Chapter 1

Introduction

1.1 Motivation

Germline mutation is the ultimate source of evolutionary change and disease-causing variants. Understanding the rates and patterns of human mutation can help us learn about their molecular origins, uncover our evolutionary history and improve our ability to identify the genetic causes of human disease. With the advent of exome and genome data sets of parent-offspring trios we have an unprecedented opportunity to characterize mutations at an individual level and to harness the increasing sample sizes to identify disease-causing mutations.

1.2 Mutational processes

1.2.1 Types of mutation

Mutations can cause small or large changes to DNA. When a mutation causes a single base pair change this leads to a single nucleotide variant (SNV). Occasionally a single mutational event can create multiple base pair changes in close proximity (typically within 20bp). This is referred to as a multi-nucleotide variant (MNV). When SNVs and MNVs fall within protein coding sequences they can lead to changes in the amino acid sequence which can alter the protein product (missense variants) or lead to a premature truncation of the protein (protein truncating variants (PTVs)).

Mutations can also result in insertions or deletions (indels). Indels range widely in size. They can affect just one or two base pairs up to tens of thousands of base pairs which can affect entire genes. When small indels fall into coding sequences they can shift the reading frame, leading to a frameshift or if they are divisible by three will lead to an inframe indel.

Structural variants encompass larger types of genetic variation including large insertions, deletions, inversions or duplications .

1.2.2 Origins of mutation

Mutations arise primarily from errors in DNA replication and from chemical damage to DNA. During replication, misincorporated nucleotides occasionally escape detection by proofreading mechanisms and can lead to single base changes. These single base mutations can be classified as either transitions or transversion and the rates of transitions is twice that of tranversions [129]. In addition to single base pair changes, small indels can be created by slippage of the polymerase during replication, typically in repeat regions.

Chemical damage to DNA can be induced by both endogenous and exogenous sources. DNA is vulnerable from alkylation and oxidations and can also incur spontaneous damage from hydrolysis and deamination. A common endogenous source of mutation is a result of hydrolytic deamination of cytosine which spontaneously deaminates to uracil. Uracil is easily identified as an unnatural base and can be efficiently repaired by the base excision pathway. However 5-methylcytosine can undergo deamination to Thymine which is repaired by less efficient pathways[37]. In humans, the $5'$ C in a CpG context is usually methylated and has a mutation rate that is higher than any other context[45]. The most commonly occurring endogenous nucleotide base lesion has been shown to be due misincorporation of ribonucleotides which can lead to genome instability and subsequently mutation if insufficiently repaired[176, 108].

Exogenous sources of damage to DNA can be due to various environmental mutagens. There are several well characterised examples. For example, exposure to UV radiation can create DNA lesions which then results in mutation. Ionising radiation is a well known mutagen that causes double strand breaks in DNA. Benzo(a)pyrene is a known carcinogen found in tobacco smoke which induces mutations after forming covalent DNA adducts.

1.2.3 DNA damage tolerance and repair

Repair and tolerance mechanisms exist to counteract DNA damage and to correct or mitigate the impact of damage. There are five primary DNA repair pathways: base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR) and non-homologous end joining (NHEJ). These pathways protect against specific types of damage. For example BER and NER correct damage that has occurred to single bases. To remove this damage BER typically excises a single base while NER removes a patch of nucleotides. For specific subsets of base damage, such as UV photolesions and

alkylated bases, there are also enzymes that can directly reverse the associated DNA damage. MMR targets base mismatches and insertion/deletion mismatches that arise as a result of DNA replication errors. HR and NHEJ are both focused on repair of double strand breaks (DSBs). Homologous recombination repairs DSBs by utilising DNA sequence information from homologous sequences while NHEJ repairs DSB by directly joining the broken ends. Translesion synthesis (TLS) is an example of a DNA damage tolerance mechanism. TLS is conducted by specific polymerases that allow DNA replication to proceed past aberrant DNA lesions.

These repair processes are crucial in preserving genetic stability however in certain instances can also create their own mutations. For example repair of DSBs by NHEJ typically introduces errors, often indels, at the repair site. Additionally translesion synthesis polymerases (such as REV1, POLζ, POLη,POLκ and POL ι) bypass DNA lesions in order to continue replication but frequently introduce an incorrect base in the process [96, 132, 216, 160].

Defects in DNA repair pathways can lead to an increase in mutation rate, which in somatic tissue can lead to cancer. Many known cancer driver mutations fall within genes involved in these repair processes [217, 28]. Germline variants in certain DNA repair genes can predispose individuals to cancer. For example germline variants in MMR genes are known to cause Lynch syndrome, a form of hereditary colorectal cancer, and variants in BRCA1/2, which are involved in homologous recombination, are known to predispose carriers to breast cancer[167]. While the link between defects in DNA repair and somatic hypermutation is well established, the effect in germline tissue is still not clear.

Different mutational mechanisms can leave distinct mutational patterns. These combinations of mutation types are referred to as 'mutational signatures' [159]. For example, UV light is a source of DNA damage that can crosslink adjacent pyrimidine bases. If this pyrimidine dimer is not repaired, DNA polymerases will typically insert two adenines opposite the dimer resulting in a predominantly C to T or CC to TT mutational signature. More complex mutational signatures can be extracted computationally and many of these have been well characterised. There are currently >100 somatic mutational signatures that have been identified in a wide variety of cancers [4]. In cancer these can provide evidence of what gene may be defective; approximately half of these signatures have been attributed to endogenous mutagenic processes or specific mutagens [165, 5, 4].

1.3 Estimating human germline mutation rates

1.3.1 Early strategies to detect mutations

Before the ability to observe *de novo* mutations (DNMs) directly, researchers developed a range of strategies to detect mutations indirectly. The initial strategy depended on phenotypic markers that were easily observed.

The first estimates of germline mutation rates were not made in humans but in *Drosophila* and maize [154, 207, 155]. A major challenge in making these estimates came from the fact that mutations happen at a very low rate. In these organisms, estimates were made from breeding many individuals and then scoring large numbers of offspring. In *Drosophila* this was done by counting the number of lethal mutations, which would impact the number of viable offspring, as an approximation of the mutation rate. In maize the mutation rate was estimated using several easily observed traits. This approach is, of course, not feasible in humans and so another strategy was needed. Haldane made one of the first estimates of the human mutation rate in 1935 by using estimates of the frequency of haemophiliac men in London [72]. This approach relied on the idea that the frequency of the harmful allele was a function of the balance between the mutation rate and the resulting fitness of that allele. Subsequent estimates were calculated using other monogenic diseases [109]. These methods made several assumptions which may affect their accuracy. Firstly they had to make an estimate of the mutational target for the disease and assumed complete ascertainment of the phenotype. They were also only able to base their mutation rate estimate on coding regions of the genome.

The mouse specific-locus test was developed in the 1950s [23, 184]. This initially involved exposing a wild-type animal to a mutagen, such as ionising radiation, and crossing them with homozygotes for specific recessive genes that were easily visible (such as ear type, hair colour or certain eye traits). The first generation offspring are then examined for differences from the expected wild type as evidence of mutation[185]. This required a large number of mice however the method was relatively straightforward with regards to counting mice with a visible phenotype. This test was able to detect specific recessive mutations and visible dominant mutations. Detecting dominant mutations depended on the researchers' power of observation and was rather subjective. Mutations affecting certain physical traits could be obvious but mutations affecting behaviour or small physical changes could easily be missed[192]. The background and radiation-induced mutation rates for 8 specific loci were estimated in the 1950s and required ~85,000 mice in total. This allowed for a better estimate of the impact of radiation on humans than previous experiments conducted in *Drosophila*.

Another strategy for estimating mutation rates came not from direct counting of phenotypes but through evolutionary comparisons. The basis of this approach is that the mutation rate for neutral mutations is equal to the rate of evolution [106, 155]. This means that one can calculate the amount of sequence divergence between non-coding DNA sequences of two species and if one has a good estimate of the time at which these species diverged then one can use these two pieces of information to estimate the mutation rate. This strategy was applied by comparing pseudogenes between human and chimpanzees and the mutation rate was estimated to be ~2.5×10^{-8} mutations per base per generation[156, 110].

1.3.2 Methods for recent direct mutation rate estimates

Advances in available technology allowed mutation rates of specific types of genomic variation to be estimated directly. With the advent of DNA fingerprinting in 1985, the mutation rate of tandem-repetitive 'minisatellite' regions in the human genome was one of the first examples of a human mutation rate estimated directly from a pedigree [95, 93]. This was done from a dataset of 40 large families that consisted of 344 offspring which allowed for identification of *de novo* changes in the length of minisatellites. Locus specific mutation rates were estimated from PCR assays of pooled sperm DNA [94]. Two decades later, CNVs were found to be widespread in the human genome and constituted an important source of genomic variation.[194, 87]. Locus specific mutation rates were, like minisattelites, initally estimated from pooled sperm DNA [218, 227]. The development of the CNV microarray allowed for identification of *de novo* CNVs from parent-offspring trios which yielded initial direct estimates of the CNV mutation rate [193, 239, 90].

The emergence of massively parallel sequencing technologies has now allowed us to directly observe *de novo* SNVs and indel mutations in pedigrees. Recent whole-genome and exome sequencing studies have allowed us to sequence trios as well as larger family structures. These studies have estimated the mutation rate to be ~1.2×10^{-8} per base per generation for SNVs. This corresponds to ~70 DNMs per individual [33, 180, 111]. This estimate is almost half that of the mutation rate estimates derived from earlier approaches. Discrepancies between these estimates affect our understanding of the timing of human evolution and may suggest that the mutation rate has changed over time [190]. More generally, it highlights a need to understand potential sources of variability and/or error in mutation rate estimates.

1.4 Variation in the human germline mutation rate

Variation in the human germline mutation rate can be considered from different perspectives: within genomes, between individuals and between populations.

1.4.1 Variation within genomes

The rate of mutation for SNVs varies by several orders of magnitude along the genome and several factors appear to affect this (Figure 1.1). Sequence context is an important source of variability. The bases flanking either side of a nucleotide, the tri-nucleotide context, have varying mutation rates beyond the increased mutability in CpG sites [86, 16]. More recent work has also shown that this variability extends beyond a trinucleotide context and there is additional value in considering the heptanucleotide context [3, 22]. On a slightly broader scale, the surrounding context of CpGs impacts their mutation rate. GC content around the CpG appears to increase the stability of methylated cytosine. This, in turn, appears to reduce its mutation rate[80]. DNA replication timing is another source of variability. Studies have shown a higher mutation rate in late-replicating regions in the germline, perhaps due to a depletion of free nucleotides[208]. Mutations are also more likely to occur near recombination hot-spots[127]. A recent study estimated that the mutation rate is ~50 fold higher within 1 kb of crossovers. They also observed that females, but not males, have increased mutation rates up to 40 kb from crossovers, especially if these are complex [73]. Transcription also affects mutational patterns. In transcribed regions, the mutation rate of A>G and A>T substitutions is higher on the non-coding strand compared to the coding strand which is most likely due to transcription coupled repair [52, 153, 68]. Low complexity regions (LCRs) have been found increase the mutation rate in the surrounding DNA. This increase in mutation rate has been found to correlate with the distance from the LCR [126].

Chromatin and nucleosome organisation may also affect mutation rates across the genome however their role is not well established. *De novo* SNVs are more abundant in regions of closed chromatin however this could be due to the fact that open chromatin is more accessible for a range of different DNA repair activities and is associated with high transcription rates and allows transcription-coupled repair to act in these regions [138]. The role of nucleosome organisation on mutation rate is also unclear. For example high nucleosome occupancy was initially thought to reduce nearby mutation rate [146] however a more recent study found the opposite effect in a different dataset [204]. Recent work has also found elevated mutation rates around translationally stable nucleosomes[128].

Another source of genomic variation in mutation rates is dependent on their relative position to each other. Mutations may not always occur independently. Mutations appear to

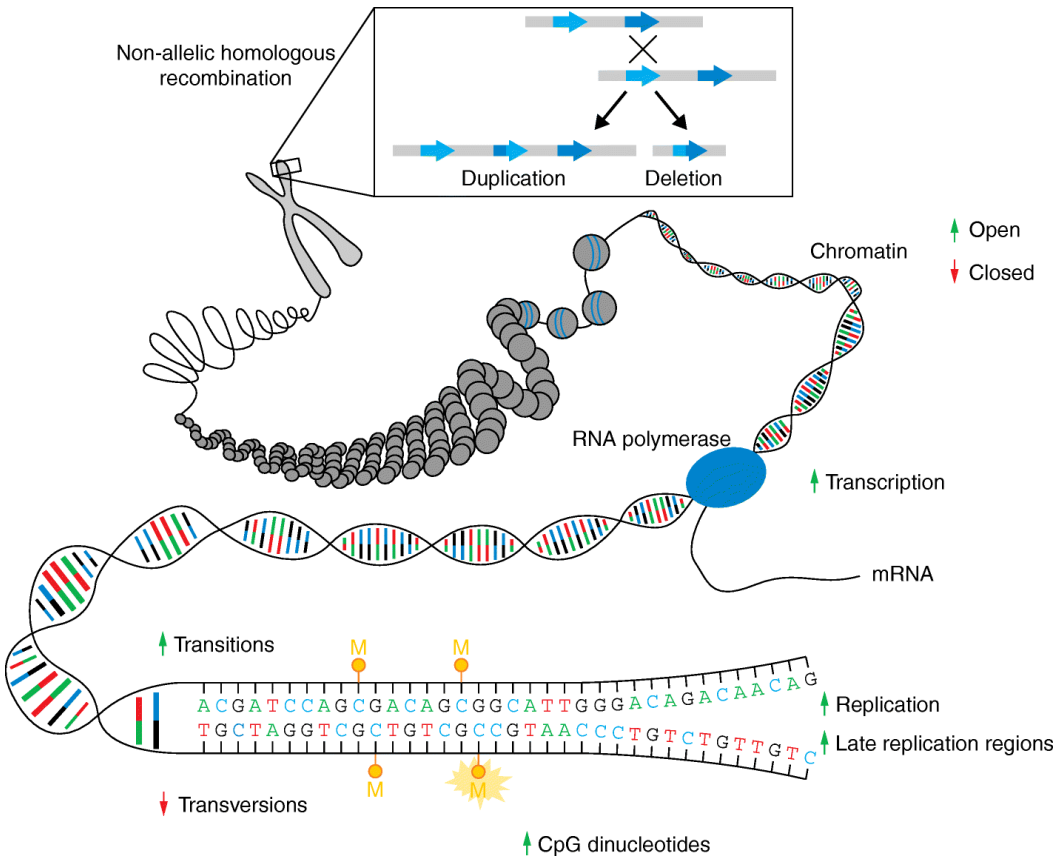

Fig. 1.1 Mechanisms of and genome level factors that influence the rate of *de novo* mutations. Figure sourced from Acuna-Hidalgo *et al* [1]

be more clustered in the genome than we would expect if they were independent [146, 196]. This has been observed for distances of up to 20 kb. These mutations do not appear to have different properties with respect to recombination rate or replication timing [52]. On a finer scale, there is an excess of pairs of mutations within 100bp that appear to be in perfect linkage disequilibrium in population samples [195, 209, 76]. A subset of these clustered mutations which fall within 20 bp of each other are referred to as multi-nucleotide variants (MNVs). The mechanisms driving MNVs are explored in more depth in Chapter 2.

Specific DNMs have been found to be highly recurrent. Moreover, the prevalence of these mutations strongly correlates with increasing paternal age. These DNMs appear to confer a selective advantage within spermatogonial stem cells which lead to clonal expansion within the testis [64, 66]. Well-known examples of these are mutations include those that fall in genes such as *FGFR2*, *FGFR3*, *RET*, *PTPN11* and *HRAS*. These mutations also lead to congenital skeletal disorders and can increase cancer risk[66, 64]. Since these mutations are positively selected for in the testis but are deleterious to the organism they have been termed

'selfish mutations' [64, 1]. The disorders caused by these 'selfish' mutations have an incidence up to 1000 times greater than expected based on the mutational target size[67, 10, 1, 65].

Aside from SNVs, variation in mutation rate of other forms of genetic variation has also been observed. Mutational hotspots have been identified for copy number variants (CNVs) which have up to a ~100 fold increase of mutation rate. These hotspots are enriched for being located near to segmental duplications likely due to these CNVs arising from non-allelic homologous recombination[175, 53]. Variation across the genome has also been observed for short tandem repeats (STRs)[49, 70]. Studies have observed a positive association between repeat number and mutation rate in humans[84]. Length of the repeat unit has also been shown to influence the STR mutation rate[210, 70]. For indels, it has been shown that reptitive regions, including STRs and homopolymer runs, increases the indel mutation rate due to the higher propensity of polymerase slippage in these regions [55, 151, 121]. The increase in the indel mutation rate in repetitive regions has shown to depend on both the size of the repeat unit and the length of the repeat tract[151].

1.4.2 Individual level variation

Mutation rates vary between individuals and several factors that influence this rate have been identified. Approximately 3-4 times as many mutations originate from the father than the mother which indicates differences in male and female germline mutation rates[71, 173, 111]. The number of DNMs observed in an individual is highly associated with paternal age. Paternal age accounts for the majority of mutation rate variation between individuals[111, 98]. It has been estimated that there is an increase of ~2 DNMs for every additional year in father's age [111, 173]. It has been proposed that the increasing number of cell divisions with age in the male germline is a likely source of this effect. In women, oocytes undergo a fixed number of cell divisions early in their life. In men, spermatogonid stem cells replicate continuously throughout their life allowing for more replicative errors (Figure1.2) . A recent study interrogating this hypothesis has examined the fraction of paternal mutations in phased DNMs and has found that this does not increase with paternal age[54]. This may suggest that replication is not driving the paternal age effect and the mutations are predominantly damage-induced. Interestingly, during the course of my PhD, a maternal age effect has also been detected with a more subtle increase of ~0.5 DNMs/year [234, 98]. This could be due to an accumulation of spontaneous mutations over time in the female germ cells. This parental age effect has also been shown to differ between families. A study of three families, each with 4 or 5 children, has shown that the paternal age effect may differ across families [173]. A more recent analysis of 33 families of three generations from Utah has confirmed significant differences between parental age effects [189].

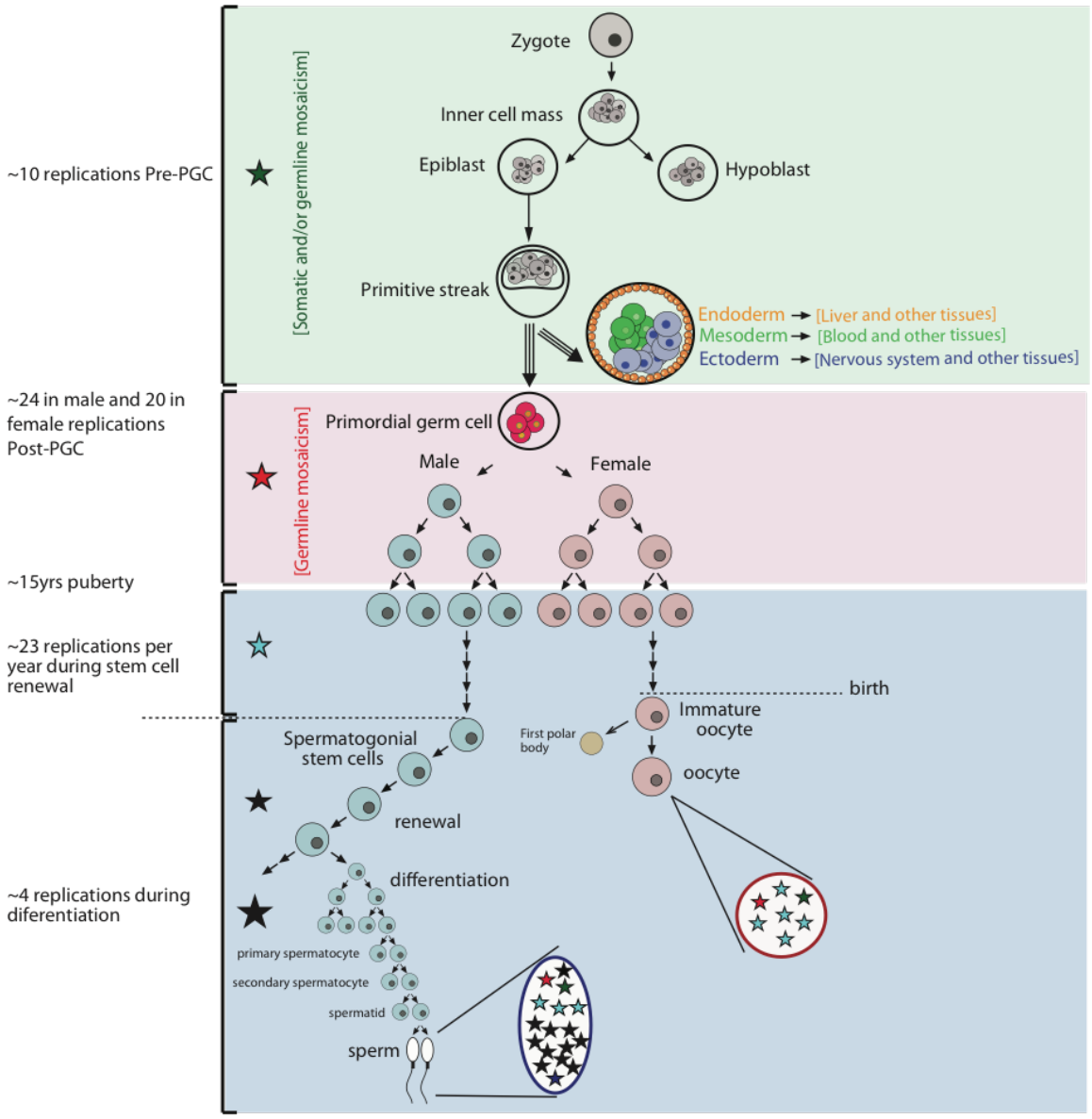

Fig. 1.2 Embryogenesis and gametogenesis. Figure sourced from Rahbari *et al* [173]

The differences in parental age effects between mothers and fathers prompts the question of what differences there may be in the mechanisms generating these mutations. Differences in the mutational spectra between maternal and paternal mutations can help inform what these could be. Subtle, but significant, differences have been observed between the overall mutational spectra[62]. Some of these differences become more pronounced with increased parental age [98, 54, 62]. The types of *de novo* maternal mutations change with the mother's age. Specifically a significant decrease in CpG>TpG mutations and an increase in C>G mutations has been observed. It has been hypothesised that this increase in C>G maternal

mutations, which does not occur in paternal mutations, is associated with double-strand breaks in aging oocytes[54, 62]. As mentioned previously, the paternal and maternal mutation rates appear to be different around crossover sites. This is likely due to the fact that the location of recombination hotspots differs between males and females and that the recombination rate differs at shared sites[15]. Beyond this, the mutational spectra of the DNMs around these sites are also significantly different [73]. It has been suggested that this may be due to differences in the sex-specific timing of meiosis in the germline development but this is still unclear. As sample sizes increase in trio-based studies a larger collection of DNMs will be crucial in examining these parental differences in more detail.

The timing of mutations is an important source of individual variation. Mutations in the germline can occur at any stage of development from zygote to gamete. When mutations occur prior to the specification of primordial germ cells (PGCs) they can result in mosaicism across somatic tissues. At least 3-4% of *de novo* mutations found in offspring have been found to be mosaic in parental somatic tissues (Figure1.2) [173, 189]. Mutations that occur around the time of PGC specification may result in germline mosaicism. These can result in an increased probability of siblings sharing the same *de novo* mutation. Mutations that occur early in the first few cell divisions in embryogenesis in the offspring can result in post-zygotic mutations (Figure1.2). The mutant allele proportion for such mutations should be lower than 50% and may be detected with incomplete sensitivity in sequencing data[99]. The contribution of mutations at each stage to the overall mutation rate is not well understood but is important in understanding mutational processes and understanding variation within families. This can be critical for estimating the risk of recurrence for families with diseases caused by damaging *de novo* mutations.

Other sources of variation in individual mutation rate could be due to genetic differences influencing mutation rate or spectra (e.g. in DNA repair pathways) or to DNA damage caused by differential exposure to mutagens. For example tobacco smoke has been shown to contribute to paternal germline mutations in mice, although a similar effect has yet to be observed in humans[242].

1.4.3 Population level variation

Mutation rates have evolved over time just as any other phenotype. This is most clearly demonstrated by the variation in mutation rates between species [152, 135]. The mutation rate in mice has been observed to be higher than in humans and differences have been seen in mutation spectra and mutation rates per cell division[35, 30, 130]. The mutation rate per spermatogonial stem cell (SSC) division is estimated to be lower in humans in mice, this is hypothesised to have evolved as a result of the larger contribution of SSC to the human

germline[130]. Since the split of hominoids and monkeys the per year mutation rates have decreased in hominoids, known as the hominoids slow down. It has been observed that evolutionary rates are faster in new world monkeys compared with old world monkey and in turn rates in old world monkeys are faster than in humans and apes[152, 63, 206]. On a shorter evolutionary time scale, mutation rates have also been suggested to differ between human populations. The mutational signature TCC->TTC is enriched in rare variation within European populations[75]. There have also been signatures shown to be private to certain Native American populations [142]. The reasons behind these differences is still elusive. It has been suggested that heterozygosity affects mutation rate and that this has impacted differences in mutational spectra across populations [240, 9]. Characterising population specific mutation rates in more detail will help to elucidate different selective pressures on the mutation rate and possible differences in underlying mechanisms.

1.5 *De novo* mutations in human disease

1.5.1 Modes of inheritance

De novo mutations are a significant cause of rare genetic disease. To understand their contribution it is important to first contextualise the possible modes of inheritance for genetic disorders in general. Diseases are referred to as 'Mendelian' when disease-causing alleles segregate according to Mendel's laws of inheritance. These disorders tend to be monogenic and are caused by rare and highly penetrant mutations. Autosomal recessive disorders only occur when an individual has two mutant alleles in the disease-associated gene. This usually occurs when an individual inherits a mutant allele from each parent therefore these disorders are unlikely to be caused by *de novo* mutations. Autosomal dominant disorders occur in individuals with only a single mutant allele in a disease-associated gene. This mutant allele can be inherited from a parent, who is likely affected, or it can arise *de novo* in the individual. X-linked disorders have slightly different inheritance patterns compared to those on the autosome. X-linked recessive disorders can occur in females when they inherit one disease allele from their father and one from their mother. For males with an X-linked recessive disorder, since they only have a single copy of the X chromosome, they always inherit this mutant allele from their mother. This means that males are much more likely than females to be affected by X-linked disorders as they only need a single mutant allele. X-linked dominant disorders are much rarer but there are a few examples. For example, Rett syndrome is a developmental disorder caused by dominant mutations in *MECP2*. This disorder almost

exclusively affects females as it appears to be embryonic lethal in males. This disorder is almost always caused by a *de novo* mutation.

1.5.2 Historical context

Research on the contribution of genetic variation to human disease has historically been focused on inherited variation. Linkage analysis was one of the first approaches used to associate regions of the genome to human disease and was fundamental in identifying Mendelian disease genes[17]. This was done by testing if a series of marker alleles co-segregated with disease status within a family or across multiple families. The emergence of microarray technology and the completion of the reference genome, through the Human Genome Project, led to the development of genome-wide association studies (GWAS). GWAS were able to test associations between allele frequencies of hundreds of thousands of SNVs with human disease across the genome. This became crucial in the progress of complex disease genetics.

Chromosomal aneuploidies were one of the the first types of observable *de novo* genetic variation that were shown to cause a disorder. The trisomy of chromosome 21, which causes Down's syndrome, was first observed through a microscope in 1959 [124]. By the 1990s a few *de novo* SNVs and large (>1MB) CNVs were shown to cause sporadic disease, however the ability to systematically study the role of smaller scale *de novo* mutations in human disease has only been possible since the development of genomic microarrays and next-generation sequencing technologies[26, 245, 134]. Early studies using Array Comparative Genomic Hybridisation (CGH) showed that *de novo* CNVs were significantly enriched in individuals with autism, epilepsy and developmental delay [193, 40]. Moreover, specific genes were associated with sporadic disease [225, 221]. The emergence of exome and whole genome sequencing has been paramount in allowing for detection of *de novo* SNVs and indels in patients with rare genetic disease. This has led to the identification of hundreds of genes associated with rare sporadic disease [1, 18, 59]. Studies have also established that *de novo* mutations are implicated in common neurodevelopmental disorders such as autism, epilepsy and intellectual disability[89, 163, 6, 40].

1.5.3 Developmental disorders

Developmental disorders (DD) encompass neurodevelopmental disorders, congenital anomalies, abnormal growth parameters and unusual behavioural phenotypes [50]. Although these disorders are individually rare, collectively they affect 2-5% of births in the UK [198, 183, 41]. DNMs are enriched in cohorts with DD and it has been estimated that ~50% of severe DD

cases are caused by a pathogenic coding DNM [140, 41]. DNMs in specific non-coding regions, such as conserved regulatory elements, are expected to explain 0.5-2% of severe DD [201]. Pathogenic DNMs in DD mostly lead to dominant genetic disorders although they can also cause recessive disorders via compound heterozygosity in conjunction with an inherited variant. There are also examples where a patient's phenotypes are explained by more than one pathogenic DNM [241, 237].

The timing of DNMs has an important impact on both the presentation and clinical impact of DDs. Post-zygotic mutations are difficult to detect but technological advances have allowed for more study into this type of mutation [20]. Post-zygotic pathogenic DNMs can result in somatic mosaicism which can lead to a less severe phenotype compared to that caused by a constitutive mutation [170, 1, 161]. There are also examples of disorders where the DNMs only appear as mosaic such as those caused by mutations in *AKT1* or *PIK3CA* [229, 113]. This may be because the mutation is lethal when constitutive [74]. Mosaic disorders also tend to be caused by activating missense mutations which result in cell proliferation and overgrowth. The clinical impact of the timing of mutation is especially important. It has been estimated that ~3% of pathogenic DNMs in DD are post-zygotic in the affected child [235]. This means that the sibling recurrence risk is very low which can be crucial information for future pregnancies. In addition, evidence of a post-zygotic mutation in a parent, which can be detected in <1% of cases, results in a much larger sibling recurrence risk [235, 97].

There are currently ~2000 genes known to be associated with developmental disorders[236]. Initially these genes were discovered using a phenotype-driven approach by aggregating patients with similar clinical presentations. However we know that the clinical manifestations of these disorders can vary widely and so in more recent years gene discovery has been complemented with a genotype-driven approach. This has involved the ascertainment of large cohorts with a diversity of related phenotypes. Novel associations between genes and DDs have then been identified by looking for genes where we observe a significantly greater number of non synonymous DNMs than expected under a null mutational model [41, 215]. Analysing DNM burden has also discovered novel gene associations for a range of different disorders aside from DDs such as autism, schizophrenia and congenital heart disease [202, 69, 157, 205].

The Deciphering Developmental Disorders (DDD) Study is an example of a large cohort that has been important in furthering our understanding of the genetic architecture of DD. The DDD study consists of 13,451 patients with a severe, but genetically undiagnosed, developmental disorder. These patients have been recruited from 24 regional genetic services across the UK and Republic of Ireland. There is a wide array of phenotypes in the dataset

however 87% have intellectual disability (Figure1.3) [215]. The patients have been systematically phenotyped, half have had array CGH and all have been been exome sequenced. For approximately 88% of these patients in the DDD the parents have also been exome sequenced. This allows for interrogation of *de novo* variation. In 2017, the DDD published results on analysis of *de novo* variants in 4,293 patients and identified 94 significant genes associated with DD, 14 of which were not previously known[41]. The exome-sequencing data generated as part of the DDD study are used in all three results chapters in this thesis.

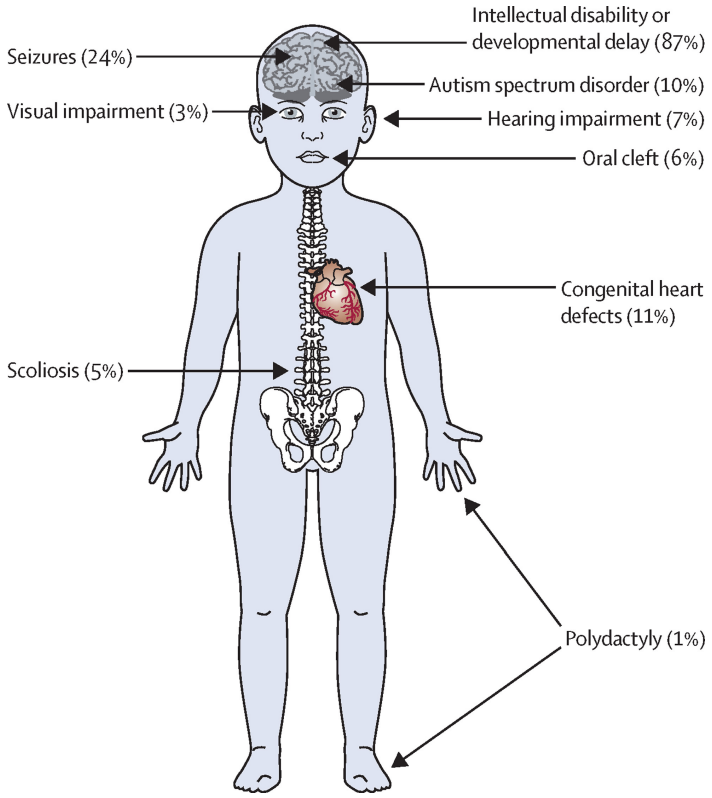

Fig. 1.3 Phenotypes in the DDD study. Figure sourced from Wright *et al*[236]

The 100,000 Genomes Project is a more recent initiative in the UK to whole-genome sequence patients with rare diseases and cancer. Genomics England Ltd was a company set up by the Department of Health and Social Care to deliver the project. As of December 2018, the milestone of 100,000 genomes was reached and Genomics England have expanded their goal to sequence 5 million genomes in the next five years. The rare disease arm of the 100,000 Genomes Project currently consists of sequencing data from ~72,000 participants which reflects ~34,000 families with a variety of family structures, as described in Table 1.1 (as of April 2020). The project includes a wide array of more than 190 rare diseases, the breakdown of the types of these diseases are described in Table 1.2. Approximately 40% of these patients are described as having a neurodevelopmental disorder. Data from the project

Rare Disease Family type	Count
Duo with Mother or Father	4,797
Duo with other Biological relative	1,286
Families with more than three participants	1,856
Singleton	12,607
Trio with Mother and Father	11,854
Trio with Mother or Father and other Biological Relationship	829
Trio with other Biological Relatives	384
TOTAL	33,613

Table 1.1 Distribution of family types within the rare disease arm of the 100,000 Genomes Project

Normalised Disease Group	Count	Proportion
Cardiovascular disorders	3,799	0.111
Ciliopathies	340	0.010
Dermatological disorders	389	0.011
Dysmorphic and congenital abnormality syndromes	591	0.017
Endocrine disorders	849	0.025
Fastroenterological disorders	118	0.003
Frowth disorders	186	0.005
Haematological and immunological disorders	913	0.027
Hearing and ear disorders	801	0.023
Infectious diseases	13	0.000
Metabolic disorders	694	0.020
Neurological and neurodevelopmental disorders	14,095	0.411
Opthalmological disorders	2,866	0.084
Psychiatric disorders	83	0.002
Renal and urinary tract disorders	3,576	0.104
Respiratory disorders	322	0.009
Rheumatological disorders	258	0.008
Skeletal disorders	833	0.024
Tumour syndromes	1,682	0.049
Ultra-rare disorders	1,847	0.054
TOTAL	34,255	

Table 1.2 Distribution of disease types in the rare disease arm of the 100,000 Genomes Project

were made available to researchers in 2019. The benefit of whole-genome sequencing over exome sequencing allows for more interrogation of the non-coding regions of the genome. From a mutational perspective it can also allow us to call *de novo* mutations across the whole genome which gives us power to look more closely at variation in mutational spectra and rates across individuals. Analyses from chapter 3 of the thesis are largely conducted using this dataset.

1.6 Outline of dissertation

The goal of this thesis was to understand sources of variation in germline mutation and the contribution of these mutations to rare developmental disorders. These sources of variation encompassed types of mutations that have been previously underrepresented in genetic research as well as individual mutation rates and spectra across individuals and parental origin.

Chapter 2 of this dissertation focuses on the mutational origins and pathogenic impact of MNVs. Here I describe how MNVs in protein-coding sequences can be more pathogenic than an SNV even when the MNV falls within a single codon. I also estimate the MNV mutation rate, explore the mutational spectra of these variant and describe the contribution of *de novo* MNVs to severe developmental disorders. This analysis was conducted on trio exome data from the DDD study.

Chapter 3 of this dissertation focuses on identifying and characterising germline hypermutators. Using whole-genome sequencing data from the DDD and GEL dataset, I identified fifteen children with an unusually large number of *de novo* mutations. Eight of these appear to be due to a paternal hypermutator. I describe analyses to try and identify a genetic cause for this hypermutation of which I found two putative paternal variants in DNA repair genes. I will also describe work focussed on whether variants in DNA repair genes impact germline mutation rates by examining a well characterised cancer somatic mutator gene and then using a broader approach across all DNA repair genes. Using the large resource of DNMs called in the GEL dataset, I also explore other sources of variation in germline mutation rate, including differences between maternal and paternal DNMs as well as the effect of parental age.

In **Chapter 4**, I describe analyses that focussed on using *de novo* mutations to identify novel genes associated with DD. This was performed with exome parent-offspring data from the DDD study pooled with trios from GeneDx, a US-based genetic diagnostic company, and trios from Radboud University Medical Center (RUMC). This chapter also describes work done to explore how these novel genes differ from those that were previously known, as well as a model-based approach to explore the likely properties of currently undiscovered genes.

Lastly, in **Chapter 5**, I summarise the main findings of these projects and what can be learnt from them. This then leads on to a discussion of how this work can be extended and what developments can advance our knowledge of human germline mutation in upcoming years.