm biopsy. Overall, the only additional segmental abnormalities affecting other chromosomes were detected in samples C9, C10 and C23 in the targeted group, affecting chromosomes 7, 8 and 9. It is worth noting that all three samples exhibited, in addition to a segmental gain or loss, five or more additional whole chromosome gains or losses. This was not the case for samples with segmental abnormality involving 6p21.3, where 5 out of 6 embryos contained none or less than three additional whole chromosome aneuploidies. Altogether, the cytogenetic analysis by low-pass WGS indicates that a significant fraction of on-target events lead to unexpected segmental abnormalities, presumably involving the target side following CRISPR-Cas9 editing in human preimplantation embryos.

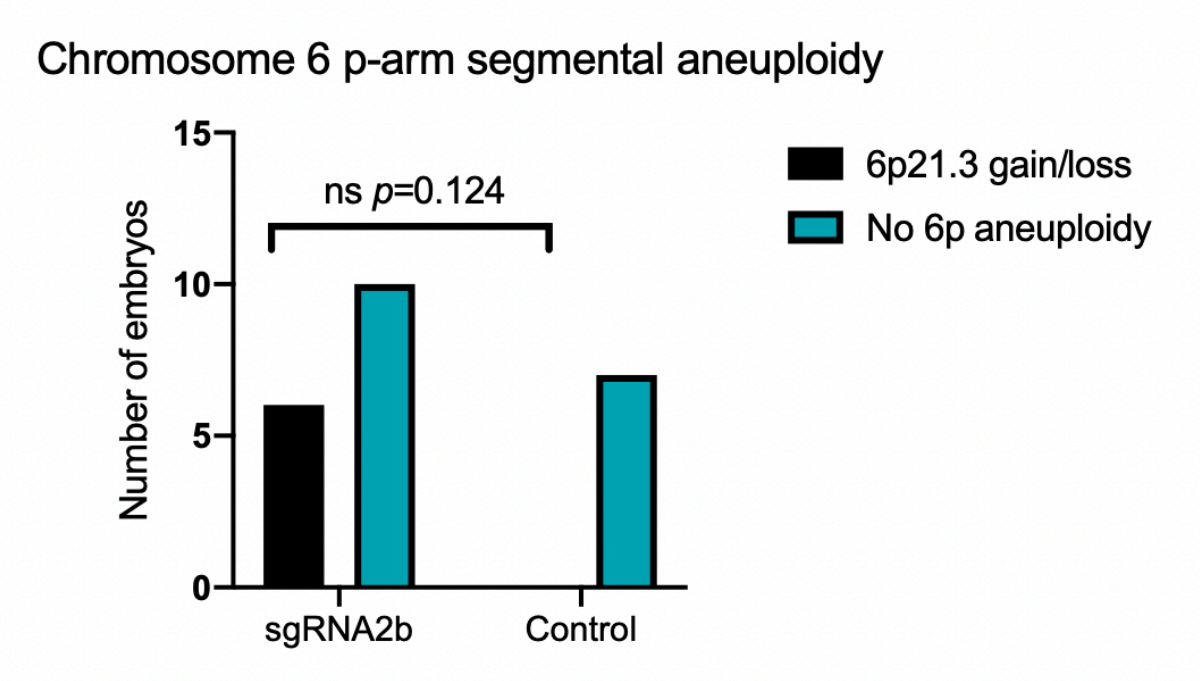


Figure 3.3: The number of control and targeted samples with segmental aneuploidy on chromosome 6p arm according to their copy number profiles generated by low-pass whole genome sequencing. The predicted *p*-value is a result of two-sided Fisher's exact test. The rate of segmental aneuploidy affecting chromosome 6p arm was not significantly higher in the targeted group (p=0.124). ns: non-significant. Figure was generated by GraphPad Prism.

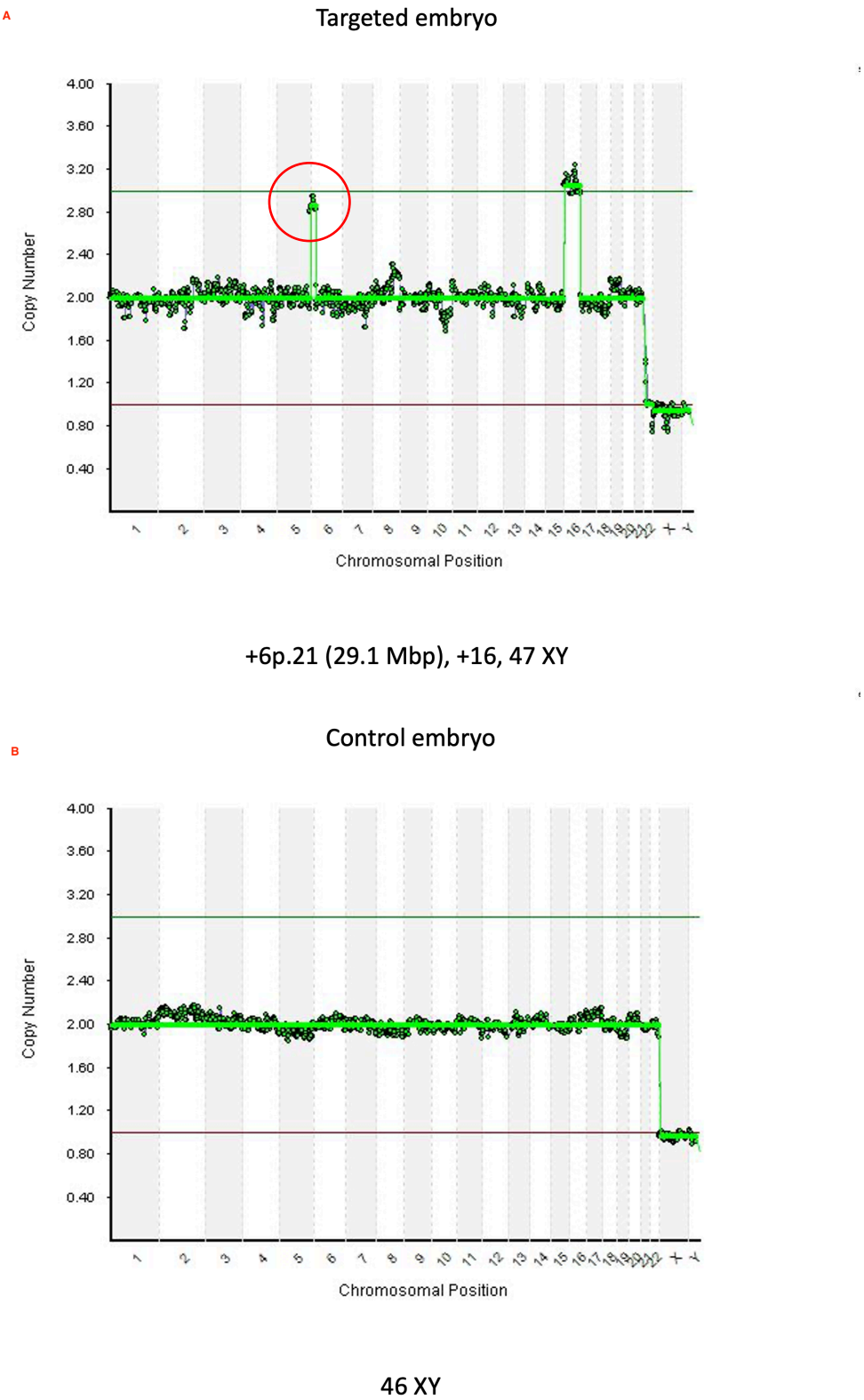


Figure 3.4: A representative chromosome copy number profile by whole genome sequencing of the embryonic DNA. A) A representative graph indicating segmental aneuploidy on chromosome 6 in the targeted embryo following sgRNA2b–Cas9 ribonucleoprotein complex microinjection. The 6p21.3 gain is circled in red. **B)** A Cas9-microinjected control embryo showing a normal copy number profile. The graphs were generated by BlueFuse Multi after sequencing using the VeriSeq PGS protocol.

3.3.3 A newly developed bioinformatic analysis of segmental aneuploidy affecting chromosome 6 reveals the breakpoints of CRISPR-Cas9 editing fall within its on-target sequence

The copy-number profiles described above with low-pass WGS data can only provide a coarse-grained cytogenetic analysis, sufficient for the analysis of whole chromosomes, but limited when it concerns the ascertainment of precise breakpoints associated with segmental aneuploidy. Developed for clinical use in PGT and PGT for chromosomal rearrangements, this methodology yields sufficient results when the exact location of the rearrangement is either known or can be approximated. Since the exact location of segmental aneuploidy affecting chromosome 6p could only be estimated from this analysis, an additional bioinformatic protocol was employed to determine whether the reported gains and losses could be localised to a single region, potentially the target of sgRNA2b-Cas9. This would support the notion that on-target CRISPR activity had led to unintended chromosome breakage, a presumed consequence of a failure to resolve DSBs due to insufficient DNA repair in preimplantation embryos.

The bioinformatic analysis developed for this purpose utilised the dataset generated from the low-pass WGS, where the reads aligning to chromosome 6 were split into 350 bins of 500 kilobases (the entire chromosome contains 171,115,067 bases and dividing it by 500,000 yields 342, the extra 8 bins were enriched around the *POU5F1* locus and scaled down to 100 kilobases). The target cut site of the sgRNA2b, located three bases adjacent to the PAM site of the sgRNA, usually at base 31,133,713, was then used as a boundary between two adjacent bins, so that in case a segmental abnormality occurs within this locus, it would be possible to observe a relative increase or decrease in the read proportion of that bin, corresponding to a sub-chromosomal gain or loss, respectively.

The results generated from this workflow confirmed the segmental aneuploidy detected in the earlier analysis in all samples except in embryos C19 and C20. It is worth noting that both of these samples also failed the targeted amplification of the *POU5F1* locus and subsequently the on-target genotypes could not be determined. In terms of the copy number profiles, both samples exhibited significant background, suggesting they might have been of poorer quality, potentially due to insufficient amplification. In the remaining six out of eight or 75% of the samples with segmental abnormalities affecting chromosome 6, the method was able to confirm the results obtained from the BlueFuse Multi analysis. In the samples affected by a gain/loss of 6p21.3pter, the method was able to locate the exact point at which the segment was either gained or lost, localised in one of the two bins adjacent to the *POU5F1* locus, thereby confirming the earlier indications that the segmental gain/loss was a direct consequence of sgRNA2b-Cas9 editing. Additionally, the profile of the control embryo with previously detected chromosome 6 q-arm abnormality was examined and confirmed the findings from the earlier analysis. This suggests that the developed methodology is capable of detecting segmental abnormalities affecting both arms of chromosome 6. Figure 3.5 represents a comparison of the results generated by the two methods (BlueFuse Multi and the bespoke bioinformatic protocol).

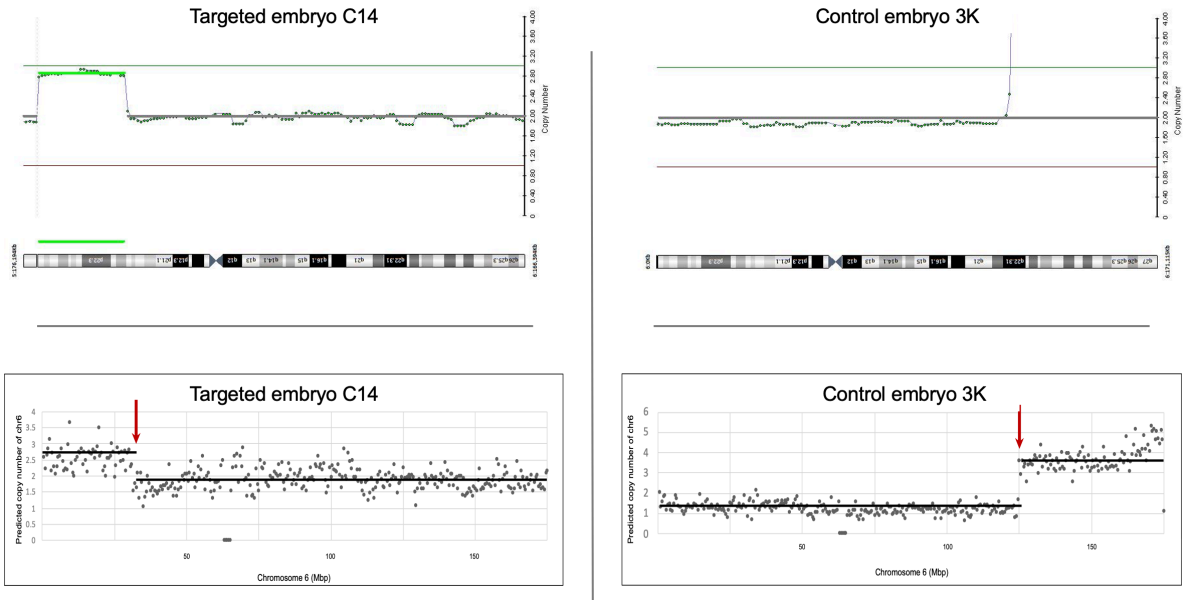


Figure 3.5: A comparison of copy number profiles generated by the two methods (BlueFuse Multi and the bespoke bioinformatic analysis of segmental aneuploidy) in CRISPPR-targeted embryo C14 and Cas9 control embryo 3K. The top two graphs represent the BlueFuse Multi analysis while the two bottom profiles represent the result obtained from the newly developed bioinformatic protocol. The results are concordant for both embryos tested. The vertical axes represent the predicted copy number whereas the horizontal axes represent the position on chromosome 6 (by aligning the graph to a graphical representation of chromosome 6 in the two top profiles and as megabasepairs [Mbp] in the two bottom profiles. The arrows indicate the detected chromosome breakpoints. In case of the edited embryo C14, both methods detected a gain in the 6p21.3 region, the sgRNA-Cas9 *POU5F1* target, extending to the end of the chromosome 6 p-arm. In the case of control embryo 3K, both methods detected an extra copy of the ch6q22.3 arm, also extending to the end of the chromosome arm, although the BlueFuse Multi result did not indicate the copy number of the gained segment.

Next, the results from the cytogenetic analysis were compared against the results obtained from the genotype analysis (described in Chapter 2), in order to investigate whether the predicted genotypes help explain segmental gains and losses affecting the 6p21.3 locus. In targeted embryos C14 and C15, 100% of the reads from the genotypic analysis contained a single mutation (23 bp deletion and 7 bp deletion, respectively) induced by CRISPR-Cas9 editing as well as a gain of 6p21.3. Furthermore, in both cleavage blastomeres a heterozygous SNP was detected in the same amplicon, precluding the possibility that only one of the two parental alleles had successfully amplified. Several possibilities exist to explain the mechanism by which the copy of the p-arm extending from the p21.3 region to the end of the p-arm was gained: 1) both parental copies were repaired by NHEJ and resulted, by chance, in the same

indel or 2) one parental copy underwent NHEJ creating the indel and the second copy used HDR, employing the other copy (now with an indel) as a template. If editing occurred after the chromosomes duplicated (after S phase and in G2), then it is possible that one sister chromatid could have failed to repair the DSB and when the sister chromatids separated, the broken chromatid malsegregated, resulting in one of the daughter blastomeres retaining an extra copy of the telomeric chromosome 6 piece that was originally destined for the second blastomere. Similarly, in the targeted embryo C10, a single on-target mutation (1 bp deletion) was detected as well as a heterozygous SNP amplified in the same amplicon, but in this case the mutation was only present in 66% of the reads, with the remaining reads containing unedited wild-type sequence. This suggests that the embryo was heterozygous for editing. Figure 3.6 proposes a mechanism by which the observed result could be explained.

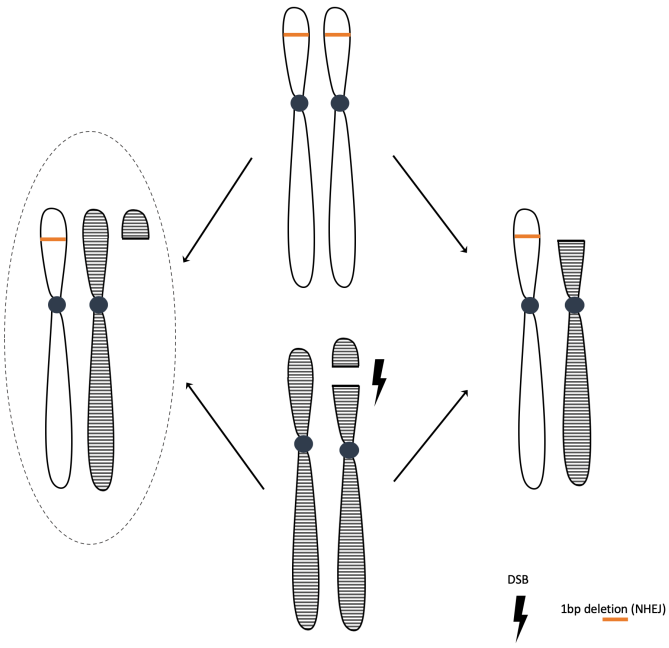
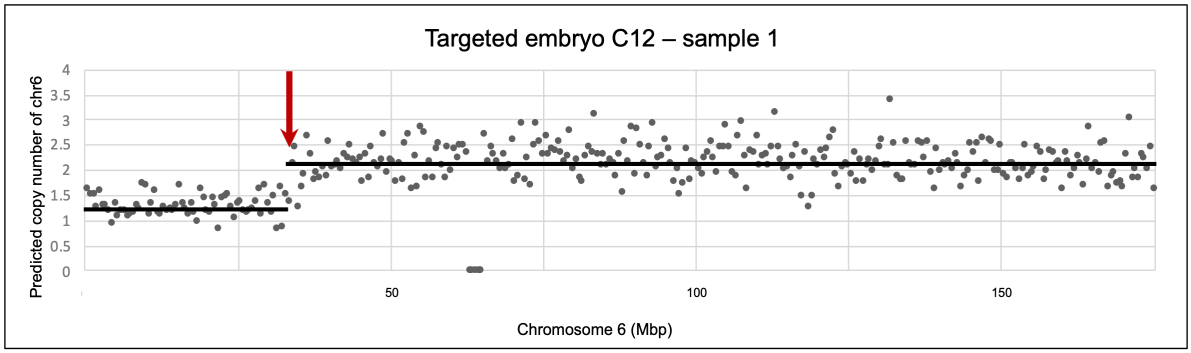
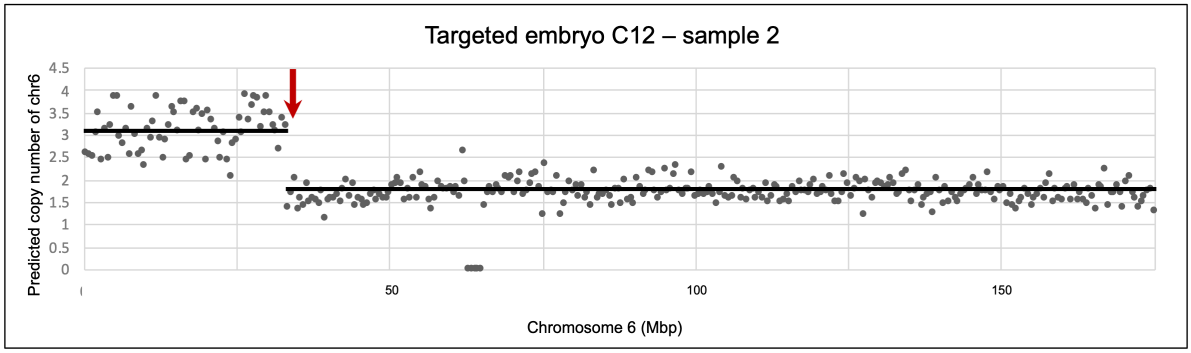


Figure 3.6: Proposed mechanism of acquiring of the 6p segmental abnormality detected in embryo C10. Editing occurred during S phase or early G2 of the first mitotic division, after chromosome duplication. One set of sister chromatids acquired a 1 bp NHEJ deletion (top). The second set of sister chromatids did not get edited but one chromatid acquired a DSB that did not get repaired (bottom). The sister chromatids then separate and the embryo divides, leaving one cell with one copy of the edited chromosome and the second with the broken 6p centromeric arm shown on the right. The second blastomere gets the second copy of the edited chromosome, one copy of the unedited chromosome and the remaining of the broken 6p telomeric arm as shown on the left in circle. The genotype analysis of the blastomere on the left reveals heterozygosity at the SNP site and at the *POU5F1* site. The cytogenetic analysis reveals an extra copy of chr6p21.3 extending to the end of the p-arm.

The most interesting result was obtained from the targeted embryo C12, where two separate trophectoderm biopsies were processed. The cytogenetic analysis identified a gain extending from 6p21.3 to the end of 6p in one sample and a reciprocal loss in the second sample of trophectoderm. Strikingly, the genotype analysis revealed that the sample with a single copy of 6p21.3 also exhibited loss of heterozygosity in the amplicon sequence. The sequence harboured a 1 bp deletion, indicating that DNA repair via NHEJ had taken place (inducing the 1 bp deletion), while the second parental sequence suffered amplification failure, presumably due to unresolved DSB occurring after CRISPR-Cas9 editing, also indicated by the loss affecting the 6p21.3 region. The second trophectoderm sample processed from this embryo contained a reciprocal gain of 6p21.3 as well as a heterozygous SNP, indicating that both parental sequences were present. The results of the segmental analysis combined with the visual representation of the genotypes in these two samples obtained from embryo 12 are presented in Figure 3.7.





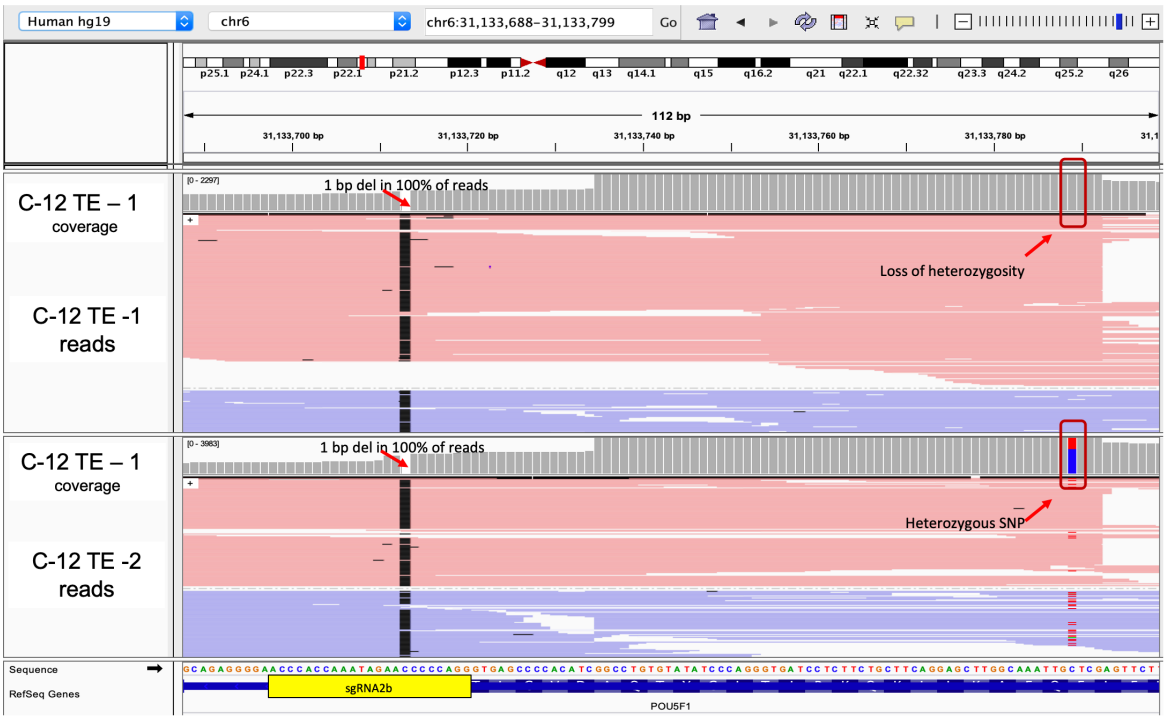


Figure 3.7: On-target and chromosome 6 analysis of the sgRNA2b-Cas9 microinjected embryo C12: comparison of copy number profiles and on-target genotypes generated from the two trophectoderm samples. The top two profiles represent the result obtained from the newly developed bioinformatic protocol, indicating a loss in the 6p21.3 region, the sgRNA-Cas9 *POU5F1* target, extending to the end of the chromosome 6 p-arm in one samples while the second biopsy profile indicates a reciprocal gain. The horizontal axes represent the position on chromosome 6 as megabases (Mbp). The arrows indicate the chromosome breakpoints. The bottom IGV (Integrative Genomics Viewer, Broad Institute) image represents the coverage of the on-target *POU5F1* region with sgRNA2b binding. The black bars indicate a 1 bp deletion that was detected in 100% of all reads in both biopsied samples. In red rectangles is the heterozygous variant, detected in the second TE sample but undetected in the TE-1 where loss of heterozygosity was observed, consistent with the loss of chr6p.21.3 region detected from the cytogenetic analysis of chromosome 6.

3.3.4 Fluorescent in situ hybridisation of chromosome 6 p-arm telomeres and centromere in CRISPR-edited human embryonic stem cells

Following the discovery that CRISPR-Cas9 targeted human preimplantation embryos suffer sub-chromosomal gains and losses induced by the editing, as evidenced by the high proportion of embryos harbouring segmental aneuploidy confined to the on-target region, we sought to investigate whether similar effects could also be observed in the hESCs induced by sgRNA targeting the same exon of the *POU5F1* gene, or whether segmental aneuploidy is a consequence of editing unique to human embryos. Unlike human zygotes, the hESCs are an unlimited resource, allowing large numbers of cells to be assessed. In this study, the copy numbers of chromosome 6 centromeres and 6p telomeres were evaluated in untreated control cells (H9) and three distinct hESC clones (induced by sgRNA1.2 with cells collected on day-2 post-induction, sgRNA1.2 with cells collected on day-4 post-induction and asRNA2b with cells collected on day-2 post-induction). The results from the transfected lines were then compared with the results obtained from the control cells for significance. In each category, centromeric (red) and telomeric (green) signals were counted in a minimum of 500 cells. In theory, if CRISPR-induced DNA strand breaks led to loss or duplication of the chromosomal segment telomeric to the breakpoint, as had been observed in the samples from human embryos, then there should be a disparity in the copy number of telomeres with respect to centromeres (fewer telomeres in the case of a segmental loss and more in the case of a segmental gain). The complete results are summarised in Table 3.2. Overall, no statistically significant differences were observed between the number of cells with abnormal number of 6p telomeres and centromeres (represented by anything other than two red signals and two green signals in the same cell) between the CRISPR treated cells and controls (Figure 3.8). The analysis revealed that the great majority of cells in each category (above 93%) contain two

copies of the chromosome 6p telomere and two copies of the centromere. Thus, CRISPR editing does not appear to increase the likelihood of segmental aneuploidy in this cell type, strengthening the argument that the cells of human preimplantation embryos may be unusually susceptible to this type of error following induction of a double strand DNA break.

Table 3.2: Summary of FISH results. The targeted ES cells were induced with sgRNA-Cas9 targeting exon 2 of *POU5F1* (identical to the exon targeted in human zygotes) and cultured for several days prior to fixation (on day-2 or -4 post-induction). The H9 control ES cells were untreated. The p-arm telomeric and centromeric signals on chromosome 6 were counted under fluorescent microscope in a minimum of 500 cells and placed into the following categories according to how many signals were detected. Ch6 C: centromeric signal on chromosome 6, ch6p: telomeric signal on chromosome 6. The probe efficiency was estimated from the number of individual signals detected in 531 untreated control cells.

Cell clone	2 ch6 C / 2 ch6p	% of cells with 2 ch6 centromeres / 2 ch6p telomeres	1 ch6 C / 1 ch6p	2 ch6 C / 1 ch6p	1 ch6 C / 2 ch6p	3 ch6 C / 2 ch6p	2 ch6 C / 3 ch6p	2 ch6 C / 0 ch6p	3 ch6 C / 3 ch6p	1 ch6 C / 3 ch6p
H9 Control	502	94.54	5	8	15	1	0	0	0	0
Clone 2.1 (Day 2)	519	93.68	11	7	14	2	1	0	0	0
Clone 2.1 (Day 4)	538	93.40	3	11	20	2	1	0	1	0
Clone 2b-1 (Day 2)	543	94.27	3	14	10	0	1	1	2	2

Probe efficiency		
CEP Orange ch6	510/531*100	96.04%
Ch6p Green	518/531*100	97.55%

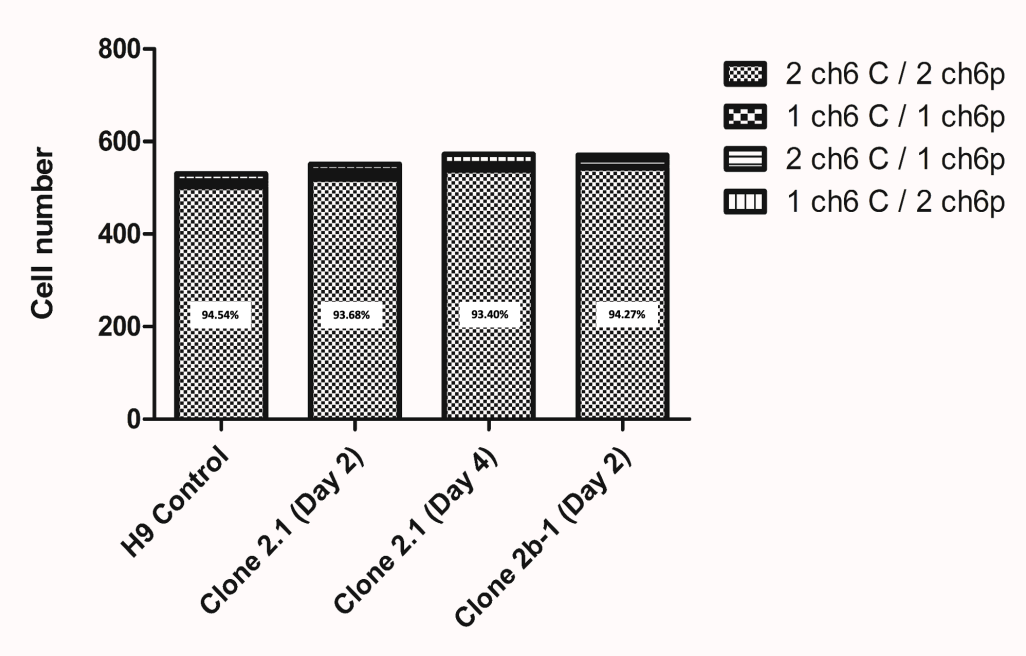


Figure 3.8: Quantification of centromeric and telomeric signals detected in targeted ES cells and untreated controls. The cells were collected following sgRNA2.1 induction for 2 and 4 days (n=554, n=576, respectively) sgRNA2b induction for 2 days (n=576) and uninduced H9 human ES control cell clone (n=531). Differences in numbers of putative segmental aneuploidies in sgRNA-induced and uninduced ES cell clones were non-significant (p=0.12, Chi-square test). Ch6 C: centromeric signal on chromosome 6, ch6p: telomeric signal on chromosome 6.

Figure 3.9 shows microscope images of the control H9 cells generated by the Metafer analysis software.

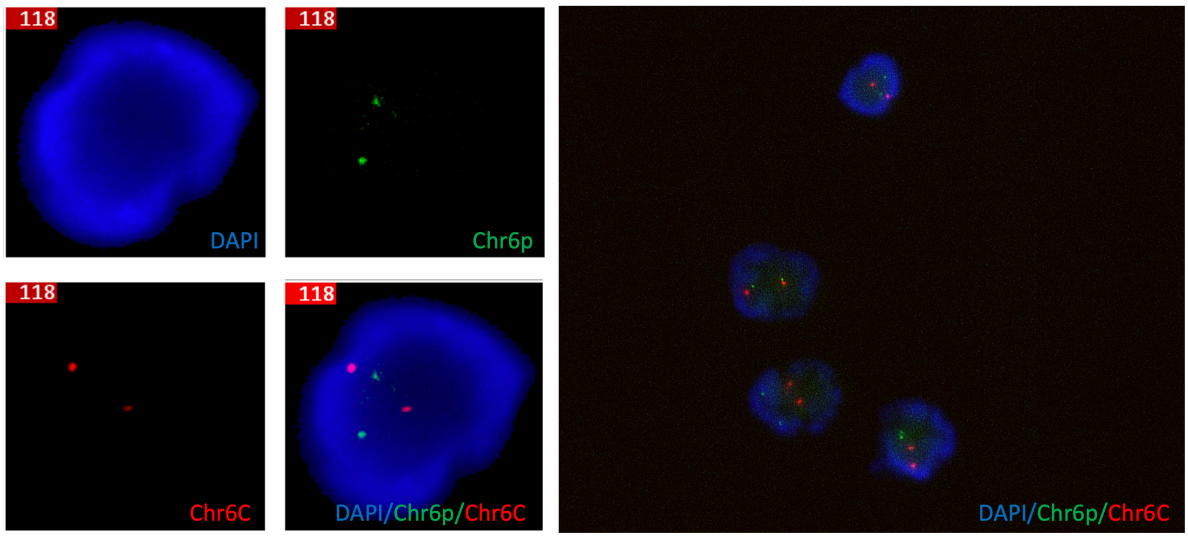


Figure 3.9: Microscope images of the H9 control cells analysed by the Metafer software. The cells were hybridised with fluorescently labelled chromosome 6 centromeric and 6p telomeric probes (in red and green, respectively), using the red and green filter set. In the upper left corner is an image of a nucleus with DNA stained by DAPI, taken using a DAPI filter set in the ultraviolet spectrum. The image on the right and bottom right is a joined overlaid image. All four cells shown here were classified as normal (with 2 signals from each probe).

Finally, an example of a cell from each category (untreated control, and three targeted ES cells clones) where two centromeric signals and two 6p telomeric signals were detected is presented in Figure 3.10.

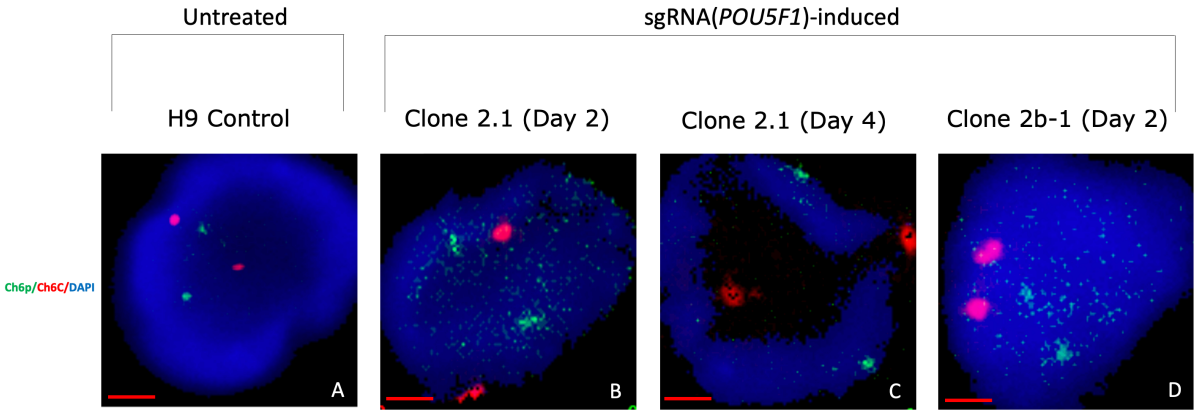


Figure 3.10: Microscope image of fluorescence *in situ* hybridisation of chromosome 6 centromeres (in red) and chromosome 6p telomeres (in green) in *POU5F1*-targeted hESC clones and untreated controls A) untreated control cells; B) 2 days post-sgRNA2.1 induction; C) 4 days post-sgRNA2.1 induction D) 2 days post-sgRNA2b induction. DAPI nuclear staining (blue). Images were analysed in the Metafer software (version 3.9.4) scale bars: 2 µm.

3.4 DISCUSSION

3.4.1 Unresolved chromosome breakage after CRISPR-Cas9 editing of human zygotes

The aim of this chapter was to investigate the extent to which human preimplantation embryos repair their DNA following CRISPR-Cas9-based genome editing. This question was based on a hypothesis that human embryos prior to the activation of their genome are at risk of persistent DNA damage due to a diminished DNA repair capacity. If DNA repair is, indeed, deficient at this stage of development, there may be an increased risk that DSBs remain unresolved, potentially leading to genetically abnormal cells, mitotic arrest and/or apoptosis. It is likely that any such outcomes will ultimately be deleterious to embryo survival. If correct, this would undoubtedly carry significant implications for those considering the use of GE technologies for the treatment of inherited disorders.

The results presented in this chapter reveal that a complex spectrum of on-target mutations can be observed after CRISPR-Cas9 editing of human zygotes. The results showed no significant increase in the rate of whole chromosome aneuploidy in the targeted embryos. However, 37.5% of all embryos subjected to CRISPR were shown to have suffered segmental gains or losses affecting 6p21.3-pter. The novel bioinformatic protocol developed during the course of this study was able to confirm that the segmental aberrations were associated with breakpoints that map precisely to the targeted *POU5F1* locus, suggesting they are a direct consequence of the editing. Most commonly, gains/losses were associated with an altered proportion of reads beginning with the bin immediately adjacent to the *POU5F1* cut site and extending to the end of the short arm of chromosome 6. Chromosome instability, including a high incidence of segmental aneuploidy, is well documented in human preimplantation embryos (Babariya et al.,

2017; Vanneste et al., 2009). In this dataset, in contrast to Cas9-microinjected control embryos, the sgRNA2b-Cas9-targeted embryos exhibited an elevated frequency of segmental gains and losses, with the increase entirely attributable to chromosome breakage occurring at the targeted site (representing 53.8% of all detected segmental errors in the CRISPR treated group). These segmental anomalies were observed in genetically distinct embryos, derived from different parents, further supporting the notion that they arise as an unintended consequence of CRISPR-Cas9.

Taken together, the results reported in this chapter indicate that human preimplantation embryos are highly susceptible to induction of DNA strand breaks by CRISPR-Cas9. This suggests that genetic instability during the first few days of embryonic development extends beyond the well-established high frequency of mitotic chromosome malsegregation and includes a sensitivity to DNA damage, as evidenced by an extremely high rate of successful targeting when compared to other cell types. Furthermore, strong evidence is presented in support of a hypothesis that human preimplantation embryos, at least those prior to EGA, have a diminished ability to repair DSBs.

3.4.2 CRISPR-Cas9 editing may lead to complex structural variations

The segmental errors identified by cytogenetic analysis presented in this thesis most likely point to the occurrence of complex genomic rearrangements in the sgRNA2b-Cas9 targeted samples, such as chromosomal translocations or end-to-end fusions, as derivative chromosomes lacking telomeres are unlikely to be stable through multiple cell divisions, while those lacking centromeres cannot be captured by the mitotic spindle and would eventually be lost (Capper et al., 2007; Lo et al., 2002; Van Steensel et al., 1998). In order to evaluate this possibility, traditional cytogenetic approaches such as G-banding, possibly supplemented by

FISH-based methods such as chromosome painting, could reveal the status of individual chromosomal regions. However, given the highly restricted numbers of cells available and the difficulties inducing individual cells to form metaphase plates suitable for cytogenetic analyses, strategies employing NGS might be more appropriate for the characterisation of any chromosome rearrangements.

As evident from the FISH analysis carried out in this work, the increased frequency of segmental abnormalities resulting from Cas9 editing does not appear to be present in hESC, the closest 'relative' of cells of the preimplantation embryo that can be easily assessed and in large numbers. In hESCs there was a normal number of chromosome 6p-arms and centromeres in over 93% of the assessed cells, essentially the same as controls. Unfortunately, carrying out FISH on human embryos comes at a cost, since the cell (or whole the embryo) would cease to be available for any other type of testing interrogating embryonic DNA. One alternative possibility is to use the newly developed Cas9 enrichment protocol, followed by long read sequencing and structural variant discovery. This approach, as already discussed in detail in Chapter 2, has a great potential for evaluation of all possible consequences of Cas9 editing, defining on-target genotype complexity (small indels, HDR repair, loss of heterozygosity and large deletions), off-target activity, as well as permitting the detection and characterisation of structural variation.

3.4.3 On-target complexity after CRISPR-Cas9 editing in human preimplantation embryos

The existence of segmental aneuploidy associated with on-target CRISPR editing is also supported by the observed loss of heterozygosity (LOH) in the sample in which low-pass NGS had identified a deletion of 6p21.3-pter. The LOH was detected using an independent

genotyping technique (described in detail in Chapter 2). Interestingly, the second trophectoderm sample of the same embryo revealed a reciprocal gain of the same region. Given the findings of the current study, it is clear that segmental abnormalities observed in any one cell analysed from an embryo are not always representative of the entire CRISPR-Cas9 targeted embryo, and may only reflect a subset of cells in some instances (Alanis-Lobato et al., 2020). Future studies involving germline genome editing should investigate the extent to which rates of segmental mosaicism affect CRISPR-treated embryos in comparison with the reported values from untreated embryos undergoing preimplantation genetic testing for aneuploidy. Altogether, the results further support the need to develop robust methodologies for testing of cells and embryos for a diverse set of genetic modifications following any attempt at genome editing. In this thesis, this need was addressed by testing for a spectrum of on-target mutations, ranging from small indels (several base pairs in size) to larger deletions (up to 23 bp deletion) as well as more expansive sub-chromosomal aberrations. Furthermore, the analysis of additional cells from embryos belonging to the same dataset but processed by our colleagues at the Francis Crick Institute identified a larger on-target deletion of 330 bp in size in one targeted embryo (Norah M. E. Fogarty et al., 2017). Most recently, the same dataset obtained from the low-pass whole genome sequencing was interrogated for genome-wide SNP analysis by collaborators at the Crick, revealing regions several kilobases in size affected by loss of heterozygosity (LOH), extending from and/or spanning through the on-target site (this work is currently under review for publication at the Proceedings of the National Academy of Sciences). Not unexpectedly, some of the samples affected by LOH belonged to the embryos that had previously failed to yield PCR products when targeted amplification had been attempted or where only one allele was detected at the CRISPR targeted site. The fact that parental DNA was not available for parallel analysis further complicated the interpretation of the results in some cases, as one can cannot exclude the possibility that a lack of heterozygosity

was simply due to these embryos inheriting a homozygous genotype. Therefore, the presence of a heterozygous SNP in at least one additional cell from the same embryo (confirming inheritance of heterozygosity at the time of conception) was required in order to call putative LOH events. Our colleagues interrogated SNPs in 24 kilobases extending upstream and downstream from the target site, where each identified heterozygous SNP was amplified in a separate amplicon from the WGA product. The SNP profiles identified three different patterns in which LOH events could manifest in the targeted embryos: 1) LOH at the on-target site only; 2) 'bookended' LOH where the size of the region with LOH can be estimated from the nearest heterozygous SNP detected in that sample on either side of the affected region because it was flanked by SNPs where heterozygosity had been maintained and thus its maximum limits were defined and 3) 'open ended' LOH where all potentially heterozygous sites are affected by LOH and the centromeric and telomeric limits of the region remain unknown. Of note, to confirm the 'bookended' deletions, one could try PCR using a forward primer from an upstream amplicon that is certain not to be affected by LOH (due to successful detection of a heterozygous SNP) in combination with a reverse primer from a downstream amplicon (which also had a heterozygous SNP. If the LOH is real, then the PCR should generate a product that is significantly truncated. The reduced size of the fragment could be visualised quickly and cheaply using simple agarose gel electrophoresis, provided that the expected size of the PCR product does not exceed the fragment length of the WGA template.

3.4.4 DNA damage response following CRISPR-Cas9 editing may be partly mediated by p53

The mechanisms of DNA repair after the introduction of DSBs specific to human preimplantation embryos remain largely unknown. There are questions whether such

mechanisms are fully active and concerning the extent to which embryo cells can cope with the DNA strand breaks induced by GE techniques. Studies in cell lines have demonstrated that the efficiency of the HDR is greatly enhanced by tumour suppressor p53 inhibition (Haapaniemi et al., 2018; Ihry et al., 2018). Haapaniemi et al. (2018) reported that CRISPR-Cas9 induces a p53-mediated DNA damage response, leading to cell cycle arrest and apoptosis in retinal pigment epithelial cell line and there is a selection against the cells with functional p53 pathways and increased editing efficiency and HDR rates after p53 disruption. The results of Ihry at al. (2018) indicated that human pluripotent stem cells are capable of achieving >80% indel frequencies if they acquire *TP53* mutations. However, they concluded that DSBs induced by CRISPR-Cas9 are toxic and kill most of the cells, and this effect is apparent when transfection efficiency is high. The toxicity presents a challenge for both the high-throughput CRISPR screens used for genome engineering and for CRISPR-based cell therapies, including those proposed targeting the germline. Whether the same applies to human preimplantation embryos is not known, but if there was a need to inhibit p53 in embryos in order to achieve efficient genome editing, this would likely be problematic for clinical application. The *TP53* gene is a master regulator of the cell cycle, DNA repair and apoptosis and its disruption, even if only temporary, might well have serious consequences for embryonic development. To account for this possibility, p53 function should be monitored when developing cell-based therapies utilising CRISPR–Cas9. More recently, Cullot et al. (2019) observed dramatic megabase-scale chromosomal truncations in CRISPR-targeted human embryonic kidney (HEK) cells at 10% increased frequency compared to the controls, and found that the frequency of edits correlate with abnormal karyotypes and p53 function, which may impact their DNA damage response machinery. The analysis of human fibroblasts did not generate similar effects, although knocking out *TP53* increased the frequency of large deletions 10-fold. Future

experiments utilising CRISPR-Cas9 technology in human embryos, should, therefore, also screen the embryos for p53 function.

3.4.5 The use of CRISPR-Cas9 editing to study early human development

Although the work presented in this chapter focused mostly on the consequences of CRISPR-Cas9 editing at the cytogenetic and DNA sequence levels, and on the capacity of the DNA repair machinery in embryos prior EGA, the dataset analysed was part of a larger study in collaboration with our colleagues at The Francis Crick Institute, whose principal aim was to apply germline genome editing to investigate the regulation of early human embryogenesis. The laboratory of Dr Kathy Niakan selected the *POU5F1* gene, encoding OCT4, because the OCT4 protein it produces plays an essential role in the maintenance of pluripotency within the inner cell mass, inhibiting the acquisition of trophectoderm fate. Although the function of the murine homologue of *POU5F1* had been thoroughly investigated using traditional transgenic approaches, its precise role in human embryogenesis remained unclear. Using the CRISPR-Cas9 system to disrupt the coding region of *POU5F1*, leading to a premature stop codon, they showed that the function of OCT4 is not fully conserved between humans and mice. While mouse *POU5F1*-edited embryos formed blastocysts consisting of only extra-embryonic trophectoderm cells, as evidenced by expression of the trophectoderm marker CDX2, the ability to form and maintain blastocysts was compromised in the human targeted embryos. Transcriptomic analysis revealed that *POU5F1*-null cells downregulated not only extra-embryonic trophectoderm genes, such as *CDX2* and *GATA2*, but also regulators of the pluripotent epiblast, including *NANOG* (Norah M. E. Fogarty et al., 2017). While OCT4 protein expression was downregulated to undetectable levels in most CRISPR-edited embryos

arrested at the cleavage stage, the analysis of blastocysts revealed at least one cell expressing nuclear OCT4, suggesting that only embryos with partial OCT4 expression are able to progress to the blastocyst stage (Norah M. E. Fogarty et al., 2017). Interestingly, a recent study of Daigneault et al. (2018) identified a similar phenotype after targeting *POU5F1* in bovine embryos to that observed in human ones. The mis-expression of genes associated with all three blastocyst lineages (the trophectoderm, the epiblast, and the primitive endoderm) in targeted human blastocysts further indicates that OCT4 may have an essential function before reaching this stage, potentially prior to the onset of EGA. Future experiments should investigate whether *POU5F1* mutations alter gene expression before reaching the blastocyst stage, which may explain the failure of blastocyst formation. Alternatively, targeting *POU5F1* in human embryos later in development, after the onset of EGA, may bypass its earlier critical role (if any) and thereby delineate its specific function in the fully formed blastocyst (Norah M. E. Fogarty et al., 2017).

The main limitation of this study was that, at the time, it was not possible to conclusively demonstrate that the inability to form a blastocyst is a direct consequence of the OCT4 loss, and not at least in part due to unresolved DNA damage, leading to developmental arrest. Furthermore, some of the embryos have loss or duplication of a large number of genes on chromosome 6p due to segmental aneuploidy. Ideally only cytogenetically normal embryos should be considered when trying to ascertain the function of OCT4. Therefore, targeting non-essential genes, or genes that have later developmental functions, could help supplement the understanding of the OCT4 phenotype. Studies like the one presented here illustrate the importance of carrying out research in human embryos. Although studies in rodents have transformed the understanding of mammalian embryogenesis, there are limits to what can be learned from model organisms. It appears that important differences exist in the underlying biology between different species and these aspects may only be appropriately addressed by

performing appropriate investigations in human embryos. For this purpose, CRISPR-Cas9 technology is a powerful method to study gene function in previously inaccessible contexts.

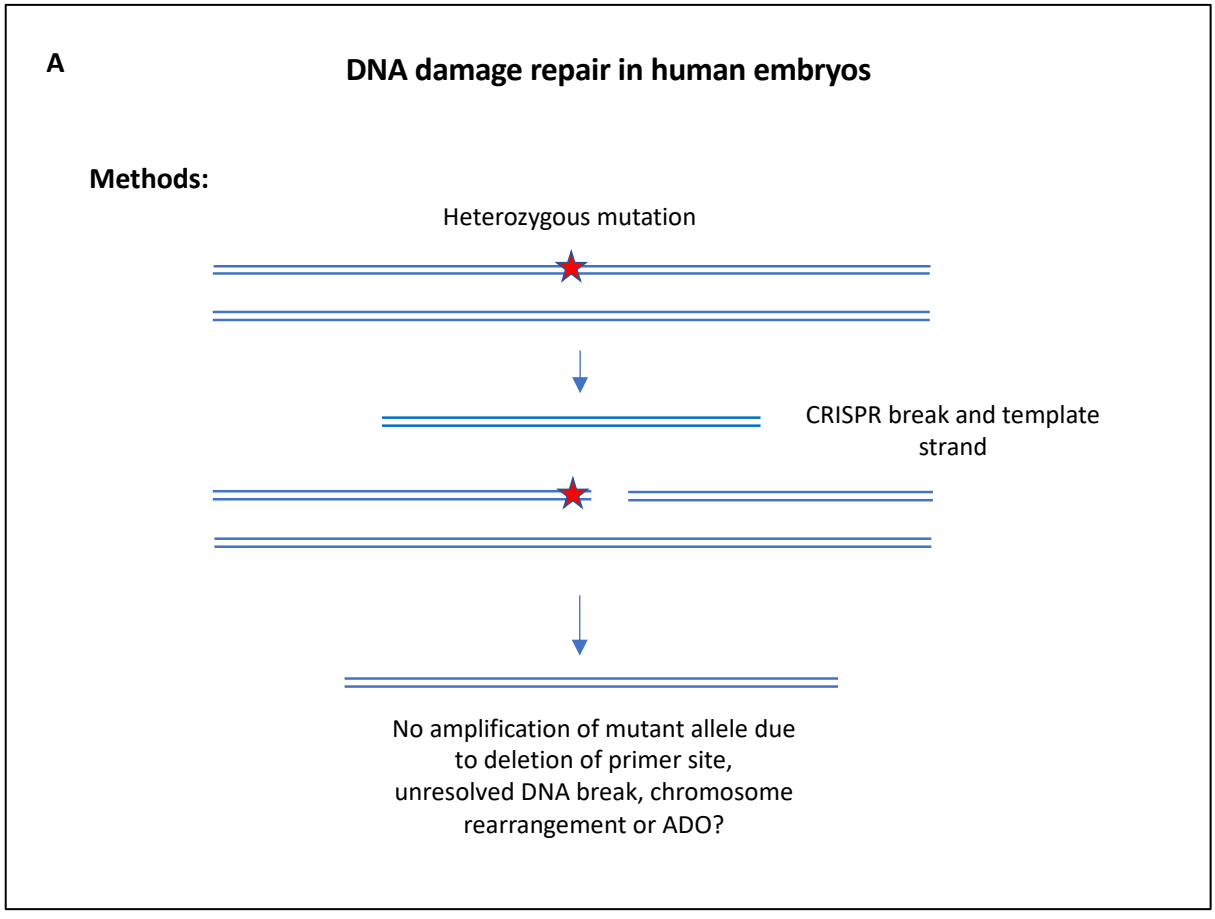
3.4.6 Future perspectives

Germline genome editing has been democratised by the use of CRISPR-Cas9. The technology is a powerful tool with tremendous potential to investigate basic regulation of human development and function of genes. Germline genome editing has also been proposed as a tool for correcting genetic defects responsible for human disease at the preimplantation stage, as an alternative strategy to preimplantation genetic testing – potentially a 'cure' rather than a strategy of diagnosis and exclusion. However, as discussed in detail in this thesis, the application of genome editing to human embryos is controversial for two principal reasons. Firstly, any induced alteration at an early embryonic stage is likely to affect the germline of the resulting individual, leading to a heritable change to the human genome. This is seen as ethically challenging. Secondly, the safety and efficacy aspects of genome editing in human preimplantation embryos have not been adequately assessed.

The work described in this chapter has indicated that the use of CRISPR-Cas9 to produce breaks in the DNA of human preimplantation embryos may be problematic due to a deficiency of DNA repair in blastomeres prior to embryonic genome activation. Although application of CRISPR-Cas9 to oocytes at fertilisation or at the pronuclear stage has been seen as an attractive strategy, capable of ensuring delivery of genome editing components into all cells of the future individual, there is now evidence to suggest that it may also be the time when embryos are most vulnerable to DNA damage. The failure to appropriately resolve DNA breaks has profound implications for the use of CRISPR/Cas9 as a therapeutic tool applied to gametes and preimplantation embryos. In order to verify the findings presented in this thesis, as well to build

on the body of work that has already been published, future experiments should seek to further elucidate the kinds of DNA repair mechanisms that are present in the early embryo. Studies harnessing CRISPR-Cas9 and related technologies should ideally induce DNA damage in embryo cells in a highly controlled manner, allowing a critical evaluation of DNA repair capacity. For investigation of the cellular response to CRISPR-induced DNA damage, it would also be beneficial for the targeted region to be situated in a non-coding area of the genome, since this would ensure that any resulting phenotype can be attributed to the editing alone and not due to loss of function of a critical gene. Critical evaluation of response to DNA damage would also benefit from a priori knowledge of the sequence of the parental genomes. This would allow identification of regions of the male and female genomes containing a contiguous string of variants where the two parents are homozygous for opposite alleles. Such sites are of value because the genotype at multiple positions in the vicinity of the targeted site can be predicted (in the absence of any cytogenetic abnormality all resulting embryos would inevitably be heterozygous at all interrogated loci). Having several heterozygous sites in close proximity to the targeted site could, with certainly, detect all instances of allele drop-out (evident from the loss of heterozygosity) and distinguish between different kinds of DNA repair mechanisms (Figure 3.11). For example, a heterozygous variant could be targeted and an exogenous DNA template carrying base modifications supplied for HDR. If HDR occurs, the presence or absence of base modifications in the resulting sequence reveal whether the exogenous template or the endogenous non-targeted allele were used for repair. The inability to distinguish between these two possibilities was the principal cause of criticism of the pioneering study of Ma et al. (2017) targeting human zygotes with an HDR template to eliminate a dominant form of cardiomyopathy. In fact, none of the studies conducted up to this date, to our knowledge, have taken into consideration parental haplotypes. Combining the analysis of haplotypes together with long-read sequencing using the Cas9 enrichment protocol

as well as in-depth cytogenetic analysis should allow explanation of all possible resulting genotypes. Figure 3.11 compares the two approaches, one (A) where there is no knowledge of the parental haplotypes, producing ambiguous results, which are often difficult to interpret, and the second (B) where information concerning the DNA sequence of the parents/donors have been acquired prior to CRISPR-Cas9 targeting, enabling detection and interpretation of all possible outcomes.



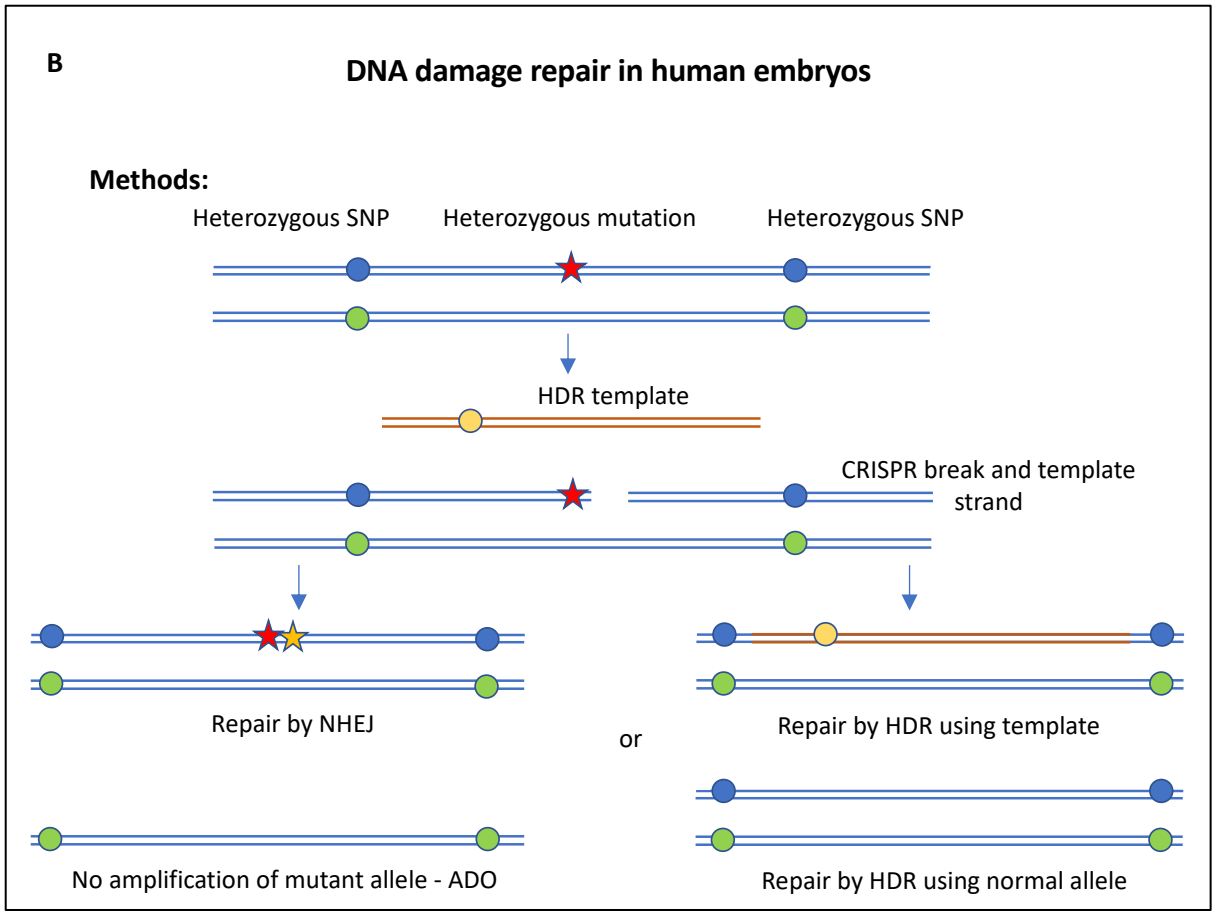


Figure 3.11: Schematic representation of possible outcomes after CRISPR-Cas9 editing. A) When there is no prior knowledge of the parental DNA. Targeting a heterozygous mutation in human embryos causes a DSB, which can either be repaired by HDR using the template strand or NHEJ. If HDR is performed using the normal wild-type allele, the resulting genotype is homozygous. The homozygous genotype is indistinguishable from ADO as well as from the genotypes involving failed amplification of the allele due to an unresolved DSB, structural rearrangement or a larger deletion causing a loss of primer annealing site. **B)** Parental haplotypes are acquired prior to targeting of human embryos. Targeting a heterozygous mutation in human embryos causes a DSB, which can either be repaired by the HDR or NHEJ. The two possible HDR outcomes, based on whether the supplied exogenous HDR template or the other normal allele was used as a template, can be differentiated from different genotypes produced. Allele drop-out (ADO) is detected from the loss of heterozygosity in the affected region.

Once a suitable combination of parents/donors is identified, their oocytes and sperm could be used to create embryos using standard assisted reproductive techniques. CRISPR-Cas9 editing should be attempted at the time of fertilisation for some oocytes and after the onset of EGA approximately 3 days later for other embryos, to allow for comparison of DNA repair capacity prior to and after EGA. The resulting embryos would need to be cultured for several divisions prior to disaggregation into individual blastomeres. All blastomeres should then be processed

and analysed separately, in order to determine the exact chromosomal make-up as well as to evaluate all possible consequences of genome editing, including mosaicism. Given the controversial report of babies born in China following the application of CRISPR-Cas9, there is an urgent need for quality research into the safety of these technologies. If further studies confirm that human embryos are sensitive to genotoxic damage, this will also have significant implications for assisted reproductive treatments, potentially helping to guide the formulation new embryo culture systems, with a focus on preserving DNA integrity and supporting DNA repair. Current IVF embryo culture techniques take little or no account of such considerations, a fact that might partially explain why the transfer of a euploid blastocyst only yields a viable pregnancy in about two-thirds of cases. It is anticipated that such studies will provide various translational opportunities and, in the longer term, the findings are expected to be biologically insightful as well as clinically impactful. Areas that stand to benefit are the fields of assisted reproductive treatment, genome editing, and other disciplines where embryonic cells are utilised or where genome instability is relevant, including stem cell and cancer research.

Concluding remarks

Since the introduction of PGT, the methods available for genetic analysis have undergone rapid and sustained evolution. Today, these advances promise to provide universal protocols for the simultaneous detection of all inherited defects caused by gene mutations or chromosome abnormalities. Such methods could provide PGT at lower costs, with shorter patient waiting times and deliver enhanced accuracy. The first chapter of this thesis focused on development, validation and clinical implementation of a universal PGT-M protocol for the detection of ß-globin mutations. The combination of minimal work-up and high throughput provided by this protocol resulted in an economical test. The issue of cost is of great relevance in this particular case, given the fact that many regions of the world where *HBB* mutations are of high prevalence are relatively resource poor. The experience from the present study confirms that NGS can provide a rapid, accurate and streamlined solution for couples seeking to use PGT to avoid genetic disease transmission. It is expected that, in the future, additional protocols similar to the one described here will be developed for the testing of other single gene disorders where mutation heterogeneity leads to problems for conventional PGT. The arrival of a new generation of inexpensive sequencers, such as the MinION pioneered by Oxford Nanopore Technologies and the more recent Flongle adapter (retailed at a fraction of the cost compared to a full-sized flow cell), hold great promise for allowing further cost reductions. The use of such equipment and consumables could reduce costs per-sample by avoiding the high capital costs of setting up a modern molecular biology laboratory, while the introduction of novel molecular methods (e.g. isothermal amplification and lateral flow assays) may mean that other analytical equipment, and the highly skilled staff needed to use it, become less of a necessity. If combined with non-invasive methods of sampling genetic material from preimplantation

embryos, costs could fall further still and risks to the embryos of the specific conditions tested (already low) would be essentially eliminated.

The easy access to genomic technologies and the continuous decrease in their cost is accelerating the development of more comprehensive testing platforms. Recently, a new form of testing has been implemented clinically, prioritising preimplantation embryos for uterine transfer according to the polygenic disease risk prediction. Although this could be considered as desirable and as a valuable addition to the important health-related data revealed during the course of PGT, with powerful new genetic technologies also comes new clinical and ethical questions. How will an embryo with an increased risk for one condition be weighed against a sibling embryo with an elevated risk for a different disease. Expansion of embryo testing to cover multiple genetic loci (or even entire genomes) that together confer a degree of risk, sometimes interacting with each other and with the environment in ways that are not fully understood, represents a significant deviation from the traditional application of PGT. As we enter an era of whole genome sequencing, we must ask how much information we really want to obtain from each embryo. If we learn all aspects of an embryo's genetics, how will patients be counselled and what elements should be considered appropriate for embryo selection.

Recently, there has been excitement over the development of new genome editing methods. Such techniques offer the hope that mutations causing inherited conditions could one day be corrected during the preimplantation stage, although considerable public consultation, thorough ethical debate, advancements in regulation and oversight, as well as assessment of safety still need to take place before such methods enter clinical practice. After the exploration of new PGT protocols employing modern technologies that was the major theme of the first chapter of this thesis, the second experimental chapter considered the possibility of a transition from the diagnosis and exclusion of affected preimplantation embryos to their treatment and 'cure'. The chapter focused on the technical feasibility of genome editing based on the

CRISPR-Cas9 system in the context of the earliest stages of human life (fertilisation and preimplantation development). The analytical methodology developed permitted a comprehensive review of the performance of CRISPR-Cas9 in preimplantation embryos, including characterisation of mutational spectra at the targeted site and the detection of any off-target activity. In particular, the use of an enrichment protocol based upon CRISPR-Cas9 and long-read sequencing represents a significant advance in the way in which the outcome of genome editing can be evaluated. Taken together, the strategies presented represent an excellent methodological framework for pre-clinical phase studies and necessary technology validation.

Finally, the third chapter of this thesis evaluated the repair outcomes after CRISPR-Cas9-induced double-stranded DNA breaks in human preimplantation embryos. The results showed that a significant number of embryos exhibit segmental aneuploidy associated with genome editing, as evidenced by the fact that chromosomal breakpoints were localised to the on-target site, leading to gains and losses of the distal portion of the affected chromosomal arm. This striking unintended consequence of CRISPR-Cas9 should serve as a cautionary note to those considering the technique for clinical use. Since gross cytogenetic lesions have not been a frequent observation following genome editing applied to other cell types, these results suggest that DNA repair may be compromised in the early human embryos, particularly prior to the onset of embryonic genome activation. Initially, the induction of a double strand DNA break during the genome editing process does not seem to interfere with the progression of the cell cycle, in fact affected embryos appear to be tolerant of segmental anomalies up to the cleavage stage (and in some cases the blastocyst stage). Ultimately, however, the presence of segmental aneuploidy is expected negatively impact viability, since previous studies have documented lower frequencies of sustained development and reduced implantation potential for embryos affected by such abnormalities. Furthermore, it is not known whether accumulated, unresolved

DNA damage increases the risk of complex structural rearrangements, potentially generating a disease phenotype if embryonic development is allowed to carry on and a child is born. Together, the results presented in Chapters 2 and 3 revealed some of the limitations in our understanding of the basic biology of human development, as well as uncertainty about how DSBs induced by CRISPR-Cas9 are processed within the cells of preimplantation embryos and their ultimate impact on the embryonic genome. It is clear from these findings that clinical utilisation of genome editing should not be considered until more research has been undertaken to solidify knowledge in areas where important deficiencies currently exist. It is imperative that the eventual application of genome editing in a clinical context is able to ensure the birth of healthy, disease-free children without any potential long-term complications resulting from the editing procedure. For use in human reproduction, safety and efficacy are paramount, and applying similar rigor as has been done in handling the validation and clinical introduction of mitochondrial replacement therapy, also a form in heritable editing, is crucial for determining the therapeutic potential of genome editing during the preimplantation stage of development. Nonetheless, as a powerful research tool, genome editing is an excellent technique for the study of human development and gene function that will undoubtfully continue to expand our knowledge in this area.

Publications arising from this work

Articles published in international peer-reviewed journals:

Elinati, E., Zielinska, A., McCarthy, A., **Kubikova, N.**, Maciulyte, V., Mahadevaiah, S., Sangrithi, M.N., Ojarikre, O., Wells, D., Niakan, K.K., Schuh, M., Turner, J.M.A., 2020. The BCL-2 pathway preserves mammalian genome integrity by eliminating recombination-defective oocytes. ***Nature Communications***. 1-10. doi.org/10.1038/s41467-020-16441-z

Kubikova, N., Babariya, D., Sarasa, J., Spath, K., Alfarawati, S., Wells, D., 2018. Clinical application of a protocol based on universal next-generation sequencing for the diagnosis of beta- thalassaemia and sickle cell anaemia in preimplantation embryos. ***Reproductive Biomedicine Online*** 37. 136–144. doi:10.1016/j.rbmo.2018.05.005

Fogarty, N.M.E., McCarthy, A., Snijders, K.E., Powell, B.E., **Kubikova, N.**, Blakeley, P., Lea, R., Elder, K., Wamaitha, S.E., Kim, D., Maciulyte, V., Kleinjung, J., Kim, J.-S., Wells, D., Vallier, L., Bertero, A., Turner, J.M.A., Niakan, K.K., 2017. Genome editing reveals a role for OCT4 in human embryogenesis. ***Nature.*** 67-73. doi:10.1038/nature24033

Manuscript currently under consideration at Proceedings of the National Academy of Sciences (PNAS), available as preprint:

Alanis-Lobato, G., Zohren, J., McCarthy, A., Fogarty, N.M.E., **Kubikova, N**., Hardman, E., Greco, M., Wells, D., Turner, J.M.A., Niakan, K.K., 2020. Frequent loss-of-heterozygosity in

CRISPR-Cas9-edited early human embryos. bioRxiv 2020.06.05.135913. doi:10.1101/2020.06.05.135913

Book chapter:

Book title: Human Reproductive Genetics, 1st Edition

Editors: Garcia-Velasco, J and Seli, E

Publisher: Elsevier

Chapter Title: **Future Technologies for (Pre-)Implantation Genetic Applications**

Authors: Nada Kubikova and Dagan Wells

In press, Paperback ISBN: 9780128165614, Published on 1st May 2020

Published peer-reviewed abstracts from international meetings:

Cutts, G; Simpkins, M; Spath, K; **Kubikova, N**; Alfarawati, S; Wells, D; Fragouli, E., Patient characteristics influence the outcome of embryo aneuploidy testing cycles, *Human Reproduction,* 32, 425-426, 2017.

Fragouli, E; Alfarawati, S; Simpkins, M; Cutts, G; Spath, K; Babariya, D; **Kubikova, N**; Rubistello, L; Munne, S; Wells, D., Factors affecting embryonic mosaicism, *Human Reproduction*, 32, 49-50, 2017.

Kubikova, N; Spath, K; Babariya, D; Simpkins, M; Cutts, G; Alfarawati, S; Fragouli, E; Wells, D., 'Batching' of multiple IVF cycles combined with preimplantation genetic testing for aneuploidy: how does the number of cycles undertaken affect outcomes?, *Human Reproduction*, 32, 430-431, 2017.

Wells, D; Babariya, D; Alfarawati, S; Spath, K; **Kubikova**, **N**; Munne, S; Fragouli, E., Frequency and clinical relevance of mosaic segmental aneuploidy in blastocyst stage human embryos, *Human Reproduction*, 32, 50-51, 2017.

Simpkins, M; Alfarawati, S; Cutts, G; Spath, K; **Kubikova, N**; Wells, D; Fragouli, E., The relationship between blastocyst morphology, euploidy, aneuploidy and mosaicism, *Human Reproduction*, 32, 50-50, 2017.

Spath, K; **Kubikova, N**; Whitney, M; Vaid, M; Rozis, G; Couchman, V; Glynn, K; Batha, S; Alfarawati, S; Fragouli, E., Development, validation and first clinical application of a novel ultra-rapid comprehensive chromosome screening technique utilising an array based nano scale quantitative DNA amplification technology., *Human Reproduction*, 32, 16-16, 2017.

Babariya, D; Fragouli, E; Alfarawati, S; Spaeth, K; Raberi, A; Taylor, S; **Kubikova, N**; Wells, D., The clinical significance of segmental aneuploidy in human oocytes and preimplantation embryos, *Human Reproduction*, 31, 13-14, 2016.

Kubikova, N; Sarasa, J; Wells, D., Development of a universal method for the preimplantation diagnosis of beta-thalassemia and sickle-cell anaemia using a novel next-generation sequencing approach: a new paradigm for PGD, *Human Reproduction*, 31, 402-403, 2016.

Wells, D; Alfarawati, S; Taylor, S; **Kubikova, N**; Spath, K; Turner, K; Hickman, C; Fragouli, E., Evidence that differences between embryology laboratories can influence the rate of mitotic errors, leading to increased chromosomal mosaicism, with significant implications for IVF success rates, *Human Reproduction*, 31, 25-26, 2016.

Babariya, D; Fragouli, E; Alfarawati, S; Spath, K; Raberi, A; Taylor, S; **Kubikova**, N; Wells, D., Prevalence and clinical significance of segmental aneuploidy in human oocytes and preimplantation embryos, *Fertility And Sterility*, 106, 3, e18, 2016.

Bibliography

Adhikari, P., Poudel, M., 2020. CRISPR-Cas9 in agriculture: Approaches, applications, future perspectives, and associated challenges. Malaysian J. Halal Res. 3, 6–16. doi:10.2478/mjhr-2020-0002

Adikusuma, F., Piltz, S., Corbett, M.A., Turvey, M., McColl, S.R., Helbig, K.J., Beard, M.R., Hughes, J., Pomerantz, R.T., Thomas, P.Q., 2018. Large deletions induced by Cas9 cleavage. Nature. E5-E16. doi:10.1038/s41586-018-0380-z

Alanis-Lobato, G., Zohren, J., McCarthy, A., Fogarty, N.M.E., Kubikova, N., Hardman, E., Greco, M., Wells, D., Turner, J.M.A., Niakan, K.K., 2020. Frequent loss-of-heterozygosity in CRISPR-Cas9-edited early human embryos. bioRxiv 2020.06.05.135913. doi:10.1101/2020.06.05.135913

Allen, F., Crepaldi, L., Alsinet, C., Strong, A.J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., Bassett, A.R., Harding, H., Galanty, Y., Muñoz-Martínez, F., Metzakopian, E., Jackson, S.P., Parts, L., 2019. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat. Biotechnol. 64-72. doi:10.1038/nbt.4317

Ashley-Koch, A., Yang, Q., Olney, R.S., 2000. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am. J. Epidemiol. 151, 839–845. doi:10.1093/oxfordjournals.aje.a010288

Baart, E.B., Martini, E., van den Berg, I., Macklon, N.S., Galjaard, R.J.H., Fauser, B.C.J.M., Van Opstal, D., 2006. Preimplantation genetic screening reveals a high incidence of aneuploidy and mosaicism in embryos from young women undergoing IVF. Hum. Reprod. 223-233. doi:10.1093/humrep/dei291

Babariya, D., Fragouli, E., Alfarawati, S., Spath, K., Wells, D., 2017. The incidence and origin of segmental aneuploidy in human oocytes and preimplantation embryos. Hum. Reprod. 223-233. doi:10.1093/humrep/dex324

Backenroth, D., Zahdeh, F., Kling, Y., Peretz, A., Rosen, T., Kort, D., Zeligson, S., Dror, T., Kirshberg, S., Burak, E., Segel, R., Levy-Lahad, E., Zangen, D., Altarescu, G., Carmi, S., Zeevi, D.A., 2019. Haploseek: a 24-hour all-in-one method for preimplantation genetic diagnosis (PGD) of monogenic disease and aneuploidy. Genet. Med. 1390-1399. doi:10.1038/s41436-018-0351-7

Balakier, H., MacLusky, N.J., Casper, R.F., 1993. Characterization of the first cell cycle in human zygotes: Implications for cryopreservation. Fertil. Steril. 359-365. doi:10.1016/S0015-0282(16)55678-7

Barrangou, R., Doudna, J.A., 2016. Applications of CRISPR technologies in research and beyond. Nat. Biotechnol. 34, 933–941. doi:10.1038/nbt.3659

Bäumer, C., Fisch, E., Wedler, H., Reinecke, F., Korfhage, C., 2018. Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. Sci. Rep. 1-10. doi:10.1038/s41598-018-25895-7

Bazrgar, M., Gourabi, H., Yazdi, P.E., Vazirinasab, H., Fakhri, M., Hassani, F., Valojerdi, M.R., 2014. DNA repair signalling pathway genes are overexpressed in poor-quality pre-implantation human embryos with complex aneuploidy. Eur. J. Obstet. Gynecol. Reprod. Biol. 152-156. doi:10.1016/j.ejogrb.2014.01.010

Ben-Nagi, J., Wells, D., Doye, K., Loutradi, K., Exeter, H., Drew, E., Alfarawati, S., Naja, R., Serhal, P., 2017. Karyomapping: a single centre's experience from application of methodology to ongoing pregnancy and live-birth rates. Reprod. Biomed. Online 35, 264–271. doi:10.1016/j.rbmo.2017.06.004

Braude, P., Bolton, V., Moore, S., 1988. Human gene expression first occurs between the four-

and eight-cell stages of preimplantation development. Nature. 459-461. doi:10.1038/332459a0

Canny, M.D., Moatti, N., Wan, L.C.K., Fradet-Turcotte, A., Krasner, D., Mateos-Gomez, P.A., Zimmermann, M., Orthwein, A., Juang, Y.C., Zhang, W., Noordermeer, S.M., Seclen, E., Wilson, M.D., Vorobyov, A., Munro, M., Ernst, A., Ng, T.F., Cho, T., Cannon, P.M., Sidhu, S.S., Sicheri, F., Durocher, D., 2018. Inhibition of 53BP1 favors homology-dependent DNA repair and increases CRISPR-Cas9 genome-editing efficiency. Nat. Biotechnol. 95-102. doi:10.1038/nbt.4021

Cao, A., Galanello, R., 2010. Beta-thalassemia. Genet. Med. 12, 61-76. doi:10.1097/GIM.0b013e3181cd68ed

Capecchi, M.R., 2005. Gene targeting in mice: Functional analysis of the mammalian genome for the twenty-first century. Nat. Rev. Genet. 507-512. doi:10.1038/nrg1619

Capmany, G., Taylor, A., Braude, P.R., Bolton, V.N., 1996. The timing of pronuclear formation, DNA synthesis and cleavage in the human 1-cell embryo. Mol. Hum. Reprod. 299-306. doi:10.1093/molehr/2.5.299

Capper, R., Britt-Compton, B., Tankimanova, M., Rowson, J., Letsolo, B., Man, S., Haughton, M., Baird, D.M., 2007. The nature of telomere fusion and a definition of the critical telomere length in human cells. Genes Dev. 2495-2508. doi:10.1101/gad.439107

Chakrabarti, A.M., Henser-Brownhill, T., Monserrat, J., Poetsch, A.R., Luscombe, N.M., Scaffidi, P., 2019. Target-Specific Precision of CRISPR-Mediated Genome Editing. Mol. Cell. 242-243. doi:10.1016/j.molcel.2018.11.031

Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. Mol. Biol. Evol. 24, 621–631. doi:10.1093/molbev/msl191

Chen, C.-K., Yu, H.-T., Soong, Y.-K., Lee, C.-L., 2014. New perspectives on preimplantation genetic diagnosis and preimplantation genetic screening. Taiwan. J. Obstet. Gynecol. 53,

146–150. doi:10.1016/j.tjog.2014.04.004

Chen, H.F., Chang, S.P., Wu, S.H., Lin, W.H., Lee, Y.C., Ni, Y.H., Chen, C.A., Ma, G.C., Ginsberg, N. a., You, E.M., Tsai, F.P., Chen, M., 2014. Validating a rapid, real-time, PCR-based direct mutation detection assay for preimplantation genetic diagnosis. Gene 548, 299–305. doi:10.1016/j.gene.2014.07.039

Chen, S.-C., Xu, X.-L., Zhang, J.-Y., Ding, G.-L., Jin, L., Liu, B., Sun, D.-M., Mei, C.-L., Yang, X.-N., Huang, H.-F., Xu, C.-M., 2016. Identification of PKD2 mutations in human preimplantation embryos in vitro using a combination of targeted next-generation sequencing and targeted haplotyping. Sci. Rep. 6, 25488. doi:10.1038/srep25488

Chiang, T., Schultz, R.M., Lampson, M.A., 2012. Meiotic Origins of Maternal Age-Related Aneuploidy1. Biol. Reprod. doi:10.1095/biolreprod.111.094367

Cho, S.W., Kim, S., Kim, J.M., Kim, J.S., 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat. Biotechnol. 230-232. doi:10.1038/nbt.2507

Costa-Borges, N., Spath, K., Miguel-Escalada, I., Mestres, E., Balmaseda, R., Serafín, A., Garcia-Jimenéz, M., Vanrell, I., González, J., Rink, K., Wells, D., Calderón, G., 2020. Maternal spindle transfer overcomes embryo developmental arrest caused by ooplasmic defects in mice. Elife. 594-606. doi:10.7554/eLife.48591

Cullot, G., Boutin, J., Toutain, J., Prat, F., Pennamen, P., Rooryck, C., Teichmann, M., Rousseau, E., Lamrissi-Garcia, I., Guyonnet-Duperat, V., Bibeyran, A., Lalanne, M., Prouzet-Mauléon, V., Turcq, B., Ged, C., Blouin, J.M., Richard, E., Dabernat, S., Moreau-Gaudry, F., Bedel, A., 2019. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. Nat. Commun. 1-14. doi:10.1038/s41467-019-09006-2

Daigneault, B.W., Rajput, S., Smith, G.W., Ross, P.J., 2018. Embryonic POU5F1 is Required for Expanded Bovine Blastocyst Formation. Sci. Rep. doi:10.1038/s41598-018-25964-x

De Rycke, M., Goossens, V., Kokkali, G., Meijer-Hoogeveen, M., Coonen, E., Moutou, C., 2017. ESHRE PGD Consortium data collection XIV-XV: Cycles from January 2011 to December 2012 with pregnancy follow-up to October 2013. Hum. Reprod. 32, 1974–1994. doi:10.1093/humrep/dex265

Deleye, L., De Coninck, D., Christodoulou, C., Sante, T., Dheedene, A., Heindryckx, B., Van Den Abbeel, E., De Sutter, P., Menten, B., Deforce, D., Van Nieuwerburgh, F., 2015. Whole genome amplification with SurePlex results in better copy number alteration detection using sequencing data compared to the MALBAC method. Sci. Rep. 5, 1–13. doi:10.1038/srep11711

Dever, D.P., Bak, R.O., Reinisch, A., Camarena, J., Washington, G., Nicolas, C.E., Pavel-Dinu, M., Saxena, N., Wilkens, A.B., Mantri, S., Uchida, N., Hendel, A., Narla, A., Majeti, R., Weinberg, K.I., Porteus, M.H., 2016. CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. Nature 539, 384–389. doi:10.1038/nature20134

Egli, D., Zuccaro, M., Kosicki, M., Church, G., Bradley, A., Jasin, M., 2017. Inter-homologue repair in fertilized human eggs? e181255. doi:10.1101/181255

Elinati, E., Zielinska, A.P., McCarthy, A., Kubikova, N., Maciulyte, V., Mahadevaiah, S., Sangrithi, M.N., Ojarikre, O., Wells, D., Niakan, K.K., Schuh, M., Turner, J.M.A., 2020. The BCL-2 pathway preserves mammalian genome integrity by eliminating recombination-defective oocytes. Nat. Commun. 1-10. doi:10.1038/s41467-020-16441-z

Fan, H.C., Gu, W., Wang, J., Blumenfeld, Y.J., El-Sayed, Y.Y., Quake, S.R., 2012. Erratum: Non-invasive prenatal measurement of the fetal genome. Nature 489, 326–326. doi:10.1038/nature11423

Fiorentino, F., Biricik, A., Bono, S., Spizzichino, L., Cotroneo, E., Cottone, G., Kokocinski, F., Michel, C.E., 2014a. Development and validation of a next-generation sequencing-based protocol for 24-chromosome aneuploidy screening of embryos. Fertil. Steril. 101,

1375-1382.e2. doi:10.1016/j.fertnstert.2014.01.051

Fiorentino, F., Bono, S., Biricik, A., Nuccitelli, A., Cotroneo, E., Cottone, G., Kokocinski, F., Michel, C.-E., Minasi, M.G., Greco, E., 2014b. Application of next-generation sequencing technology for comprehensive aneuploidy screening of blastocysts in clinical preimplantation genetic screening cycles. Hum. Reprod. 29, 2802–2813. doi:10.1093/humrep/deu277

Fiorentino, F., Magli, M.C., Podini, D., Ferraretti, a P., Nuccitelli, a, Vitale, N., Baldi, M., Gianaroli, L., 2003. The minisequencing method: an alternative strategy for preimplantation genetic diagnosis of single gene disorders. Mol. Hum. Reprod. 9, 399–410. doi:10.1093/molehr/gag046

Fogarty, N.M. E., McCarthy, A., Snijders, K.E., Powell, B.E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaitha, S.E., Kim, D., Maciulyte, V., Kleinjung, J., Kim, J.-S., Wells, D., Vallier, L., Bertero, A., Turner, J.M.A., Niakan, K.K., 2017. Genome editing reveals a role for OCT4 in human embryogenesis. Nature. 67-73. doi:10.1038/nature24033

Forman, E.J., Hong, K.H., Ferry, K.M., Tao, X., Taylor, D., Levy, B., Treff, N.R., Scott, R.T., 2013. In vitro fertilization with single euploid blastocyst transfer: A randomized controlled trial. Fertil. Steril. 100, 100-107. doi:10.1016/j.fertnstert.2013.02.056

Gasiunas, G., Barrangou, R., Horvath, P., Siksnys, V., 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. U. S. A. 2579-2586. doi:10.1073/pnas.1208507109

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., Liu, D.R., 2017. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature. 464-471. doi:10.1038/nature24644

Giménez, C., Sarasa, J., Arjona, C., Vilamajó, E., Martínez-Pasarell, O., Wheeler, K., Valls, G., Garcia-Guixé, E., Wells, D., 2015. Karyomapping allows preimplantation genetic

diagnosis of a de-novo deletion undetectable using conventional PGD technology. Reprod. Biomed. Online 31, 770–775. doi:10.1016/j.rbmo.2015.08.017

Gu, B., Posfai, E., Rossant, J., 2018. Efficient generation of targeted large insertions by microinjection into two-cell-stage mouse embryos. Nat. Biotechnol. 632-637. doi:10.1038/nbt.4166

Gutiérrez-Mateo, C., Sánchez-García, J.F., Fischer, J., Tormasi, S., Cohen, J., Munné, S., Wells, D., 2009. Preimplantation genetic diagnosis of single-gene disorders: experience with more than 200 cycles conducted by a reference laboratory in the United States. Fertil. Steril. 1544-1556. doi:10.1016/j.fertnstert.2008.08.111

Haapaniemi, E., Botla, S., Persson, J., Schmierer, B., Taipale, J., 2018. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. Nat. Med. 927-930. doi:10.1038/s41591-018-0049-z

Handyside, A.H., 2010. Preimplantation genetic diagnosis after 20 years. Reprod. Biomed. Online 21, 280–282. doi:10.1016/j.rbmo.2010.07.007

Handyside, A.H., Harton, G.L., Mariani, B., Thornhill, A.R., Affara, N., Shaw, M.-A., Griffin, D.K., 2010. Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. J. Med. Genet. 47, 651–658. doi:10.1136/jmg.2009.069971

Handyside, A.H., Kontogianni, E.H., Hardy, K., Winston, R.M., 1990. Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. Nature 344, 768–770. doi:10.1097/00006254-199107000-00024

Handyside, A.H., Lesko, J.G., Tarín, J.J., Winston, R.M., Hughes, M.R., 1992. Birth of a normal girl after in vitro fertilization and preimplantation diagnostic testing for cystic fibrosis., The New England journal of medicine. doi:10.1136/jmg.29.12.927-b

Hardy, K., Handyside, A.H., Winston, R.M.L., 1989. The human blastocyst: Cell number,

death and allocation during late preimplantation development in vitro. Development. 597-604.

Harper, J.C., Coonen, E., Handyside, A.H., Winston, R.M.L., Hopman, A.H.N., Delhanty, J.D.A., 1995. Mosaicism of autosomes and sex chromosomes in morphologically normal, monospermic preimplantation human embryos. Prenat. Diagn. 41-49. doi:10.1002/pd.1970150109

Harper, J.C., Wells, D., Piyamongkol, W., Abou-Sleiman, P., Apessos, A., Ioulianos, A., Davis, M., Doshi, A., Serhal, P., Ranieri, M., Rodeck, C., Delhanty, J.D. a, 2002. Preimplantation genetic diagnosis for single gene disorders: Experience with five single gene disorders. Prenat. Diagn. 22, 525–533. doi:10.1002/pd.394

Harton, G.L., De Rycke, M., Fiorentino, F., Moutou, C., Sengupta, S., Traeger-Synodinos, J., Harper, J.C., 2011. ESHRE PGD consortium best practice guidelines for amplification-based PGD. Hum. Reprod. 26, 33–40. doi:10.1093/humrep/deq231

Hattori, M., Yoshioka, K., Sakaki, Y., 1992. High-sensitive fluorescent DNA sequencing and its application for detection and mass-screening of point mutations. Electrophoresis 13, 560–565.

Horlbeck, M.A., Witkowsky, L.B., Guglielmi, B., Replogle, J.M., Gilbert, L.A., Villalta, J.E., Torigoe, S.E., Tjian, R., Weissman, J.S., 2016. Nucleosomes impede cas9 access to DNA in vivo and in vitro. Elife. e12677. doi:10.7554/eLife.12677

Human genome editing: Science, ethics, and governance, 2017. , Human Genome Editing: Science, Ethics, and Governance. doi:10.17226/24623

Hye, J.K., Lee, H.J., Kim, H., Cho, S.W., Kim, J.S., 2009. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome Res. 1279-1288. doi:10.1101/gr.089417.108

Ihry, R.J., Worringer, K.A., Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S.,

Chen, J., Sondey, M., Ye, C., Randhawa, R., Kulkarni, T., Yang, Z., McAllister, G., Russ, C., Reece-Hoyes, J., Forrester, W., Hoffman, G.R., Dolmetsch, R., Kaykas, A., 2018. P53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat. Med. 939-946. doi:10.1038/s41591-018-0050-6

Isaac, R.S., Jiang, F., Doudna, J.A., Lim, W.A., Narlikar, G.J., Almeida, R., 2016. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. Elife. e13450. doi:10.7554/eLife.13450

Ji, X., Zhang, Z., Shi, J., He, B., 2019. Clinical application of NGS-based SNP haplotyping for the preimplantation genetic diagnosis of primary open angle glaucoma. Syst. Biol. Reprod. Med. 258-263. doi:10.1080/19396368.2019.1590479

Jiang, B., Tan, A.S.C., Chong, S.S., 2012. Molecular strategies for pre-implantation genetic diagnosis of single gene and chromosomal disorders. Best Pract. Res. Clin. Obstet. Gynaecol. 26, 551–559. doi:10.1016/j.bpobgyn.2012.06.007

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 816-821. doi:10.1126/science.1225829

Kang, X., He, W., Huang, Y., Yu, Q., Chen, Y., Gao, X., Sun, X., Fan, Y., 2016. Introducing precise genetic modifications into human 3PN embryos by CRISPR/Cas-mediated genome editing. J. Assist. Reprod. Genet. 581-588. doi:10.1007/s10815-016-0710-8

Kennedy, E.M., Cullen, B.R., 2015. Bacterial CRISPR/Cas DNA endonucleases: A revolutionary technology that could dramatically impact viral research and treatment. Virology. 128-135. doi:10.1016/j.virol.2015.02.024

Khosravi, S., Salehi, M., Ramezanzadeh, M., Mirzaei, H., Salehi, R., 2016. Novel Multiplex Fluorescent PCR-Based Method for HLA Typing and Preimplantational Genetic Diagnosis of β-Thalassemia. Arch. Med. Res. 47, 293–298.

doi:10.1016/j.arcmed.2016.07.006

Kiessling, A.A., Bletsa, R., Desmarais, B., Mara, C., Kallianidis, K., Loutradis, D., 2010. Genome-wide microarray evidence that 8-Cell human blastomeres over-express cell cycle drivers and under-express checkpoints. J. Assist. Reprod. Genet. 265-276. doi:10.1007/s10815-010-9407-6

Kiessling, A.A., Bletsa, R., Desmarais, B., Mara, C., Kallianidis, K., Loutradis, D., 2009. Evidence that human blastomere cleavage is under unique cell cycle control. J. Assist. Reprod. Genet. 187-195. doi:10.1007/s10815-009-9306-x

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.-I., Kim, J.-S., 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat. Methods 12, 237–243. doi:10.1038/nmeth.3284

Kim, J.S., 2016. Genome editing comes of age. Nat. Protoc. doi:10.1038/nprot.2016.104

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., Liu, D.R., 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 533, 420–424. doi:10.1038/nature17946

Kono, N., Arakawa, K., 2019. Nanopore sequencing: Review of potential applications in functional genomics. Dev. Growth Differ. 61, 316–326. doi:10.1111/dgd.12608

Konstantinidis, M., Prates, R., Goodall, N.-N., Fischer, J., Tecson, V., Lemma, T., Chu, B., Jordan, A., Armenti, E., Wells, D., Munné, S., 2015. Live births following Karyomapping of human blastocysts: experience from clinical application of the method. Reprod. Biomed. Online 31, 394–403. doi:10.1016/j.rbmo.2015.05.018

Kosicki, M., Tomberg, K., Bradley, A., 2018. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nat. Biotechnol. 765-771. doi:10.1038/nbt.4192

Kubikova, N., Babariya, D., Sarasa, J., Spath, K., Alfarawati, S., Wells, D., 2018. Clinical

application of a protocol based on universal next-generation sequencing for the diagnosis of beta- thalassaemia and sickle cell anaemia in preimplantation embryos. Reprod. Biomed. Online 37, 136–144. doi:10.1016/j.rbmo.2018.05.005

Kubikova, N., Wells, D., 2020. Chapter 15 - Future technologies for preimplantation genetic applications, in: García-Velasco, J.A., Seli, E.B.T.-H.R.G. (Eds.), . Academic Press, pp. 255–269. doi:https://doi.org/10.1016/B978-0-12-816561-4.00016-8

Kuliev, A., Rechitsky, S., Verlinsky, O., Ivakhnenko, V., Evsikov, S., Wolf, G., Angastiniotis, M., Georghiou, D., Kukharenko, V., Strom, C., Verlinsky, Y., 1998. Preimplantation diagnosis of thalassemias, in: Journal of Assisted Reproduction and Genetics. 219-225. doi:10.1023/A:1022571822585

Lea, A.R., Niakan, K., 2019. Human germline genome editing. Nat. Cell Biol. 1479-1489. doi:10.1038/s41556-019-0424-0

Li, G., Liu, Y., Zeng, Y., Li, J., Wang, L., Yang, G., Chen, D., Shang, X., Chen, J., Huang, X., Liu, J., 2017. Highly efficient and precise base editing in discarded human tripronuclear embryos. Protein Cell. 776-779. doi:10.1007/s13238-017-0458-7

Liang, P., Ding, C., Sun, H., Xie, X., Xu, Y., Zhang, X., Sun, Y., Xiong, Y., Ma, W., Liu, Y., Wang, Y., Fang, J., Liu, D., Songyang, Z., Zhou, C., Huang, J., 2017. Correction of β-thalassemia mutant by base editor in human embryos. Protein Cell 8, 811–822. doi:10.1007/s13238-017-0475-6

Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J., Xie, X., Chen, Y., Li, Y., Sun, Y., Bai, Y., Songyang, Z., Ma, W., Zhou, C., Huang, J., 2015. CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. Protein Cell. 363-372. doi:10.1007/s13238-015-0153-5

Lin, S., Staahl, B.T., Alla, R.K., Doudna, J.A., 2014. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife. 1-13.

doi:10.7554/eLife.04766

Lo, A.W.I., Sprung, C.N., Fouladi, B., Pedram, M., Sabatier, L., Ricoul, M., Reynolds, G.E., Murnane, J.P., 2002. Chromosome Instability as a Result of Double-Strand Breaks near Telomeres in Mouse Embryonic Stem Cells. Mol. Cell. Biol. 4836-4950. doi:10.1128/mcb.22.13.4836-4850.2002

Long, C., McAnally, J.R., Shelton, J.M., Mireault, A.A., Bassel-Duby, R., Olson, E.N., 2014. Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. Science (80-.). 345, 1184–1188. doi:10.1126/science.1254445

Lovell-Badge, R., 2019. CRISPR babies: a view from the centre of the storm. Development. dev175778. doi:10.1242/dev.175778

Ma, H., Marti-Gutierrez, N., Park, S.-W., Wu, J., Lee, Y., Suzuki, K., Koski, A., Ji, D., Hayama, T., Ahmed, R., Darby, H., Van Dyken, C., Li, Y., Kang, E., Park, A.-R., Kim, D., Kim, S.-T., Gong, J., Gu, Y., Xu, X., Battaglia, D., Krieg, S.A., Lee, D.M., Wu, D.H., Wolf, D.P., Heitner, S.B., Belmonte, J.C.I., Amato, P., Kim, J.-S., Kaul, S., Mitalipov, S., 2017. Correction of a pathogenic gene mutation in human embryos. Nature 548, 413–419. doi:10.1038/nature23305

Maddalo, D., Manchado, E., Concepcion, C.P., Bonetti, C., Vidigal, J.A., Han, Y.-C., Ogrodowski, P., Crippa, A., Rekhtman, N., de Stanchina, E., Lowe, S.W., Ventura, A., 2014. In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. Nature 516, 423–427. doi:10.1038/nature13902

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., Church, G.M., 2013. RNA-guided human genome engineering via Cas9. Science (80-.). 823-826. doi:10.1126/science.1232033

Martín, J., Cervero, A., Mir, P., Conejero Martinez, J.A., Pellicer, A., Simón, C., 2013. The impact of next-generation sequencing technology on preimplantation genetic diagnosis

and screening. Fertil. Steril. 99. 1054-1061. doi:10.1016/j.fertnstert.2013.02.001

Maurer, M., Ebner, T., Puchner, M., Mayer, R.B., Shebl, O., Oppelt, P., Duba, H.C., 2015. Chromosomal Aneuploidies and early embryonic developmental arrest. Int. J. Fertil. Steril. 346-353. doi:10.22074/ijfs.2015.4550

Moleirinho, A., Seixas, S., Lopes, A.M., Bento, C., Prata, M.J., Amorim, A., 2013. Evolutionary constraints in the β-globin cluster: The signature of purifying selection at the δ-globin (HBD) locus and its role in developmental gene regulation. Genome Biol. Evol. 5, 559–571. doi:10.1093/gbe/evt029

Moutou, C., Goossens, V., Coonen, E., De Rycke, M., Kokkali, G., Renwick, P., Sengupta, S.B., Vesela, K., Traeger-Synodinos, J., 2014. ESHRE PGD Consortium data collection XII: Cycles from January to December 2009 with pregnancy follow-up to October 2010. Hum. Reprod. 29, 880–903. doi:10.1093/humrep/deu012

Munné, S., Grifo, J., Wells, D., 2016. Mosaicism: "survival of the fittest" versus "no embryo left behind." Fertil. Steril. 105, 1146–1149. doi:10.1016/j.fertnstert.2016.01.016

Munné, S., Wells, D., 2017. Detection of mosaicism at blastocyst stage with the use of high-resolution next-generation sequencing. Fertil. Steril. 1085-1091. doi:10.1016/j.fertnstert.2017.03.024

Natesan, S. a., Handyside, A.H., Thornhill, A.R., Ottolini, C.S., Sage, K., Summers, M.C., Konstantinidis, M., Wells, D., Griffin, D.K., 2014. Live birth after PGD with confirmation by a comprehensive approach (karyomapping) for simultaneous detection of monogenic and chromosomal disorders. Reprod. Biomed. Online 29, 600–605. doi:10.1016/j.rbmo.2014.07.007

Natesan, S. a, Bladon, A.J., Coskun, S., Qubbaj, W., Prates, R., Munne, S., Coonen, E., Dreesen, J.C.F.M., Stevens, S.J.C., Paulussen, A.D.C., Stock-Myer, S.E., Wilton, L.J., Jaroudi, S., Wells, D., Brown, A.P.C., Handyside, A.H., 2014. Genome-wide

karyomapping accurately identifies the inheritance of single-gene defects in human preimplantation embryos in vitro. Genet. Med. 16, 1–8. doi:10.1038/gim.2014.45

Orthwein, A., Fradet-Turcotte, A., Noordermeer, S.M., Canny, M.D., Brun, C.M., Strecker, J., Escribano-Diaz, C., Durocher, D., 2014. Mitosis inhibits DNA double-strand break repair to guard against telomere fusions. Science (80-.). 189-193. doi:10.1126/science.1248024

Ottolenghi, S., Lanyon, W.G., Williamson, R., Weatherall, D.J., Clegg, J.B., Pitcher, C.S., 1975. Human globin gene analysis for a patient with β 0/δβ 0 thalassemia. Proc Natl Acad Sci U S A 72, 2294–2299. doi:10.1073/pnas.72.6.2294

Owens, D.D.G., Caulder, A., Frontera, V., Harman, J.R., Allan, A.J., Bucakci, A., Greder, L., Codner, G.F., Hublitz, P., McHugh, P.J., Teboul, L., de Bruijn, M.F.T.R., 2019. Microhomologies are prevalent at Cas9-induced larger deletions. Nucleic Acids Res. 7402– 7417. doi:10.1093/nar/gkz459

Park, J., Lim, K., Kim, J.S., Bae, S., 2017. Cas-analyzer: An online tool for assessing genome editing results using NGS data. Bioinformatics. 286-288. doi:10.1093/bioinformatics/btw561

Piel, F.B., Patil, A.P., Howes, R.E., Nyangiri, O. a, Gething, P.W., Williams, T.N., Weatherall, D.J., Hay, S.I., 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat. Commun. 1, 104. 107. doi:10.1038/ncomms1104

Piyamongkol, W., Bermúdez, M.G., Harper, J.C., Wells, D., 2003. Detailed investigation of factors influencing amplification efficiency and allele drop-out in single cell PCR: implications for preimplantation genetic diagnosis. Mol. Hum. Reprod. 9, 411–420. doi:10.1093/molehr/gag051

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 841-842. doi:10.1093/bioinformatics/btq033

Ray, P.F., Handyside, a H., 1996. Increasing the denaturation temperature during the first cycles of amplification reduces allele dropout from single cells for preimplantation genetic diagnosis. Mol. Hum. Reprod. 2, 213–218.

Rechitsky, S., Strom, C., Verlinsky, O., Amet, T., Ivakhnenko, V., Kukharenko, V., Kuliev, a., Verlinsky, Y., 1998. Allele dropout in polar bodies and blastomeres. J. Assist. Reprod. Genet. 15, 253–257. doi:10.1023/A:1022532108472

Rees, D.C., Williams, T.N., Gladwin, M.T., 2010. Sickle-cell disease. Lancet 376, 2018–2031. doi:10.1016/S0140-6736(10)61029-X

Ren, Y., Zhi, X., Zhu, X., Huang, J., Lian, Y., Li, R., Jin, H., Zhang, Y., Zhang, W., Nie, Y., Wei, Y., Liu, Z., Song, D., Liu, P., Qiao, J., Yan, L., 2016. Clinical applications of MARSALA for preimplantation genetic diagnosis of spinal muscular atrophy. J. Genet. Genomics 43, 541–547. doi:10.1016/j.jgg.2016.03.011

Renwick, P.J., Trussler, J., Ostad-Saffari, E., Fassihi, H., Black, C., Braude, P., Ogilvie, C.M., Abbs, S., 2006. Proof of principle and first cases using preimplantation genetic haplotyping – a paradigm shift for embryo diagnosis. Reprod. Biomed. Online 13, 110–119. doi:10.1016/S1472-6483(10)62024-X

Scott, R.T., Upham, K.M., Forman, E.J., Hong, K.H., Scott, K.L., Taylor, D., Tao, X., Treff, N.R., 2013a. Blastocyst biopsy with comprehensive chromosome screening and fresh embryo transfer significantly increases in vitro fertilization implantation and delivery rates: A randomized controlled trial. Fertil. Steril. 100, 697–703. doi:10.1016/j.fertnstert.2013.04.035

Scott, R.T., Upham, K.M., Forman, E.J., Zhao, T., Treff, N.R., 2013b. Cleavage-stage biopsy significantly impairs human embryonic implantation potential while blastocyst biopsy does not: A randomized and paired clinical trial. Fertil. Steril. 624-630. doi:10.1016/j.fertnstert.2013.04.039

Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., Sherwood, R.I., 2018. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature. 646-651. doi:10.1038/s41586-018-0686-x

Sherlock, J., Cirigliano, V., Petrou, M., Tutschek, B., Adinolfi, M., 1998. Assessment of diagnostic quantitative fluorescent multiplex polymerase chain reaction assays performed on single cells. Ann. Hum. Genet. 62, 9–23. doi:10.1046/j.1469-1809.1998.6210009.x

Sint, D., Raso, L., Traugott, M., 2012. Advances in multiplex PCR: Balancing primer efficiencies and improving detection success. Methods Ecol. Evol. 3, 898–905. doi:10.1111/j.2041-210X.2012.00215.x

Song, B., Fan, Y., He, W., Zhu, D., Niu, X., Wang, D., Ou, Z., Luo, M., Sun, X., 2015. Improved Hematopoietic Differentiation Efficiency of Gene-Corrected Beta-Thalassemia Induced Pluripotent Stem Cells by CRISPR/Cas9 System. Stem Cells Dev. 24, 1053–1065. doi:10.1089/scd.2014.0347

Spath, K., Wellsreferences, D., 2015. Deep impact: sequencing embryo biopsy specimens at increasing depth. Reprod. Biomed. Online 31, 1–3. doi:10.1016/j.rbmo.2015.05.008

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., Sermon, K., 2006. Whole-genome multiple displacement amplification from single cells. Nat. Protoc. 1, 1965–1970. doi:10.1038/nprot.2006.326

Spits, S., 2009. PGD for monogenic disorders: aspects of molecular biology. Prenat. Diagn. 26, 980–984. doi:10.1002/pd

Strom, C.M., Rechitsky, S., Wolf, G., Cieslak, J., Kuliev, a., Verlinsky, Y., 1998. Preimplantation diagnosis of autosomal dominant retinitis pigmentosum using two simultaneous single cell assays for a point mutation in the rhodopsin gene. Mol. Hum. Reprod. 4, 351–355. doi:10.1093/molehr/4.4.351

Swain, J.E., Carrell, D., Cobo, A., Meseguer, M., Rubio, C., Smith, G.D., 2016. Optimizing

the culture environment and embryo manipulation to help maintain embryo developmental potential. Fertil. Steril. 571-587. doi:10.1016/j.fertnstert.2016.01.035

Taber, B.J., 2013. Journal of Career Assessment 21(2). 200-209. doi:10.1177/1069072712466722

Takayama, K., Igai, K., Hagihara, Y., Hashimoto, R., Hanawa, M., Sakuma, T., Tachibana, M., Sakurai, F., Yamamoto, T., Mizuguchi, H., 2017. Highly efficient biallelic genome editing of human ES/iPS cells using a CRISPR/Cas9 or TALEN system. Nucleic Acids Res. 5198–5207. doi:10.1093/nar/gkx130

Tang, L., Zeng, Y., Du, H., Gong, M., Peng, J., Zhang, B., Lei, M., Zhao, F., Wang, W., Li, X., Liu, J., 2017. CRISPR/Cas9-mediated gene editing in human zygotes using Cas9 protein. Mol. Genet. Genomics. 525-533. doi:10.1007/s00438-017-1299-z

Taylor, T.H., Gitlin, S.A., Patrick, J.L., Crain, J.L., Wilson, J.M., Griffin, D.K., 2014. The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. Hum. Reprod. Update. 571-581. doi:10.1093/humupd/dmu016

Telenti, A., Pierce, L.C.T., Biggs, W.H., Di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., Brewerton, S.C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B.A., Och, F.J., Turpaz, Y., Venter, J.C., 2016. Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U. S. A. 11901- 11906. doi:10.1073/pnas.1613365113

Treff, N.R., Fedick, A., Tao, X., Devkota, B., Taylor, D., Scott, R.T., 2013a. Evaluation of targeted next-generation sequencing–based preimplantation genetic diagnosis of monogenic disease. Fertil. Steril. 99, 1377-1384.e6. doi:10.1016/j.fertnstert.2012.12.018

Treff, N.R., Fedick, A., Tao, X., Devkota, B., Taylor, D., Scott, R.T., 2013b. Evaluation of targeted next-generation sequencing-based preimplantation genetic diagnosis of

monogenic disease. Fertil. Steril. 99, 1377-1384.e6. doi:10.1016/j.fertnstert.2012.12.018

Treff, N.R., Forman, E.J., Scott, R.T., 2013c. Next-generation sequencing for preimplantation genetic diagnosis. Fertil. Steril. 99, e17–e18. doi:10.1016/j.fertnstert.2013.02.034

Treff, N.R., Zimmerman, R., Bechor, E., Hsu, J., Rana, B., Jensen, J., Li, J., Samoilenko, A., Mowrey, W., Van Alstine, J., Leondires, M., Miller, K., Paganetti, E., Lello, L., Avery, S., Hsu, S., Melchior Tellier, L.C.A., 2019. Validation of concurrent preimplantation genetic testing for polygenic and monogenic disorders, structural rearrangements, and whole and segmental chromosome aneuploidy with a single universal platform. Eur. J. Med. Genet. 103647- 103652. doi:10.1016/j.ejmg.2019.04.004

Van der Aa, N., Esteki, M.Z., Vermeesch, J.R., Voet, T., 2013. Preimplantation genetic diagnosis guided by single-cell genomics. Genome Med. 5, 71. doi:10.1186/gm475

Van Steensel, B., Smogorzewska, A., De Lange, T., 1998. TRF2 protects human telomeres from end-to-end fusions. Cell. 401-413. doi:10.1016/S0092-8674(00)80932-0

Vanneste, E., Voet, T., Le Caignec, C., Ampe, M., Konings, P., Melotte, C., Debrock, S., Amyere, M., Vikkula, M., Schuit, F., Fryns, J.P., Verbeke, G., D'Hooghe, T., Moreau, Y., Vermeesch, J.R., 2009. Chromosome instability is common in human cleavage-stage embryos. Nat. Med. 577-583. doi:10.1038/nm.1924

Verlinsky, Y., Rechitsky, S., Schoolcraft, W., Strom, C., Kuliev, A., 2001. Preimplantation diagnosis for fanconi anemia combined with hla matching. J. Am. Med. Assoc. 3130-3133. doi:10.1001/jama.285.24.3130

Vilarino, M., Suchy, F.P., Rashid, S.T., Lindsay, H., Reyes, J., McNabb, B.R., van der Meulen, T., Huising, M.O., Nakauchi, H., Ross, P.J., 2018. Mosaicism diminishes the value of pre-implantation embryo biopsies for detecting CRISPR/Cas9 induced mutations in sheep. Transgenic Res. 525-537. doi:10.1007/s11248-018-0094-x

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., Qiu, J.-L., 2014. Simultaneous

editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat. Biotechnol. 32, 947–951. doi:10.1038/nbt.2969

Weatherall, D.J., 2001. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat. Rev. Genet. 2, 245–255. doi:10.1038/35066048

Wells, D., Alfarawati, S., Fragouli, E., 2008. Use of comprehensive chromosomal screening for embryo assessment: Microarrays and CGH. Mol. Hum. Reprod. 14, 703–710. doi:10.1093/molehr/gan062

Wells, D., Bermúdez, M.G., Steuerwald, N., Malter, H.E., Thornhill, A.R., Cohen, J., 2005. Association of abnormal morphology and altered gene expression in human preimplantation embryos. Fertil. Steril. 343-355. doi:10.1016/j.fertnstert.2005.01.143

Wells, D., Kaur, K., Grifo, J., Glassner, M., Taylor, J.C., Fragouli, E., Munne, S., 2014. Clinical utilisation of a rapid low-pass whole genome sequencing technique for the diagnosis of aneuploidy in human embryos prior to implantation. J. Med. Genet. 51, 553–562. doi:10.1136/jmedgenet-2014-102497

Wells, D., Sherlock, J.K., 1998. Strategies for preimplantation genetic diagnosis of single gene disorders by DNA amplification. Prenat. Diagn. 18, 1389–1401. doi:10.1002/(SICI)1097-0223(199812)18:13<1389::AID-PD498>3.0.CO;2-6

Wells, D., Sherlock, J.K., Handyside, A.H., Delhanty, J.D., 1999. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. Nucleic Acids Res. 27, 1214–1218.

Wells, D., Vermeesch, J.R., Simpson, J.L., 2019. Current Controversies in Prenatal Diagnosis 3: Gene editing should replace embryo selection following PGD. Prenat. Diagn. doi:10.1002/pd.5442

Wu, H., Shen, X., Huang, L., Zeng, Y., Gao, Y., Shao, L., Lu, B., Zhong, Y., Miao, B., Xu, Y., Wang, Y., Li, Y., Xiong, L., Lu, S., Xie, X.S., Zhou, C., 2018. Genotyping single-sperm

cells by universal MARSALA enables the acquisition of linkage information for combined pre-implantation genetic diagnosis and genome screening. J. Assist. Reprod. Genet. 1071-1078. doi:10.1007/s10815-018-1158-9

Yan, L., Huang, L., Xu, L., Huang, J., Ma, F., Zhu, X., Tang, Y., Liu, M., Lian, Y., Liu, P., Li, R., Lu, S., Tang, F., Qiao, J., Xie, X.S., 2015. Live births after simultaneous avoidance of monogenic diseases and chromosome abnormality by next-generation sequencing with linkage analyses. Proc. Natl. Acad. Sci. 112, 15964–15969. doi:10.1073/pnas.1523297113

Yang, Z., Liu, J., Collins, G.S., Salem, S. a, Liu, X., Lyle, S.S., Peck, A.C., Sills, E., Salem, R.D., 2012. Selection of single blastocysts for fresh transfer via standard morphology assessment alone and with array CGH for good prognosis IVF patients: results from a randomized pilot study. Mol. Cytogenet. 5, 24. 1-8. doi:10.1186/1755-8166-5-24

Yin, H., Xue, W., Chen, S., Bogorad, R.L., Benedetti, E., Grompe, M., Koteliansky, V., Sharp, P.A., Jacks, T., Anderson, D.G., 2014. Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. Nat. Biotechnol. 32, 551–553. doi:10.1038/nbt.2884

Zamani Esteki, M., Dimitriadou, E., Mateiu, L., Melotte, C., Van der Aa, N., Kumar, P., Das, R., Theunis, K., Cheng, J., Legius, E., Moreau, Y., Debrock, S., D'Hooghe, T., Verdyck, P., De Rycke, M., Sermon, K., Vermeesch, J.R., Voet, T., 2015. Concurrent Whole-Genome Haplotyping and Copy-Number Profiling of Single Cells. Am. J. Hum. Genet. 894-912. doi:10.1016/j.ajhg.2015.04.011

Zhang, J., Liu, H., Luo, S., Lu, Z., Chávez-Badiola, A., Liu, Z., Yang, M., Merhi, Z., Silber, S.J., Munné, S., Konstandinidis, M., Wells, D., Huang, T., 2017. Live birth derived from oocyte spindle transfer to prevent mitochondrial disease. Reprod. Biomed. Online. 361-368. doi:10.1016/j.rbmo.2017.01.013

Zheng, H., Jin, H., Liu, L., Liu, J., Wang, W.-H., 2015. Application of next-generation

sequencing for 24-chromosome aneuploidy screening of human preimplantation embryos. Mol. Cytogenet. 8, 38. 1-9. doi:10.1186/s13039-015-0143-6

Zheng, Y., Wang, N., Li, L., Jin, F., 2011. Whole genome amplification in preimplantation genetic diagnosis. J. Zhejiang Univ. Sci. B 12, 1–11. doi:10.1631/jzus.B1000196

Zhou, C., Zhang, M., Wei, Y., Sun, Yidi, Sun, Yun, Pan, H., Yao, N., Zhong, W., Li, Y., Li, W., Yang, H., Chen, Z. jiang, 2017. Highly efficient base editing in human tripronuclear zygotes. Protein Cell. 772-775. doi:10.1007/s13238-017-0459-6

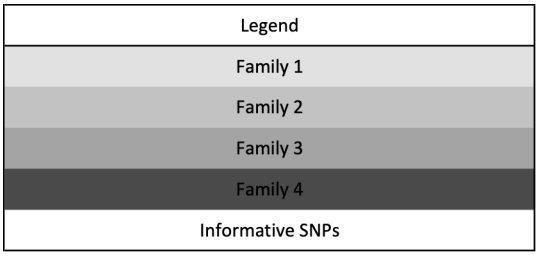
Appendix 1

Distribution of genotypes present in four tested families at positions of SNPs, which were informative for at least one family. Read depth = sequence coverage.

	Additional SNPs identified within the HBB gene					Amplicon 9	Amplicon 11
SNP ID	rs1609812	rs7480526	rs713040	rs63750628	rs12574989	rs3813727	rs4910543
Position	5247141	5247733	5248243	5248282	5246514	5255912	5258827
Sample	HBB locus	HBB locus	HBB locus	HBB locus	HBB locus		
F1 Male	A/A	A/C	G/G	G/G	G/G	G/G	C/C
read depth	274	106	127	124	57	138	208
F1 Female	A/A	A/C	G/G	G/G	G/G	G/G	C/C
read depth	197	51	84	83	46	99	158
F1 Son	A/A	A/C	G/G	G/G	G/G	G/G	C/C
read depth	232	65	61	60	32	111	154
F2 Male	A/G	A/A	A/G	G/G	C/T	G/G	C/C
read depth	85	24	35	35	18	50	54
F2 Female	A/A	A/A	G/G	G/G	C/C	A/G	C/G
read depth	160	54	95	90	49	111	141
F2 Son	A/G	A/A	A/G	G/G	C/T	G/G	C/C
read depth	132	38	89	87	46	115	133
F3 Male	A/A	A/C	G/G	G/G	G/G	G/G	C/C
read depth	181	55	90	88	31	99	127
F3 Female	A/A	C/C	G/G	G/G	G/G	A/G	C/C
read depth	128	44	56	52	35	72	101
F3 Daughter	A/A	A/C	G/G	G.G	G/G	G/G	C/C
read depth	251	76	62	62	32	106	128
F4 Male	A/G	A/C	A/G	A/G	G/T	A/G	C/G
read depth	102	22	45	44	29	62	28
F4 Female	A/G	A/C	A/G	G/G	G/T	A/G	G/G
read depth	115	52	70	68	37	106	117
F4 Daughter	A/G	A/C	A/G	A/G	G/T	A/G	C/G
read depth	110	33	78	76	22	79	122

Legend
Family 1
Family 2
Family 3
Family 4
Informative SNPs

	Amplicon 11	Amplicon 11	Amplicon 12	Amplicon 13	Amplicon 15	Amplicon 16	Amplicon 17
SNP ID	rs4910735	rs4910544	rs4910736	rs2105819	rs11036364	rs7936823	rs6578588
Position	5258852	5258856	5258989	5259727	5249004	5250168	5252251
Sample							
F1 Male	A/A	A/A	A/A	C/C	A/G	A/A	C/C
read depth	215	216	227	374	17	26	92
F1 Female	A/A	A/A	A/A	C/C	A/G	A/A	C/C
read depth	166	167	123	275	11	24	62
F1 Son	A/A	A/A	A/A	C/C	A/G	A/A	C/C
read depth	162	165	141	271	10	14	39
F2 Male	A/A	T/T	A/A	C/C	A/G	A/A	C/C
read depth	61	61	76	134	10	21	44
F2 Female	A/G	A/T	A/C	C/G	G/G	A/G	C/T
read depth	149	149	151	296	7	25	75
F2 Son	A/A	T/T	A/A	C/C	A/G	A/A	C/C
read depth	139	139	134	295	11	27	79
F3 Male	A/A	A/A	A/A	C/C	A/G	A/A	C/C
read depth	127	130	146	328	27	28	64
F3 Female	A/A	A/A	A/A	C/C	A/A	A/G	C/C
read depth	109	111	98	162	12	19	53
F3 Daughter	A/A	A/A	A/A	C/C	A/A	A/A	C/C
read depth	136	137	128	233	27	4	34
F4 Male	A/G	A/T	A/C	C/G	A/G	A/G	C/T
read depth	43	43	63	127	5	11	34
F4 Female	G/G	T/T	C/C	G/G	A/A	G/G	T/T
read depth	120	119	126	233	31	21	66
F4 Daughter	A/G	A/T	A/C	C/G	A/A	A/G	C/T
read depth	127	128	105	187	23	13	58



	Amplicon 18	Amplicon 19	Amplicon 20	Amplicon 21	Amplicon 21	Amplicon 21	Amplicon 24	Amplicon 24
SNP ID	rs7945118	rs34220818	rs10837620	rs12364872	rs10837626	rs10837628	rs10837631	rs7110263
Position	5236417	5236851	5243559	5244144	5244299	5244404	5246356	5246512
Sample								
F1 Male	G/G	T/T	A/A	A/G	A/T	A/G	A/T	G/G
read depth	36	300	282	28	38	25	266	57
F1 Female	G/G	T/T	A/A	A/G	A/A	A/G	A/T	G/G
read depth	29	219	198	11	22	12	205	46
F1 Son	G/G	T/T	A/A	A/G	A/A	A/G	A/T	G/G
read depth	27	233	239	10	17	11	181	32
F2 Male	C/G	C/T	A/G	A/A	A/T	A/G	A/T	G/T
read depth	15	108	98	11	18	12	88	31
F2 Female	G/G	T/T	A/A	A/A	A/A	G/G	A/A	G/G
read depth	20	255	209	24	37	20	205	50
F2 Son	C/G	C/T	A/G	A/A	A/T	A/G	A/T	G/T
read depth	30	214	222	21	25	12	202	46
F3 Male	G/G	C/T	A/G	A/G	A/A	A/G	A/T	G/G
read depth	30	247	252	22	31	17	169	31
F3 Female	G/G	T/T	G/G	G/G	A/A	A/A	T/T	G/G
read depth	25	156	151	13	19	13	148	35
F3 Daughter	G/G	T/T	A/G	A/G	A/A	A/G	A/T	G/G
read depth	20	169	202	10	23	14	185	32
F4 Male	A/G	A/A	G/G	A/G	A/T	A/A	T/T	G/G
read depth	11	71	98	10	15	10	106	29
F4 Female	A/G	A/A	G/G	A/G	A/T	A/A	T/T	G/G
read depth	28	188	227	16	24	10	106	37
F4 Daughter	A/G	A/A	G/G	A/G	A/T	A/A	T/T	G/G
read depth	32	165	194	16	27	17	189	22

Legend
Family 1
Family 2
Family 3
Family 4
Informative SNPs

Appendix 2

Genotypes at single nucleotide polymorphism and HBB mutation positions in the tested embryos in the three clinical cases.

Case 1	rs4910736		rs7945118		rs2105819		rs12364872		rs7936823		rs34220818		rs6578588		rs7110263		rs10837626		c.92+6T>C		c.118C>T	
	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %
Mother	A/A	100	C/G	47/53	C/C	100	A/G	43/57	A/A	100	C/T	48/52	C/C	100	T/G	71/29	T/A	44/56	G/A	52/48	G/G	100
Father	A/C	51/49	G/G	100	C/G	50/50	G/G	100	A/G	53/47	T/T	100	C/T	53/47	G/G	100	A/A	100	A/A	100	G/A	49/51
E1	A/A	100	C/G	91/9	C/C	100	A/G	93/7	A/A	100	C/T	90/10	C/C	100	T/G	90/10	T/A	96/4	G/A	97/3	G/G	97/3
E2	A/C	20/80	C/G	90/10	C/G	7/93	A/G	91/9	A/G	97/3	C/T	88/12	C/T	98/2	T/G	88/12	T/A	88/12	G/A	96/4	G/A	65/35
E3	A/C	84/16	C/G	99/1	C/G	85/15	A/G	70/30	A/G	54/46	C/T	2/98	C/T	42/58	T/G	47/53	T/A	71/29	G/A	61/39	G/A	69/31
E4	A/C	55/45	G/G	100	C/G	56/44	G/G	100	A/G	35/65	T/T	100	C/T	38/62	G/G	100	A/A	100	A/A	100	G/A	86/14
E5	A/C	66/34	C/G	3/97	C/G	64/36	A/G	2/98	A/G	84/16	C/T	4/96	C/T	82/18	T/G	1/99	T/A	3/97	G/A	2/98	G/A	4/96
E8	A/A	100	C/G	66/34	C/C	100	A/G	29/71	A/A	100	C/T	65/35	C/C	100	T/G	14/86	T/A	29/71	G/A	23/76	G/G	100
E10	A/A	100	C/G	51/49	C/C	100	A/G	20/80	A/A	100	C/T	49/51	C/C	100	T/G	9/91	T/A	17/83	G/A	84/16	G/G	100
E11	A/A	100	G/G	100	C/C	100	G/G	100	A/A	100	T/T	100	C/C	100	G/G	100	A/A	100	A/A	100	G/G	100
E12	A/A	100	G/G	100	C/C	100	G/G	100	A/A	100	T/T	100	C/C	100	G/G	100	A/A	100	A/A	100	G/G	100
E14	A/C	20/80	C/G	44/56	C/G	19/81	A/G	21/79	A/G	29/71	C/T	42/58	C/T	24/76	T/G	16/84	T/A	21/79	G/A	30/69	G/A	38/62

Case 2	rs4910543		rs4910735		rs4910544		rs4910736		rs2105819		rs11036364		rs6578588		rs10837620		rs12364872		rs10837628		rs10837631		rs7480526		rs63750628		c.93-21G>A		c.316-106C>G	
	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %
Mother	C/G	48/51	A/G	49/51	A/T	49/51	A/C	55/45	C/G	49/51	A/G	62/38	C/T	51/49	G/A	53/47	G/A	51/49	A/G	50/50	T/A	52/48	C/A	50/50	G/G	100	T/C	45/55	G/G	100
Father	C/C	100	A/A	100	A/A	100	A/A	100	C/C	100	A/A	100	C/C	100	G/G	100	A/A	100	A/A	100	T/T	100	A/A	100	G/A	50/50	C/C	100	G/C	75/24
E4	C/G	32/68	A/G	29/71	A/T	30/70	A/C	35/65	C/G	35/65	A/G	46/54	C/T	38/61	A/G	37/63	A/A	100	A/G	61/39	A/T	100	A/A	100	G/A	66/34	C/C	100	G/C	83/17
E7	C/C	100	A/A	100	A/A	100	A/A	100	C/C	100	A/A	100	C/C	100	G/G	100	G/A	54/46	A/A	100	T/T	100	C/A	51/49	G/G	100	T/C	48/51	G/G	100

Case 3	rs713040		rs1609812		rs10837631		rs7110263		rs12574989		rs10837626		N/A		N/A		N/A		rs7945118		c.27_28insG		c.92+5G>C	
	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %	Genotype	Read %
Mother	G/G	100	A/A	100	A/T	46/54	G/G	100	C/C	100	A/A	100	G/A	50/50	C/T	50/50	G/A	50/50	G/G	100	A+C	69/31	C/C	100
Father	G/A	51/49	A/G	49/51	A/A	100	G/T	37/63	C/T	37/63	A/T	54/47	G/G	100	C/C	100	A/A	100	G/C	51/49	A	100	C/G	51/49
TE1	G/G	100	A/A	100	A/A	100	G/G	100	C/C	100	A/A	100	G/G	100	C/C	100	G/A	25/75	G/G	100	A+C	80/20	C/C	100
TE2	G/A	23/77	A/G	25/75	A/A	100	G/T	18/82	C/T	18/82	A/T	30/70	G/G	100	C/C	100	G/A	35/65	G/C	34/66	A+C	82/18	C/G	24/76
TE4	G/A	65/35	A/G	66/34	A/A	100	G/T	57/43	C/T	57/43	A/T	73/27	G/G	100	C/C	100	G/A	69/31	G/C	70/30	A+C	61/39	C/G	65/35
TE5	G/A	83/17	A/G	83/17	T/A	24/76	G/T	76/24	C/T	76/24	A/T	89/11	G/A	15/85	C/T	15/85	A/A	100	G/C	85/15	A	100	C/G	84/16

Appendix 3

Complete script used to generate analysis and results of the Cas9 enrichment of *POU5F1* in human embryonic stem cells.

```
---
title: "Evaluation of read-mapping characteristics from a Cas-mediated PCR-free enrichment"
date: "Report created: 2020-02-14"
output:
  html_document:
    keep_md: yes
    number_sections: yes
    self_contained: yes
    theme: default
    highlight: null
    css: Static/ont_tutorial.css
    toc: yes
    toc_depth: 2
    toc_float:
      collapsed: yes
      smooth_scroll: yes
    df_print: paged
link-citations: yes
bibliography: Static/Bibliography.bib
always_allow_html: yes
---

<div            style="position:absolute;top:0px;right:0px;padding:15px;background-color:gray;width:45%;">
![](https://nanoporetech.com/themes/custom/nanopore/images/ont-logo.svg?tutorial=cas9enrichment)<!-- -->
</div>



## What will this workflow produce?

* A **`rich HTML`** format report containing summary statistics and figures highlighting performance of the enrichment protocol
* **`Microsoft Excel`** format files containing coordinates and summary statistics for on-target regions
* **`Fastq`** format file containing reads that correspond to each of the target regions specified
* **`Microsoft Excel`** format files containing summary statistics for sequence regions defined as showing off-target enrichment
```

* **`Fastq`** format sequence file containing reads that were classified as off-target enrichment
* **`Coordinates`** and instructions for reviewing candidate genomic regions with **`IGV`** the Integrated Genomics Viewer

Methods utilised include:

* **`conda`** for management of bioinformatics software installations
* **`snakemake`** for managing the bioinformatics workflow
* **`minimap2`** for mapping sequence reads to reference genome
* **`samtools`** for SAM/BAM handling and mapping statistics
* **`RSamtools`** and **`GenomicAlignments`**; R software for parsing BAM files
* **`seqtk`** for writing out the subseq of sequence reads that map to the target region
* **`IGV`** for visualising mapping characteristics at specific genomic regions

Analysis of the fastq format sequence data

Mapping sequence reads to the reference genome

The first step for the analysis of the Cas9 enrichment strategy is to assess the distribution and regional coverage of sequence reads across the whole genome. The **`fastq`** sequences produced during the DNA sequencing are mapped to the reference genome using the **`Minimap2`** software (@minimap22018). Results from the mapping analysis are passed to the **`samtools`** software (@samtools2009). **`Samtools`** is used to (1) filter out the unmapped sequence reads, (2) convert the uncompressed **`Minimap2`** SAM format output into the compressed BAM format and to (3) sort the sequences in the BAM file by their mapping coordinates. Further indexing the BAM file (again, using Samtools) enables efficient access to BAM entries that correspond to specific genomic locations.

Definition of background and off-target regions of the genome

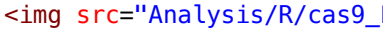
The Cas enrichment protocol depletes off-target DNA therefore enriching for the region of interest. In this tutorial all reads are aligned to the reference genome but not all of the reads sequenced during a Cas9 enrichment experiment align to the region of interest. All reads can be classed into four different mutually exclusive groups:

* **`On Target`** - reads that align to the regions of interest provided in the **`BED`** format coordinate file (**RawData/enrichment_targets.bed**)
* **`Target Proximal`** - reads that align to the regions immediately upstream or downstream of the region of interest (this regions is defined as 10000 bases)
* **`Off Target`** - Each crRNA in a panel should allow Cas9 to cut genomic DNA at sequence complementary sites with perfect alignment. Cas9 may also cut genomic DNA

at complementary sites with multiple mismatches. Such regions are classified as off-target if the depth of coverage is > **20X** over the mean background level
* **`Background`** - Reads that align to the reference genome but are not included in any of the categories above

The identification of the genomic regions corresponding to these mapping groups was performed using the **`R`** software. The **`GenomicRanges`** and **`GenomicAlignments`** packages (@granges2013) were used for genome geometry methods and the **`Rsamtools`** package (@R-rsamtools) and **`GenomicAlignments`** (@granges2013) packages were used to summarise the depth-of-coverage information used to identify the **`off-target`** genomic intervals.

Executive Summary



The information presented above summarises key metrics for benchmarking the performance of a DNA sequencing run following the Cas-mediated PCR-free enrichment protocol. The expected values below are for a 24hr MinION/GridION run

* Output will be lower following a Cas-mediated enrichment protocol compared to an average Nanopore sequencing experiment (0.5-3.5 Gb depending on the number of gene-targets and number of pooled-samples that are included in the sequencing run)
* 1-10% of the sequenced data should be on target
* The mean coverage per target should be >200X
* A 3000X depletion of non-target DNA should be observed

All these metrics are variable between experiments and depend on the size of the region of interest and the experimental set up. For further information on how to optimise these numbers please refer to the protocol.

Mapping characteristics by genomic segments

<table class="table table-striped table-condensed" style="margin-left: auto; margin-right: auto;">
<caption>Table summarising global mapping characteristics ranked by on-target, target-flanking and off-target</caption>
 <thead>
<tr>
<th style="border-bottom:hidden" colspan="1"></th>
<th style="border-bottom:hidden; padding-bottom:0; padding-left:3px;padding-right:3px;text-align: center; " colspan="1"> Background </th>

```
<th   style="border-bottom:hidden;   padding-bottom:0;   padding-left:3px;padding-
right:3px;text-align:  center;  "  colspan="1"><div  style="border-bottom:  1px  solid
#ddd; padding-bottom: 5px; ">Off-Target</div></th>
<th   style="border-bottom:hidden;   padding-bottom:0;   padding-left:3px;padding-
right:3px;text-align:  center;  "  colspan="1"><div  style="border-bottom:  1px  solid
#ddd; padding-bottom: 5px; ">Target-flanking</div></th>
<th   style="border-bottom:hidden;   padding-bottom:0;   padding-left:3px;padding-
right:3px;text-align:  center;  "  colspan="1"><div  style="border-bottom:  1px  solid
#ddd; padding-bottom: 5px; ">On-Target</div></th>
</tr>
  <tr>
   <th style="text-align:left;">   </th>
   <th style="text-align:left;">   </th>
   <th style="text-align:left;">   </th>
   <th style="text-align:left;">   </th>
   <th style="text-align:left;">   </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> total sequence reads<sup>*</sup> </td>
   <td style="text-align:left;"> 91,225 </td>
   <td style="text-align:left;"> 4,898 </td>
   <td style="text-align:left;"> 35 </td>
   <td style="text-align:left;"> 23 </td>
  </tr>
  <tr>
   <td style="text-align:left;"> mapped reads (primary)<sup>†</sup> </td>
   <td style="text-align:left;"> 24,478 </td>
   <td style="text-align:left;"> 4,898 </td>
   <td style="text-align:left;"> 35 </td>
   <td style="text-align:left;"> 23 </td>
  </tr>
  <tr>
   <td style="text-align:left;"> bases sequenced </td>
   <td style="text-align:left;"> 261,931,539 </td>
   <td style="text-align:left;"> 22,427,864 </td>
   <td style="text-align:left;"> 131,780 </td>
   <td style="text-align:left;"> 56,526 </td>
  </tr>
  <tr>
   <td style="text-align:left;"> bases mapped </td>
   <td style="text-align:left;"> 118,820,466 </td>
   <td style="text-align:left;"> 22,427,864 </td>
   <td style="text-align:left;"> 131,780 </td>
   <td style="text-align:left;"> 56,526 </td>
  </tr>
  <tr>
   <td style="text-align:left;"> Fraction of genome (%) </td>
   <td style="text-align:left;"> 99.293% </td>
   <td style="text-align:left;"> 0.693% </td>
```

```
   <td style="text-align:left;"> 0.012% </td>
   <td style="text-align:left;"> 0.002% </td>
  </tr>
  <tr>
   <td style="text-align:left;"> Mean coverage (primary)<sup>‡</sup> </td>
   <td style="text-align:left;"> 0.15 </td>
   <td style="text-align:left;"> 7.68 </td>
   <td style="text-align:left;"> 1.76 </td>
   <td style="text-align:left;"> 16.63 </td>
  </tr>
</tbody>
<tfoot><tr><td style="padding: 0; border: 0;" colspan="100%">
<span style="font-style: italic;">please note: </span> <sup>*</sup> fastq bases are calculated from the qwidth field of the mapped sequences and from the sequence length of unmapped sequences <sup>†</sup> this table presents only primary sequence mappings <sup>‡</sup> depth of coverage based only on primary mapping reads</td></tr></tfoot>
</table>
```

* Background reads result from the incomplete dephosphorylation of the genomic DNA followed by a non-specific ligation of the adapter sequence
* Off target reads result from the Cas9 protein cutting the DNA at a genomic location outside of the target region. Further graphs to show the location and distribution of off target regions are presented later in the report. If the number of off target regions and reads is higher than desired, please review the probe design to assess possible SNPs and candidate sequence mismatches
* Comparing the number of bases or reads classified as target-flanking relative to on-target values shows the efficiency of the probe design. A high number of reads/bases classified as target-flanking indicates read-through; it would be recommended to review the probe design for the crRNA probe that appears to “leak”

Evaluation of individual target performance

To gain the best insight on the performance of the Cas-mediated PCR-free enrichment protocol it is preferable to consider the performance of each discrete target separately. The table below highlights the characteristics for the different target regions defined within the starting BED file.

```
<table class="table table-striped table-condensed" style="margin-left: auto; margin-right: auto;">
<caption>Table summarising target mapping for pre-defined regions of interest</caption>
 <thead>
  <tr>
   <th style="text-align:left;"> Target Gene </th>
   <th style="text-align:left;"> Target size (nt) </th>
   <th style="text-align:left;"> Mean coverage </th>
   <th style="text-align:left;"> Read count<sup>*</sup> </th>
   <th style="text-align:left;"> Bases<sup>†</sup> </th>
```

```
   <th style="text-align:left;"> Mean readLength </th>
   <th style="text-align:left;"> Mean readQuality </th>
   <th style="text-align:left;"> Mean mapQuality </th>
   <th style="text-align:left;"> Reads on FWD(%)<sup>‡</sup> </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> POU5F1 </td>
   <td style="text-align:left;"> 4,081 </td>
   <td style="text-align:left;"> 16.63 </td>
   <td style="text-align:left;"> 23 </td>
   <td style="text-align:left;"> 56,526 </td>
   <td style="text-align:left;"> 2,915 </td>
   <td style="text-align:left;"> 9.79 </td>
   <td style="text-align:left;"> 60 </td>
   <td style="text-align:left;"> 43.48 </td>
  </tr>
</tbody>
<tfoot><tr><td style="padding: 0; border: 0;" colspan="100%">
<span style="font-style: italic;">please note: </span> <sup>*</sup> Reads are counted as all sequence reads where the SAM start location is located within the target interval. This does not correct for sequences on the reverse strand. <sup>†</sup> Bases are counted as the sum of nucleotides from all reads where the SAM start location is within target region; some of these bases will overlap the flanking region <sup>‡</sup> reads are assessed for strand of mapping; here reads on + strand are summarised as percentage of all</td></tr></tfoot>
</table>
```

* The mean coverage per target should be >200x
* Reads on FWD(%) indicates the percentage of sequence reads that map to the forward strand. If this value is not in the region of 50% then one of the probes is not working effectively
* A perfect mean map quality should be 60. A value of 60 indicates that reads are mapping to a single location in the genome (the target location). Lower mapping qualities may indicate either fragmented mapping (blocks of sequence interspersed by regions of no mapping at a single genomic location) or multi-mapping (the sequences can be mapped to multiple locations in the genome) leading to off-target effects
* Comparison of target read lengths may be used to identify the targets (and their probes) that either allow read-through. The ratio between the mean read length and target size should also be considered.

If the values in the table above are not ideal then please check the probe design advice and input requirements in the Cas-mediated PCR-free enrichment protocol.

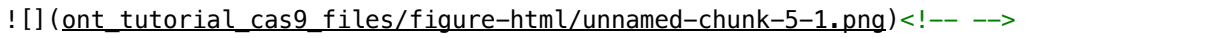
The output files prepared for the on-target analysis include

* The list of on-target read Ids can be found in the file **`Analysis/OnTarget/cas9_FAK75845.<TARGETNAME>.mappedreads `**
* The **`fastq`** sequence file containing the raw sequence reads corresponding to these Ids can be found in the file **`Analysis/OnTarget/cas9_FAK75845.<TARGETNAME>.fastq `**
* The coordinate information for the off-target regions can be found in the file **`Analysis/OnTarget/cas9_FAK75845_ontarget.xlsx `**

Graphical review of depth-of-coverage for target genes

The tables presented in the previous two sections have provided a summary of general mapping characteristics and on-target statistics. Plotting depth of coverage across the target regions also allows for an assessment of the performance of the crRNA guide used. The plots in this section review the depth of coverage, strandedness of mapping and leakiness of sequence coverage beyond the boundaries of the target region.

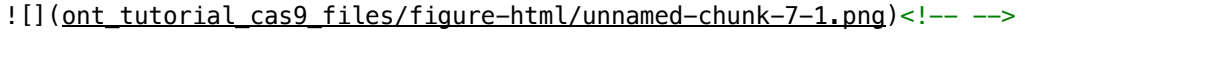
```r
singlePlot(names(ontargetUniverse)[1], aggregatedGR)
```



```r
# to plot the figure for a target called ¨Example¨ we could specify
# singlePlot('Example', aggregatedGR)
```

The figure above shows the depth-of-coverage around a target region. The on-target region is located within the vertical red-bars and is flanked by the target-proximal regions. The horizontal bar shows the threshold at which an off-target feature would be defined. This plot is for the **` POU5F1 `** target used in this tutorial.

```r
strandedPlot(names(ontargetUniverse)[1], aggregatedGR)
```



```r
# to plot the figure for a target called ¨Example¨ we could specify
# strandedPlot('Example', aggregatedGR)
```

The figure above presents the depth of coverage but is shaded by the strand (forward or reverse) to which the reads are mapped. This figure can be used to observe deviations from the expected 50:50 distribution of mapping between the + and - strands. Sequences that extend from the target regions and into the target-proximal regions may indicate suboptimal performance of a crRNA guide sequence.

Off-target mapping

Having assessed on-target characteristics, it makes sense to also consider what has been mapped to off-target regions of the genome.

The **`ideogram`** below presents a description of the off-target mapping locations split by chromosome. Each shaded region (or bar) corresponds to an off-target region. There are in total **1032** genomic regions that satisfy the mean depth-of-coverage threshold of **4.04**

The coordinates for these off-target regions have been written to an accompanying CSV file that may be imported into Excel for further analysis. The top 10 regions, ranked by mean depth-of-coverage, are presented in the table below.

<table class="table table-striped table-condensed" style="margin-left: auto; margin-right: auto;">
<caption>Table summarising the location and characteristics for the off-target regions with the highest depth-of-coverage</caption>
 <thead>
  <tr>
   <th style="text-align:left;"> chrId </th>
   <th style="text-align:left;"> start </th>
   <th style="text-align:left;"> end </th>
   <th style="text-align:right;"> width </th>
   <th style="text-align:right;"> mean coverage </th>
   <th style="text-align:right;"> reads in segment </th>
   <th style="text-align:left;"> mean read length </th>
   <th style="text-align:right;"> %FWD reads </th>
   <th style="text-align:right;"> mean readQ </th>
   <th style="text-align:right;"> mean MAPQ </th>
  </tr>
 </thead>
<tbody>
  <tr>
   <td style="text-align:left;"> 6 </td>
   <td style="text-align:left;"> 104,489,401 </td>
   <td style="text-align:left;"> 104,495,400 </td>

```
 <td style="text-align:right;"> 6000 </td>
 <td style="text-align:right;"> 26 </td>
 <td style="text-align:right;"> 42 </td>
 <td style="text-align:left;"> 6,087 </td>
 <td style="text-align:right;"> 50.00 </td>
 <td style="text-align:right;"> 9.70 </td>
 <td style="text-align:right;"> 1.13 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 40,236,901 </td>
 <td style="text-align:left;"> 40,241,700 </td>
 <td style="text-align:right;"> 4800 </td>
 <td style="text-align:right;"> 22 </td>
 <td style="text-align:right;"> 55 </td>
 <td style="text-align:left;"> 5,545 </td>
 <td style="text-align:right;"> 53.45 </td>
 <td style="text-align:right;"> 9.51 </td>
 <td style="text-align:right;"> 3.42 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 28,621,201 </td>
 <td style="text-align:left;"> 28,624,300 </td>
 <td style="text-align:right;"> 3100 </td>
 <td style="text-align:right;"> 21 </td>
 <td style="text-align:right;"> 37 </td>
 <td style="text-align:left;"> 5,842 </td>
 <td style="text-align:right;"> 62.16 </td>
 <td style="text-align:right;"> 9.65 </td>
 <td style="text-align:right;"> 2.31 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 78,926,401 </td>
 <td style="text-align:left;"> 78,929,000 </td>
 <td style="text-align:right;"> 2600 </td>
 <td style="text-align:right;"> 21 </td>
 <td style="text-align:right;"> 31 </td>
 <td style="text-align:left;"> 5,235 </td>
 <td style="text-align:right;"> 47.37 </td>
 <td style="text-align:right;"> 10.08 </td>
 <td style="text-align:right;"> 2.65 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 139,210,501 </td>
 <td style="text-align:left;"> 139,216,600 </td>
 <td style="text-align:right;"> 6100 </td>
 <td style="text-align:right;"> 20 </td>
 <td style="text-align:right;"> 43 </td>
```

```
 <td style="text-align:left;"> 5,502 </td>
 <td style="text-align:right;"> 48.98 </td>
 <td style="text-align:right;"> 9.54 </td>
 <td style="text-align:right;"> 1.45 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 2,584,401 </td>
 <td style="text-align:left;"> 2,589,500 </td>
 <td style="text-align:right;"> 5100 </td>
 <td style="text-align:right;"> 18 </td>
 <td style="text-align:right;"> 46 </td>
 <td style="text-align:left;"> 4,194 </td>
 <td style="text-align:right;"> 46.00 </td>
 <td style="text-align:right;"> 9.70 </td>
 <td style="text-align:right;"> 3.09 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 106,404,601 </td>
 <td style="text-align:left;"> 106,411,000 </td>
 <td style="text-align:right;"> 6400 </td>
 <td style="text-align:right;"> 18 </td>
 <td style="text-align:right;"> 42 </td>
 <td style="text-align:left;"> 5,917 </td>
 <td style="text-align:right;"> 50.00 </td>
 <td style="text-align:right;"> 10.00 </td>
 <td style="text-align:right;"> 2.15 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 169,371,801 </td>
 <td style="text-align:left;"> 169,372,900 </td>
 <td style="text-align:right;"> 1100 </td>
 <td style="text-align:right;"> 17 </td>
 <td style="text-align:right;"> 27 </td>
 <td style="text-align:left;"> 4,095 </td>
 <td style="text-align:right;"> 48.48 </td>
 <td style="text-align:right;"> 9.64 </td>
 <td style="text-align:right;"> 2.26 </td>
</tr>
<tr>
 <td style="text-align:left;"> 6 </td>
 <td style="text-align:left;"> 28,305,801 </td>
 <td style="text-align:left;"> 28,310,900 </td>
 <td style="text-align:right;"> 5100 </td>
 <td style="text-align:right;"> 17 </td>
 <td style="text-align:right;"> 40 </td>
 <td style="text-align:left;"> 6,396 </td>
 <td style="text-align:right;"> 56.82 </td>
 <td style="text-align:right;"> 9.65 </td>
```

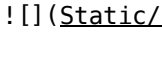
<td style="text-align:right;"> 3.23 </td>
</tr>
<tr>
<td style="text-align:left;"> 6 </td>
<td style="text-align:left;"> 110,539,201 </td>
<td style="text-align:left;"> 110,540,900 </td>
<td style="text-align:right;"> 1700 </td>
<td style="text-align:right;"> 16 </td>
<td style="text-align:right;"> 23 </td>
<td style="text-align:left;"> 5,485 </td>
<td style="text-align:right;"> 57.14 </td>
<td style="text-align:right;"> 9.73 </td>
<td style="text-align:right;"> 3.30 </td>
</tr>
</tbody>
<tfoot><tr><td style="padding: 0; border: 0;" colspan="100%">
please note: [*] This table has been prepared using only read mapping information that corresponds to a primary map [†] The reads in segment column describes the number of sequences that start within this genomic interval (using SAM start coordinate only) [‡] mean read length is the mean sequence read length for the mapping reads identified; their strandedness is summarised in %FWD reads (the number of sequences that appear on the forward strand) and the mapping quality is summarised in mapq</td></tr></tfoot>
</table>

The output files prepared from the off-target analysis include

* The list of off-target read Ids can be found in the file **`Analysis/OffTarget/cas9_FAK75845.OffTarget.mappedreads`**
* The **`fastq`** sequence file containing the raw sequence reads corresponding to these Ids can be found in the file **`Analysis/OffTarget/cas9_FAK75845.OffTarget.fastq `**
* The coordinate information for the off-target regions can be found in the file **`Analysis/OffTarget/cas9_FAK75845_offtarget.xlsx `**

Read Mapping Visualisation using IGV

The Integrative Genomics Viewer (**`IGV`**) is a tool (@10.1093/bib/bbs017, @Robinsone31), that has been installed by **`conda`**, for the visualisation of genomics data. The software provides functionality for the display of sequence mapping data from BAM files that can subsequently be overlaid with "tracks" of information that can include depth-of-coverage for mapping data and gene annotations.



The figure above presents a screenshot from the **`IGV`** software. The coordinates for an off-target region have been selected (see the top bar of the figure for the coordinates) and the display has been zoomed-in so that the quality and mapping strand can be observed.

`IGV` can be started from the command line by using the command `igv`.

In this tutorial we recommend that you instead encourage **`IGV`** to display sequence information around a target of interest. During the analysis presented in this tutorial, we have been reviewing the enrichment of sequences around the **`HTT`** gene.

We would like to explore the read mapping information around this gene. The command below will open the **`IGV`** browser to the appropriate genome coordinates - in this example we are using the feature with coordinates `[Chr 4] Start=3072436, Stop=3079444` - this corresponds to the **`HTT`** gene.

```
igv -g ./ReferenceData/Homo_sapiens.GRCh38.dna.chromosome.4.fa \
./Analysis/Minimap2/cas9_FAK76554.bam,./RawData/enrichment_targets.bed \
4:49091201-49156600
```

** please note that if you are using your own reference genome and BED file that the parameters above may require modifications **

Reproducible research - produce your own report

This report has been created using **`Rmarkdown`**, publicly available **`R`** packages, and the \LaTeX document typesetting software for reproducibility. For clarity the **`R`** packages used, and their versions, are listed below.

\fontsize{8}{12}

```
R version 3.6.2 (2019-12-12)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: macOS Catalina 10.15.2

Matrix products: default
BLAS/LAPACK: /Users/nadakubikova/anaconda3/envs/ont_tutorial_cas9/lib/libopenblasp-
r0.3.7.dylib

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

loaded via a namespace (and not attached):
  [1] colorspace_1.4-1         showtext_0.7-1           biovizBase_1.34.0
  [4] htmlTable_1.13.3         base64enc_0.1-3          dichromat_2.0-0
  [7] rstudioapi_0.11          farver_2.0.3             showtextdb_2.0
 [10] bit64_0.9-7              AnnotationDbi_1.48.0     xml2_1.2.2
 [13] splines_3.6.2            knitr_1.28               Formula_1.2-3
```

```
 [16] cluster_2.1.0           dbplyr_1.4.2             png_0.1-7
 [19] graph_1.64.0            BiocManager_1.30.10      readr_1.3.1
 [22] compiler_3.6.2          httr_1.4.1               backports_1.1.5
 [25] assertthat_0.2.1        Matrix_1.2-18            lazyeval_0.2.2
 [28] formatR_1.7             acepack_1.4.1            htmltools_0.4.0
 [31] prettyunits_1.1.1       tools_3.6.2              gtable_0.3.0
 [34] glue_1.3.1              GenomeInfoDbData_1.2.2   rappdirs_0.3.1
 [37] Rcpp_1.0.3              vctrs_0.2.2              rtracklayer_1.46.0
 [40] xfun_0.12               stringr_1.4.0            proto_1.0.0
 [43] rvest_0.3.5             lifecycle_0.1.0          ensembldb_2.10.0
 [46] gtools_3.8.1            XML_3.99-0.3             zlibbioc_1.32.0
 [49] BSgenome_1.54.0         VariantAnnotation_1.32.0 ProtGenerics_1.18.0
 [52] hms_0.5.3               RBGL_1.62.1              AnnotationFilter_1.10.0
 [55] curl_4.3                memoise_1.1.0            gridExtra_2.3
 [58] biomaRt_2.42.0          rpart_4.1-15             reshape_0.8.8
 [61] latticeExtra_0.6-29     stringi_1.4.5            RSQLite_2.2.0
 [64] checkmate_2.0.0         GenomicFeatures_1.38.0   rlang_0.4.4
 [67] pkgconfig_2.0.3         bitops_1.0-6             evaluate_0.14
 [70] lattice_0.20-38         purrr_0.3.3              labeling_0.3
 [73] htmlwidgets_1.5.1       bit_1.1-15.2             tidyselect_1.0.0
 [76] GGally_1.4.0            plyr_1.8.5               magrittr_1.5
 [79] R6_2.4.1                Hmisc_4.3-1              DBI_1.1.0
 [82] pillar_1.4.3            foreign_0.8-75           withr_2.1.2
 [85] survival_3.1-8          RCurl_1.98-1.1           nnet_7.3-12
 [88] crayon_1.3.4            OrganismDbi_1.28.0       BiocFileCache_1.10.0
 [91] rmarkdown_2.1           sysfonts_0.8             jpeg_0.1-8.1
 [94] progress_1.2.2          grid_3.6.2               data.table_1.12.8
 [97] blob_1.2.1              digest_0.6.24            webshot_0.5.2
[100] openssl_1.4.1           munsell_0.5.0            viridisLite_0.3.0
[103] askpass_1.1
```

\fontsize{10}{14}

It is also worth recording the versions of the software that have been used for the analysis.

\fontsize{8}{12}

```
# packages in environment at /Users/nadakubikova/anaconda3/envs/ont_tutorial_cas9:
#
# Name                    Version                   Build  Channel
bioconductor-rsamtools    2.2.0            r36h6de7cb9_0    bioconda
igv                       2.4.9                        1    bioconda
minimap2                  2.17                hfbae3c0_1    bioconda
r-rstudioapi              0.11             r36h6115d3f_0    conda-forge
rstudio                   1.1.456             h04f5b5a_1    r
samtools                  1.10                h457b48f_2    bioconda
seqtk                     1.3                 h2573ce8_2    bioconda
```

```
snakemake-minimal         5.10.0                     py_0    bioconda
```

\fontsize{10}{14}

\pagebreak

Customise the tutorial template

The **`R`** code to prepare your own report is included in the distributed **`Rmarkdown`** file. The **`Rmarkdown`** file can be loaded, viewed, and edited in the **`RStudio`** software. Within your **`conda`** environment (and within your tutorial folder), simply type

\fontsize{8}{12}
```
rstudio ont_tutorial_cas9.Rmd
```
\fontsize{10}{14}

Final thoughts. Behind this **`Rmarkdown`** file is a modest amount of **`R code`** - please explore the **`Rmarkdown`** template; modify it and run with your own samples.

To extract the whole set of **`R code`** from the **`Rmarkdown`**, use the **`purl`** command - this will extract the R code into its own file.

```
knitr::purl("ont_tutorial_cas9.Rmd", quiet=TRUE)
```

Further explore the mapping data

Running the tutorial Rmarkdown script saves the R data objects, presentation methods, and results in a sessionFile. This saved session can be opened directly and it is possible to further explore the data in a dynamic fashion.

To load the data

```
rstudio
```

Once the R console has loaded re-load the saved session with

```
library(session)
```

```
restore.session("Analysis/Results/enrichment.Rdata")
```

There are a number of R objects that could be of immediate interest for further exploration of the data

* **`ontargetUniverse`** - a **`GenomicRanges GRanges`** object containing the genomic coordinates and summary information for the on-target regions of the genome
* **`offtargetUniverse`** - a **`GenomicRanges GRanges`** object containing the genomic coordinates and summary information for the off-target regions of the genome
* **`backgroundUniverse`** - a **`GenomicRanges GRanges`** object containing the genomic coordinates and summary information for the background regions of the genome
* **`aggregatedGR`** - a fine resolution **`GenomicRanges GRanges`** object giving depth of coverage information over the target region(s) and the proximal sequence.

Glossary of terms

* __BAM__ is a compressed and binary version of the __SAM__ file format; please see __SAM__

* __BED__ is a tab-delimited file format and acronym for Browser Extensible Data. In this tutorial BED files are used to define genomic regions and the columns describe respectively [chromosome, start, end, name]

* __knit__ is the command to render an Rmarkdown file. The knitr package is used to embed code, the results of R analyses and their figures within the typeset text from the document.

* __L50__ describes the number of sequences (or contigs) that are longer than, or equal to, the N50 length and therefore include half the bases of the assembly

* __N50__ describes the length (read length, contig length etc) where half the bases of the sequence collection are contained within reads/contigs of this length or longer

* __SAM__ is a file format and acronym for Sequence Alignment/Map file format. SAM files are a tab-delimited file and store information on the sequence reads that can be mapped to a genome and the confidence that the mapping is correct.

* __QV__ the quality value, -log10(p) that any given base is incorrect. QV may be either at the individual base level, or may be averaged across whole sequences

* __Rmarkdown__ is an extension to markdown. Functional R code can be embedded in a plain-text document and subsequently rendered to other formats including the PDF format of this report.

Appendix 4

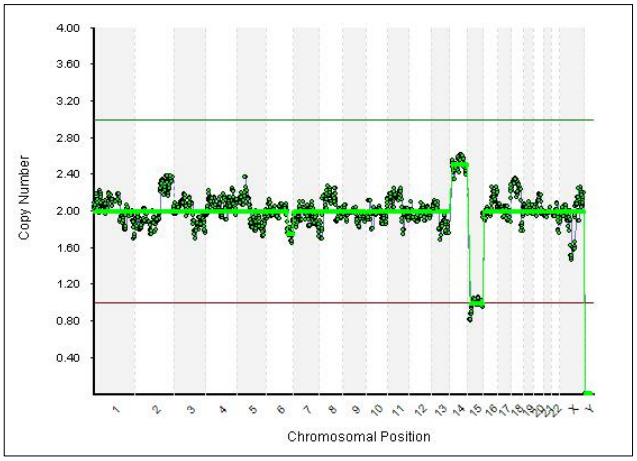
Example script developed to generate read counts in bins used for chromosome copy number analysis. The script can be used on DNA sequence files of any organism with adjusting the reference and bed files (containing the bin specification) accordingly.

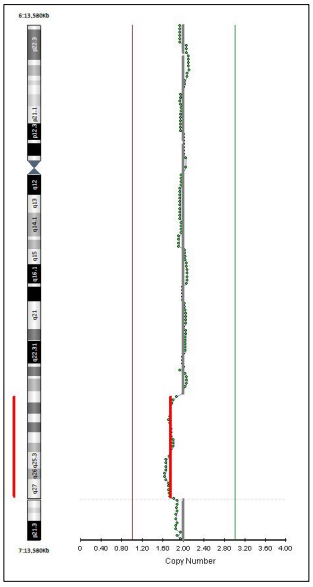
```
## What will this workflow produce? ##

* **`Coordinates`** and instructions for reviewing candidate genomic regions with corresponding read counts


## Methods utilised include:

* **`minimap2`** for mapping sequence reads to reference genome
* **`samtools`** for SAM/BAM handling and mapping statistics
* **`GATK`** for writing out the sequence reads that map to the target genome



Samples

#!/bin/bash
for r in *.bam
do
    sample=${r/.bam/}
    echo "processing sample: "$sample

    echo "adding read groups sample: "$sample
    ./gatk  AddOrReplaceReadGroups  -I  $sample".bam"  -O  $sample"_arg.bam"  -LB Experiment -PL Platform -PU 001 -SM Species

    echo "indexing BAM file: "$sample
    samtools index $sample"_arg.bam"

    echo "collecting read counts: "$sample
    ./gatk CollectReadCounts -I $sample"_arg.bam" -O $sample"_arg.tsv" -L input.bed -imr OVERLAPPING_ONLY --format TSV

done
```

Appendix 5

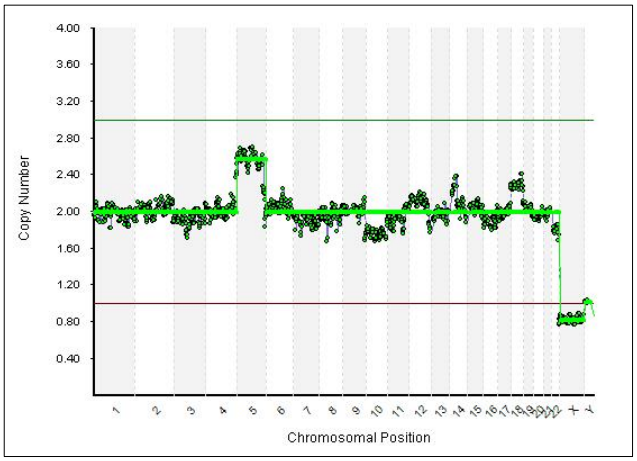
Whole chromosome copy number profiles (left) and chromosome 6 copy number profile (right) generated by BlueFuse Multi software after low-pass whole genome sequencing of targeted and control human embryonic DNA.

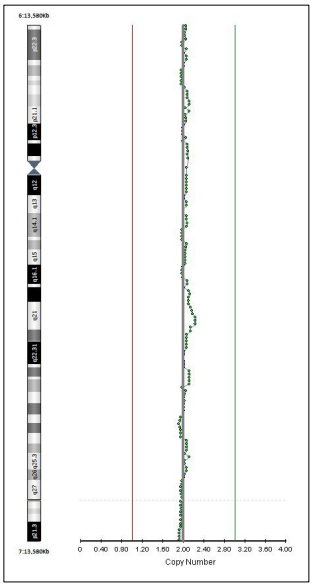
Embryo C1



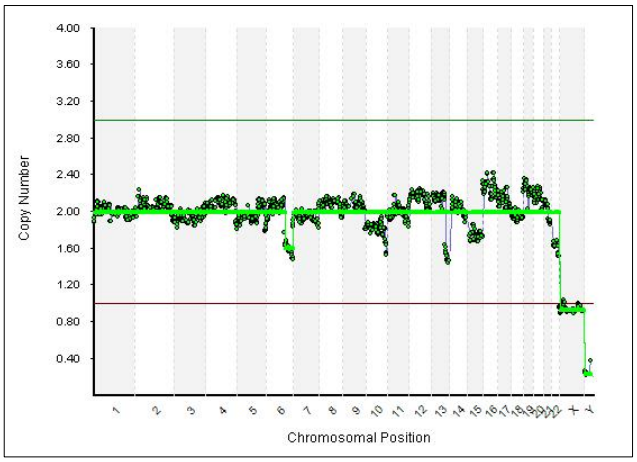


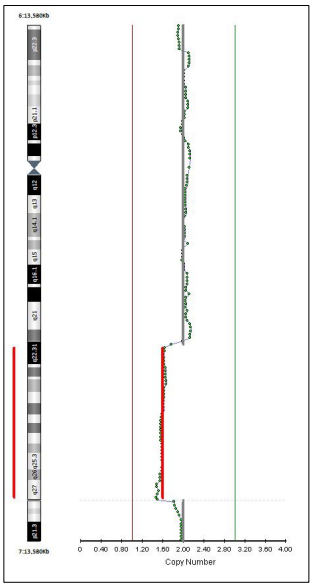
Embryo C2



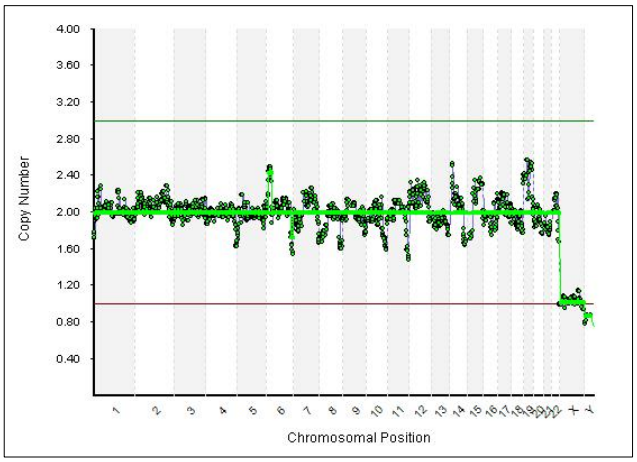


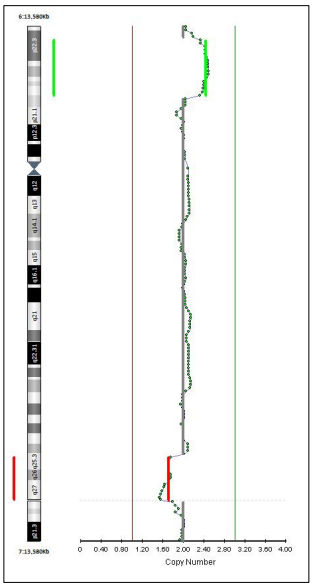
Embryo C3



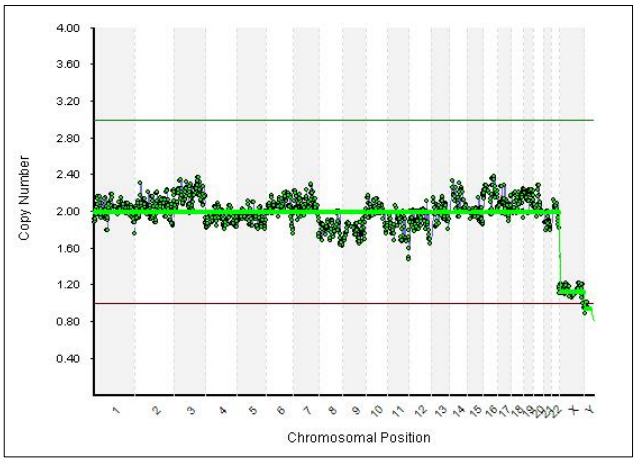


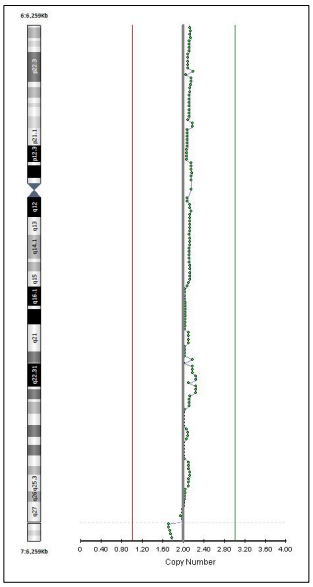
Embryo C4



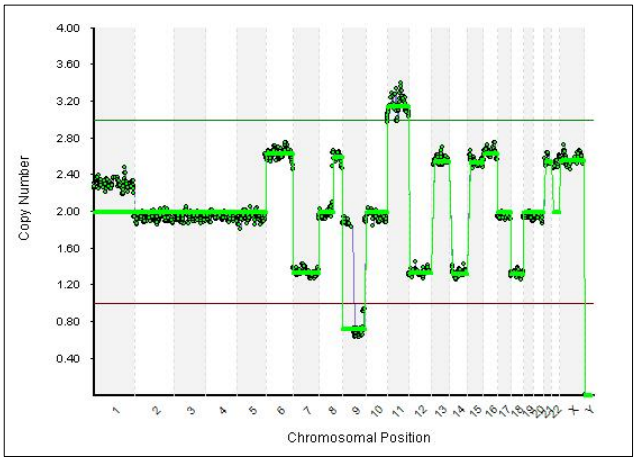


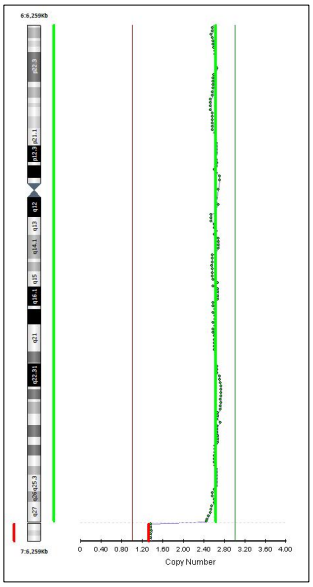
Embryo C8



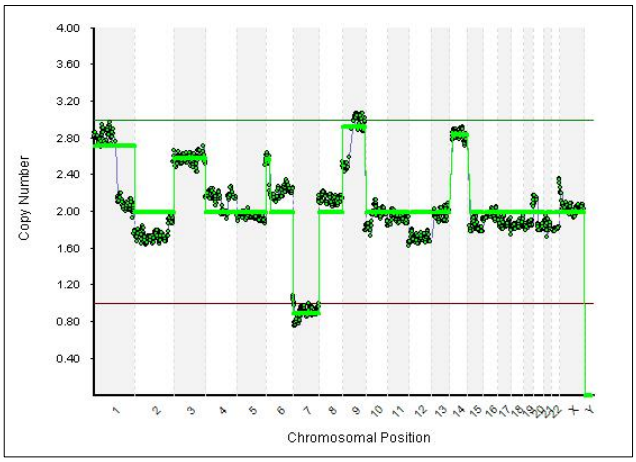


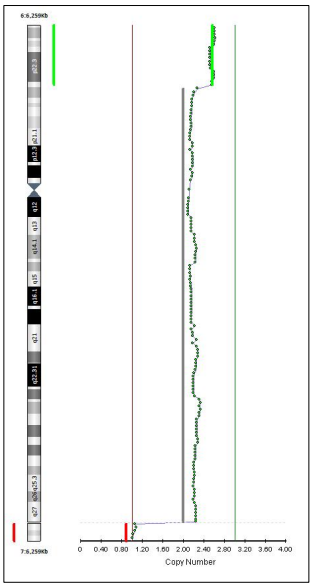
Embryo C9



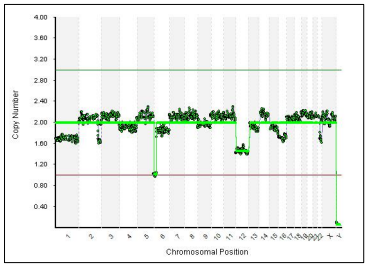


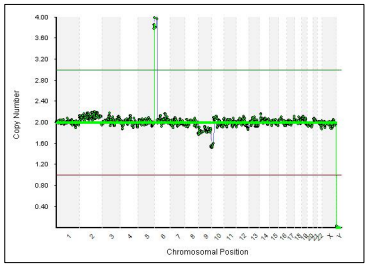
Embryo C10

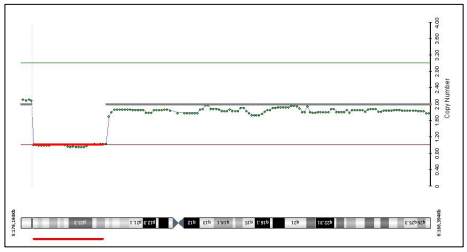


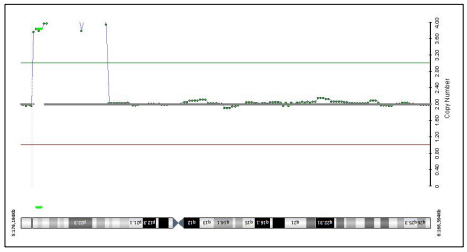


Embryo C12

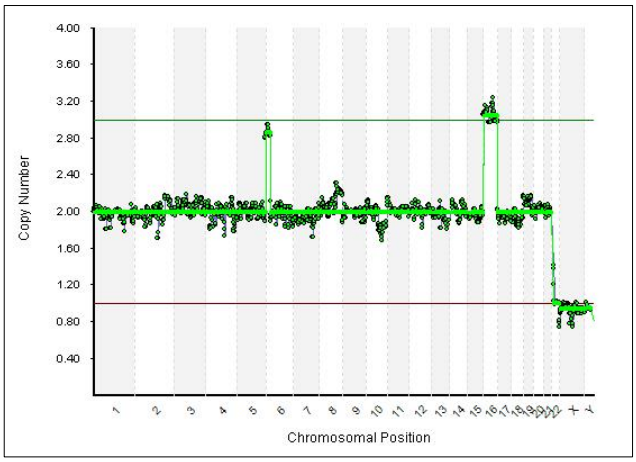


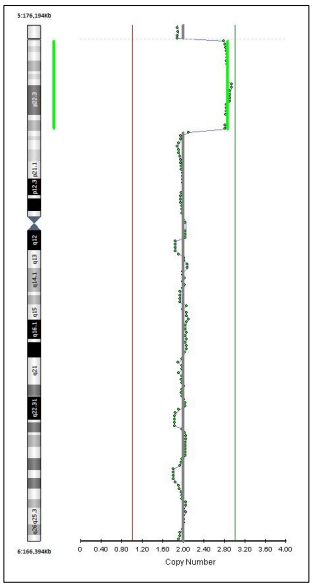




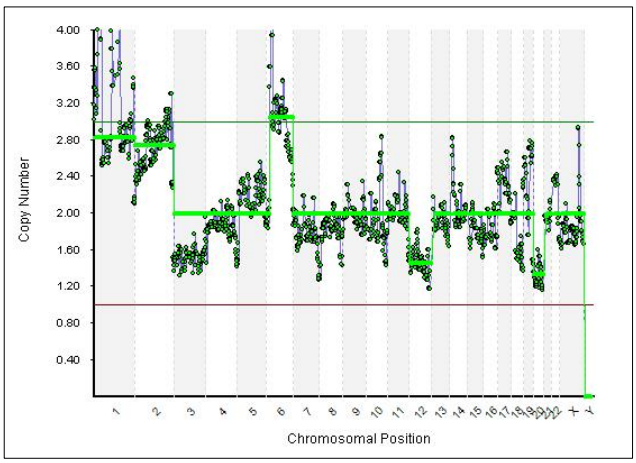


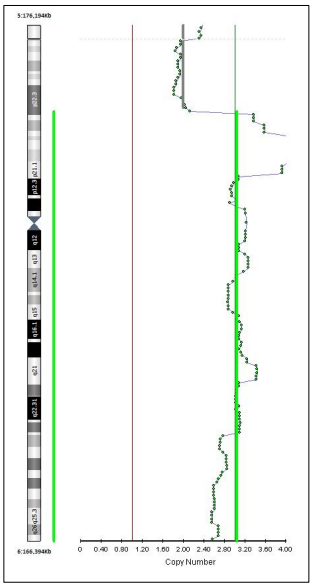
Embryo C14



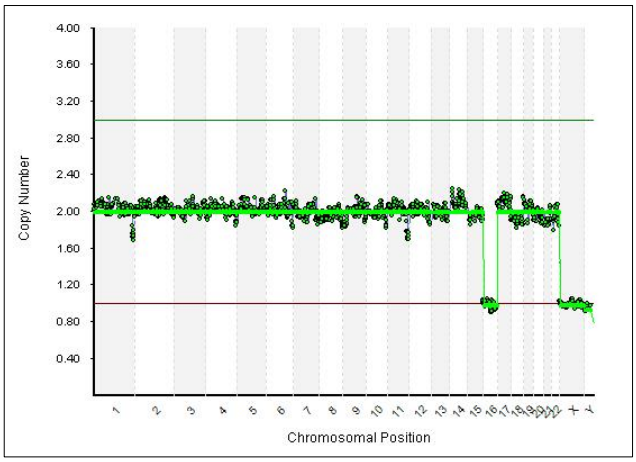


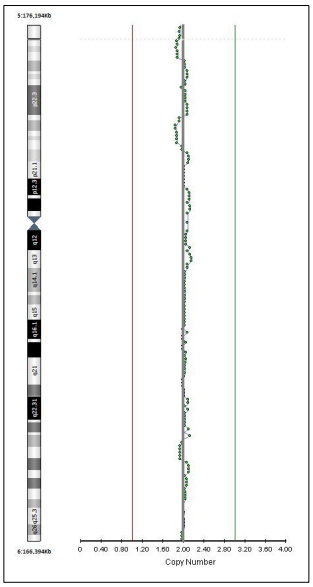
Embryo C15



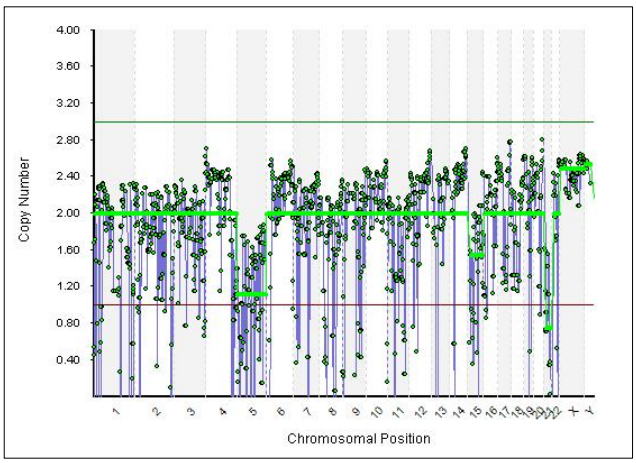


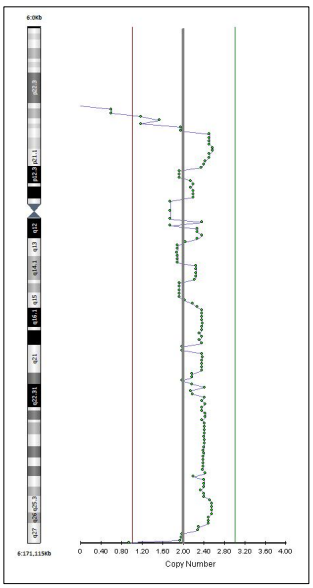
Embryo C16



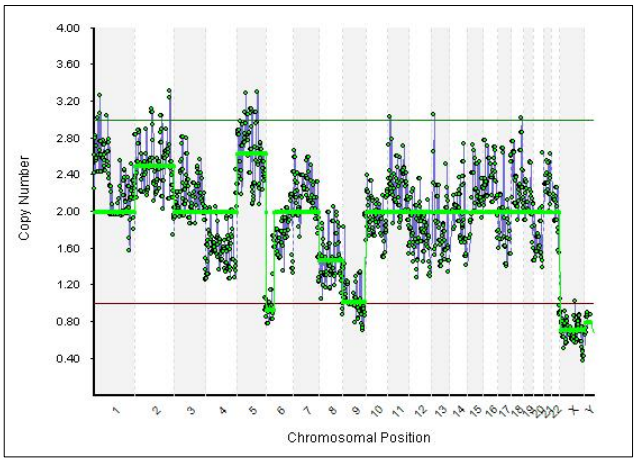


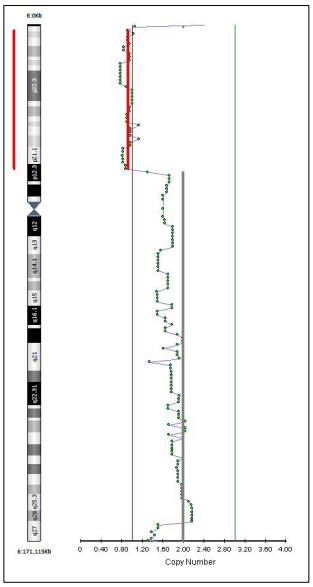
Embryo C19



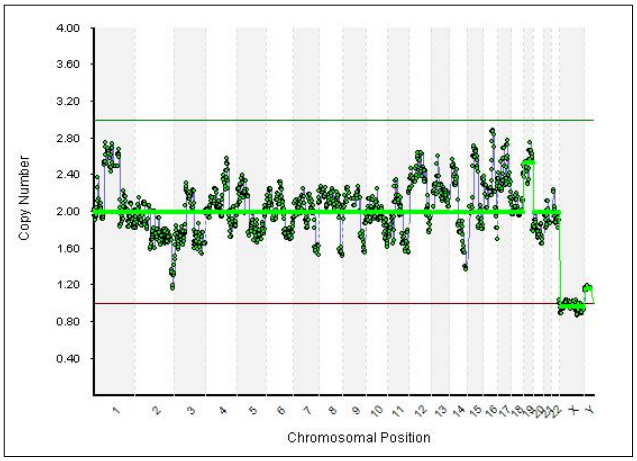


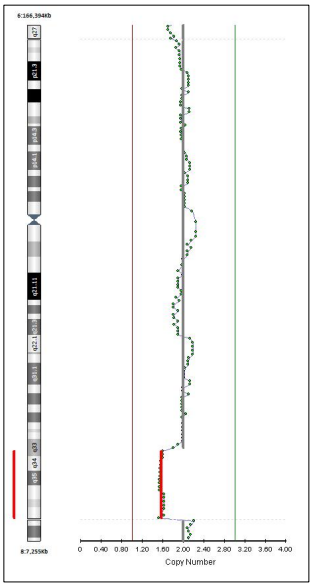
Embryo C20



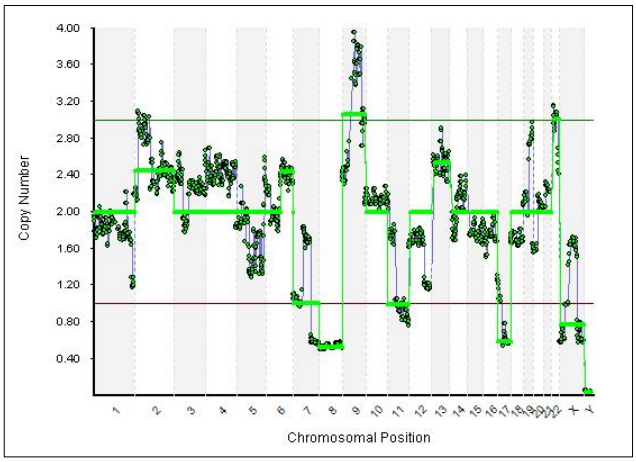


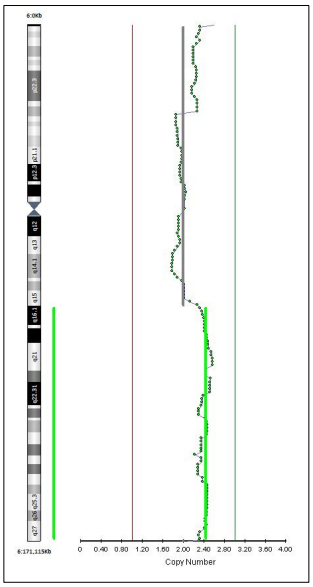
Embryo C22



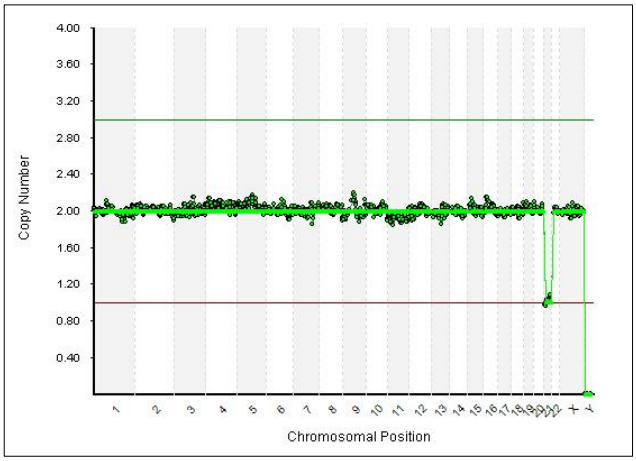


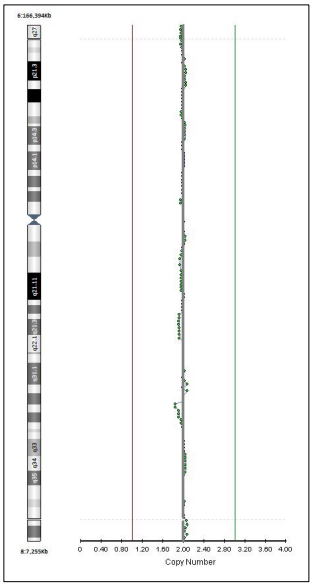
Embryo C23



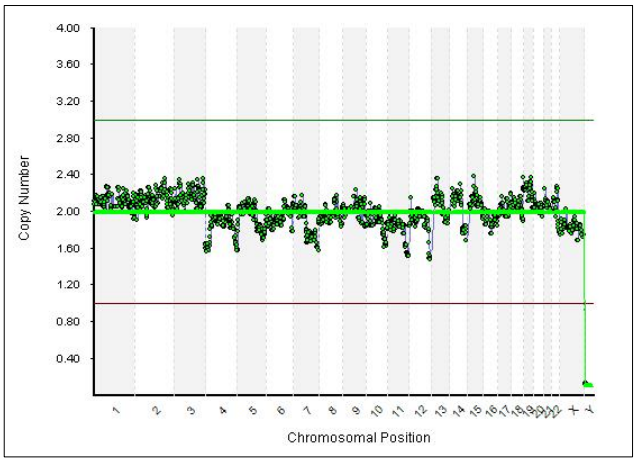


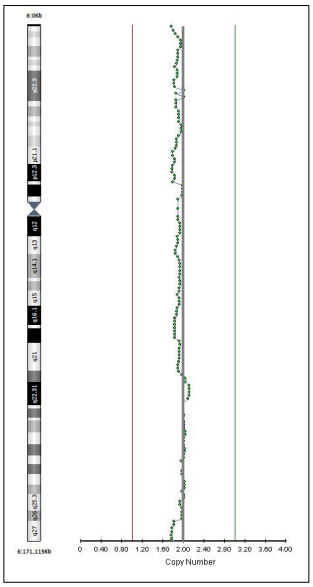
Embryo C24



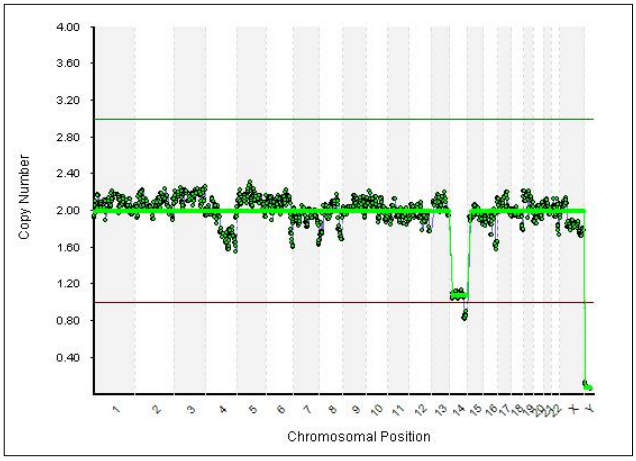


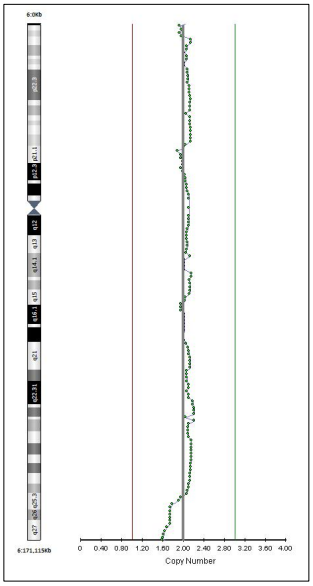
Control Embryo 5K



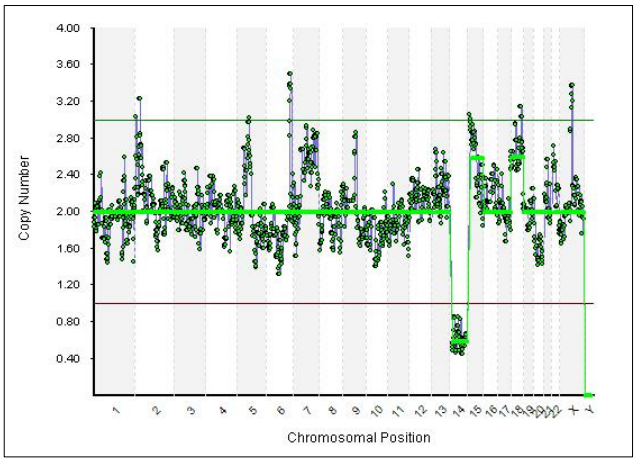


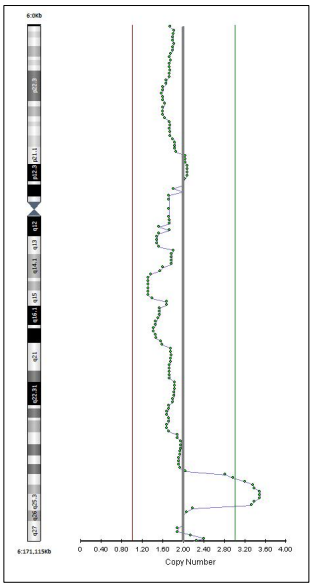
Control Embryo 7K



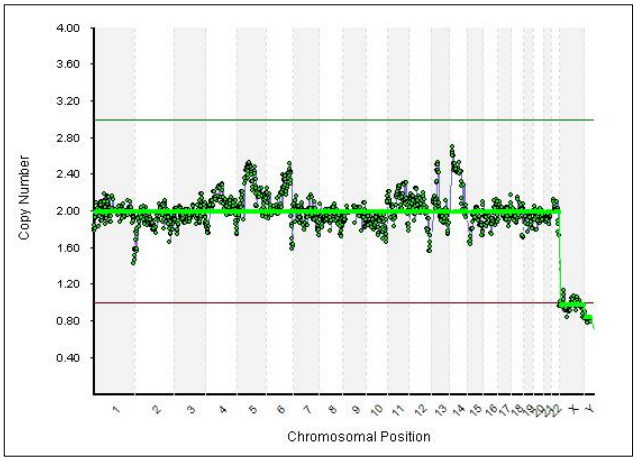


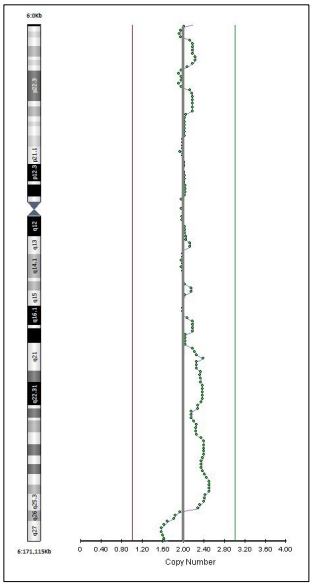
Control Embryo 8K



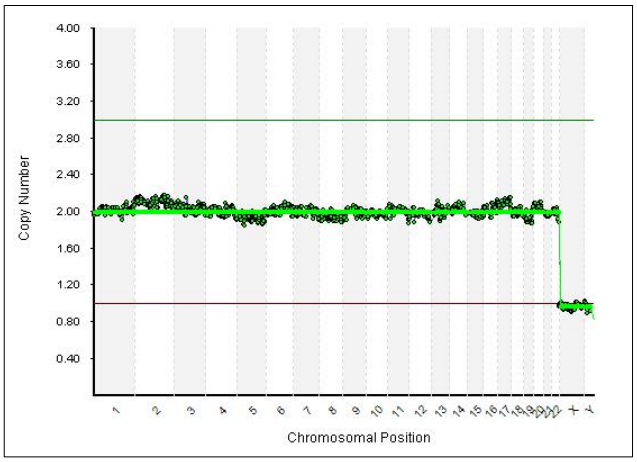


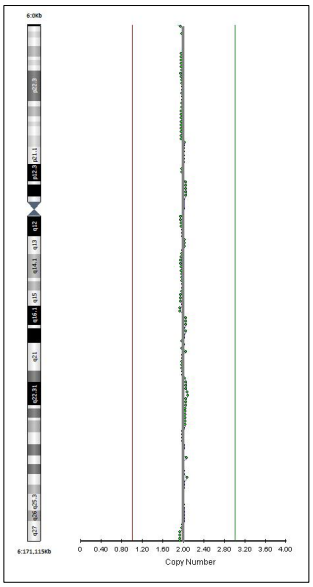
Control Embryo 1K



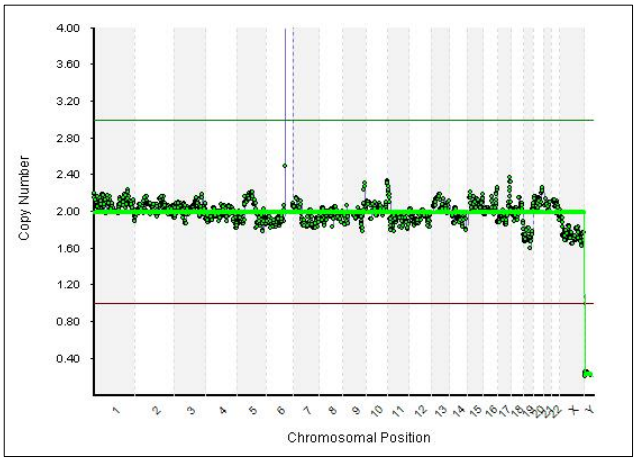


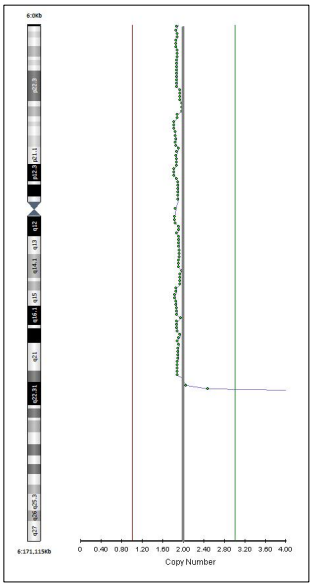
Control Embryo 2K



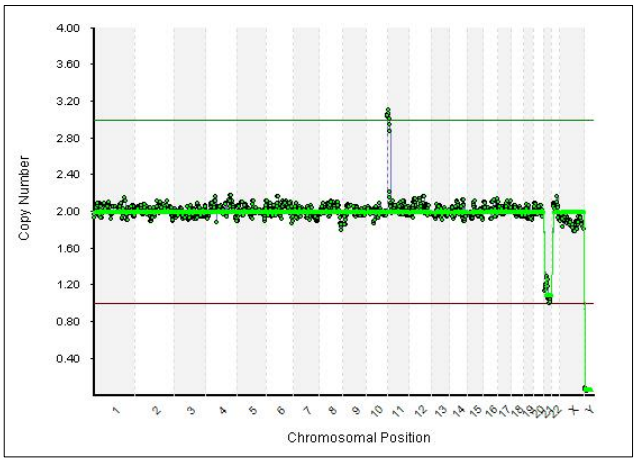


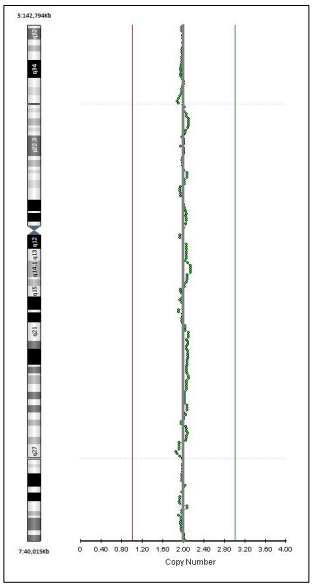
Control Embryo 3K





Control Embryo 4K





Control Profile