Chapter 5

Discussion

5.1 Summary of Findings

Understanding variation and consequences of germline mutation is paramount for understanding evolutionary process, biological mechanisms of mutation and causes of genetic disease. In this thesis I have described three distinct projects that have attempted to explore different aspects of germline mutation.

In Chapter 2 I examined the mutational origins and pathogenic consequences of MNVs, an important source of genetic variation. I found that the most frequent type of MNV was at adjacent nucleotides and that more than half of these were likely due to a single mutational event. I then confirmed observations from previous studies that there appear to be several mutational processes creating different types of MNVs. I estimated the MNV mutation rate to be 1.6% that of SNVs. Most population genetics models assume that mutations arise from independent events and ignoring clustered mutations can lead to incorrect inferences about positive selection. Understanding the mutational spectra and mutation rate of MNVs will help inform future refinements to these models. The presence of a possible MNV mutator phenotype that I observed in individuals in the DDD study may further complicate these refinements. I found that MNVs in protein coding sequences are on average more pathogenic than SNVs. Even when the MNV falls within a single codon, they can create a larger physiochemical change and have a greater functional impact than an SNV in the same codon. I identified 10 pathogenic *de novo* MNVs within the DDD study. My findings demonstrate that MNVs constitute a unique class of variant in both mutational origin and functional impact. This has implications in how variants are annotated in the future and correct annotation is important in furthering our understanding of their role in evolution and disease.

In Chapter 3 I describe my work on the identification and characterisation of germline hypermutators. I identified fifteen individuals, 1 from the Deciphering Developmental Disorders (DDD) study and 14 from the 100,000 Genomes Project (100kGP), with a significantly increased number of *de novo* SNV mutations (DNMs) that ranged from a 2 to 7 fold enrichment compared to the expected number. The DNMs in these individuals exhibited distinctive mutational spectra, some of which mapped on to known somatic mutational signatures identified in cancers as well currently unknown mutational signatures. I found that for the majority of these individuals the excess mutations were paternally derived implicating the father as a possible germline hypermutator. In two of these fathers I identified rare nonsynonymous homozygous variants in genes known to be associated with DNA repair, *MPG* and *XPC*. *MPG* is known to be involved in the base excision repair pathway while *XPC* is associated with the nucleotide excision repair pathway. The variant in *XPC* is a pathogenic PTV known to cause xeroderma pigmentosum with which the father has previously been diagnosed. This finding suggests that other *XPC* carriers may have a similar mutator phenotype and should be followed up with additional pedigree WGS studies. Germline hypermutation accounted for 7% of variation in the germline mutation rate in 100kGP despite only affecting 14 individuals. I estimated that parental age accounted for ~70% of germline mutation rate variation which leaves ~20% of variance unaccounted for. These findings suggest that defects in DNA repair genes can dramatically increase the germline mutation rate. However, I found that the presence of a nonsynonymous variant across different subsets of DNA repair genes did not significantly impact the number of DNMs per person across the cohort, even those variants known to increase the somatic mutation rate. This suggest that at most a subset of variants in DNA repair genes associated with cancer also affect the germline and possibly that hypermutation is more often a recessive phenotype and heterozygous variants may not increase mutation rates. This is also demonstrated by my findings on the impact of germline PTVs in the gene *MBD4*. I found that these variants, that are known to be associated with a somatic mutator phenotype, have no detectable effect on germline mutation rate. These analyses have provided new insights into how genetic variation can impact individual germline mutation rate and that hypermutation accounts for a substantial fraction of variance in germline mutation rates. The remaining unexplained variance calls attention to the need to investigate additional sources of variation such as polygenic or environmental contributions. This will likely require being able to cheaply and accurately assay germline mutation rates to be able to perform a genome-wide association study.

My final project, described in Chapter 4, shifts away from investigating variation in the types and rates of DNMs to their role in causing rare disease. I integrated health care and research exome sequences and analysed *de novo* mutations in a cohort of 31,058 parent

offspring trios with developmental disorders. I developed an improved statistical framework that increased power to identify gene-specific enrichments. I applied this on the cohort and identified 285 genes significantly associated with DD, including 28 that have not previously been robustly associated with DDs. Despite detecting more DD-associated genes, ~50% of the excess of DNMs in protein-coding genes remained unaccounted for. I performed a down-sampling analysis and found that the discovery of DD-associated genes has not plateaued and that increasing samples sizes should result in discovery of many novel DD-associated genes. I modelled the likelihood of the observed distribution of both missense and protein-truncating DNMs and my results suggest that over 1,000 novel DD-associated genes await discovery. I found that the undiscovered genes are likely to be less penetrant than the currently known genes and that pre/peri-natal death is reducing power to detect novel DD-associated genes. This has important implications for how the field approaches the discovery of the remaining genes and that this will require improved analytical methods as well as increased sample sizes. A substantial role for incomplete penetrance suggests that combining inherited and *de novo* variation may aid discovery of novel DD-associated genes. I also identified a set of highly recurrent mutations which are strong candidates for novel germline selection mutations. These need to be examined further as they may result in a much higher disease incidence than expected which would increase with paternal age.

5.2 Limitations and future directions

There are several limitations in the work I have described in this thesis and these can help highlight future avenues of research to further our understanding of germline mutation and its role in rare disease.

This work highlights the deficiencies of current mutational models in addressing mutation heterogeneity and different types of genetic variation. In Chapter 4, I found that some of the initially significant DD-associated genes were enriched for synonymous variants in gnomAD and this could be evidence of mutation rate heterogeneity across genes. To improve the mutational model one can start to include additional annotation of features known to influence mutation rate. For example, recently developed models have started to incorporate methylation status which impacts CpG mutation rates[102, 22]. Other factors known to influence mutation rate such as GC content, proximity to recombination hotspots or replication timing could also be incorporated. Somatic mutational models have started to include these and a plethora of other local genomic features ([13, 120, 171, 141]). However these models have the benefit of being able to be trained on a large number of somatic mutations in normal tissue which is not currently tractable with germline mutations. The

average number of DNMs per person is ~70 and require whole genome sequencing of both the child and two parents. Currently most large collections of DNMs are in disease cohorts, such as for DD or autism, where DNMs are enriched and due to clinical ascertainment the mutations would not be distributed as expected under a null germline mutational model. A large dataset of *de novo* mutations in healthy individuals would be useful in training a null germline mutational model. These DNMs would also help to inform how we can include MNVs in a null mutational model and help us further investigate individual mutation rate variation. DNMs in healthy individuals would represent a combination of both mutation and negative selection, to examine mutation alone sequencing of gametes is needed however this is very difficult to amass. Trio sequencing of a healthy population would also take time to build and a more feasible approach currently would be to train a mutational model on rare variation in large healthy population cohorts as an approximation for germline mutation.

In all three projects I focused primarily on SNVs and expanding these to include indels would be informative. MNVs which include multiple indels, or possibly include an SNV together with an indel, are likely to have a larger functional impact compared to multiple SNVs and it would be interesting to explore if there are specific mutational mechanisms that may create these. When investigating germline hypermutation in Chapter 3, there was some evidence that some individuals have an excess of *de novo* indels despite having an expected number of SNVs. Since the indel calling was less sensitive than SNVs these candidate *de novo* indels would need to be carefully examined to ensure they are real. In Chapter 4, I was not able to estimate the positive predictive value of being pathogenic directly for indels and I used weights from missense and nonsense variants as proxies for inframe and frameshift mutations respectively. This was due to the fact that I did not have a mutational model for indels that could reliably calculate the expected number of these mutation types within the exome. Developing a mutational model for indels would be an important development in the field. Indels are inherently more complex to create a mutational model than SNVs as there are added dimensions of the length of the indel as well as whether an insertion/deletion is created. Training such a model requires a large dataset of high quality indels which is difficult to compile since indel calling tends to be less accurate than SNVs and indels have a much lower mutation rate compared to SNVs.

Sample size is a limitation in all three of the projects I describe and building larger parent-offspring trio sequencing datasets will be key in addressing further unanswered questions from this thesis. In the context of developmental disorders, despite pooling over 31,000 exome-sequenced trios from two research and one healthcare generated datasets in Chapter 4, I found that larger sample sizes of 100,000 trios would be needed to detect additional genes associated with developmental disorders. Achieving sample sizes of this size is only possible

through collaboration with healthcare generated data. GeneDx continues to grow their dataset through clinical genomics testing. In the UK the creation of the NHS Genomics Medicine service means that children with rare diseases, as well as certain rare disease and cancer in adults, will be eligible for whole genome sequencing. The 100,000 Genomes Project has now been expanded and the next aim is to sequence 5 million genomes via research and industry partners with the NHS and UK biobank contributing to 1 million of these genomes. These datasets will include a large number of parent-offspring trios and, as well as helping our study of rare genetic disease, could also be useful in understanding properties of germline mutation. Larger sample sizes are needed to untangle the impact of genetic variation on germline mutation rate. These efforts could focus on a curated set of variants in genes already known to be involved in DNA repair or be more agnostic and look across all genes. These studies could also help us to understand the differences between variants that impact somatic and germline mutation differently. Inclusion of families with more than one child or even large families in these datasets will also allow us to further investigate how germline mutation rate can vary between, and within, families.

A limitation to much of my work in this thesis is that it has been focussed on coding variation. The increasing sample sizes of WGS datasets will allow more interrogation of the non-coding genome. In Chapter 4 the study of DNMs in DD was limited to coding regions of the genome. Previous work conducted in the Hurles group started to address the role of DNMs in highly conserved fetal brain regulatory elements, however it was concluded that much larger sample sizes would be needed to have the power to identify specific elements associated with DD[201]. Short et al. also highlighted the the importance of improved pathogenicity annotations in the non-coding genome. There has been several methods developed recently to predict pathogenicity by combining many different genomic features that prioritise non-coding variants[107, 230, 85, 246, 203, 199, 88]. As WGS datasets increase in size and there are a sufficient amount of DNMs, these annotations could be included in the DeNovoWEST framework by incorporating them into the weights. Improvements in pathogenicity annotation will also be key in furthering other work from this thesis. When searching for possible germline mutator variants in hypermutated individuals in Chapter 3 I was also restricted to coding variants but some of these variants of interest may lie in regulatory regions. My study of MNVs in Chapter 2 was restricted to the exome and even with large WGS we do not currently have the power to detect the functional impact of non-coding MNVs. Better annotation of regulatory regions may allow us to see differing levels of enrichment in disease cohorts of MNVs compared to SNVs.

In addition to large scale whole genome sequencing of families, single-cell sequencing of sperm will be an important tool to investigate different properties of male germline

mutation. As mentioned in Chapter 3, current estimates of individual germline mutation rate are restricted to averaging the number of DNMs across offspring however this is often based on a single observation and is a combination of the maternal and paternal germline mutation rates. The development of single-cell sequencing and its application to sperm will allow more accurate estimation of individual mutation rates. These technologies should also allow examination of within individual mutation rates as well as clonal dynamics occurring in the testes and could help identify further mutations that undergo germline selection. In the context of germline hypermutation, sequencing of sperm would confirm if hypermutation occurs in all sperm and explore whether some mutator variants might be mosaic, specific to the germline and thus undetectable in soma-derived DNA.

My thesis was focused on germline mutation and so in Chapter 4 I only assessed the contribution of DNMs to developmental disorders. However, these analyses suggested that reduced penetrance of mutations was impacting the power to detect novel genes associated with DD and that incorporating inherited variation together with DNMs will be an important next step in identifying these genes. Pre/peri-natal lethality also reduces power to detect DD-associated genes and expanding pre-natal fetal sequencing cohorts may be helpful in discovering these genes.

5.3 Concluding remarks

The advent of exome and whole genome sequencing has allowed for close examination of germline mutation in the last decade. Direct identification of DNMs from families has led to improved estimation of mutation rates and has provided important insights into factors that can influence this rate at a genome, individual and population level. These advances have enabled the construction of a mutational model which has aided identification of genes associated with rare genetic disease and led to countless diagnoses for families which may pave the way for possible treatments. However we have only started to scratch the surface of understanding germline mutation. The increase in sample size of sequenced cohorts in the last few years has revealed additional complexities in sources of variation such as those discussed in this thesis; from the impact of non-independent mutations to the existence of hypermutators. In the future, more detailed characterisation of individual environment and geospatial analyses of mutation rates may also uncover the impact of environmental mutagens. The increase in sample size has also started to uncover the deficiencies in current mutational models which can confound our detection of disease associations. Incorporation of what we have learnt about the process of germline mutation into improved mutational models is paramount, both for evolutionary studies and to improve how we assess the contribution of

DNMs to disease. Even with improvements to the mutational model, integration of data from vast numbers of families will be needed to achieve a more complete picture of the genes involved in rare disease. Global collaboration between healthcare systems and research initiatives will be key in amassing the amounts of data needed and this will further not only our understanding of rare disease but also of germline mutation as a whole.