1 INTRODUCTION

Developmental disorders (DDs) include congenital anomalies, neurodevelopmental disorders, and abnormalities in growth and behaviour (Figure 1-1)[1]. DD can be relatively mild, presenting, for example, as an isolated learning disability, or severe. Severe DD is generally characterised as one many rare, often neurodevelopmental diseases, usually appearing within the first few years of life[2], and is the focus of this dissertation.

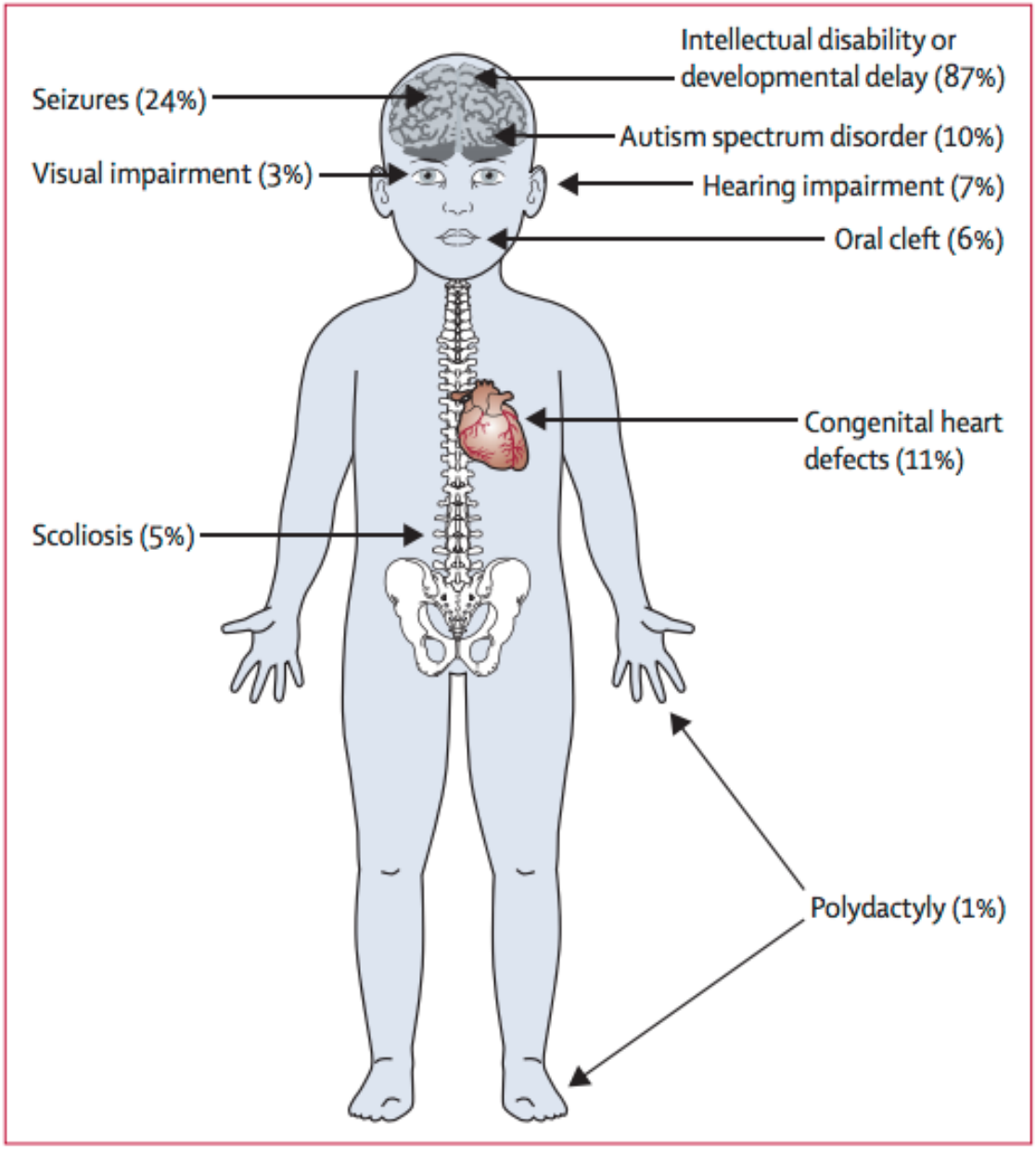

Figure 1-1 Prevalence distribution of phenotypes observed in a recent large study of severe DD[3].

Understanding the aetiology of DD in a child is crucial for management, prognosis, and family planning. In the absence of an identified environmental insult (e.g. teratogens, gestational problems, or child neglect), and especially in the presence of specific syndromic or familial features, the presumed cause is genetic. This dissertation addresses the detection and implication of uniparental disomy and mosaic forms of large-scale variation in DD genetics. To frame the context of this work this introduction describes the detection and implication of large-scale variation in DD, and the new methods I developed to enable detection of large-scale variation from DNA sequencing-based assays.

Several recent advances are improving the diagnostic yield of genetic testing: the increasing availability of exome sequence analysis as a assay platform in clinical diagnostic testing[4], the application of proband-parent trio studies for the detection of autosomal dominant *de novo* and compound heterozygous mutations[5], and the development and implementation of new algorithmic approaches. The Deciphering Developmental Disorders (DDD) study[6], exemplifies this paradigm; it is a large trio-based study of children with undiagnosed DDs that studies the genetic architecture of rare disease using primarily exome-sequencing data, with the implementation of existing and development of new algorithmic approaches.

Despite recent progress in delineating the genetic causes of DD, the detection of mutations that are definitively explanatory of the disorder (i.e. 'causative') is possible in fewer than half of children investigated postnatally for DD[7]. Identifying the underlying genetic basis of DD is challenging for many reasons, such as 1) extensive genetic heterogeneity, as over 1,000 genes are associated with DD[8], and a substantial fraction of children with DD have one of thousands of rare monogenic diseases[9]; 2) the functional role for most genes in the genome is still not known[10]; and 3) clinical diagnostic testing in the UK is usually limited to the detection of non-mosaic ('constitutive') chromosomal abnormalities and mutations in specific genes of interest[11], despite many additional classes of genomic variation also implicated in DD[8,12].

Due to the many different mechanisms by which mutations are generated and detected, it can be useful to stratify how genomes vary between individuals by three criteria: 1) size of the genetic variant, from small-scale (point and insertions and deletions (indels)) variation to large-scale (structural) variation[13]; 2) copy number: distinguishing balanced (copy neutral; loss of heterozygosity (LOH), uniparental disomy[14], translocation, inversion)[15] and unbalanced (copy number; deletion or duplication)[15]; and 3) clonality, in which assayed cells exhibit genetic homogeneity (constitutive variation) or heterogeneity (mosaicism or chimerism)[16]. Decades of genetic analyses have yielded insights into the diversity of mutations underlying DD, implicating all combinations of constitutive and mosaic small-scale and large-scale abnormalities.

This dissertation will address the detection and impact of large-scale variation and mosaicism on children with DD.

1.1 Strategies for detecting structural variation

The historical timeline of detecting large-scale variation in the genome can be classified into the following technological eras: optical cytogenetics, molecular cytogenetics, and next-generation sequencing.

1.1.1 Optical cytogenetics

Cytogenetics is the study of chromosome structure and function, and was originally performed optically, using light microscopy. In the first half of the 20th century, visualisation of the chromosomes was unreliable and the human chromosome number was thought to be 48, a belief sustained for nearly 40 years[17]. A cascade of discoveries in the mid 20th century revolutionised cytogenetics: the discovery of the Barr body in the interphase nuclei in females[18], enabling cytological determination of sex[18]; the discovery of hypotonic solution for cell preparation[19], allowing the separation of the chromosomes; advances in culture medium[20], permitting cell survival for analysis; and the use of colchicine in condensing metaphase chromosomes, permitting karyotyping (Figure 1-2)[21]. As a result of these advances, the chromosome number was corrected to 46 and numerical differences between chromosomes could be discriminated.

Figure 1-2 The first human karyotype, adapted from Levan *et al.*[21].

The development of chromosome banding techniques, in which segments of euchromatin and heterochromatin are differentially stained, facilitated the delineation the chromosomes and enabled the identification of sub-chromosomal "structural" changes to the chromosomes (Figure 1-2). The most common chromosomal banding technique, G-banding, uses Giemsa staining (methylene blue, eosin, and Azure), originally used for microbial staining in 1904[22], to stain approximately 128 bands[23] per genome, an average of one band per ~24 Mb. High resolution G-banding was invented by Yunis *et al*. in 1978[24] and enabled the detection one band per ~5-10 Mb, which remains today the typical resolution for optical genetics. Thus, optical cytogenetics can be used to identify structural changes to chromosomes that are at least 5-10 Mb in size and is used for clinical diagnostic testing in many centres. Additionally, cytogenetics can be used to detect large inversions and translocations, but copy neutral LOH is not visible.

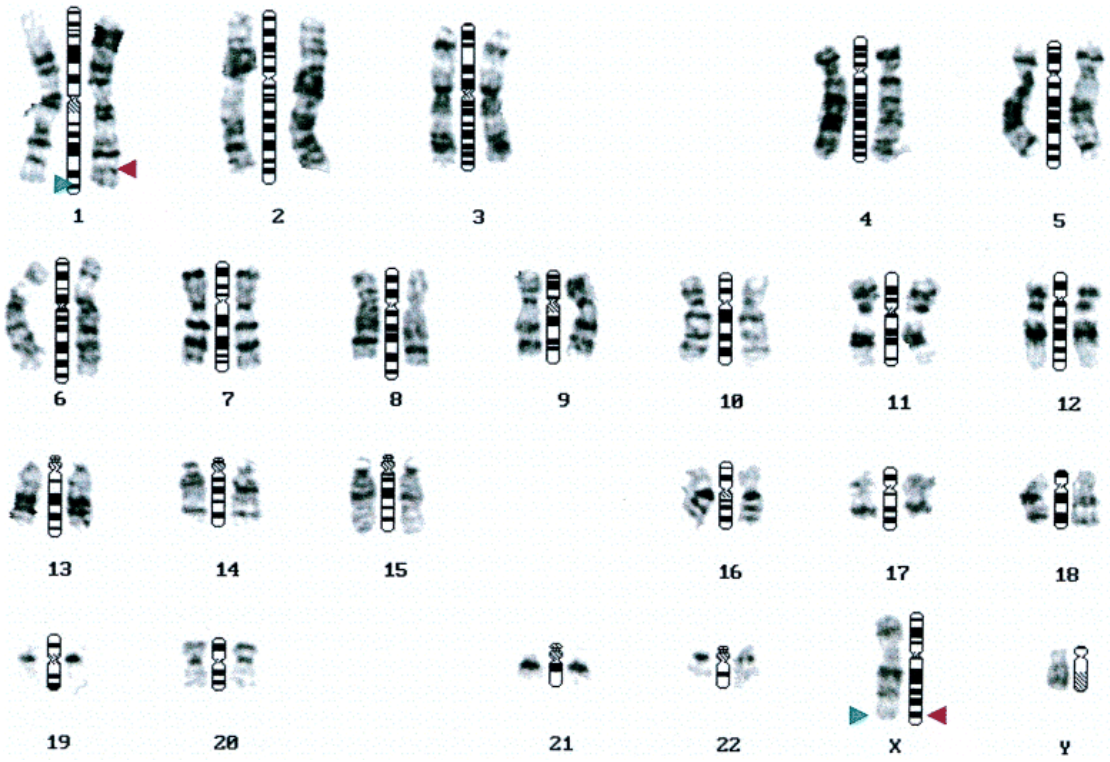

Figure 1-3 Banded karyotypes. In this case, a translocation of material between chromosomes 1 and X, adapted from Mattei *et al.*[25].

Karyotyping through optical genetics can detect mosaic structural abnormalities by identifying a proportion of cells from the same individual with a distinct structural complement. However, this process is labour-intensive because, for example, 14 cells must be examined per individual to exclude 10% mosaicism with 95% confidence[26].

Optical genetics established numerical and structural variation as important genetic components of DD. Nevertheless, banded karyotyping has several limitations: assay resolution is coarse at 5-10 Mb; results require subjective interpretation[27]; the preparation of chromosome banding and of multiple cells per sample to assay mosaicism is labour-intensive; cell culture is required and requires one week of preparation time, which delays diagnosis and is not always successful (in the case of macerated foetal tissue, for example); and, lastly, it is blind to copy neutral loss of heterozygosity. Many of these limitations would be overcome in the molecular cytogenetics era.

1.1.2 Molecular cytogenetics

Molecular cytogenetics is characterised by the adhesion (hybridisation) of DNA molecules ('probes') to a DNA sample using complementary base pairing. Probes can be constructed to hybridise to a specific region of interest. Resolving power is related to the size of the probes, which has substantially decreased with time, initially from hundreds of kb (yeast & bacterial artificial chromosomes), to tens of kb (fosmid probes), to hundreds of base pairs (synthesised oligonucleotides)[28].

The first implementation of molecular cytogenetics was the extension of karyotyping with DNA hybridisation. This technology, *in situ* hybridisation (ISH), originally used probes with radioactive labels[29] but fluorescent labels (FISH)[30] are now mainstream. FISH offers improved resolution compared to karyotyping and interphase FISH can be performed without cultured cells. Metaphase FISH enables simultaneous visualisation of a structural abnormality and the chromosomes, but is culture-dependent. Interphase and metaphase FISH are still used today to detect unbalanced abnormalities, whilst metaphase FISH is used to examine suspected translocations. FISH is used in this dissertation to validate structural abnormalities detected by orthogonal methods.

The second implementation of molecular cytogenetics is hybridisation to microarrays. This involves a set of imaging techniques that, instead of visualising the chromosomes themselves, quantitate the intensity and frequency of light emitted by fluorescent probes hybridised to a DNA sample. Microarray cytogenetics has several advantages compared to karyotyping in that cell culture is not required, mosaicism is more easily identified because thousands of cells are assayed simultaneously, and quantitative data can be statistically analysed and objectively interpreted. DNA probes

can be designed to target loci throughout the genome, thus providing a high-throughput genome-wide molecular assay.

There are two formats of microarray commonly used today: 1) comparative genomic hybridisation (CGH), invented in the early 1990s[31] for copy number analysis of tumours, which gave rise to modern array-based CGH (aCGH)[32]; and 2) single nucleotide polymorphism (SNP) microarray, also known as genotyping microarray[33], designed as a high throughput assay of single nucleotide polymorphism but in recent years has also been used for the detection of large-scale abnormalities[34].

There are advantages and disadvantages for both types of microarray in the detection of large-scale abnormalities. Traditionally, aCGH has been preferred in diagnostic labs for more sensitive CNV detection performance and design flexibility. However, SNP microarray additionally enables detection of runs of homozygosity (useful for finding loss of heterozygosity and consanguinity), and is more sensitive for mosaicism. SNP microarray has been increasingly used for diagnostic testing[35,36] and recently, integrated microarray array chips combining both aCGH and SNP probes have been created to combine the benefits of both technologies[37]. Many of the analyses presented in this dissertation used SNP microarray as a detection platform.

SNP microarray methodology uses fluorescent tags (red and green) to label each allele, and an imaging system is used to detect the colour and signal intensity. The ratio of red to green light colour frequency reflects the sample's allele frequency. The fraction of the less-common allele, the b allele frequency (BAF), is an important metric used for genotyping and mosaicism detection. The light intensity, 'r value', is compared to the light intensity seen for this SNP from a pool of reference samples, and is recorded as a log r ratio (LRR)[38] (Figure 1-4).

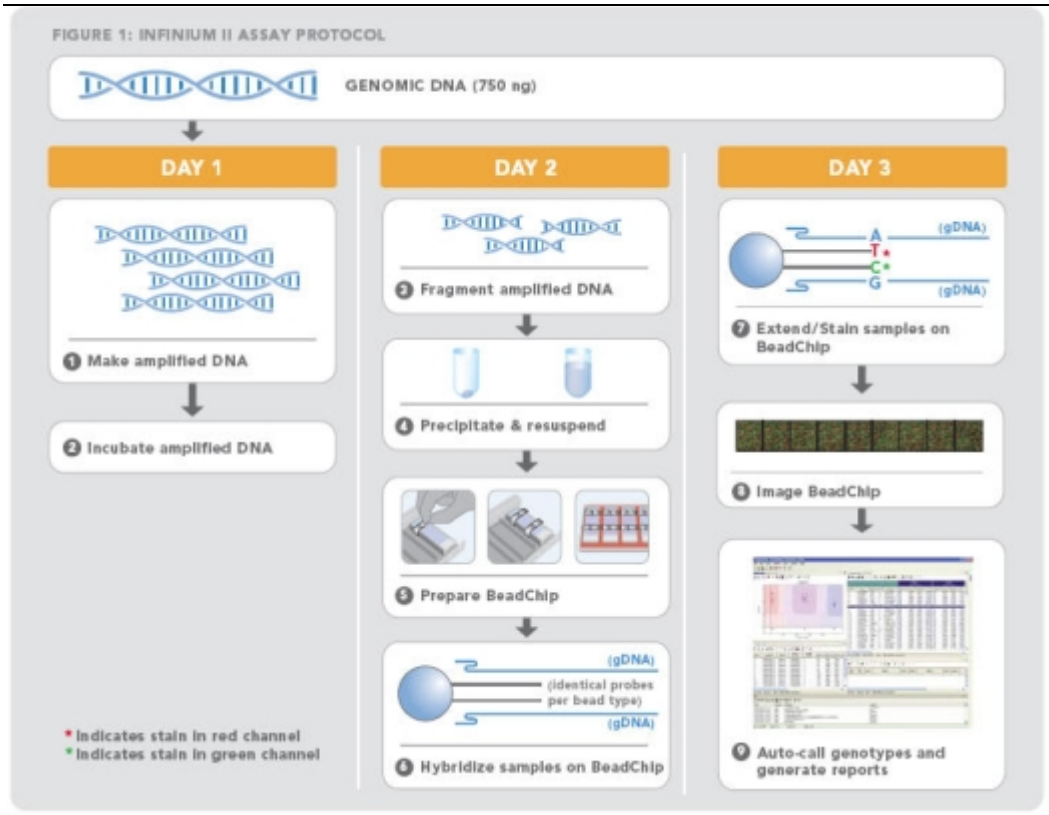

Figure 1-4 Illumina BeadArray technology, adapted from Illumina documentation[39].

Copy number data from aCGH is also measured using probe light intensity but in aCGH, the light intensity from both test and reference are measured in the experiment and they are compared using the $\log_2$ ratio. In aCGH, the $\log_2$ data provide signal for detection of copy number while in SNP data, both BAF and LRR probe metrics can be used for analysis. The detection of structural abnormalities can be cast as a segmentation problem with abnormalities as unusual segments in an otherwise normal chromosome. Several statistical methods can be used for detecting copy number analysis. While wavelets[40], penalised-least squares[41], and piecewise-constant vectors[42], primarily identify segments different from the norm (reject the null hypothesis of no difference from their surroundings), other methods, such as Bayesian methods[42,43], and hidden Markov models[44,45] directly assess the null hypothesis and a strong expectation of an alternate (constitutive) hypothesis. In genome alteration detection analysis (GADA)[42] segmentation is performed in three steps: genomic segments are represented in computationally-efficient piecewise constant vectors, then sparse Bayesian learning finds the most likely location of the breakpoints given a prior estimate of the number of segments, and lastly a backward elimination procedure adjusts the number of segments based upon a statistical threshold. Because of the speed and accuracy of GADA it has

become one of the most popular packages for the detection of copy number from aCGH data.

SNP microarray can additionally be used as a genome-wide screen for constitutive copy-neutral LOH. The first use of SNP data in this manner was for the detection of isodisomy in cancer research[46]. An important type of LOH in children is called uniparental disomy (UPD) and is canonicaly due to the inheritance of a chromosome in which both homologues originate from the same parent. The appreciation of UPD as a disease mechanism in children spurred the implementation of SNP microarray for clinical diagnostic testing of UPD[47]. In chapter 2 I describe the software tools available for detecting constitutive UPD and how their limitations motivated my development of a new UPD-detection algorithm.

Techniques differ in the use of SNP data for the detection of constitutive and mosaic abnormalities. In non-mosaic tissue, an allele is present in exactly 0, 1, or 2 discrete copies (on the autosomes), which can be precisely recorded using one of three genotype categories (AA, AB, BB). In contrast, mosaicism represents a locus with a genetically heterogeneous cell population. BAF, as a quantitative measure, is an inherently more sensitive measure compared to genotype to denote the relative contribution of the underlying allele mixture. Therefore, whilst constitutive abnormalities may be identified using alteration of genotype, mosaic methods require more sensitive methods and frequently employ deviation in BAF, as described further below.

Compared to the detection of constitutive large-scale variation, fewer software tools exist for *mosaic* copy number and UPD from SNP data. Illumina states that its proprietary algorithm, cnvPartition, can detect mosaic copy number variation in tumour samples[48], but does not specify how it does this. The open-source tool MAD[49] identifies mosaic copy number and UPD by segmenting deviations in BAFs from SNP data with GADA segmentation (Figure 1-5). The MAD algorithm was recently chosen for the study of 50,000 samples with SNP chip data[50]. When SNP data are available from trios, a different software tool, triPOD[51] can leverage haplotype structure and BAF deviation to identify strings of inheritance imbalance from the same parent, thereby increasing the sensitivity and specificity of mosaicism calls.

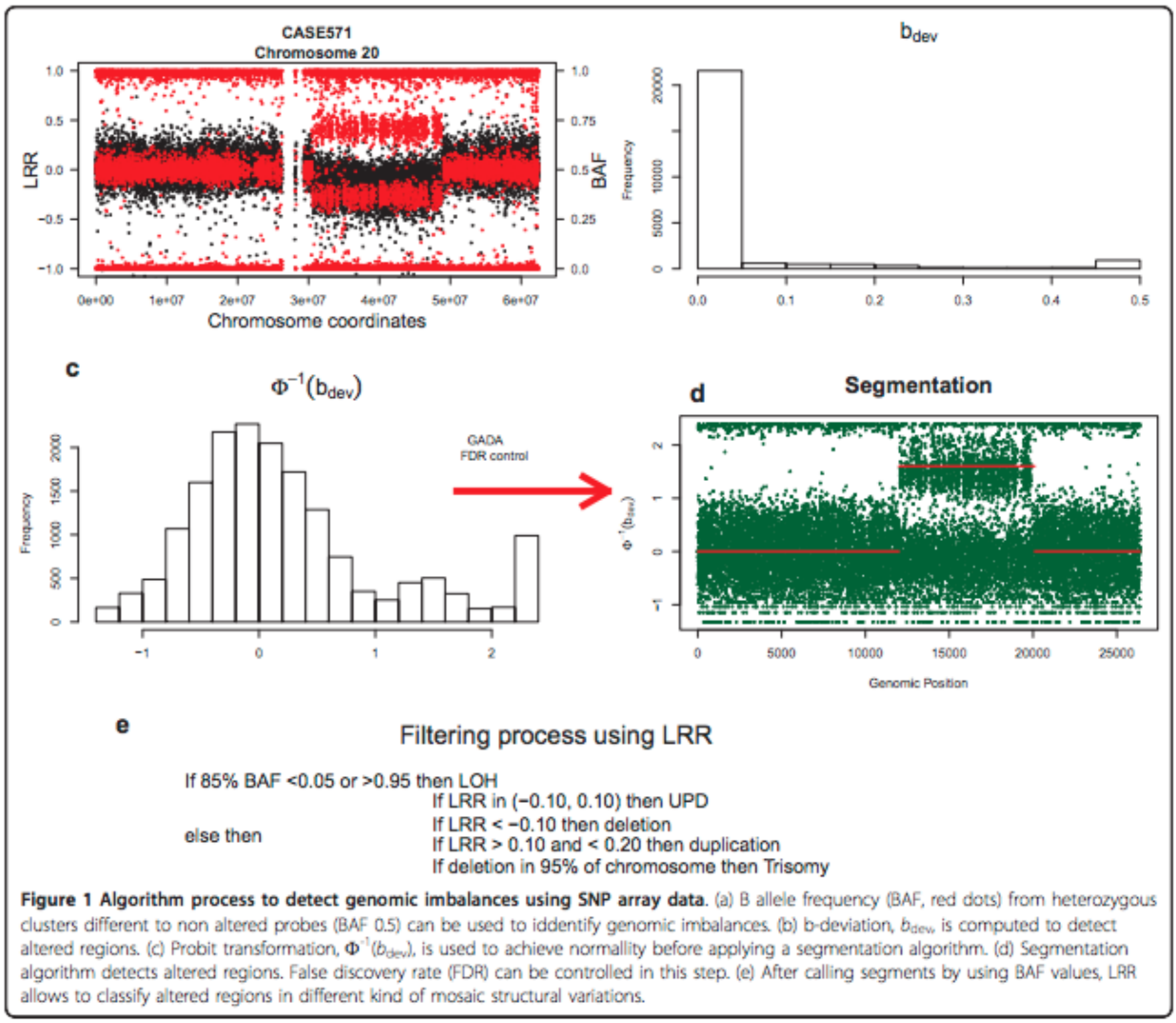

Figure 1 Algorithm process to detect genomic imbalances using SNP array data. (a) B allele frequency (BAF, red dots) from heterozygous clusters different to non altered probes (BAF 0.5) can be used to iddentify genomic imbalances. (b) b-deviation, b_{dev}, is computed to detect altered regions. (c) Probit transformation, $\Phi^{-1}(b_{dev})$, is used to achieve normallity before applying a segmentation algorithm. (d) Segmentation algorithm detects altered regions. False discovery rate (FDR) can be controlled in this step. (e) After calling segments by using BAF values, LRR allows to classify altered regions in different kind of mosaic structural variations.

Figure 1-5 Illustration of the MAD method, adapted from Gonzalez *et al.* [49] Note that MAD begins by calculating the deviation in BAF from genotype-expected BAF (B_{dev}).

In chapter 3, I will demonstrate a comparative analysis of MAD and triPOD of mosaic copy number and copy neutral genomic variations in children with and without DD.

1.1.3 DNA sequencing

DNA sequencing is the process of determining the identity and order of DNA nucleotides in a DNA molecule. The early sequencing technique used radioactively-labelled[52], later fluorescently-labelled[53] nucleotides, incorporated into a DNA molecule. The DNA molecules were size-separated (typically by capillary electrophoresis) and the labelled bases were imaged. This process, called capillary or Sanger sequencing, has

been widely used and is still used as an inexpensive approach to assay targeted genetic variation.

Sanger sequencing can identify the DNA sequence of up to approximately 1,000 bases from a single DNA molecule[54]. Next-generation (2nd generation) sequencing approaches entail sequencing numerous, typically shorter, DNA molecules in parallel to increase throughput. This has allowed for assessment of the 'mappable' genome, which is the accessible, non-repetitive, well-characterised regions of genomes[55]. Third generation sequencing[56] involves the massively parallel sequencing using long 'single-molecule' sequence reads. These technologies are in development, and potentially offer benefits for the study of genomes where the reference is unknown or very repetitive[56] but are not routinely used for rare disease studies in humans and are not considered further here.

The second-generation platform used for the analyses described below is that of Illumina®, mainly the HiSeq™ 2000 and HiSeq™ 2500 sequencing machines[55]. The Illumina sequencing approach begins with fragmentation of DNA and selection of fragments approximately 500 bp long. The sequencing procedure uses a glass substrate ('flow-cell'[55]) with adhered oligonucleotides that bind fragments of DNA. Bound fragments undergo an amplification step (bridge amplification) that generates many clones of fragments. Fragments are denatured so they are single stranded and imaging techniques capture growing strands and record strings of bases, known as "reads". Each sequence read contains bases from a location in the genome.

The Human Genome Project was an international collaboration that used first-generation capillary and 'shot-gun' sequencing of large-insert clones to determine the sequence of DNA bases of the chromosomes of *Homo sapiens*. Subsequent 'resequencing' of the genome uses the reference sequence determined by the HGP as a haploid scaffold, upon which short (~100bp) DNA sequence reads from next generation sequencing can be aligned ('mapped') to the reference, commonly performed using the Burrows-Wheeler Algorithm[57]. Sufficient sequencing coverage of the genome is essential to assess both chromosome homologues, to account for allelic sampling, errors in sequencing, and to produce accurate genotypes. A widely used genotyping approach, SAMtools, makes a prediction of the genotype based on which genotype is most likely given the bases and qualities of aligned reads[58]. The proportion of reads supporting each allele is a measurement of allele fraction, analogous to the theta value calculated from

SNP microarrays. Sequencing coverage at a given position is referred to as 'read depth' and is an analogous measure of the r value.

Whilst the cost of next-generation sequencing has declined precipitously[59], it is still too expensive to sequence a whole-genome to high depth for most applications. The Human Genome Project observed that much of the genome appears to be repetitive, low-complexity sequence, and that only approximately 1-2% includes protein-coding (exon) sequence[60]. Therefore, in order to maximize the yield from limited sequencing resources, it has been a common strategy to restrict sequencing to all the known protein-encoding exons (the 'exome') of the genome. Exome sequencing entails enrichment for DNA molecules overlapping the (approximately 180,000) exons of the genome, followed by sequencing of this enriched library of molecules. In 2009, the first exome paper demonstrating the clinical utility of exome sequencing was published, and correctly identified the known genetic cause of a rare autosomal dominant disorder, Freeman-Sheldon syndrome[60]. Since then, genetic causes of many rare diseases have been discovered using exome sequencing[61].

Initially, exome analysis focused on the detection of smaller genetic variation but various efforts have been used recently to harness sequence reads to detect copy number variants. Estimating copy number from exome data can be challenging, as sequence read depth is sparsely clustered and non-evenly distributed across the genome, and because measured read depth is a biased estimate of the underlying sample copy number[62] (since enrichment efficiency, sequencing efficiency and mapping efficiency vary considerably among targeted regions). Nevertheless, several approaches have been developed to calculate copy number from read depth by accounting for these biases. One approach is to consider these biases as covariates, and another is to normalise coverage to an empirical distribution of expected coverage based upon a pool of samples. Accordingly several software tools are available to detect copy number using read-depth coverage[63-66]. Additionally, other approaches have been used, including paired-end approaches[67,68], and split-reads[69,70]. The DDD study has used Convex[6]; this software tool normalises sequence coverage in a proband exome based upon a pool of exomes and in addition accounts for biases in the enrichment capture (melting temperature, GC content, and delta free energy of hybridisation). These tools are not optimised for detecting mosaic copy-number variation as mosaicism leads to an intermediate deviation in log2r, which is difficult to distinguish from stochastic

sampling variation. Incidentally, compared to Bayesian and HMM approaches, which model discrete copy number states, Convex segmentation, based on the Smith-Waterman algorithm[71] may be less prone to problems with mosaicism.

Recent progress in detecting mosaic copy-number from sequence data has come from efforts to detect foetal aneuploidy prenatally using circulating placental foetal DNA by whole genome sequencing of maternal plasma-derived DNA. At one trimester of gestational age, approximately 10% of circulating cell-free DNA in maternal plasma is of foetal origin[72]. The detection of foetal aneuploidy from maternal plasma sequencing has been based on 'relative chromosome dosage', the concept that foetal trisomy will result in a statistically significant increase of sequence reads[73,74]. A recent theoretical framework to identify sub-chromosomal foetal *de novo* CNVs from maternal plasma uses whole genome sequencing to recover parental haplotypes, then combines information from parent-specific allele imbalance and depth of coverage as metrics of detection[75]. Whilst this introduces a framework for the detection of mosaic CNVs, the generation of whole genome sequence data is still expensive for practical widespread clinical application and this method requires the availability of paternal DNA.

The lack of an exome-based approach to detect mosaic copy-number is a major limitation given the popularity of exome-based analyses in rare-disease genetics. In addition, copy-neutral structural variation does not result in changes to read depth and cannot be detected this way. These limitations motivated the development of a sequencing-based mosaic structural variation tool capable of detecting mosaic copy-number and LOH mosaicism from exome or whole-genome sequencing data, described in detail in chapter 4.

1.2 Structural variation in developmental disorders

1.2.1 Copy-number variation in DD

Despite the resolution of optical cytogenetics, limited to only multi-megabase chromosomal abnormalities, this technology was revolutionary in improving our understanding of large CNVs as a cause of DD. Discovery of the first copy-number events was followed from the discovery of the Barr body, the inactive copy of the X-chromosome in cells of females. Thus, the first copy-number abnormalities identified were gonosomal aneuploidies in individuals with syndromic sexual dysfunction: XXY,

Kleinfelter syndrome (Figure 1-6)[76] and X0, Turner syndrome[77].

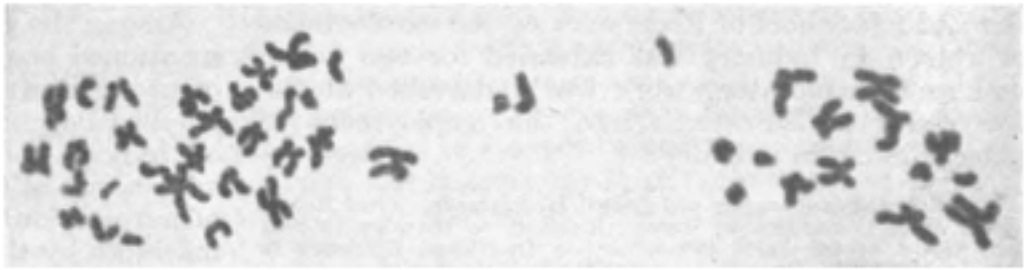

Fig. 1. Metaphase plate showing 47 chromosomes

Figure 1-6 The first published aneuploidy, Kleinfelter syndrome, adopted from Jacobs *et al.*[76]

The breakthroughs of gonosomal disease and advances in karyotyping led quickly to insights of autosomal aneuploidy and DD, beginning with the trisomy syndromes: Down Syndrome in 1959[78], Patau Syndrome in 1960[79], and Edwards Syndrome in 1960[80]. Studies from this period showed that aneuploidy occurs in 53% of spontaneous abortuses[66,81], cementing the importance of aneuploidy in diseases of development.

In addition to numerical abnormalities, copy-number structural abnormalities were also associated with DD. The first association of a sub-chromosomal copy number event associated with DD was found in 1963[82], as a large chromosome 5 deletion in a child with cri du chat syndrome. Subsequent use of banded cytogenetics was used systematically in the 1980s and 1990s to study structural variation in prenatal diagnostics and postnatal incidence studies. These experiments showed that cytogenetic evaluation of children with developmental delay by karyotyping could identify numerical or structural abnormalities in 9.5% of children[83]. Studies of consecutive live-births using cytogenetics identified abnormalities in 0.16%[84] (without routine banding) and 0.63%[85] (with banded chromosomes). The rate of mosaicism detected in live-births was 0.16% (3 in 1,830), the three detections including one mosaic chromosome 21, and two ‘supernumerary small metacentric marker chromosome with satellites on both ends’ whose origin chromosome was not specified[84].

In the last 15 years, microarray technology has provided a higher-resolution assay of CNVs compared with karyotyping. Seminal papers in selected individuals[86,87] and across human populations[88] have revolutionised our appreciation of constitutive CNVs as a common form of genomic variation, finding that CNVs are ubiquitous among humans and account for a nearly ten-fold greater proportion of variation in the

genome compared to SNPs[89]. CNVs account for about 18% of the genetic variation in gene expression[90]. Some CNVs are pathogenic, driven, for example, by disturbances in gene dosage[91], imbalances in protein networks[92], disrupting long range (regulatory) effects[93], and by gene interruption or gene fusion products[94].

Comparison of the performance of aCGH and karyotyping has shown that whilst aCGH misses some balanced rearrangements and triploidy, it yields a net increase of diagnoses compared to karyotyping because it can detect smaller unbalanced mutations that are missed by karyotyping[95-98]. Genetic evaluation of children with DD by microarray (using 50 kb median spacing) identified numerical or structural abnormalities in 19% of children[99], approximately twice the rate of karyotyping. aCGH microarray has at least equivalent sensitivity for diagnosis of common aneuploidies, and has increased sensitivity for smaller diagnostic CNVs (but not balanced arrangements)[100]. A study of over 36,000 children with idiopathic mental retardation and multiple chromosomal abnormalities demonstrated that the rate of diagnoses by microarray is twice that of karyotyping, and that karyotyping would identify those balanced rearrangements to only yield an additional one percent of diagnoses[99]. As of 2010, microarray is the recommended primary genetic test for children with DD[101].

In 2011, Cooper *et al.* reported a copy-number variation DD burden analysis, comparing 15,767 children with intellectual disability and congenital anomalies to 8,329 controls[102] for copy-number anomalies using microarray with 300 kb resolution. The results of this study showed a 14% burden of CNVs at least 400 kb in size in children with DD compared to controls (25.7% of cases compared to 11.5%), that increases in CNV length correlate with a greater excess of CNV enrichment in children with DD, and that larger CNVs were more often associated with syndromic malformations.

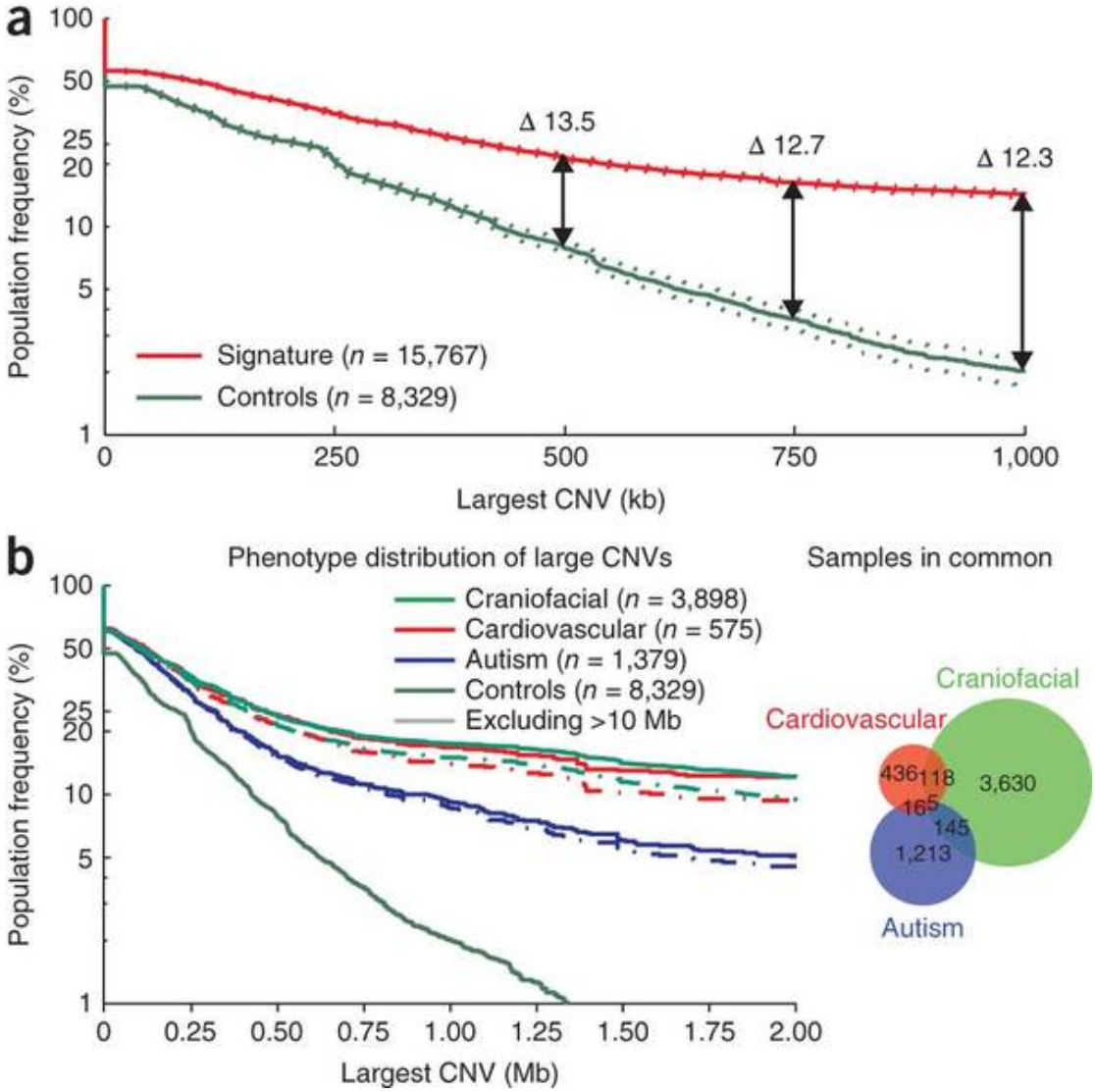

Figure 1-7 Cooper [102] showed that larger CNVs were correlated with pathogenic burden and syndromic phenotypes

Whilst this study identified an overall aggregate burden of CNVs in children with DD, interpreting the pathogenicity of individual copy number variants is more challenging. A deductive understanding of CNVs and phenotype is difficult because it would require considerable knowledge about underlying gene function and the effect of dosage on gene function for the genetic region overlapped (and perhaps bordered) by the CNV. Therefore, the most common method of identifying disease association is empiric, based on observation of shared phenotypes among multiple children containing overlapping CNVs. As an aid for interpretation, various paper[103] and electronic resources[104] have compiled lists of regions recurrently mutated with CNVs and, when available, the phenotypes found in children with such CNVs. These techniques allowed for the association of multiple genomic disorders with unbalanced abnormalities. These

resources are used in this dissertation to assist interpretations of pathogenicity of structural abnormalities found in children with their phenotypes.

1.2.2 Copy-neutral loss of heterozygosity (uniparental disomy) in DD

Uniparental disomy (UPD) is a balanced chromosomal abnormality, generally resulting from a defect of inheritance, in which both chromosomes of a homologous chromosome pair originate from a single parent. The UPD chromosome can be characterized in four ways: 1) extent: affecting the whole chromosome (complete) or a portion of the chromosome (segmental), the latter a hallmark of post-zygotic (somatic) recombination; 2) zygosity: affecting all cells (constitutive) or a proportion of cells (mosaic); 3) by homologue segregation: whether the centromeric regions are identical (isodisomy), resulting from an error in meiosis II or post-zygotic duplication, or represent both grandparental homologues (heterodisomy), resulting from an error in meiosis I; and 4) by parental-origin: maternal or paternal (Figure 1-8).

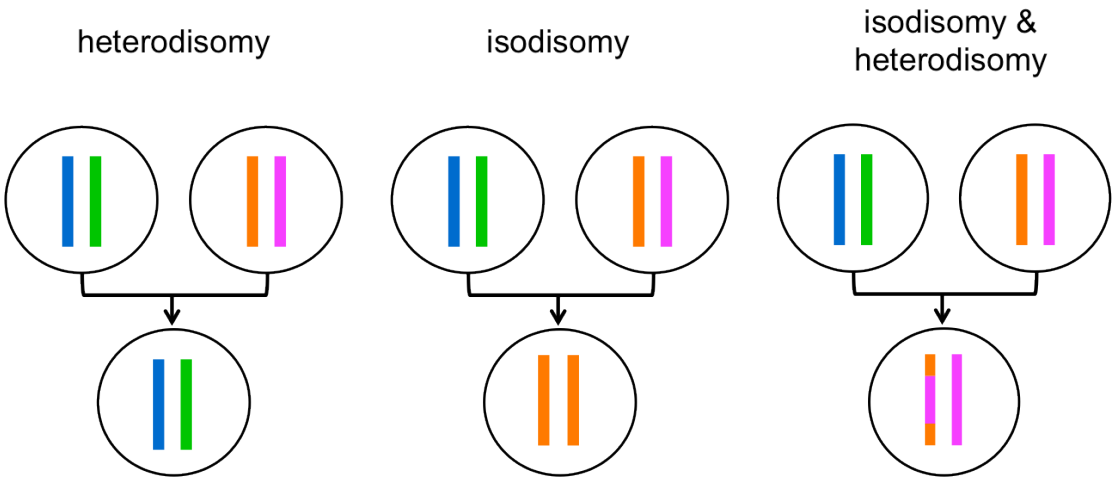

Figure 1-8 Types of uniparental disomy

UPD has three important mechanisms of disease causation: 1) imprinting disease, by disrupting the inheritance of essential parent-specific epigenetic modifications[105]; 2) recessive disease, by converting deleterious alleles bequeathed from a heterozygous parent to a homozygous state[106]; and 3) residual trisomy mosaicism, by its relationship to incomplete trisomy rescue[107]. UPD contributes to rare genetic diseases and its identification is an important part of the search for disease-causing variations.

Uniparental disomy is a balanced chromosomal rearrangement imperceptible to karyotype analysis or to aCGH, and because genome-wide screening of zygosity was not possible until widespread utilisation of SNP microarray in the early 2000s, the

earliest cases of UPD were difficult to recognise. However, before they were identified *in vivo*, such events were predicted on a theoretical basis.

In the 1970s, karyotype screening of spontaneous abortuses showed that half of first trimester abortuses were aneuploid[108]. In a paper replete with foresight, Eric Engel in 1980 deducted that, given this frequency of aneuploidy, the rare but nonetheless 'statistically likely' fusion of two aneuploid gametes, one nullisomic and one disomic for the same chromosome, might provide the compensatory complementation to rescue euploidy and result in a viable zygote; this zygote would have a homologous chromosome pair solely derived from a single parent, a phenomenon he neologised as *uniparental disomy* (UPD)[109]. Furthermore, he postulated several complications of UPD, suggesting, for example, the long regions of homozygosity created by isodisomy would predispose to recessive diseases, and that UPD could result in the unusual endowment of recessive disease from a single carrier parent. Engel calculated on the basis of per-chromosome aneuploidy frequency that the rate of uniparental disomy might be approximately 3 in 10,000. Indeed, these above predictions would be verified experimentally with time. Notably, however, imprinting (parent-specific inheritance of gene expression) disorders, were not yet discovered in humans and thus were not discussed as a complication of UPD in Engel's earliest work, but are now recognised as an important clinical complication of UPD on some chromosomes.

The earliest detections of UPD in humans describe a loss of heterozygosity in cancer that is acquired post-zygotically, also called acquired UPD. Investigators in the early 1980s, using polymorphic enzyme phenotypes, observed that cultured cancer cell lines had less heterozygosity than the general population, a phenomenon called 'loss of heterozygosity'[110]. In 1987, Yokota *et al*., using the newly developed restriction fragment length polymorphism (RFLP) assay on fresh tumour samples found that LOH was ubiquitous in lung cancers, and suggested that such events may be 'critical in the genesis of tumour rather than a secondary event'[111]. These findings were of great interest to the cancer community because they provided an explanation for loss of tumour suppressor genes and further evidence of the ubiquity of structural variation in cancer.

The first published example of UPD in a child with DD appears to be the 1984 finding of loss of heterozygosity on chromosome 11 in three children with unusual, rare cancers and Beckwith-Wiedemann syndrome[112]. Nevertheless, it does not appear that

this study alerted interest in the DD community, as UPD as the genetic basis of imprinting disorders was not discussed until 1989. The first clinical report of UPD was by Spence *et al*. in 1988, in which a child with cystic fibrosis was found to have homozygosity of a pathogenic maternal mutation due to maternal isodisomy[113]. Shortly after, Nicholls *et al*. reported the first case of clinical heterodisomy in Prader-Willi syndrome[114] and suggested that Angelman and Prader-Willi syndrome may be due to disruption of different parental alleles, a conjecture substantiated by Schinzel *et al.*[115], thereby giving rise to the field of imprinting disorders in humans. That same year, Vidaud *et al.*[116] reported transmission of haemophilia, a sex-linked-recessive condition, from the child's father, due to uniparental heterodisomy of the gonosomes.

In 1991, Engel suggested[117], based upon the finding of segmental UPD in Drosophila, that the distribution of UPD events across the chromosomes in humans could locate imprinting vulnerability regions that cause disease when disrupted. The first effort to derive an imprinting map in humans was made in 1995[118] and provided definitive evidence for imprinting on four chromosomes.

In 1992, Robinson *et al.* showed that among 120 children with maternal UPD15 (causing Prader-Willi syndrome), the most common cause was due to meiosis I errors (71%), while post-zygotic duplication (16%) and meiosis II errors (13%) were less frequent[119]. An early UPD study found that there was an exponential increase of the frequency of UPD15 with maternal age[119]. Two years later, Field *et al.* presented several reports of UPD on chromosome 1 with no apparent effects, which suggested "in the absence of isodisomy for recessive deleterious genes, UPD for chromosomes that do not harbour imprinted loci may be quite harmless[120]". Two years later, Robinson *et al.* calculated, based on the frequency of UPD15 (1/80,000), the frequency of UPD in live births to 1 in 3,500[121], close to Engel's original estimate of 3 in 10,000.

In 2001, the first guidelines from the American College of Medical Genetics on diagnostic testing for UPD were published[122] and specified that RFLP analysis should be used on child, mother, and father, when prenatally-detected mosaicism for imprinting-susceptible chromosomes was found or if the patients had features of known imprinting disorders. Similar to the interpretation of specific CNVs in children, understanding the pathogenesis of UPD events in children has been advanced from empiric findings. Using paper[123] and online catalogues[124], collections of UPD regions can be compiled, enabling identification of recurrent phenotypes among children with UPD, from which new UPD disease associations can be established. By these means,

instances of all but three of the 44 possible uniparental autosomal pairs have been reported, with imprinting disorders resulting from maternal disomy of chromosomes 7, 14, and 15 and from paternal disomy of chromosomes 6, 11, 14, and 15[122].

	UPD and other molecular alterations	Frequency	Chromosomal region	Mosaicism and UPD[a]	Clinical features
ID					
Transient neonatal diabetes mellitus (TNDM)	UPD(6)pat dup(6q) *PLAGL1* hypomethylation	41% 29% 30%	6q24	2× 47,XN,+6/46,XN	Prenatal and postnatal growth retardation, transient diabetes with dehydration, hyperglycemia without ketoacidosis, macroglossia, umbilical hernia
Silver–Russell syndrome	UPD(7)mat	7–10%	7	Single cases with 47,XN,+7 on CVS and postnatal UPD(7)mat, single case with postnatal 47,XN,+7/46,XN(UPD)	Prenatal and postnatal growth retardation, relative macrocephaly with triangular face, hemihypertrophy
Silver–Russell syndrome	UPD(11)mat dup(11p15)mat ICR1 hypomethylation dup(11p15)mat *CDKN1C* mutations	$n = 1$ Single cases >38% Single cases $n = 1$	11p15.5	Mosaicism unknown	
Beckwith–Wiedemann syndrome	UPD(11)pat dup(11p15)pat ICR1 hypermethylation ICR2 hypomethylation *CDKN1C* mutations	20% Single cases 4% 50% 5%		(Segmental) isodisomies, normally mosaicism 46,XN(bip)/46,XN(UPD)	Prenatal and postnatal overgrowth, organomegaly, macroglossia, omphalocele, neonatal hypoglycemia, hemihypertrophy, increased tumor risk
Temple syndrome [UPD(14)mat]	UPD(14)mat del14q32 *MEG3* hypomethylation	78.4% 9.8% 11.7%	14q32	6× 47,XN,+14/46,XN	Prenatal and postnatal growth retardation, small hands and feet, obesity, muscular hypotonia with feeding difficulties, early puberty
Kagami–Ogata syndrome [UPD(14)pat]	UPD(14)pat del14q32 *MEG3* hypermethylation	65.4% 19.2% 15.4%		1× 47,XN,+14/46,XN	Polyhydramnios, abdominal wall defects, bell-shaped thorax with coat-hanger rib sign
Angelman syndrome	UPD(15)pat del15q11q13 aberrant methylation *UBE3A* mutations	1–2% 75% ~3% 5–10%	15q11q13	Rare	Microcephaly, ataxia, seizures, restlessness, frequent unmotivated laughing, mental retardation, no speech
Prader–Willi syndrome	UPD15mat del15q 11q13 aberrant methylation	25–30% 70–75% ~1%		2× UPD(15)mat/biparental 46,XN cell lines 1× UPD(15)mat/47,XN,+15	Muscular weakness, initially feeding difficulties, followed by hyperphagia and obesity, growth retardation, mental retardation
Pseudohypoparathyroidism Ib (PHPIb)	UPD(20)pat aberrant methylation	Unknown	20q13	1× 47,XX,+20/45,XY,psu dic(20;20)/46,XX,psu dic(20;20)	Isolated parathormone resistance
UPD-associated disorder					
Genome-wide paternal UPD (BWS-like phenotype)	UPiD(AC)pat[c]	Unknown	All chromosomes	Viable only as mosaicism	BWS phenotype is predominant, massively increased tumor risk

Figure 1-9 Summary of UPD disorders, from Eggermann et al.[125]. Imprinting syndromes are caused by defects in methylation. For some imprinting syndromes, such as Temple syndrome, UPD is the most common imprinting-disruption mechanism. For others, such as Angelman syndrome, other mechanisms are more common.

Isodisomy can be detected by identifying long strings of homozygous genotypes in probands. Collectively, more than 10,000 children have been studied across three experiments and identified a rate of isodisomy of approximately 0.2%[35,37,126]. Unlike the identification of isodisomy, detecting heterodisomy directly requires trio data. Due to the dearth of large research studies with trio SNP data, very little was known regarding the prevalence of heterodisomy in children with DD. In addition, the absence of software to detect UPD directly from exome sequence data, which are now routinely generated in rare disease genetics, motivated my development of UP*D*io, a sequence-based UPD detection tool. I applied UP*D*io on exome data from several thousand trios recruited for developmental disorder to detect isodisomy and heterodisomy in children with DD and this analysis is described in chapter 2.

1.2.3 Mosaic structural rearrangements and DD

Mosaic abnormalities are more difficult to detect than constitutive abnormalities because mosaic events are present in only a proportion of cells. As explored in detail in chapters 3 and 4, mosaicism can only be detected if the abnormality is present in the tissue type assayed and in sufficient clonality to be perceptible to the platform used.

The first example of mosaic aneuploidy was discovered in the very early years of cytogenetics in a patient with Klinefelter syndrome and XY/XXY mosaicism[127]. However, large-scale study of structural mosaicism during the cytogenetics era was immature, as the detection resolution was limited and prenatal screening rarely assayed sufficient numbers of metaphases to make reliable data on mosaic frequency. Even so, attempts have been made to aggregate data for mosaicism from cytogenetics. Meta-analysis of nearly 180,000 prenatal diagnostic cases for the assessment of *mosaic* structural abnormalities has observed a rate of 0.3%[128].

Instead of attempting to measure multiple metaphases, SNP microarray provides a platform to assay multiple cells simultaneously using techniques discussed in detail in chapter 3. Several recent studies have studied SNP microarray to better understand the frequency and consequence of structural mosaicism. The timing and origin of UPD was reviewed extensively in reviews by Kotzot in 2001 and 2008, highlighting several important insights: mosaic aneuploidy and UPD frequently co-occur; trisomy often precedes UPD; incomplete monosomy and trisomy rescue could result in combinations of aneuploidy and UPD; the origin of UPD often includes meiotic nondisjunction followed by a mitotic rescue event[129], but crossing-over of homologues, mis-segregation of translocated chromosomes, association with marker chromosomes, and other complex events, are possible[107] (Figure 1-10).

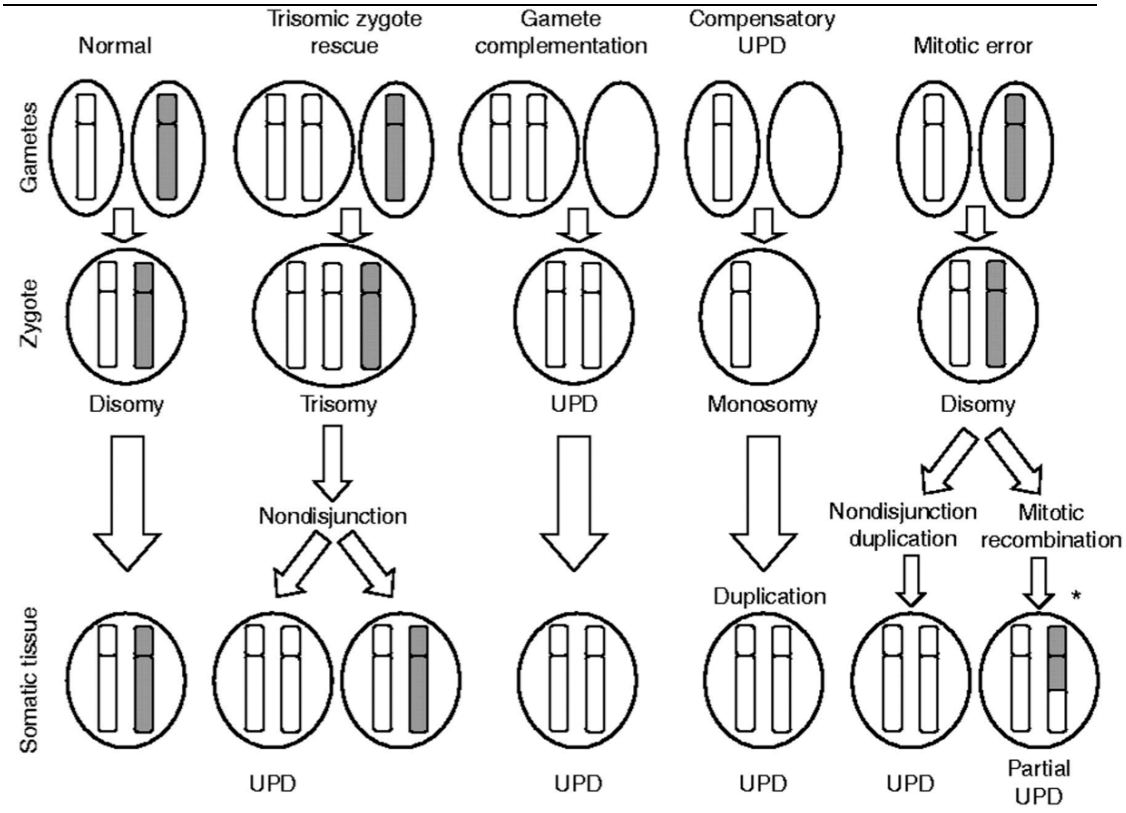

Figure 1-10 Some common mechanisms of UPD formation (adapted from Kotzot 2001[129]). More complex mechanisms of UPD formation are also possible, see Kotzot 2008[107].

Several studies have investigated the rate of structural mosaicism in children ascertained for genetic testing. In 2010, Conlin *et al.* examined blood from 2,019 children with pervasive developmental delay or congenital abnormalities, identifying 12 with mosaic aneuploidy (0.6%) and eight with UPD. Of these eight UPD events, four were from trisomy rescue, two were from monosomy rescue, and two were mitotic in origin. Mosaicism was only detected in the two mitotic cases. The origin of the other six UPD events was inferred from the allele fraction patterns. Of the 12 aneuploidies, 9 were monosomies, and all of these monosomies arose from mitotic non-disjunction (and therefore post-zygotically), suggesting that early stage (inherited) monosomy is lethal, whilst half of the trisomies arose by meiotic non-disjunction. In addition, one of the children with a mosaic abnormality was chimeric. Chimerism is similar to mosaicism in that it represents a mixture of genetically distinct cells in an organism, but unlike mosaicism in which the genetic divergence originates post-zygotically, the cells lines in chimerism originate from two zygotes that then fuse into one organism. In the chimeric identified in the Conlin *et al.* study, the heterogeneity of genetic components was best explained by the early fusion of an XY cell line with a parthenogenic, diploid XX cell line[36]. Other studies include Bruno *et al.* which investigated 5,000 children referred for

clinical diagnostic testing and identified 12 with mosaicism (0.24%) and Pham *et al.* which examined 10,362 children recruited for diagnostic testing with high-resolution aCGH and identified mosaicism in 57 (0.55%), of which 12 were smaller events detected by exon-focussed probes.

Studies of structural mosaicism using SNP data in adults[50,130,131] have shown that mosaicism increases with age and predisposes to haematological cancer. However, the incidence and burden of structural mosaicism in children is not well ascertained because of the limited number of generally healthy children analysed by SNP microarray for the detection of mosaicism. Additionally, the absence of large studies of tissue other than blood-derived tissue, for example, of buccal epithelium, hinders assessment of tissue-limited mosaicism, a concept revisited in chapters 3 and 4.

Estimating the pathogenic potential of mosaic structural variation can be difficult. Whilst resources like DECIPHER[104] and the Liehr UPD database[132] assist the interpretation of constitutive CNV and UPD, less is known about the pathogenic impact of mutations across the continuum of clonality, across different cell types. Additionally, unlike the burden analysis performed by Cooper *et al.* for constitutive CNV in DD, the lack of studies investigating the rates of UPD and structural mosaicism in healthy children (indeed, of multiple cell types from healthy children) hinders the assessment of mosaic burden, undermining attribution of mosaicism as a pathogenic class of genetic variation. These deficits motivated the third chapter of this dissertation, in which pre-existing software tools are used to calculate the rate of structural mosaicism from SNP chip data in healthy children. The lack of software tools to identify structural mosaic abnormalities from exome or whole-genome sequencing data, motivated the fourth chapter of this dissertation.

1.3 Clinical diagnostic testing of developmental disorders

Developmental abnormalities may present at any stage of development. Common indications that trigger diagnostic evaluation include abnormal prenatal screening results, dysmorphic features observed post-partum, failure to attain developmental milestones, and learning disabilities observed during school-age years. The assessment of a child with the features above is performed by a paediatrician and often in collaboration with a clinical geneticist. Assessment of the child will vary depending on the age of the child but often includes family history, gestational history, patient history,

physical examination with anthropometrics, neurological examination, behavioural examination, and genetic testing.

The genetic tools available to clinicians for clinical diagnostic testing vary by local institution. Historically, (and in many centres today) genetic diagnosis has been performed using karyotyping. Indeed, as seen above, cytogenetics has a long history of detecting DD and the large number of children studied by karyotyping has left a legacy on our current understanding of aneuploidy and structural variation in DD. However, despite prior investigation with karyotype, telomeric FISH, and targeted gene testing, the discovery of the underlying genetic cause is successful in only half of children with cognitive delay[7].

Current guidelines for genetic diagnostic testing of "patients with intellectual disabilities, autism and/or congenital anomalies" now recommend microarray, and ideally, a combined aCGH and SNP microarray, as the first-tier test[133]. In the UK, standard genetic tests available in most referral centres include karyotypic analysis, microarray, and targeted gene testing. These tests can identify aneuploidy, structural mutations, and mutations in specific disease genes of interest based on the child's phenotype. Genetic diagnosis of children with non-monogenic, non-syndromic disorders, like ADHD or autism is even more challenging[134].

In the last few years, DNA sequencing of the patient's exonic (protein coding) regions, so-called exome sequencing, has yielded unprecedented throughput and resolution to the genomes of children with DD. Whilst pedigree study designs have proven helpful in elucidating the genetic causes of many recessive diseases, the trio study design has yielded important contributions of *de novo* variation to rare disease and has enabled the identification of previously unknown disease-causing genes. A framework integrating high-throughput sequencing, trio sample recruitment, and computational development requires substantial resources. A collaborative paradigm combining patient recruitment in hospitals with the technical analysis in research institutions has enabled patient access to state-of-the-art genetic analysis. In the UK, whilst exome sequencing is not yet available for diagnostic testing of DD as a local test in most hospitals, it is possible through participation in the Deciphering Developmental Disorders study.

1.3.1 Deciphering Developmental Disorders study

The DDD study is an on-going collaborative medical research project aimed to determine the underlying genetic basis of disease in children with severe DD (Table 1-1) in the UK, for whom prior investigation has yielded no definitive diagnosis. The study consists of approximately 12,000 patient-parent trios, who have been recruited by physicians at hospitals across the UK and Ireland. Several data are collected, including a gestational history, prenatal and postnatal history. Each child is given a thorough examination, including an assessment of developmental milestones, with phenotypic abnormalities recorded using a standardised vocabulary, the Human Phenotype Ontology (HPO)[135]. DNA is extracted from sampled saliva & blood from probands and from the saliva of parents. Genetic assays and computational tool development and analysis are primarily performed at the Wellcome Trust Sanger Institute (WTSI). Clinical geneticists at WTSI, led by Helen Firth, perform clinical assessment of the predicted pathogenic potential of discovered genetic variation. Their findings are relayed to the clinical geneticist who recruited the child into the study. Variants of interest are presented using a strength of confidence ontology developed by Plon, *et al.*[136]. In this 5-tiered scheme, class 3 variants are considered to be pathogenic with 5% - 94.9% probability ('uncertain'), class 4 variants have 95% – 99% probability ('likely pathogenic'), and class 5 variants have above 99% probability ('definitely pathogenic').

Table I: The Deciphering Developmental Disorders (DDD) study is recruiting children with severe and extreme developmental phenotypes

Inclusion criteria for the DDD study
Neurodevelopmental disorder
Congenital anomalies
Abnormal growth parameters (height, weight, occipitofrontal circumference)
Dysmorphic features
Unusual behavioural phenotype
Genetic disorder of significant impact for which the molecular basis is currently unknown

Table 1-1 DDD Inclusion Criteria, adapted from Firth *at al.*[1]

The genetic assays conducted include exome sequencing for all three members of each trio, high-resolution aCGH for each proband, and SNP microarray analysis for 4,000 trios. Genetic results are agglomerated across probands to identify genetic

similarities among patients that may indicate a shared underlying disease. Likely diagnostic findings from the study are returned to clinicians who confer diagnostic interpretation to the families.

Analysis of the first 1,133 trios[3,6] has recently been completed and yielded new monogenic disease associations for 12 genes, based on enrichment of *de novo* mutations. These associations enabled a 10% relative increase in the fraction of children for whom the molecular diagnosis could now be identified, yielding a total of approximately 350 new diagnoses in this set. The most common mutational category underlying new diagnoses was *de novo* point mutations followed by *de novo* CNVs. In addition, other large-scale abnormalities, including constitutive UPD and mosaic structural variants, were also identified using analytical approaches and software tools I developed. This dissertation will describe in detail the detection and discovery of these elements.

1.4 Summary

This dissertation presents an analysis of non-inherited structural variation among the first 5,000 trios from the DDD study. The main components of this work are descriptions of: a new method for detecting uniparental disomy from exome trio data (chapter 2); a burden analysis of mosaic structural variation and the clinical consequences of mosaic structural variation in children with DD (chapter 3); a new method for the detection of mosaic structural variation using next generation sequence data (chapter 4); a recapitulation of the main findings and a discussion of this research in broader context (chapter 5).