Uniparental disomy and mosaic structural variation in developmental disorders

Daniel Alexander King, M.D.

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2015

DECLARATION OF ORIGINALITY

This dissertation describes my research performed between January 2012 and June 2015 under the supervision of Dr. Matthew Hurles. This work is my own and includes nothing which is the outcome of work done in collaboration except where explicitly stated in the text. It has not been previously submitted for any qualification, and it complies with the prescribed word limit set by the Degree Committee for the Faculty of Biology.

ABSTRACT

Developmental disorders (DDs) are diseases of impaired childhood development and include congenital anomalies, neurodevelopmental disorders, and abnormalities in growth and behaviour. Determining the genetic causes underlying DD is a major goal of contemporary medical research and the recent entrance of exome sequencing data into the rare-disease field has been transformative in uncovering the importance of *de novo* point mutations as a major source of DD-associated mutations. Recent efforts have successfully harnessed exome sequencing data to detect constitutive copy-number variation, a form of large-scale structural abnormality. However, at the inception of my doctoral work, no software tools had yet been developed to identify, from exome sequencing data, uniparental disomy (UPD), a form of copy-neutral variation, nor large-scale ('structural') abnormalities, which have long been implicated as important contributors to DD. The research underlying this work aimed to fill this void.

This dissertation addresses the development of new software tools, UP*D*io and MrMosaic, which have extended the diagnostic reach of sequencing data to identify UPD and structural mosaicism, and have been made freely available. Simulation analyses show that these tools can detect the large-scale abnormalities identified by karyotyping or microarray in standard clinical testing. Implementation on nearly 5,000 children with undiagnosed diseases demonstrated that UPD and structural mosaicism are enriched in children with developmental disorders compared with healthy children and suggested that most of the detected abnormalities are likely to be pathogenic. Investigation of the clinical impact of the detected events identified several disease-causing mechanisms, including UPD-associated imprinting and recessive diseases, and genomic disorders associated with large mosaic deletions and duplications.

The five chapters of this dissertation are: 1) an introduction, to describe the context of this doctoral work; 2) a description of UP*D*io, a new method for detecting uniparental disomy from exome trio data; 3) a burden analysis of mosaic structural variation and the clinical consequences of mosaic structural variants found in children with DD; 4) a description of MrMosaic, a new method for the detection of mosaic structural variation using next generation sequence data; and lastly, 5) a discussion that recapitulates the results of these analyses, describes their limitations, and considers future directions.

ACKNOWLEDGEMENTS

Great thanks are owed to many. In chronological order, my parents deserve credit, overdue and too understated, for teaching me how to walk, talk, and create (thanks mom) and inspiring my passions in science (thanks dad). Josh and Jason provided a sibling rivalry that was a great cure of laziness and I look forward to one day raising our families together.

One of my first patients in medical school was a middle-aged dad, experiencing new, severe headaches, and died from glioblastoma a few months later. Thank you for allowing me to follow your case as the experience led me to pursue medical research at the NIH. There, Dr. Les Biesecker nurtured my interests in programming, advocated for me, and encouraged me to become a physician scientist. Jamie Teer was a patient and forgiving scripting coach and Larry Singh taught me that statistics is really not so boring at all.

I am sincerely grateful to my Ph.D. supervisor, Matt Hurles, with whose intuition this dissertation research has greatly benefited, and with whose guidance the pleasure of research was conspicuous and reinforcing. One of my favourite phrases of yours is ‘it’s a testable hypothesis’, a maxim demonstrating your perpetual loyalty to evidence and methodical experimentation. I am grateful to have pursued with your mentorship and collaboration science as it evolved in front of us, yielding findings often unexpected and fascinating, and hope that such discovery can continue in partnership for many years to come.

Thank you to everyone in Team 29 for putting up with my cheeky birthday surprises and for your friendship. Saeed, Manu and Raheleh, thanks for being good coffee-break companions. Ray, Alejandro, Jeremy, and Tom, I enjoyed our brainstorming sessions. Thank you to the DDD laboratory and informatics teams for performing so many upstream analyses. The work presented here was not possible without the participation of thousands of children and their families; thank you for joining DDD. Wellcome Trust kindly funded this research. Annabel & Christina, and Carol, much appreciation for organising the logistics for all my academic activities.

Thank you Tanya for always being there for me, and for enabling my life inside and outside the lab these Ph.D. years to be adventurous and fulfilling.

CONTENTS

PUBLICATIONS

King, D.A., Sifrim, A.S., Fitzgerald, T.W. & Hurles, M.E. Detection of structural mosaicism from targeted and whole-genome sequencing data. *In review.*

King, D.A., Jones, W.D., Crow, Y.J., Dominiczak, A.F., *et al.* Mosaic structural variation in children with developmental disorders. *Human molecular genetics* **24**, 2733-2745 (2015).

Carvalho, C.M.B., Pfundt, R., King, D.A., Lindsay, S.J., *et al.* Absence of heterozygosity due to template switching during replicative rearrangements. *The American Journal of Human Genetics* **96**, 555-564 (2015).

Akawi, N., McRae, J., Ansari., M, Balasubramanian, M., et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature genetics* **47**, 1363-1369 (2015).

Wright, C.F., Fitzgerald, T.W., Jones W.D., McRae, J.F., *et al*. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet* **385**, 1305-1314 (2015).

TDDD Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **12**, 223-228 (2014).

King, D.A., Fitzgerald, T.W., Miller, R., Canham, N., *et al*. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome research* **24**, 673-687 (2014).

LIST OF TABLES

LIST OF FIGURES

TABLE OF ABBREVIATIONS AND ACRONYMS

1000G	1000 Genomes study
aCGH	array comparative genomic hybridisation
ALSPAC	Avon Longitudinal Study of Parents and Children
AUC	area under the curve
BAC	bacterial artificial chromosome
BAF	b allele frequency
B_{dev}	BAF deviation
C_{dev}	copy-number deviation
CNV	copy number variation
DD	developmental disorders
DDD	Deciphering Developmental Disorders
DECIPHER	DatabasE of genomiC varIation & Phenotype in Humans using Ensembl Resources
FISH	fluorescent *in situ* hybridisation
GADA	genome alteration detection analysis (software)
GRC	genome reference consortium
GRCh37	genome reference consortium, human genome reference 37
GWAS	genome-wide association studies
HGP	human genome project
HMM	hidden Markov model
HPO	human phenotype ontology
indels	insertions and deletions
LOH	loss of heterozygosity
LRR	log r ratio
MAD	mosaic alteration detection
MAF	minor allele frequency

Mb	megabases
MrMosaic	Mosaic Rearrangements by Merging Of Sequenced Alleles using their Identity and Counts
OMIM	online Mendelian inheritance in man
QC	quality control
RFLP	restriction fragment length polymorphism
ROH	region of homozygosity
SFHS	Scottish Family Health Study
SNP	single nucleotide polymorphism
SV	structural variation
TEDS	Twins Early Development Study
UK10K	United Kingdom 10,000 Genomes Project
UPD	uniparental disomy
VCF	variant call format
WES	whole-exome sequencing
WGS	whole-genome sequencing

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.

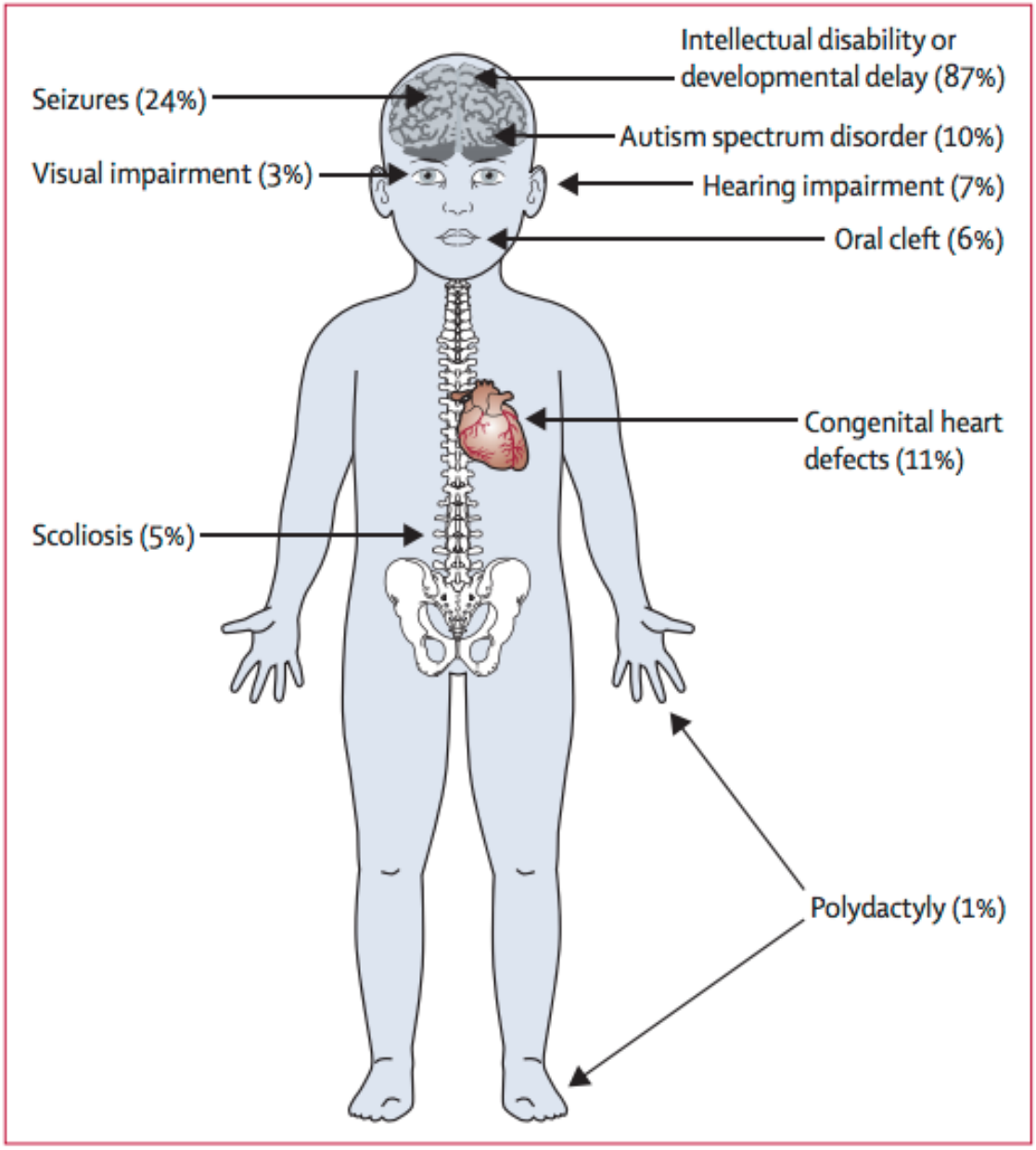

Figure 1-1 Prevalence distribution of phenotypes observed in a recent large study of severe DD[3].

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.

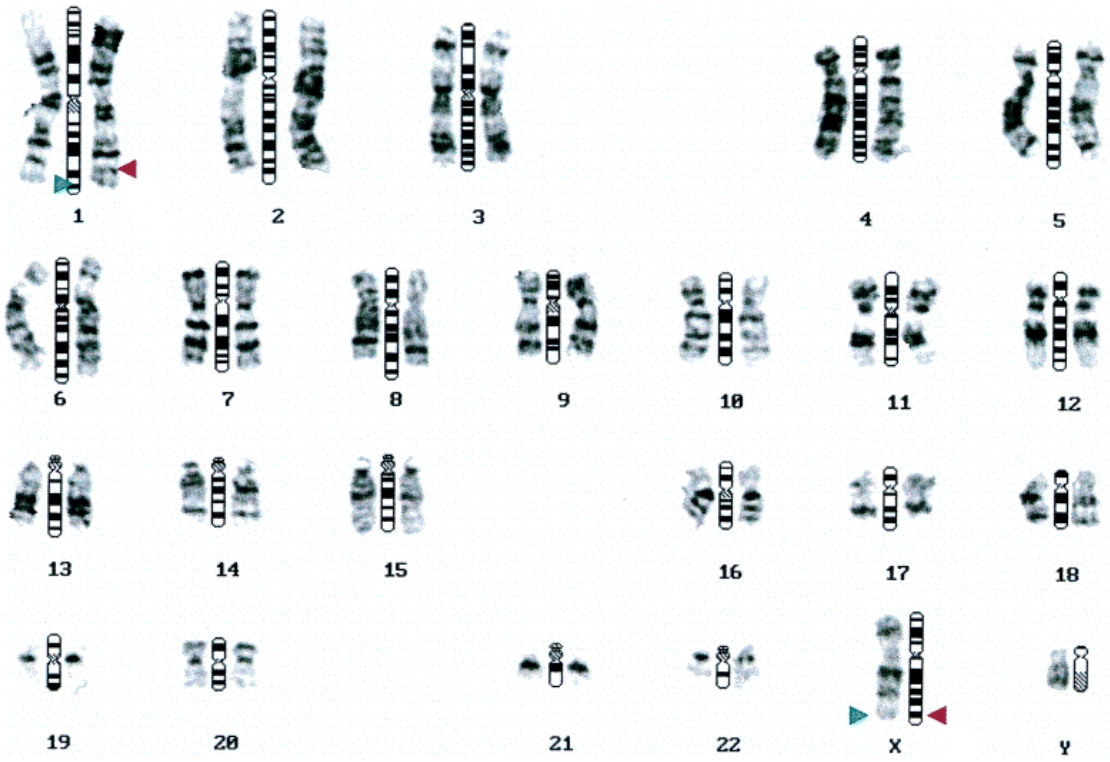

Figure 1-3 Banded karyotypes. In this case, a translocation of material between chromosomes 1 and X, adapted from Mattei *et al.*[25].

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).

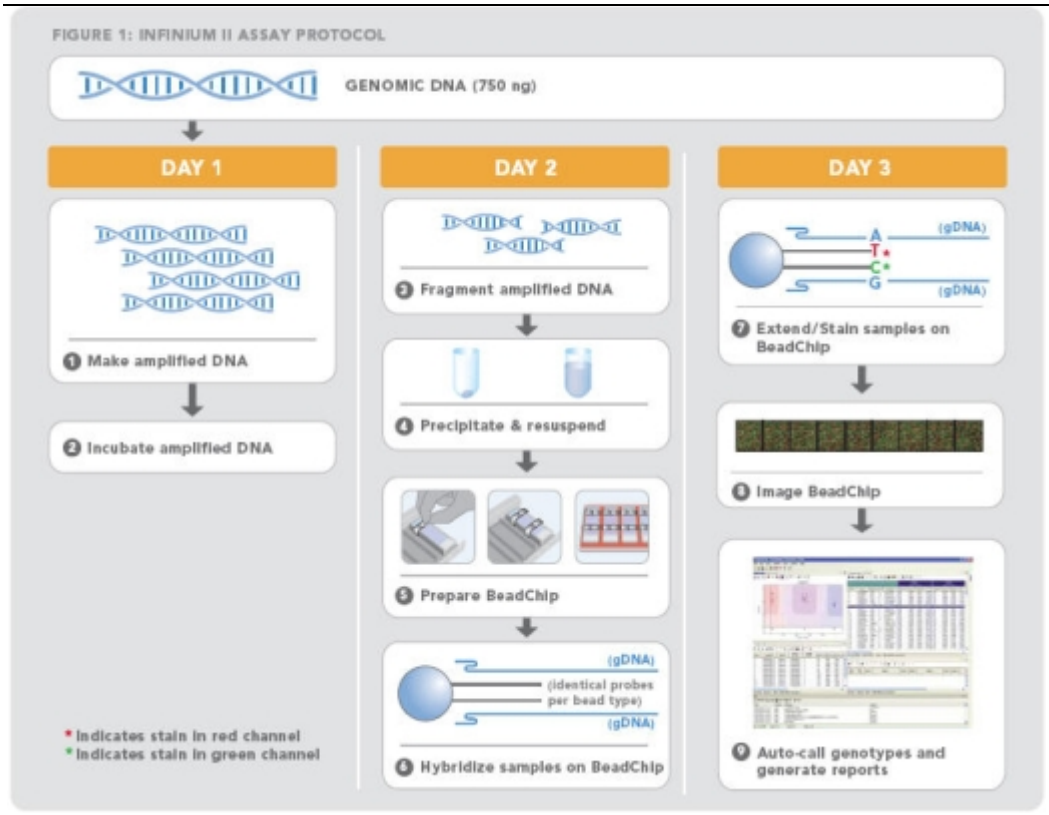

Figure 1-4 Illumina BeadArray technology, adapted from Illumina documentation[39].

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.

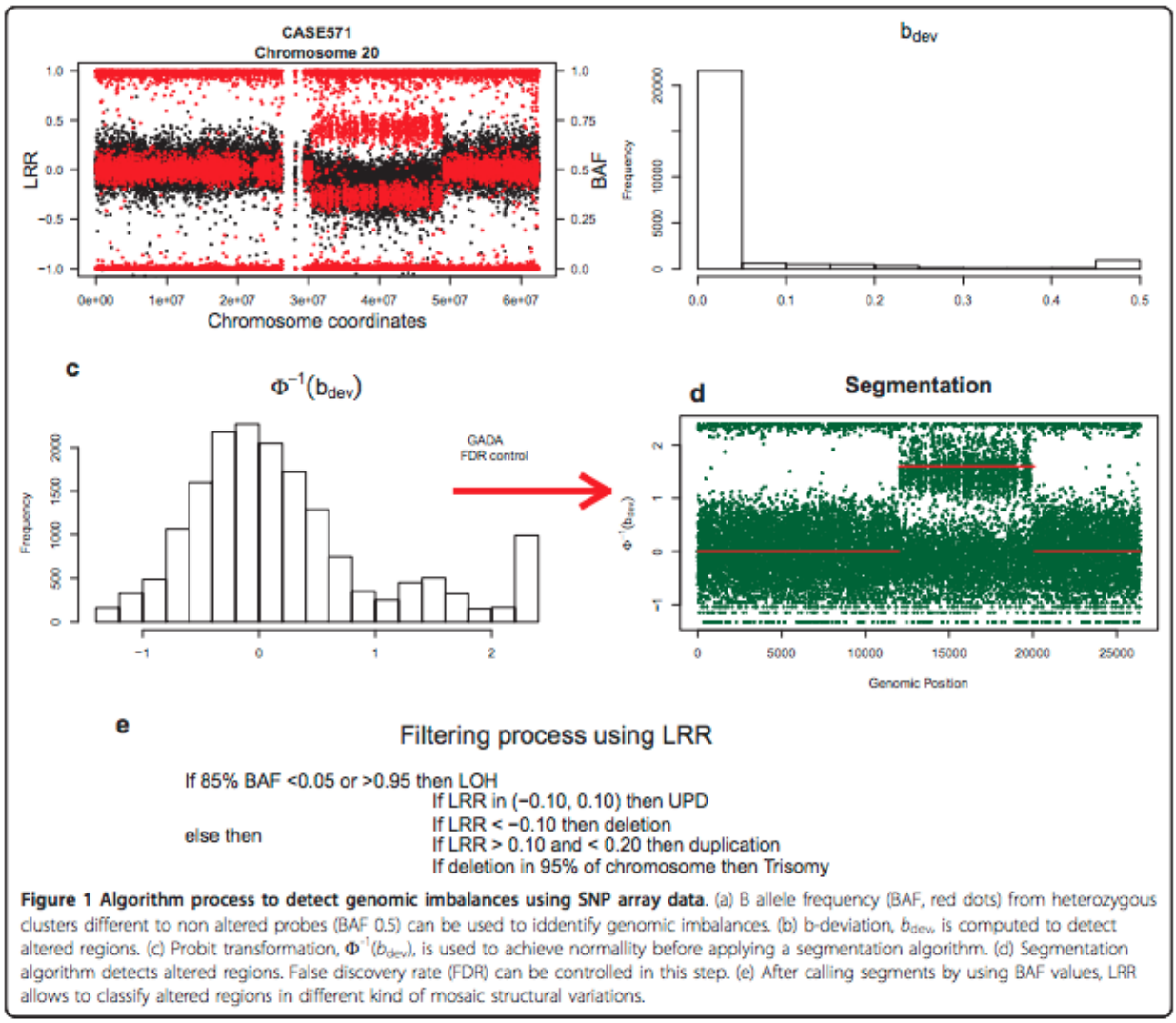

Figure 1 Algorithm process to detect genomic imbalances using SNP array data. (a) B allele frequency (BAF, red dots) from heterozygous clusters different to non altered probes (BAF 0.5) can be used to iddentify genomic imbalances. (b) b-deviation, b_{dev}, is computed to detect altered regions. (c) Probit 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].

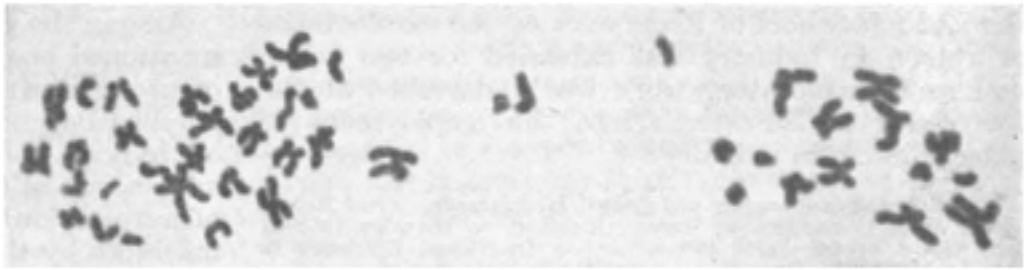

Fig. 1. Metaphase plate showing 47 chromosomes

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.

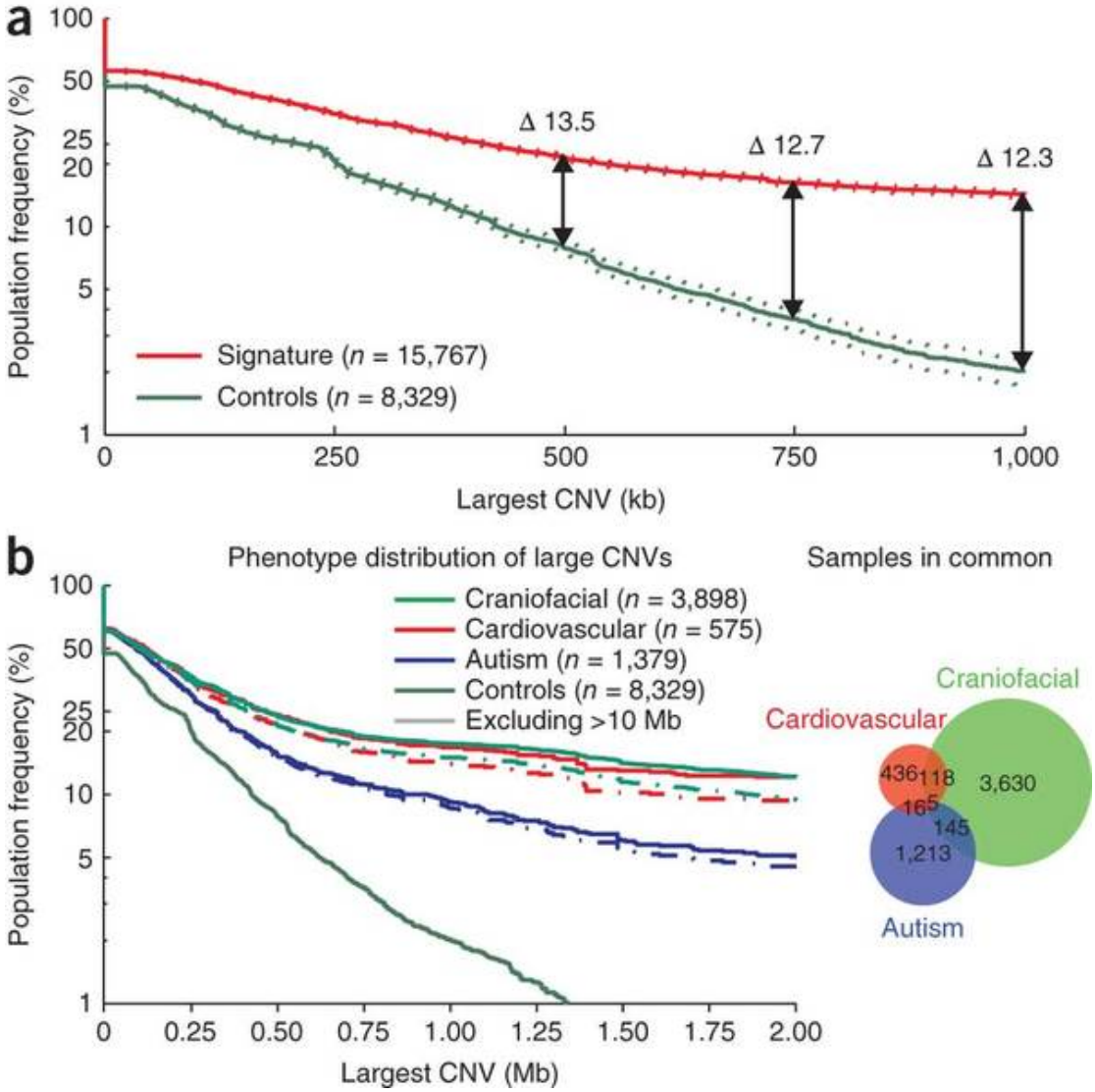

Figure 1-7 Cooper [102] showed that larger CNVs were correlated with pathogenic burden and syndromic phenotypes

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).

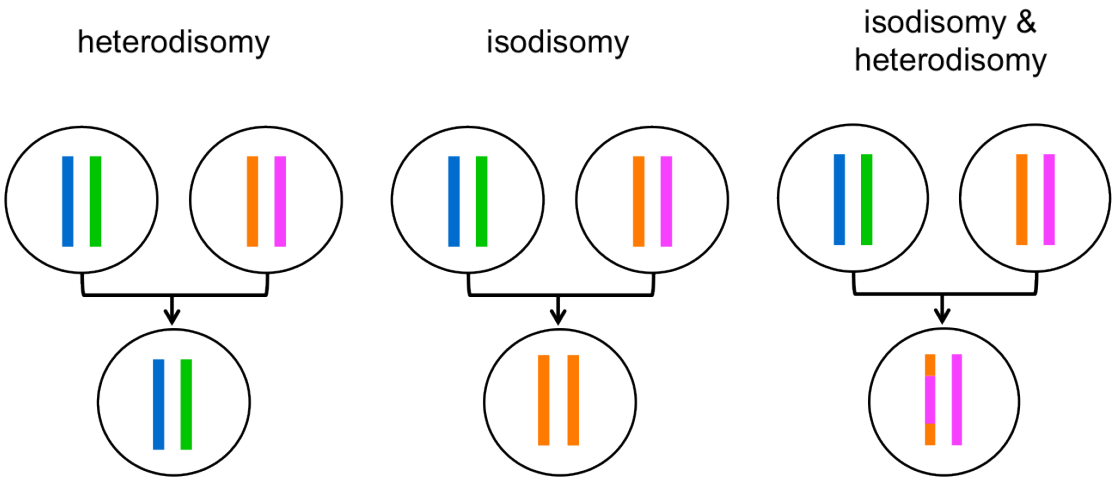

Figure 1-8 Types of uniparental disomy

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).

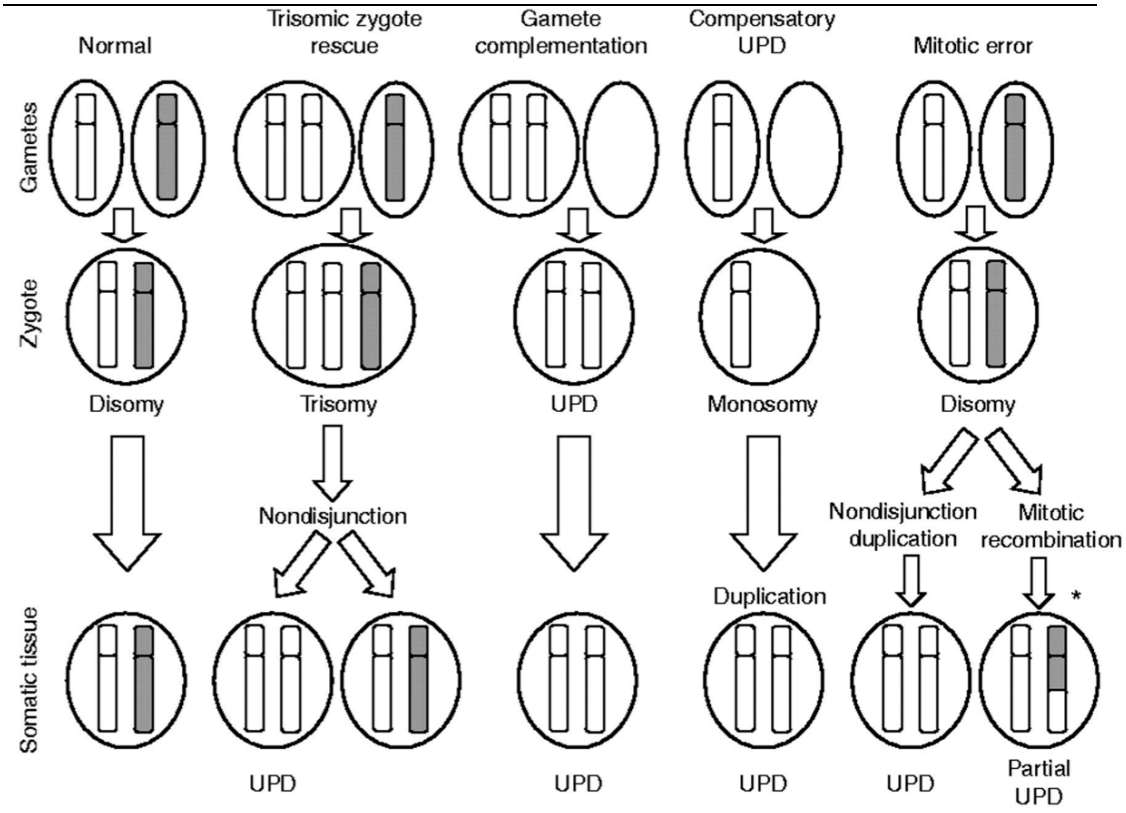

Figure 1-10 Some common mechanisms of UPD formation (adapted from Kotzot 2001[129]). More complex mechanisms of UPD formation are also possible, see Kotzot 2008[107].

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).

2 UNIPARENTAL DISOMY

2.1 Publication Note

Most of the work described in this chapter was previously published in 2014[137]. Sections describing the second stage of analysis contain unpublished results. Unless explicitly stated otherwise, the analysis described herein is the work I performed myself, under the supervision of Matthew Hurles.

2.2 Introduction

A review of definitions: uniparental disomy (UPD) is a type of copy-neutral structural variation, characterised as the same-parent origin of both chromosomes of a homologous chromosome pair. Isodisomy reflects a single parental homologue transmitted in duplicate, resulting in homozygosity, whilst heterodisomy reflects both chromosome homologues from a single parent. Due to meiotic recombination, the inherited UPD chromosome often contains a mixture of heterodisomic and isodisomic regions (mixed UPD). UPD can be constitutive or mosaic. Constitutive UPD is evident using genotype data and is the subject of this chapter. In contrast, mosaic UPD is not easily detected from genotype and alternative methods to detect mosaic UPD will be addressed in chapters 3 and 4.

As stated in the previous chapter, UPD is a known contributor to DD. The three pathogenic mechanisms of UPD are imprinting disorders, residual trisomy mosaicism, and recessive diseases. With regard to the last, isodisomy, like the autozygosity (identity by descent) resulting from consanguineous unions, provides a rich source of candidate recessive variants. For example, complete isodisomy of

chromosome 4 (191 Mb) in a proband reflects homozygosity of 6.4% of the 3 Gb-genome, which is a nearly the same proportion of homozygosity expected among offspring of first-cousin marriages (1/16, ~6.3%). Multiple mechanisms may act simultaneously; for example, isodisomy of an imprinted chromosome may lead to an imprinting disorder as well as a recessive disease. In children with DD, isodisomy is found in 0.2% of children with DD[35,37,126], whilst the frequency of heterodisomy is not well ascertained.

Isodisomy and autozygosity result in large regions of homozygosity, but the former is usually present on only a single chromosome and in a region of homozygosity larger than 10 Mb[138] or 13.5[126] Mb. Early attempts at detecting isodisomy relied on the detection of a large stretch of homozygosity in probands; however, analysing proband data in isolation may misclassify autozygosity as isodisomy, may misclassify segmental UPD as complete mixed UPD, and is blind to heterodisomy (as this type of UPD does not produce homozygous genotypes). Therefore, comprehensive and accurate UPD detection requires a different approach than using proband genotypes alone.

Alternatively, UPD can be detected from genotypes in a proband and both parents, a parent-offspring trio, by searching for an enrichment of genotypes that are only compatible with uniparental inheritance. Important advantages of this approach include the discrimination of isodisomy from inherited homozygosity, greater resolution of UPD detection, and detection of heterodisomy. Software tools have been developed for detecting UPD from SNP microarray trio data. SNPtrio is a webtool published in 2007 that accepts as input Illumina® BeadStudio or Affymetrix® CNAT SNP data and uses a test to identify statistically unlikely runs of contiguous UPD-informative genotypes[139]. A different software, UPDtool, detects non-Mendelian errors from tab-separated-value custom genotype files and classifies chromosomes with a given number of UPD-identifying genotypes as UPD chromosomes[140]. These tools share similar drawbacks: they requires inputs limited to SNP microarray software outputs or custom TSV files, they do not avoid copy number deleted regions in the proband (hemizygosity is a frequent source of false segmental isodisomy), and they use statistical approaches inherently sensitive to platform genotyping density and quality.

The genotype data used for trio genotypes can derive from SNP microarray array or sequencing data. Exome sequencing is becoming routine in rare disease studies and the variant call format (VCF[141]) is the *de facto* standard for storing sequence-

derived genotype data. Genotyping data can be stored in single-sample format, which generally records only the genomic loci that differ from the reference ('variants'), while the multi-sample format records genotypes for all samples in which any one sample varies from the reference. Combining single-sample VCF files into a multi-sample VCF file, necessary for assaying trio genotypes, can be problematic, in that a locus absent in one file but present in others may reflect a position where 1) read-data are absent (no data) or 2) read-data are available but the genotype matched the reference, and thus may be informative for UPD detection. Thus, combining single-sample VCFs requires additional data to support the inference that absence from the VCF file implies homozygous reference data (and not absence of read-data), such as accepting this inference at 1) loci overlapping target regions, which are more likely to have adequate read-coverage and 2) polymorphic positions, which have a higher prior probability for being variant in the sample. Multi-sample VCFs should theoretically be higher in genotyping accuracy as multi-sampling genotype prediction avoids the inference step (and the potential of inference errors), and may gain additional accuracy from multi-sample genotype prediction.

The sensitivity and resolution of UPD detection is inherently determined by the density, distribution, and accuracy of genotyped sites. The trio-based strategy of using informative genotypes as a signal for uniparental disomy can be polluted by hemizygous or erroneous genotypes that mimic uniparental signatures. Thus, the removal of regions overlapped by copy-number deletions could improve detection power by reducing the number of hemizygous genotypes. Maps of copy-number polymorphisms are available[142] and software tools now exist to detect CNVs from SNP microarray and exome data[6,62,143-145] for sample-specific CNV detection. Therefore, it should be possible to include CNV data to reduce the noise floor of inaccurate genotype combinations.

In order to determine whether children with DD have a burden of UPD events, a frequency estimate of UPD in generally healthy children is needed. However, the best estimate available for this rate, 1 in 3500, is based on extrapolation from the rate calculated at a single locus[121] and had not been measured empirically. In addition, knowledge of UPD frequency in children with DD is sparse because no large trio-based studied had yet been undertaken to measure both isodisomy and heterodisomy accurately in children. These considerations, as well as the hope of detecting pathogenic

UPD events that could lead to diagnosis in children in DDD motivated the development of a new UPD detection tool, UP*D*io.

UP*D*io accepts VCF-formatted trio genotypes and compares the allelic composition of proband genotypes with parental genotypes. Unlike the previously developed methods that identify consecutive runs of UPD-genotypes, this method aggregates UPD signatures on a whole-chromosomal basis, with subsequent inspection to refine the extent of the UPD. This per-chromosome binomial test can detect UPD events accurately from genotyping platforms of variable density, such as WES data, SNP data, and WGS data, without extensive platform-specific parameter manipulation. This method also avoids copy-number regions via the filtering of common CNV and sample-specific (when such data are available) CNVs, to increase statistical power. I applied UP*D*io on exome data from several thousand trios recruited for developmental disorders, in two stages. The first stage consisted of a simulation-based evaluation of the method, an implementation on 1,057 trios, and a burden analysis of UPD frequency in children with DD compared to children in the WTCCC study lacking imprinting disorders and used here as a control group. Simulations of SNP and exome data at the default p value threshold demonstrated high accuracy at detecting whole-chromosomal UPD and segmental UPD above 1 Mb for SNP data and 10 Mb for exome data. The UPD detection rate in the first stage was 0.57% (6 in 1,057; 5 complete and 1 segmental), a significant burden compared to the frequency (~0.04%) measured in healthy children. The second stage consisted of UPD detection implemented in a separate and larger set of children with DD and the detection rate in this analysis was 0.46% (15 in 3,263; 13 complete and 2 segmental). Phenotypic interpretation of the detected UPD events for each child from both stages identified UPD-associated imprinting disorders, recessive diseases, and pathogenic rearrangements.

2.3 Methods

2.3.1 Genotype segregation and statistical analysis

A site genotyped in parents and proband is considered 'informative' if it is diagnostic for uniparental or biparental inheritance.

Parent 1	Parent 2	Child	I Inheritance Type	Symbol
AA	BB	AB	Biparental	BPI
AA	BB	AA or BB	Uniparental – Ambiguous	UA
AA	AB	BB	Uniparental – Isodisomic	UI

Table 2-1 Informative genotypes for UPD analyses. Sites at which parents are opposing homozygotes and the child is heterozygous are diagnostic of biparental inheritance. Uniparental inheritance combinations include those that result only from isodisomy (UI), and those that may result from either heterodisomy or isodisomy (UA) as the proband alleles may have arisen from a duplication of one parental homologue, or may present both homologues.

Some genotype configurations supporting UPD are definitive for isodisomy (uniparental–isodisomic, i.e. UI), while others could reflect isodisomy or heterodisomy (uniparental–ambiguous, i.e. UA). That is, one class of uniparental genotype configuration is specifically informative for isodisomy (UI, uniparental–isodisomic), and the other class does not distinguish heterodisomy from isodisomy (UA uniparental–ambiguous). Heterodisomic events contain only UA genotypes and lack UI genotypes, while isodisomic events contain mixtures of UA and UI genotypes. These configurations can be further classified by maternal or paternal inheritance, reflecting a total of four uniparentally inherited signatures: $\in$ = {UI_M, UI_P, UA_M, UA_P}. Genotype configurations may also be supportive only of eudisomy, i.e., normal biparental inheritance (BPI). Note that genotyping errors can raise the 'noise-floor' by creating apparent UA and UI configurations in non-UPD chromosomes, and can obfuscate real UPD by creating BPI configurations within UPD. Additionally, copy-number deletions create blocks of hemizygosity and genotype prediction programs genotype such regions as homozygous; this results in genotype configurations that mimic UPD, and segments of such configurations can result in false UPD detections. The method filters hemizygous regions using copy number data.

The number of informative genotypes arising from maternal or paternal origin was counted for each chromosome. A binomial test was used to compare the proportion

of genotypes supporting each of the four types of UPD on each chromosome to the genome-wide average proportion for that UPD type. Those chromosomes harbouring an enrichment of UPD-type proportions were classified as UPD if they were statistically unlikely. The threshold of statistical significance used (p value of 0.000568) was based on a Bonferroni correction of an initial 0.05 alpha based on 88 tests (four different types of UPD event possible on each of 22 autosomes), a threshold demonstrated through simulation to be a sensitive and specific calibration.

2.3.2 Samples analysed

In the DDD study, proband DNA and parental DNA are genotyped genome-wide using SNP microarray and/or exome sequencing, and copy-number profiled in the proband using aCGH. The data in the first stage consisted of 1,057 trios for which all probands had aCGH CNV data available and the vast majority had genome-wide genotype data available both from SNP microarrays and exome sequencing. The second data freeze was exclusive of the first; it consisted of trio exome data for an additional 3,263 samples, and 3,196 samples had CNV data available. The samples with UPD events were recruited and phenotyped by Drs. Yanick Crow, Emma Hobson, Tessa Homfray, Sahar Mansour, Sarju G. Mehta, Mohammed Shehla, Susan E. Tomkins, and Pradeep C. Vasudevan.

2.3.3 Exome processing

Exome capture was performed as described fully elsewhere[6]. In the first stage analysis, exome sequencing genotypes were available for 937 (of 1,057; 89%) of trios. The target regions defining the exome regions, were the set from the Agilent® SureSelect v.3 50-Mb bait design and augmented with 5 Mb of custom regulatory sequences (DDD v3 Plus). Di-allelic, autosomal SNVs and indels passing quality-control filters (genotype quality at least 5, variant depth below 1,200, strand bias below 10.0) were used.

In the first stage analysis, genotype prediction was executed separately for each sample. This ‘single-sample genotype calling’ procedure outputted single-sample VCF files, which, as mentioned previously, do not contain positions that are homozygous for the reference base. To include these homozygous positions (required for deducting inheritance patterns), the assumption was made that common polymorphisms in well-covered exome-targeted regions were homozygous for the reference allele if no alternate allele was genotyped at that position. Accordingly, homozygous-reference

genotypes were annotated to positions in our VCF files if the position was contained within the inner 80% of highly covered (30 median average sequence read depth) exome-targeted regions and the minor allele frequency (MAF, based on the 1000 Genomes Project Consortium[146]) of the variant was between 0.05 and 0.95. The 'noise floor' of genotyping errors was measured by calculating the median number of the four categories of uniparental informative event types and was consistently one per chromosome. During UPD detection from SNP data, a proband with a UPD event for which no exome data had been generated was observed; exome analysis was performed for this trio *post hoc* to enable confirmatory validation of this event from exome data.

In the second stage analysis, trio VCFs were extracted from a large (13,000+) multi-sample VCF file, thus avoiding the homozygous-reference imputation procedure described in the previous paragraph. Position quality-control was conducted by selecting positions in which all trio members had a read depth of at least 8 reads, and the position was present in dbSNP[147], to exclude extremely rare variants, which are enriched for artefacts. SNP microarray chip data were not used in the second stage analysis.

2.3.4 SNP microarray data processing

Genome-wide SNP array genotypes were available for 1,041 trios analysed in the first stage. The SNP microarray platform used was a custom genotyping chip, using a backbone of 733,059 HumanOmniExpress-12v1_A-b37 positions and the addition of 94,840 selected positions. Autosomal SNPs (695,829) were used. The Sanger SNP Genotyping Core performed the genotyping, using Illuminus[148], recorded in PLINK format[149], and I converted the PLINK data to VCF format using plinkseq version 0.08. Samples were rejected on the basis of a high proportion of missing genotypes, but not due to unusually high levels of genome-wide heterozygosity, to prevent exclusion of samples that may contain UPD chromosomes. Among the 1,041 trios available, 1,035 SNP trios passed sample QC and were analyzed in this study. After UPD detection was performed in exome data, it was determined that one of these QC-failed samples in the SNP data was the father of a proband with a UPD event; this trio was processed *post hoc* to enable confirmatory validation of the UPD event in the SNP data.

2.3.5 Avoiding positions in copy-number variant regions

The diploid human genome can vary locally in copy-number, through deletions and duplications of chromosomal segments. The majority of genotype prediction software,

including the one used in this study, are ignorant to changes in copy number, i.e., they assume diploidy, and interpret hemizygosity as diploid homozygosity, which can be problematic because as single-copy loci may be spuriously identified as UPD. Therefore, the software includes a copy-number filter that avoids genotyped sites present in or near (within 10 kb) deletions common in the population or present in the sample (using user-specified CNV data encoded in VCF or tab-separated-value format).

The list of common deletions was acquired by selecting copy number variable regions of greater than 1.0% population frequency from a composite of multiple studies[150,151]. Sample-specific CNV data were generated using a custom, exome-focused, 2 million probe Agilent aCGH array and the CNV prediction software tool CNsolidate[6].

2.3.6 Simulation testing

A variety of data sets were generated to evaluate the detection accuracy of UP*D*io and to compare its accuracy with two other trio-based UPD detection methods.

To evaluate sensitivity, a maternal UPD event was introduced using maternal genotypes introduced into a single chromosome of a simulated proband. Then, the three methods were implemented using each tool's default parameters to detect maternal UPD events in a trio consisting of the original parents and the modified proband.

For simulating heterodisomy, proband genotypes were substituted for both alleles of maternal genotypes in the selected regions. For simulating isodisomy, proband genotypes were substituted for homozygosity of one of the maternal alleles, chosen at random. Complete UPD as well as segmental UPD were simulated at various sizes: 1, 2, 5, 10, and 20 Mb. Simulated regions of the required length were randomly placed across autosomes and selected unless the region overhung the edge of the chromosome or greater than 25% of its length overlapped known GRC-defined 'gap' regions. For each permutation of UPD size, class, and platform, 100 trio data sets were generated. Sensitivity was defined as the proportion of these trios with detection of the simulated maternal event by the algorithm.

For assessing specificity, empirical genotype SNP and exome data were selected from trios in which the probands had no obvious UPD events at Bonferroni-corrected p values, nor contained any large (longer than 10 Mb) regions of homozygosity. The rationale for doing so was that only genotyping errors and rare

undetected CNVs would lead to false UPD detections. Specificity was then defined as the proportion of trios lacking any maternal UPD.

The procedure described above was used to calculate UP*D*io sensitivity and specificity at various p value stringencies to construct receiver operator characteristic (ROC; true positive vs. 1-false positive rate) curves. In addition, the sensitivity and specificity of all three methods using default parameters was calculated. For UP*D*io, a Bonferroni-corrected p value threshold was used. For UPDtool, the following defaults settings were used: min_mes (300), window_size (10 kb), min_mes_fraction (1%), min_hetero (90%), min_iso (85%), min_mes_paternal (80%), and max_mes_ paternal (20%). Although SNPtrio is supported as a webtool, the investigators kindly provided the source code, which I adapted to run locally. The webtool outputs and plots all events, regardless of p value significance, and, likewise, a threshold was not imposed when running this tool.

2.3.7 Assessing pathogenic variation in samples with UPD events

The survey of candidate mutations came from four sources: 1) the UPD event itself and association with imprinting disorders[14]; 2) *de novo,* recessive and compound-heterozygous variants provided by the DDD clinical reporting pipeline ('ClinFilt') developed by Dr. Jeremy McRae and others; and for isodisomic regions, detailed inspection of 3) copy number variation data, detected from the aCGH platform and 4) rare and homozygous single-nucleotide and indel variants ('RareHomIso') contained within the VCF file for each child. The last step was required because many variants in isodisomic regions fail a ClinFilt QC-check mandating Mendelian-inheritance. In addition, heightened inspection of variants in isodisomic regions was warranted, given the enrichment of UPD events observed this study as an indication of pathogenic burden.

For the RareHomIso analysis, Variant Effect Predictor (VEP)[152] version 2.6 was used to classify mutations into the categories 'functional' (missense variant, regulatory, or splice region, inframe insertion, inframe deletion) or 'loss-of function' (splice donor variant, splice acceptor variant, stop gained, frameshift variant, stop lost). Loss of function variants in all genes and functional variants in genes implicated in DD ('DDG2P genes', https://twitter.com/ddg2p) were included for analysis.

CNV data were generated by Dr. Tomas Fitzgerald and were derived from aCGH. CNVs overlapping isodisomic regions were analysed if they represented

homozygous deletions, at least 50 kb, overlapped at least one gene, and if they passed a QC-threshold (MEANLR2 / MADL2R above 10) recommended to me by Tom. The d*e novo* variants in the clinical reporting pipeline were detected by DeNovoGear[153], executed by the DDD informatics team, and subjected to stringent algorithmic filtering and experimental validation[6].

2.3.8 Using WTCCC data to estimate UPD in the general population

The Wellcome Trust Case Control Consortium (WTCCC) is a group of research studies in the UK that investigate the genetic basis for common diseases. The WTCCC1 was a study composed of 14,000 individuals having one of seven diseases, and an additional 3,000 individuals in control groups; the data were used in this study to estimate the epidemiology of UPD in a generally healthy population of children. Genotyping was conducted by Affymetrix® using their 500K-probe SNP microarray chip (http://www.wtccc.org.uk/ccc1/overview.html). Jeffrey Barrett kindly distributed the PLINK data to me. I used a 'missing genotype' quality-control metric to remove samples with more than 10% missing genotypes. Since isodisomy is expected to affect the average rate of genomic heterozygosity, samples were not filtered based on abnormal rates of heterozygosity. A total of 16,881 individuals were included for analysis. I used PLINK (v1.07)[149] to calculate runs of homozygosity that contained at least 50 homozygous positions and spanned at least 500 kb in size. I used Perl scripts to select samples with large (larger that 10 Mb) stretches of homozygosity and identify those samples containing large regions of homozygosity affecting only one chromosome.

2.3.9 Computational performance

The UP*D*io calling method uses iterators to scan VCFs line-by-line, resulting in a low memory footprint (30 Mb of RAM per trio), regardless of genotyping density. The calling speed is reasonably quick (3 min for a SNP trio), and scales linearly with number of probes. Each trio can be run independently; therefore, the number of trios that can be analyzed simultaneously is only limited by the capacity of the data centre used to drive the tool. I wrote the UPD code using Perl v5.10.0 All required Perl modules are available on CPAN. A plotting tool is included that allows the visual display of aberrant genotypes and zygosity of the proband. Plotting scripts are adapted from the R library 'quantsmooth'[154].

2.3.10 Software availability

Software for UPD detection in trios, UP*D*io, is freely available at https://github.com/findingdan/UPDio. Instructions and pre-processing scripts are included to enable users to prepare VCF input files from custom exome capture designs.

2.4 Results

The approach to identify pathogenic UPD events is composed of three steps: 1) genotype preparation, 2) UPD detection, and 3) candidate variant selection.

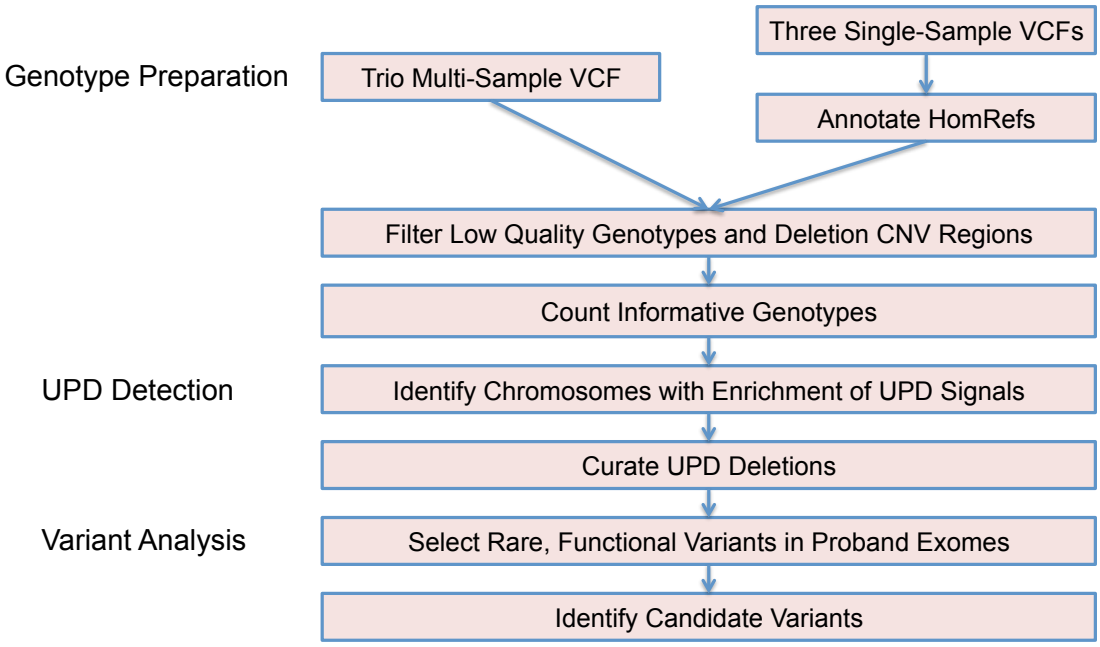

Figure 2-1 Study workflow. The study consisted of three main steps: data preparation, UPD detection, and candidate variant analysis. In the data preparation stage, informative genotypes were collected in all members of each trio. Either a multi-sample trio VCF or three single-sample VCFs can be used as input; the latter requires the annotation of homozygous reference genotypes, not usually encoded in single-sample VCF files. In the UPD detection stage, trios were selected containing a proband chromosome with an enrichment of UPD-informative genotypes. Exomes available for samples with a detected UPD event were selected for the candidate workup analysis, in which rare protein-altering variants were reported that may manifest in the proband's phenotypes.

Genotype preparation begins with pre-processing the genotype data from SNP microarray or exome sequencing data. Data pre-processing is critical and includes three steps: 1) creating trio VCF files; 2) removal of low-quality genotypes; 3) removal of genotyped sites within CNVs.

For the exome data analysed in the first stage analysis, trio VCF files were created from single-sample VCF files, and homozygous reference genotypes were imputed (see Methods Section 2.3.3). To assess imputation accuracy I assessed the correlation in genotype dosage among 1,369,049 QC-passed sites from 50 samples genotyped by SNP and exome platforms and the correlation was extremely high ($r = 0.9958$), suggesting the imputation procedure was robust to error. Among the 937 trios

analyzed by exome, the per-trio average of genotype positions in which all members of the trio were jointly genotyped was 54,394 positions, of which 3,619, on average, were informative, yielding an average density of informative exome sites per megabase of 1.2 (3,619 * 1e6 / 3e9). In the SNP microarray data, an average of 42,490 sites per trio were informative. Thus, the average density of informative SNP genotypes across one megabase was 14.2 (42,490 * 1e6 / 3e9). The median number of the four categories of uniparental informative event types was consistently zero per chromosome.

The exome trios in the second-stage analysis were generated from a large multi-sample VCF file so the homozygous reference imputation step was not required. Based on a calculation involving 100 trios, the per-trio average number of informative positions was 4,923, yielding an average density of informative exome sites per megabase of 1.6 (4,923 * 1e6 / 3e9). The median number of the four categories of uniparental informative event types was 1.5 per chromosome, a low noise-floor. The density of informative sites was 50% higher in trios extracted from the multi-sample VCF compared to combining single-sample VCFs. Thus, even though imputation was robust to accuracy, avoiding imputation recovered 50% more sites.

After pre-processing, the proband genotypes diagnostic of uniparental or biparental inheritance were counted on each chromosome. Uniparental genotypes could be quantitatively distinguished from one another by the relative proportions of the two different classes of genotype configurations that were diagnostic for uniparental inheritance (Table 2-1), or qualitatively by visualization.

2.4.1 Simulations

Simulations were used to assess the accuracy of UPD calling in UP*D*io (see Methods). The sensitivity of UPD detection was measured at a range of sizes (1, 2, 5, 10, and 20 Mb) to test detection rates of segmental UPD and chromosome-wide, to test detection of complete UPD. Simulations were performed for heterodisomy and isodisomy from data generated by exome and SNP microarray platforms (Figure 2-2).

The method was more sensitive for detecting isodisomy than heterodisomy; this was expected given that the former generates more informative sites (both UA and UI combinations). Also, the method was more sensitive at a given size using SNP microarray data than using exome data, primarily due to both the greater density of genotyped sites, with a possible minor contribution from the likely higher genotype accuracy in SNP microarrays. At Bonferroni-adjusted significance threshold (light-blue

line, p value of 0.000568), near perfect sensitivity in SNP microarrays data was observed for detecting either class of UPD event (heterodisomy or isodisomy) at 5 Mb. At 2 Mb, 98% of isodisomy and 91% of heterodisomy could be detected. Sensitivity of isodisomy detection from exome data was 99% for isodisomy and 75% for heterodisomy at 10 Mb.

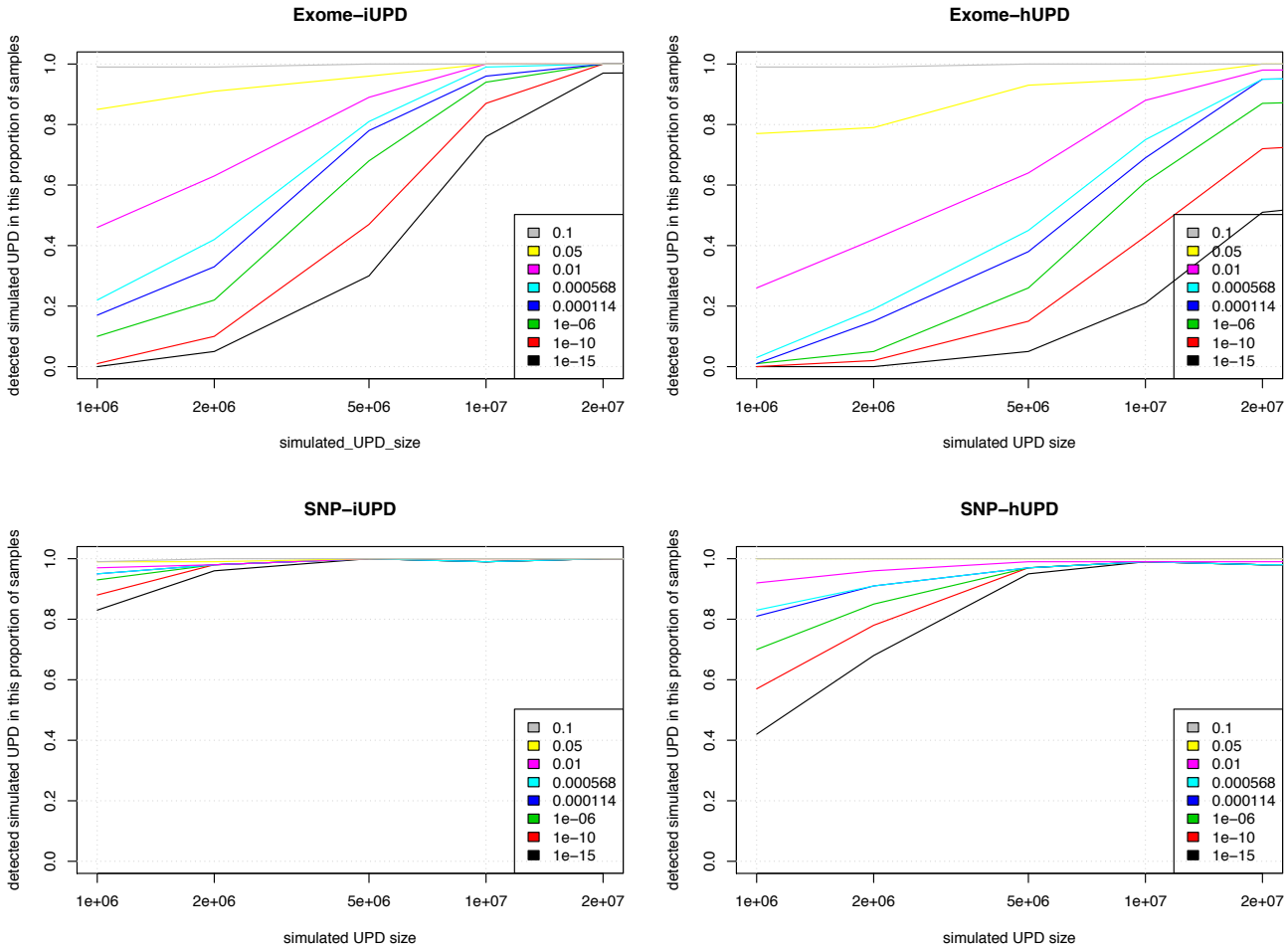

Figure 2-2 Sensitivity of UPD detection simulations. Simulations to assess sensitivity of UPD detections at different sizes, from different data sources. (iUPD) isodisomy; (hUPD) heterodisomy

Specificity was defined as the proportion of tested non-UPD trios that lacked maternal UPD calls. At the Bonferroni-adjusted p value of 0.000568, specificity was 99% for exome data and 100% for SNP data. The cause of the single false-positive UPD event was found to be due to a slight excess of genotype errors resulting in an event called with a significant p value (p value of 0.00044, close to the Bonferroni-adjusted p value cut-off).

Given that a size threshold for suspecting UPD in clinical molecular diagnostics is typically near 10 Mb[36], the successful detection of UPD of this size is of

practical utility. Indeed, even 2 Mb isodisomic events were detected accurately from SNP microarray data, a result likely due to low genotyping error rates and relatively uniform genotyping density; although at this size, the accuracy of detection of heterodisomy from SNP microarray data, and isodisomy and heterodisomy from exome data, was appreciably lower.

2.4.2 Comparing UPD detection software tools

I compared the strengths and limitations of three trio-based UPD detection tools, SNPtrio, UPDtool, and UP*D*io (Table 2-2).

	SNPtrio	**UPDtool**	**UP*D*io**
Platform Source	SNP only	Cross platform	Cross platform
Genotype Input Format	TSV from SNP software	Custom TSV	VCF
Integrated CNV filtering	No	No	Yes
Statistical Method	Binomial test per block	Sliding window over blocks of Mendelian errors	Binomial test per chromosome
Statistical Confidence Measure	p value	Fractions of event types	p value
Dynamic Platform Independent Calibration	No	No	Yes
Visualization	UPD & CNVs	Event fractions	Yes, UPD & zygosity
Accepts compressed files	No	No	Yes
Language	Perl, R	C#	Perl, R
Run Environment	Webtool	Windows & Linux	Linux
Performance	51 seconds / 265 Mb	15 seconds / 65 Mb	151 seconds* / 21 Mb

Table 2-2 Software comparisons. Comparing three trio-based UPD software tools. TSV (tab separated value). *total run time including parsing input files, CNV filtering, and UPD detection.

There are substantial differences in the interface, statistical methods, calibrations, and outputs of these three tools. One notable difference is the input format requirements. UPDtool requires the construction of custom tab-separated-value genotype files, while SNPtrio processes SNP-genotyping software output files, and UP*D*io reads VCF files, which is a platform-independent standard file format for genotype data. The underlying statistical methods vary as well. UP*D*io is the only tool that integrates CNV filtering during genotype parsing, which occurs before statistical

analysis. In terms of calling confidence, UP*D*io and SNPtrio provide a p value output measurement, while UPDtool does not provide a confidence score for its UPD detections. For threshold calibration, the webtool SNPtrio accepts a parameter 'minimum number of SNPs in an event region'; UPDtool has a list of seven adjustable parameters (min_mes, window size, min_mes_fraction, min_hetero, min_iso, min_mes_paternal and max_mes_paternal); and lastly, UP*D*io allows for user control of the p value threshold as a single parameter. Neither SNPtrio nor UPDtool parameters are recalibrated dynamically based on input data but are tuned for platforms resembling the density and noise characteristics of high-density SNP trios. In contrast, UP*D*io calculates a per-chromosome proportion-based statistic, which is innately normalized for input data of different global density and genotyping error rates.

Simulations assessed the comparative accuracy of three trio based UPD detection tools: SNPtrio, UPDtool, and UP*D*io (Figure 2-3). All three platforms were run using default parameters, on the same simulated data sets (reformatted to accommodate each tool's input requirements). Sensitivity results were tabulated as the proportion of tested samples with maternal UPD detection on the chromosome containing the simulated event.

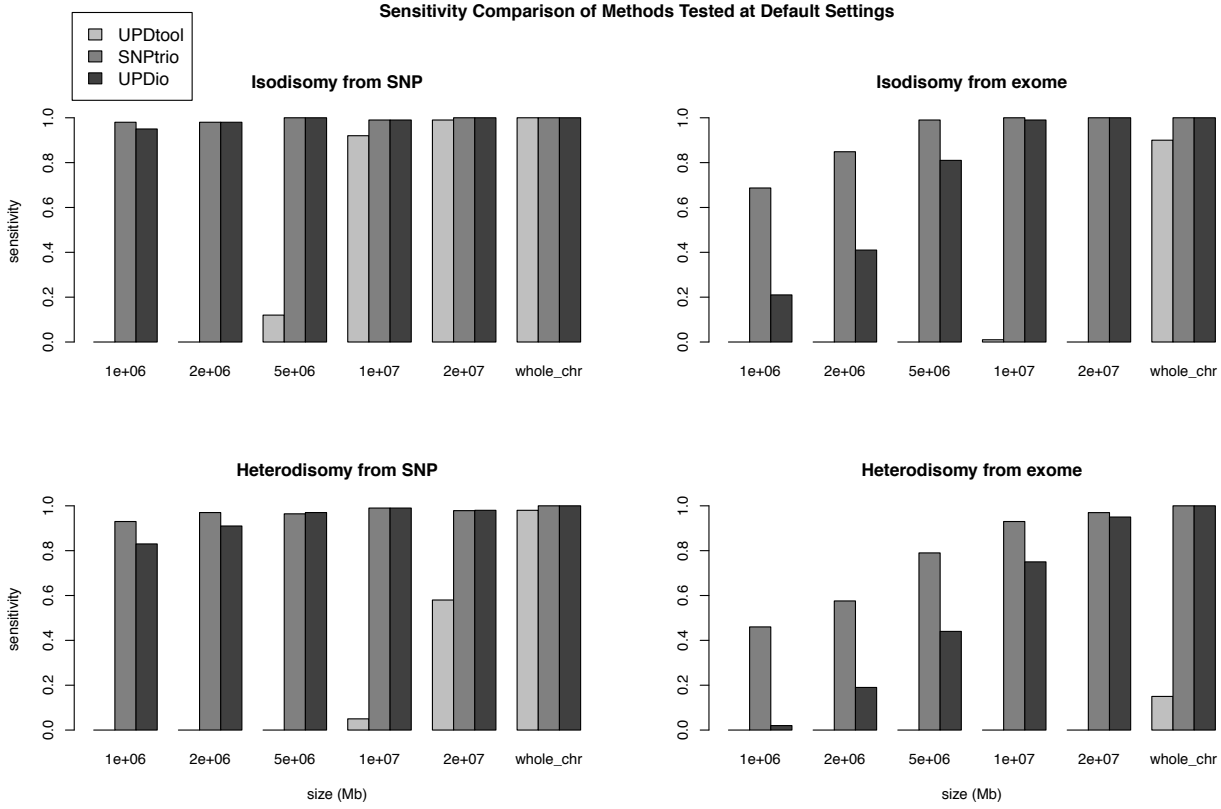

Figure 2-3 Sensitivity comparisons. Simulations were performed to measure the sensitivity of detecting introduced UPD events from SNP and exome data, ranging in size from 1 Mb to chromosomal.

Specificity was calculated as the proportion of samples not containing maternal UPD events in samples without obvious UPD events (Figure 2-4).

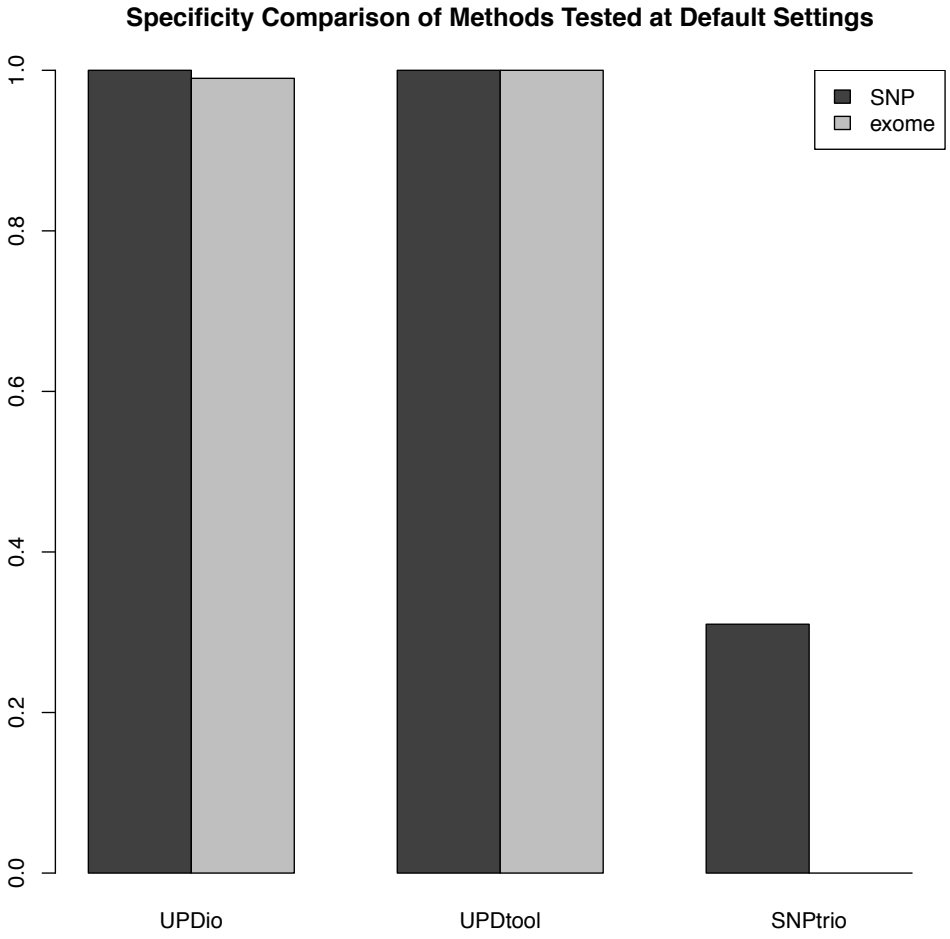

Figure 2-4 Specificity comparisons. Simulations on normal SNP and exome samples were compared to measure the proportion of samples without UPD detections.

Simulation results demonstrated that SNPtrio was the least specific algorithm (31% for SNP data and ~0% for exome data), and UPDtool was the least sensitive tool, capable of detecting only the very largest UPD events. Unsurprisingly, specificity and sensitivity were inversely related. UPDtool was 100% specific, and made no false UPD assignments in normal samples from either SNP or exome data. UP*D*io was nearly as specific as UPDtool. SNPtrio was the most sensitive, which was most evident in the detection of smaller heterodisomic events from exome data. UP*D*io was only very

slightly less sensitive than SNPtrio for events 10 Mb and greater in size in exome data and for events 1 Mb and greater in size in SNP data.

Receiver operator characteristic (ROC) curves were used to evaluate the calling performance of UP*D*io at various p value thresholds (Figure 2-5).

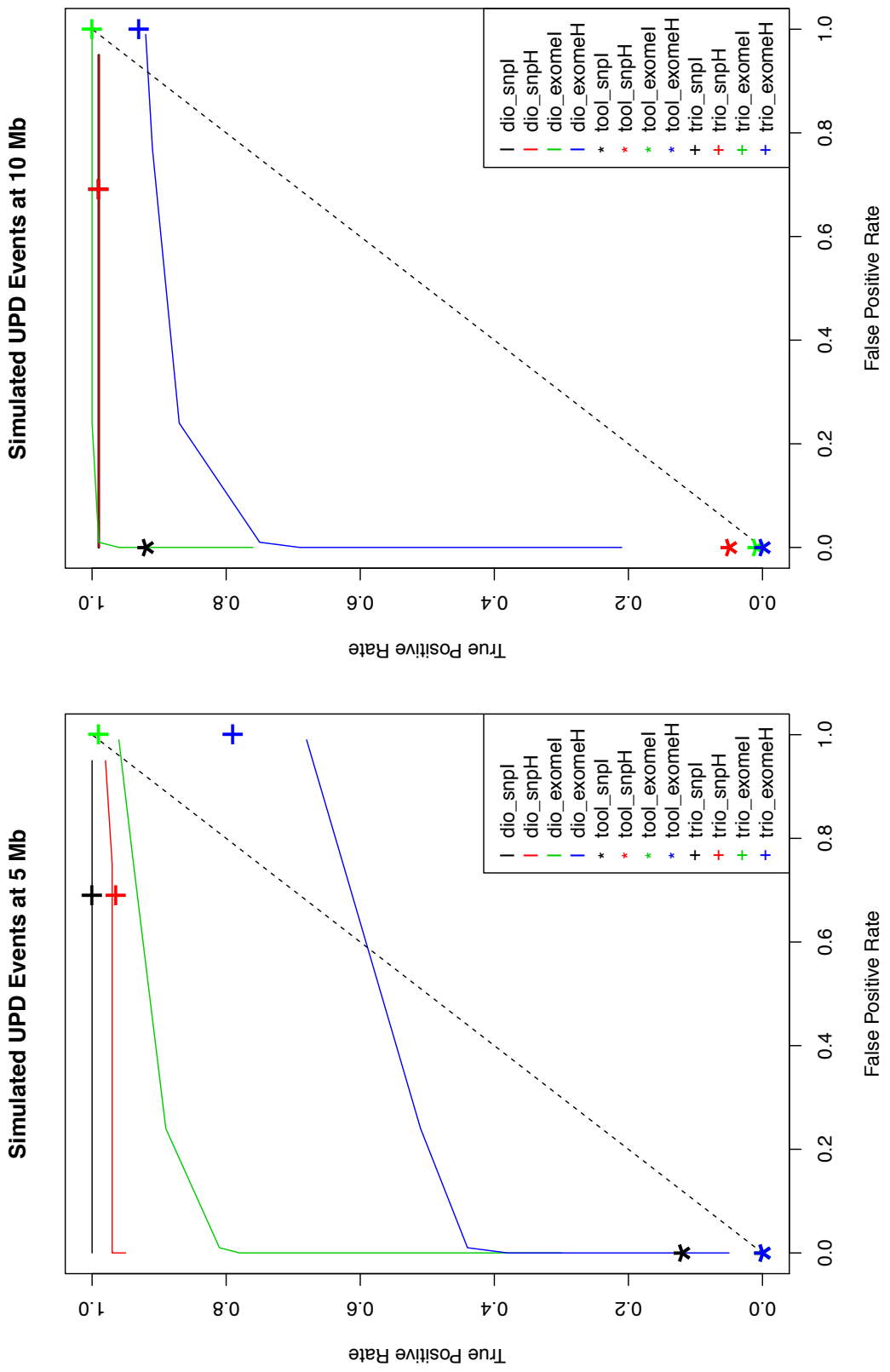

Figure 2-5 Receiver operator characteristic curve comparing UPD detection accuracy at different simulated UPD sizes. (dio) UP*D*io, (tool) UPDtool, (trio) SNPtrio.

The UP*D*io curves demonstrated excellent classification of UPD events from SNP platform at 5 Mb and 10 Mb. The classification of UPD events from exome data was noticeably weaker, especially for detection of heterodisomy at a size of 5 Mb. The Bonferroni corrected p value of 0.000568 represented a good balance of sensitivity and specificity for both data types and both classes of UPD event. Thus, this p value was used as a default parameter for UPD calling in UP*D*io.

For the two ROC curves the classification performances of UPDtool ('tool') and SNPtrio ('trio') were plotted for the calculated sensitivity and specificity of these programs at their default parameter settings. While most SNPtrio classifications demonstrated high true-positive rates, these came at the expense of very high false-positive rates that would require substantial additional downstream manual filtering such that large-scale application is inherently limited. On the other hand, UPDtool performance was characterized by low true-positive rates, near zero for most event types and platforms, with the notable exception of isodisomy from SNP data at a size of 10 Mb. In contrast, UP*D*io, using the default p value threshold, detected a substantially higher ratio of true to false events compared with the other programs under all conditions. These differences are likely to be accentuated when implementing these tools for whole-genome sequence data sets.

UP*D*io was tested on WGS HapMap child-mother-father trio (NA12878, NA12891, NA12892) and CNV data[155]. Whole-genome analysis counted an average of 278 informative genotypes per Mb, 20x greater density than our SNP platform, required 9 min and 27 Mb of memory and detected no UPD events beyond marginal significance.

2.4.3 Implementing quality control of UPD detections

In the first stage analysis, UPD detection was implemented on 1,057 unique DDD parent-offspring trios. The majority (915) of these trios were analyzed by both SNP and exome data, with slightly more trios available from SNP data (1,035) compared with exome data (937). A p value of 0.000568 was used as a statistical threshold (see section 'Genotype Segregation and Statistical Analysis' in Methods) for identifying putative UPD events for further investigation. The putative UPD events had calculated p values that were bimodal in distribution (Figure 2-6).

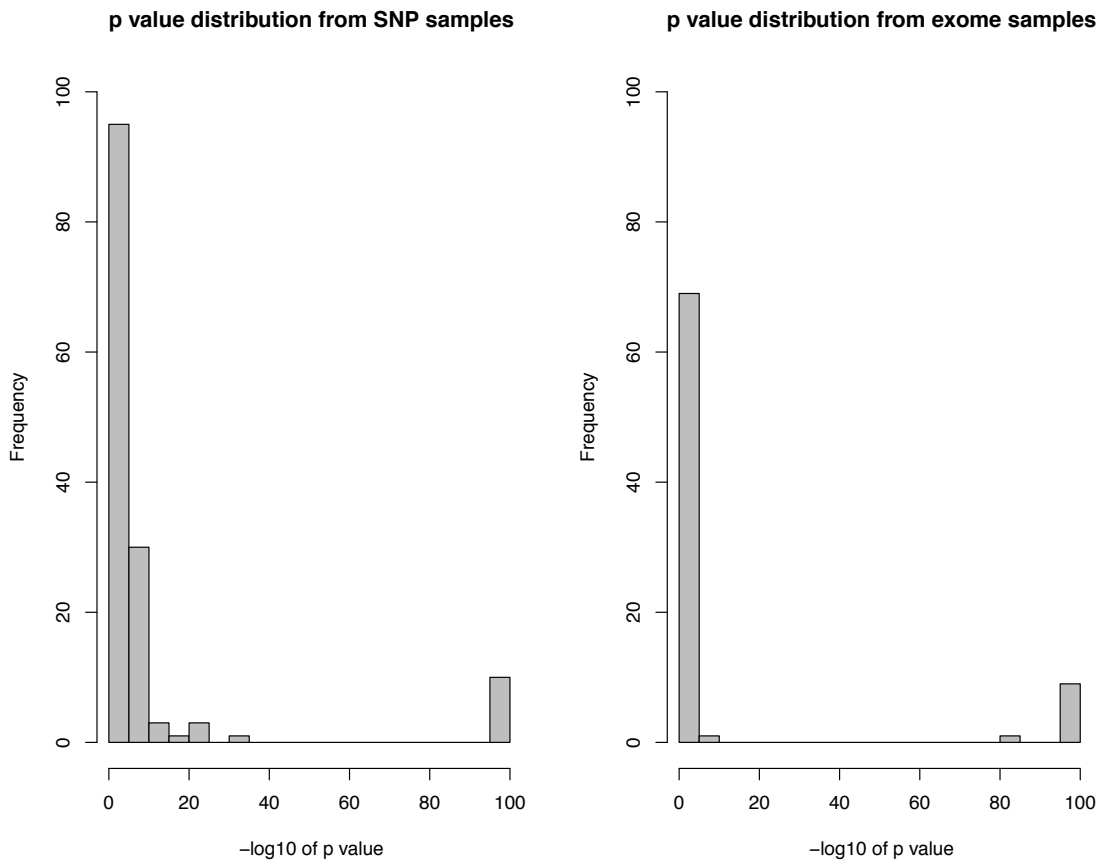

Figure 2-6 DDD UPD p value distributions. Distribution of the –log10 p values for UPD detections from different data sources, with or without CNV data. Presence of sample-specific CNV data increases the proportion of extremely significant events and decreases the proportion of events with p values less significant than 1e-10. significant events. p value minimum truncated to 1e-100.

The extremely significant events were considered authentic UPD detections on the basis of having consistent UPD signatures on a single chromosome; these were selected for further analysis, and validated, as described below.

I investigated the less-significant group of detections and observed differences between the two platforms regarding the number and underlying cause of these spurious events. The SNP data had 133 such events while the exome data had 70 such events. The underlying cause of these false detections in the SNP data usually (80% of the time) was due to misattribution of undetected (and thus unfiltered) CNV regions as isodisomy. This was especially true for the most significant events of this category; for example, a 1 Mb deletion (which escaped detection by aCGH due to low-quality array data) resulted in false signals of high significance (UI_P at 1e-31 and UA_P of 1e-22). In contrast, the underlying cause in the exome data in most (85%) cases was due to stochastic fluctuations of genotyping errors. The disparity between SNP-detected and exome-detected spurious events likely reflects underlying platform differences, namely

that the SNP platform has far greater genotyping density, especially in noncoding regions, thus is more prone to detecting hemizygous genotypes within small deletions than the exome data, while the exome data (from single sample calling) has a slightly higher genotyping error rate, and is therefore more susceptible to the random aggregation of genotyping errors.

Large UPD events have substantial numbers of both UI and UA events. Consequently, binomial tests assessing the enrichment of both event types often redundantly detect these large UPD events by both signatures. I developed a visualization tool to illustrate the distribution of informative sites along each chromosome in a trio to clarify the type and extent of these events, which may include both isodisomy and heterodisomic regions (Figure 2-7).


22
21
20
19
18
17
16
15
14
13
12
1
2
3
4
5
6
7
8
9
10
11
Chromosome
UA_P
UI_P
UA_M
UI_M
BPI
Het
Hom

Figure 2-7 Example of a UPD plot. A plot of QC-passing proband genotypes on each autosome. The position and colour reflect zygosity (homozygous, heterozygous) and informative state (biparental inheritance, maternal isodisomy, maternal heterodisomy or isodisomy, paternal isodisomy, paternal heterodisomy or isodisomy). The figure displays each chromosome ideogram. Each chromosome has an x-axis (chromosome position) and y-axis (zygosity, and informative event type). In this case, the UPD event for chromosome 2 is depicted with a mixture of dark-green points (maternal isodisomy) and light-green points (maternal isodisomy or maternal heterodisomy). The zygosity row demonstrates homozygosity along the entirety of the chromosome, reflecting the complete isodisomy.

In addition, the method provides additional output files to specify all informative genotype events comprising the UPD region.

The p values of the putative UPD detections in the second stage analysis were plotted and the shape of the distribution was bimodal, as seen in the first stage analysis (Figure 2-8). Inspection of events less significant than 1e-10 identified similar artefacts as seen in the first stage analysis. Inspection of all events with p values more significant than 1e-10 identified a small number of spurious UPD events (chance aggregation of uniparental sites on a chromosome along with BPI probes) and a single event with a p value of 1e-24, which was due to hemizygosity (an undetected deletion). All events more significant than 1e-24 were real UPD events.

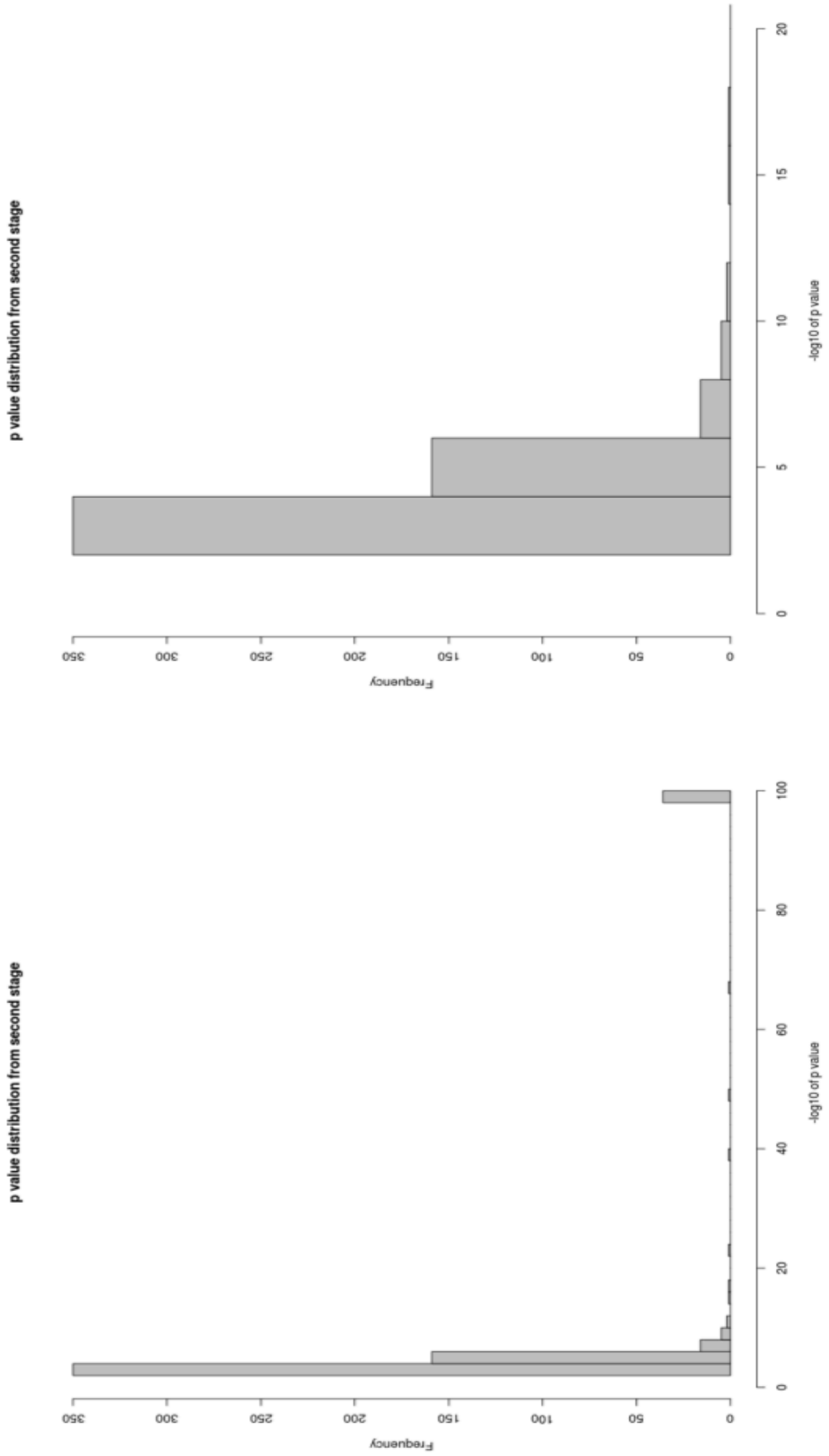

Figure 2-8 Distribution of –log10 p values for UPD detections in the second stage analysis. p value minimum truncated to 1e-100. The vast majority of candidate UPD calls are at low significance and cluster below 1e-10. The second graph depicts the events more significant than 1e-20.

2.4.4 UPD detections

UPD detection was executed in two stages and the results from both stages are provided below (Table 2-3).

In the first stage, there were six probands with UPD events. All events were cross-validated, that is, detected using both SNP data and exome sequence data. The six events comprised a variety of UPD events.

In the second stage, there were 16 probands with at least one extremely significant (more significant than 1e-12) putative event type. One event passing this level of significance, with a p value of 1e-24, was found to reflect a copy number deletion event undetected by CNV calling. The remaining 15 probands each had a single chromosome with a UPD event of 1e-40 or more significant.

The majority (16 of 21) of the detected UPD events were maternally derived. Eighteen of 21 were complete UPD. There were 11 isodisomies, 3 heterodisomies, and 7 mixed events. In 7 of 21 cases, the UPD chromosome appeared on a chromosome that has been associated with imprinting disorders and in two cases, appears on maternal chromosome 16, which is controversially associated with imprinting[125]. Of the eight UPD events detected in this study that were entirely or mostly heterodisomic, 7 of 8 were on a chromosome associated with imprinting disorders.

ID	-log10 p val	UPDchr	size	homologue-pattern	origin
258308*	323	17	complete	isodisomy	maternal
260453*	323	9	complete	isodisomy	maternal
259010*	323	2	complete	isodisomy	maternal
261229*	323	14	complete	mixed (80/20 h/i)	maternal
258370*	323	1	complete	isodisomy	paternal
257814*	313	1	segmental 12Mb	isodisomy	maternal
270667	162	1	complete	isodisomy	paternal
273472	49	1	segmental 8Mb	isodisomy	maternal
277020	179	2	complete	mixed (50/50 h/i)	maternal
266581	136	4	complete	mixed (30/70 h/i)	maternal
273401	162	7	complete	isodisomy	maternal
271037	67	11	segmental -6Mb	isodisomy	maternal
265596	248	14	complete	heterodisomy	maternal
265472	216	15	complete	heterodisomy	maternal
277316	289	15	complete	mixed (75/25 h/i)	paternal
271552	314	15	complete	mixed (75/25 h/i)	paternal
264527	226	16	complete	mixed (75/25 h/i)	maternal
271631	297	16	complete	mixed (75/25 h/i)	maternal
266931	119	17	complete	isodisomy	paternal
271839	154	22	complete	heterodisomy	maternal
264255	102	22	complete	isodisomy	maternal

Table 2-3 Summary table of first stage (samples with a *) and second stage detections. h/i: heterodisomy/isodisomy.

2.4.5 Investigating UPD frequency

Compared with the widely quoted birth prevalence of UPD (1/3,500)[121] the proportion of UPD events detected in the trio analyses (21/4,032) is significantly higher (binomial test p value 1.21e-19). The UPD rate at birth in the general population has been estimated on extrapolation from clinically relevant UPD events at a single locus, and thus is potentially susceptible to variation among chromosomes in UPD rate. To generate an empirical estimate of the population prevalence of all classes of UPD would require dense genome-wide genotypes for tens of thousands of parent–offspring trios

sampled randomly from the population; such data are not currently available. However, it is possible to estimate the rate of uniparental isodisomy from dense genome-wide genotypes on unrelated individuals since isodisomy manifests with an easily detectable signature: a long region of homozygosity. Identity by descent processes, such as consanguinity[156] or cryptic relatedness[157] similarly generate long regions of homozygosity, but are distinguishable from isodisomy because these other processes often involve multiple chromosomes and are rarely longer than 20 Mb[156].

A total of 16,881 samples from the Wellcome Trust Case Control Consortium (WTCCC) data set were used to develop an empirical estimate of the rate of complete uniparental isodisomy by observing the number of samples containing a single chromosome burden of large regions of homozygosity. First, PLINK[149] was used to identify large (>10 Mb) tracts of homozygosity for each sample, and retained samples with a large homozygous region or regions confined to a single chromosome. There were many (103) samples, which satisfied this criterion. Of these, only a single sample appeared to have whole-chromosomal isodisomy, but a further five samples had significant homozygosity that extended over at least half of the chromosome. These five samples comprised four telomeric events on chromosomes 4, 21, 22, 22, and one on chromosome 4 with two large interstitial regions of homozygosity. As the homozygosity of these events covered the majority of the chromosome and represents the only major tract of homozygosity in these genomes, these events were considered likely to reflect mixtures of isodisomy and heterodisomy and less unlikely to reflect inherited homozygosity. Under the conservative assumption that all these chromosomes reflect complete uniparental disomy of a chromosome in these individuals, this represents a frequency of 6 uniparental disomy events in 16,881 (0.036%) individuals, which is not significantly different from the reported frequency of 1 in 3,500 (0.029%, binomial test p value of 0.4934). Notably, by enforcing the same criteria to define a UPD event (the majority of the chromosome homozygous and large homozygosity confined to a single chromosome), there were twelve such UPD detections in DDD. This reflects a proportion ten times greater and significantly enriched compared with the population estimate (binomial test p value of 4e-9); additionally, this proportion is significantly enriched compared with the WTCCC data (Fisher exact test, p value of 1.5e-5).

The WTCCC data were used to investigate the prevalence of segmental UPD, however, despite stringent filtering of sub-chromosomal segments of homozygosity, the expected pattern of terminal segmental UPD events was not detected[132]. Therefore, most of the regions of segmental homozygosity in the WTCCC were not likely reflective of segmental UPD events and estimating prevalence of segmental UPD events from this data set was not undertaken. Analyses of segmental UPD, which are typically mosaic[131], are better suited to algorithms that interrogate the b allele frequency, rather than genotype data.

2.4.6 Investigating pathogenicity in children with UPD events

A fully comprehensive understanding of pathogenic variation in each child with a detected UPD event requires an in-depth analysis that is well beyond the scope of this dissertation. The genetic basis of disease in children with detected UPD events may be fully, partially, or not explained by the UPD event. Still, the enrichment of UPD observed in this study suggests that most of these UPD events are pathogenic, providing a target to focus candidate variant assessment. I analysed the UPD chromosome as a source of pathogenic variation and also included variants that were identified in the DDD clinical reporting pipeline (see Methods 2.3.7). Note that residual trisomy represents an additional source of UPD-associated pathogenicity and whilst the UPD events presented in this chapter were not later associated with mosaicism, the possibility of hidden residual mosaicism cannot be excluded. Mosaic structural variation is addressed in detail in chapters 3 and 4.

To summarise the results detailed below (Table 2-4, Table 2-5), of 4,320 children investigated, a UPD event was discovered in 21 children. In 14 cases, the UPD chromosome provided the best source of pathogenic candidates, including seven UPD events associated with imprinting syndromes. In one case, the best candidate variant was a *de novo* mutation not located on the UPD chromosome. In the remaining cases, no strong candidate variants were detected. I now describe in greater detail the genotype-phenotype associations for these 21 child patients.

2.4.6.1 UPD chromosome is the dominant source of candidate variant(s)

In three patients (1-3), UPD detection identified UPD events on imprinting-associated chromosomes for which NHS-investigation had already uncovered the UPD events and provided diagnosis. Patient 1 (ID273401) had Silver-Russel Syndrome, patient 2

(ID277316) had Angelman Syndrome, and patient 3 (ID265472) had Prader-Willi Syndrome.

For patients 4-6, the child's phenotypes were most consistent with imprinting syndromes but the child had not yet been diagnosed. Patient 4 (ID265596) had a maternal UPD of chromosome 14, a UPD event that causes Temple Syndrome. Most of the listed phenotypes listed in DECIPHER for this individual – intrauterine growth retardation (IUGR), generalised hypotonia, feeding difficulties in infancy, motor delay and frontal bossing – are consistent with Temple Syndrome[158]. There were no other genetic abnormalities detected in the child.

Patient 5 (ID261229) had maternal UPD of chromosome 14. Temple Syndrome (maternal UPD14) is the primary source for most of the child's phenotypes, including truncal obesity (weight 99th centile), moderately short stature (height first centile), and mild intellectual disability[158], while the diabetes mellitus phenotype is likely attributed to the metabolic consequences of the disorder (BMI 38; class II obesity). In addition, the child has sensorineural hearing impairment, which has not been reported as a sign of Temple Syndrome. This proband had novel compound heterozygous variants - a missense substitution inherited from the mother and a stop gained mutation inherited from the father - in the *TECTA* gene. *TECTA* encodes an extracellular matrix protein (tectorin alpha) of the tectorial membrane, the surface of the sensory epithelium of the cochlea[159], and is a well known cause of autosomal dominant (OMIM 601543) and autosomal recessive (OMIM 603629) hearing loss. Neither parent has a documented hearing disability, suggesting that the compound heterozygosity has resulted in the recessive form of hearing loss in the child. Recently, a hearing-impaired proband with normal-hearing parents was found to contain compound heterozygous variants (missense and splicing mutation leading to truncated protein) in the *TECTA* gene, which was indicated to be definitely pathogenic through *in vitro* functional characterisation[160]. Thus the phenotypes in this child are best explained by considering both the imprinting syndrome on the UPD chromosome in addition to the recessive-mediated hearing loss caused by a mutation on a different chromosome.

Patient 6 (ID271552) had a paternal UPD of chromosome 15, a UPD event causing an imprinting syndrome called Angelman syndrome. Most of the child's features -- sleep disturbance, severe developmental delay, and characteristic dysmorphic features -- are consistent with Angelman syndrome. In addition, the child has a rare

(MAF of 0.00028) homozygous splice-acceptor variant in gene *DUOX2*, a gene for which homozygous stop mutations have been associated with congenital hypothyroidism (CH)[161]. Abnormal sleep patterns and intellectual disability are seen in Angelman syndrome as well as in CH, so it is possible that CH may explain some of the child's signs. It is not clear if the child was screened for CH; if not, clinical investigation of thyroid hormone level may be warranted, and any disturbances medically treated.

For the remaining patients, the UPD events are not closely associated with imprinting syndromes. For patient 7, the UPD chromosome is related to a pathogenic rearrangement, and for patients 8-14, the best candidate mutations are recessive candidates in isodisomic regions.

Patient 7 (ID257814) had a maternal segmental UPD on chromosome 1. Investigation of copy number abnormalities in this sample identified a 12-Mb *de novo* triplication event flanking the UPD event. In collaboration with Carvalho *et. al*, we showed that the UPD and flanking triplication resulted from a replication-induced DNA repair mechanism, microhomology-mediated break-induced replication (MMBIR)[162]. This large rearrangement was considered definitely pathogenic and the finding returned to the patient and family.

For the following patients, the UPD event is considered likely pathogenic through conversion to homozygosity by isodisomy of a variant inherited from a parent who was heterozygous as this locus (a carrier). Patient 8 (ID266581) had maternal UPD of chromosome 4 with dysmorphic features and cardiac abnormalities: flat occiput, low-set ears, short philtrum, impaired ocular abduction, bilateral ptosis, overlapping fingers, deep palmer creases, short thumb, pulmonary artery stenosis, and abnormalities of the heart valves. The child had two rare homozygous mutations at isodisomic regions on the UPD chromosome, a suspected loss-of-function splice acceptor variant in the *IDUA* gene with MAF of 0.00056 and a missense variant in the *IGFBP7* gene. Hurler syndrome is a recessive disease due to loss-of-function mutations in *IDUA* and causes a severe disease, with some features that are consistent with the child's presentation although the child does not appear to have hepatosplenomegaly, which is common in this syndrome. This variants was considered uncertainly pathogenic nevertheless merits additional investiation. A biochemical assay for excess mucopolysaccharides in urine is diagnostic and may be warranted for this child pending further clinical evaluation.

Enzyme replacement therapies are currently in use for Hurler syndrome so clinical assessment should be pursued.

Patient 9 (ID258308) had UPD of chromosome 17. This child had delayed developmental milestones, growth retardation, microcephaly, and suffers from seizures intractable to medical intervention. She was found to have decreased serum magnesium and renal magnesium wasting but genetic testing for diseases of renal hypomagnesium wasting (*TRPM6* and *SCN1A* gene testing) was normal. Her seizures did not resolve after intravenous magnesium infusion and resulting restoration of blood magnesium to normal range, suggesting that hypomagnesaemia alone is not the cause of her seizures. An MRI showed grossly normal cerebral architecture. The child has three variants in DDG2P disease genes (*PGAP3*, *SCN4A*, *CCDC40*), all in isodisomic regions of chromosome 17. Two of these genes are strong candidates for follow-up. Recessive mutations in *PGAP3* result in 'hyperphosphatasia with mental retardation syndrome 1[163]' and the child has a very rare (0.0006 MAF) missense mutation in this gene. The child also has a very rare (0.0012 MAF) missense SNV in *SCN4A*, a gene that encodes a subunit of a voltage-gated sodium channel. This sodium channel is implicated in a diversity of neuromuscular disorders, such as periodic paralysis and myotonia congenita, diseases that mimic seizure disorders[164,165]. While channelopathies often follow a dominant mode of inheritance[166], recessive modes have been seen as well[167], and several genes encoding channel proteins are known to underlie severe seizure disorders, such as *KCNQ2* (Ohtahara syndrome)[168] and prologues of *SCN4A*, such as *SCN1A*[169], *SCN2A*[170], and *SCN9A*[171]. These two mutations are the best candidates in this child. In addition the child has homozygous stop-gained mutations in *CCDC40*, a gene associated with ciliary dyskinesia, but the child's phenotypes do not match this disease.

Patient 10 (ID264255) is a male patient with dyslexia and progressive pes cavus. The UPD chromosome is 22, maternally inherited, and the isodisomic interval contains a homozygous rare (MAF of 0.00012) stop-gained mutation in the *SBF1* gene. This gene is associated with a recessive form of Charcot-Marie-Tooth syndrome, type 4B3, a disease associated with pes cavus and distal neuropathy. However, this gene is not in the DDG2P set, presumably because most forms of Charcot-Marie Tooth do not appear until early adulthood. Family history reports pes cavus in the father, suggesting that the child's pes cavus may be related to an inherited paternal variant, however, the mutation was maternally inherited. Suspicion that a sample swap between parents may

have occurred was disabused after inspection of the number of mapped reads to chromosome Y showed that the labelled father and labelled mother were male and female, respectively (data not shown). The inconsistency between shared phenotypes and the origin of the *SBF1* variant raises doubt to the pathogenicity of the mutation.

Patient 11 (ID271037) has a 16.2 Mb telomeric segmental UPD of chromosome 11, of maternal disomy. The child has several abnormalities, including nystagmus and developmental delay. No known imprinting disorder arises from 3' telomeric disomy of chromosome 11. However, in the isodisomic region of chromosome 11, the child has a homozygous, rare (MAF of 0.00012) missense variant in *ROBO3*. Homozygous missense variants of this gene have been implicated in 'gaze palsy with progressive scoliosis', a condition that may be consistent with the child's nystagmus. However the child has other phenotypes, such as vesicouteral reflux, hypotelorism, joint hypermobility, and posteriorly rotated ears, which appear to represent syndromic dysmophology; therefore, the variant has uncertain pathogenicity.

The best disease candidates for patients 12 through 14 were in isodisomic intervals but the relationship between these mutations and each child's phenotypes is more tenuous. Patient 12 (ID266931) has paternally inherited disomy of chromosome 17. His phenotypes include ID, oral dysmorphology and obesity. The child "may have had 1 or 2 words at 1 year old, now none". In the isodisomic UPD region, the child has a homozygous rare (MAF of 0.0048) missense variant in *NAGS*, a gene in which frameshift mutations have been associated with N-acetylglutamate deficiency[172], a urea cycle disorder, which results in regressive phenotypes. Nevertheless, the effect of missense mutations on this gene is not well known and the variant was considered of uncertain pathogenicity.

Patient 13 (ID270667) has a uniparentally inherited disomy of chromosome 1. The child has aganglionic megacolon, microcephaly, ID, ventricular septal defect and pulmonic stenosis, and short stature. The child has several (9) homozygous missense and loss of function variants on the UPD chromosome. Notable variants include a rare (MAF of 0.00098) homozygous missense variant in *CAMTA1*, a gene which has been associated with DD and constipation, the latter, a phenotype which may be reflective of abnormalities in peristalsis. The child has a rare (MAF of 0.0002) homozygous splice region variant in *FLG*, a gene associated with a ichthyosis vulgaris, and a rare (0.003) homozygous missense variant in *ASPM*, a gene associated with microcephaly, a rare

(0.006) homozygous missense variant in *PARP1*, a gene associated with mental retardation. These variants have uncertain pathogenicity.

Patient 14 (ID260453) had complete isodisomy of chromosome 9. This is a 15-yr-old male patient with developmental delay and intellectual disability, recruited following noninformative aCGH CNV analysis. His family history was notable for having several second-degree family members with similar phenotypes. The child also has a congenital heart defect. As the clinical features were relatively common among children with congenital disorders, it was more challenging to use phenotypic matching to identify specific genetic candidates in this patient. The child has rare functional variants in four DDG2P disease genes (*CDK5RAP2, LAMC3, HNRNPU, ROBO3*), two of which (*CDK5RAP2* and *LAMC3*), lie in isodisomic regions. *CDK5RAP2* is associated with recessive microcephaly, but the child's head circumference is not grossly abnormal (5^{th} centile). *LAMC3* is associated with cortical malformations; the child had a normal MRI. Another candidate is the *de novo* missense mutation in *HNRNPU*, a gene on chromosome 1 listed in DDG2P as a 'possible DD gene'. This *de novo* variant is well supported by sequencing data (11 of 22 sequence reads in proband and absent in well-covered parents). The variant has never been seen before in the DDD study; it is exceedingly rare.

2.4.6.2 Non-UPD chromosome is the dominant source of candidate variant(s)

Patients 15 (ID277020) had a UPD event detected on chromosome 2. She exhibited short stature, microcephaly, moderate global developmental delay, delayed skeletal maturation. The child had heterozygous missense variants in five DDG2P genes (*GRHL3, POGZ, FLNB, ELN, SCN8A),* which were in the DDG2P gene list and were very rare. The best candidate mutation is the *FLNB* gene[173], a gene on chromosome 3 in which missense mutations are associated with a dominant disease of skeletal development, Larsen syndrome. According to DECIPHER, parents share a similar phenotype but it is not listed which phenotype is shared.

2.4.6.3 Variants with uncertain pathogenicity

Patient 16 (ID259010) had maternal UPD of chromosome 2. This is a 7-yr-old male patient, with a complex phenotype profile including global developmental delay, glandular hypospadias, overriding toe and bicuspid aortic valve. Recently, a female child, also with maternal UPD of chromosome 2 and complex phenotype, distinct from our patient, had been exome sequenced and many (18) candidate variants were

identified on the UPD chromosome, none reported to be likely pathogenic[174]. None of that girl's phenotypes is coincident with this patient, suggesting that an imprinting disease is not the likely cause of the diseases in these children. There were no strong candidates in this child. There were three variants in DDG2P disease genes (*EIF2AK3, AGXT, ALMS1*), all on the isodisomic UPD chromosome, were observed. *EIF2AK3* is the cause of Wolcott-Rallison Syndrome, which is not consistent with this child's phenotypes. *AGXT* is the cause of hyperoxaluria but this child does not have kidney stones. Defects in *ALMS1* are a cause of Alstrom Syndrome, but this child does not have multiorgan dysfunction.

There were two children, patients 17 (ID271631) and 18 (ID264527), with maternal UPD of chromosome 16. Both UPD events had relatively small regions of isodisomy (only about 25% of the chromosomes), and no candidate mutations were present in these isodisomic regions, which may suggest that the UPD event is pathogenic but not through recessive causation. Maternal UPD of chromosome 16 is inconsistently associated with abnormalities, although intrauterine growth retardation may be common, children with UPD maternal 16 have "variable outcome from almost normal to only growth retardation and rarely to malformation and/or mental retardation"[175]. Given the inconsistency of the phenotypes between these children and the tenuous association of imprinting abnormalities with chromosome 16, these UPD detections have uncertain pathogenicity; additionally, there were no strong recessive or *de novo* candidates in these children. Female patient 17 (ID271631) exhibited IUGR, pulmonic stenosis, GERD, drooling, talipes equinovarus, overfriendliness, and coordination abnormalities and has a *de novo* frameshift mutation in the *DDX3X* gene on the X chromosome, a gene associated with X-linked recessive mechanism of DD in males; however the consequences of a heterozygous mutation in this gene in females is not documented. Male patient 18 (ID264527) had a low birth weight (-2.14 standard deviations) suggestive of intrauterine growth retardation but had several severe phenotyepes (including autism, aphasia, global developmental delay) suggesting an underlying genetic syndrome not explained solely by the UPD event.

Patient 19 (ID271839) had a UPD on chromosome 20. He had an arachnoid cyst, clinodactyly of the 5th finger, conductive hearing impairment, epicanthus, global developmental delay, hypertelorism, rhizomelic short stature, tetralogy of fallot, triangular mouth, uplifted earlobe. The child has a *de novo* 'splice region' mutation in *SCRAP* a gene causing very rare Floating-Harbor syndrome, which also causes

clinodactyly, short stature, and some similar facial phenotypes. However, pictures were not available on DECIPHER to assess phenotypic concordance and the ‘splice region’ variant was considered as a variant of uncertain pathogenicity.

Patients 20 (ID273472) and 21 (ID258370) had UPD events on chromosomes not associated with imprinting disorders, had no homozygous variants in DDG2P genes that remained after clinical filtering, and no isodisomic variants.

ID	mut_type	chr	pos	gene	maf	gt	cq	fun
270667	UPDchr:1							
270667	RareHomIso	1	7798367	CAMTA1_yes	0.000976	1/1	missense	fn
270667	RareHomIso	1	68564392	GNG12-AS1_no,WLS_no	0.000488	1/1	frameshift	lof
270667	RareHomIso	1	92756989	GLMN_yes	0.002483	1/1	missense	fn
270667	RareHomIso	1	152287956	FLG_yes,FLG-AS1_no	0.000244	1/1	splice_region	fn
270667	RareHomIso	1	156693150	ISG20L2_no	0.000122	1/1	frameshift	lof
270667	RareHomIso	1	197060077	ASPM_yes	0.002897	1/1	missense	fn
270667	RareHomIso	1	226550829	PARP1_yes	0.006468	1/1	missense	fn
270667	RareHomIso	1	227152761	ADCK3_yes	0.001655	1/1	missense	fn
270667	RareHomIso	1	227152778	ADCK3_yes	0.003586	1/1	missense	fn
258370	UPDchr:1							
273472	UPDchr:1							
259010	UPDchr:2							
259010	RareHomIso	2	73786275	ALMS1_yes	0.000122	1/1	splice_region	fn
259010	RareHomIso	2	88883014	EIF2AK3_yes	0.005793	1/1	missense	fn
259010	RareHomIso	2	241817472	AGXT_yes	0.000488	1/1	missense	fn
277020	UPDchr:2							
277020	ClinFilt	1	24673119	GRHL3_yes	0.000854	1,0,1	missense	fn
277020	ClinFilt	1	151400289	POGZ_yes	0.000414	1,1,0	missense	fn
277020	ClinFilt	3	58118639	FLNB_yes	0.000732	1,1,0	missense	fn
277020	ClinFilt	7	73474862	ELN_yes	.	1,1,0	missense	fn
277020	ClinFilt	12	52099216	SCN8A_yes	.	1,1,0	missense	fn
266581	UPDchr:4							
266581	RareHomIso	4	994668	IDUA_yes	0.000552	1/1	splice_acceptor	lof
266581	RareHomIso	4	57976289	IGFBP7_yes	0.000138	1/1	missense	fn
260453	UPDchr:9							
260453	RareHomIso	9	123171581	CDK5RAP2_yes	0.000122	1/1	missense	fn
260453	RareHomIso	9	133932355	LAMC3_yes	0.000138	1/1	missense	fn
260453	ClinFilt	11	124745468	ROBO3_yes	0.001655	1,0,1	missense	fn
260453	ClinFilt	11	124746198	ROBO3_yes	0.007811	1,1,0	missense	fn
260453	DeNovo	1	245027192	HNRNPU_yes	0	0/1	missense	fn
271037	UPDchr:11	imprinting						
271037	RareHomIso	11	124739427	ROBO3_yes	0.000122	1/1	missense	fn
264527	UPDchr:16	imprinting						
271631	UPDchr:16	imprinting						
271631	ClinFilt	X	41205794	DDX3X_yes	.	1,0,0	frameshift	lof
271631	CNVs	16	28326710	28391016	64306	del		
271631	DeNovo	11	6652911	DCHS1_yes	0	0/1	missense	fn
271631	DeNovo	X	41205794	DDX3X_yes	0	0/1	frameshift	lof
258308	UPDchr:17							
258308	RareHomIso	17	37824754	PGAP3_yes	0.000552	1/1	missense	fn
258308	RareHomIso	17	62018952	SCN4A_yes	0.001655	1/1	missense	fn
258308	RareHomIso	17	78021155	CCDC40_yes	0.006345	1/1	stop_gained	lof
266931	UPDchr:17							
266931	RareHomIso	17	42082405	NAGS_yes	0.004828	1/1	missense	fn
271839	UPDchr:22							
271839	DeNovo	16	30745810	SRCAP_yes	0	0/1	splice_region	fn
264255	UPDchr:22							
264255	RareHomIso	22	50903104	SBF1_no	0.000122	1/1	stop_gained	lof

Uniparental Disomy

265472	UPDchr:15	imprinting						
265472	ClinFilt	8	6372298	ANGPT2_no	0.004393	2,1,1	missense	fn
257814	UPDchr:1							
257814	CNVs	1	11860126	20573006	8712880	dup		
277316	UPDchr:15	imprinting						
277316	DeNovo	4	159627433	ETFDH_yes	0	0/1	missense	fn
273401	UPDchr:7	imprinting						
261229	UPDchr:14	imprinting						
261229	ClinFilt	11	121000407	TECTA_yes	0.000122	1,0,1	stop_gained	Lof
261229	ClinFilt	11	121008311	TECTA_yes	0.000122	1,1,0	missense	Fn
271552	UPDchr:15	imprinting						
271552	RareHomIso	15	45392428	DUOX2_no	0.000276	1/1	splice_acceptor	lof
271552	ClinFilt	8	144994508	PLEC_yes	0.000138	1,0,1	missense	fn
271552	ClinFilt	8	144999571	PLEC_yes	0.000414	1,1,0	missense	Fn
265596	UPDchr:14	imprinting						

Table 2-4 Investigating candidate variants, including UPD events, de novo variants, variants passing clinical filtering, recessive variants and CNVs. Fn: functional, lof: loss-of-function. _yes and _no suffix refers to presence or absence in DDG2P gene set.

Decipher ID	Phenotypes from Decipher
257814	Cutaneous finger syndactyly, 2-3 toe syndactyly, Short nose, Epicanthus, Bilateral single transverse palmar creases, Wide intermamillary distance, Abnormality of the skin, Delayed speech and language development
258308	Seizures, Seizures, Bruxism, Global developmental delay, Delayed speech and language development, Delayed gross motor development, Renal magnesium wasting, Hypomagnesemia
258370	Short attention span, Moderately short stature, Joint hypermobility, Impaired T cell function, IgG deficiency, Slow-growing hair, High anterior hairline, Abnormality of the nasal tip, Abnormality of the skeletal system, Hypermetropia
259010	Glandular hypospadias, Overlapping toe, Bicuspid aortic valve, Global developmental delay, Meckel diverticulum, Eczema, Gastroesophageal reflux
260453	Abnormality of the heart, Global developmental delay, Specific learning disability, Abnormality of prenatal development or birth
261229	Abnormality of macular pigmentation, Truncal obesity, Intellectual disability mild, Sensorineural hearing impairment, Moderately short stature, Diabetes mellitus, Abnormality of the toenails
264255	Periventricular gray matter heterotopia, Microcephaly, Pes cavus, Abnormality of the skeletal system, Delayed speech and language development, Myopia, Specific learning disability, Generalized keratosis follicularis, Achilles tendon contracture
264527	Hemihypertrophy of lower limb, Deeply set eye, Moderate global developmental delay, Absent speech, Autism spectrum disorder, Hypospadias
265472	Delayed speech and language development, Generalized neonatal hypotonia, Moderate global developmental delay
265596	Intrauterine growth retardation, Cryptorchidism, Generalized hypotonia, Oligohydramnios, Feeding difficulties in infancy, Large fontanelles, Relative macrocephaly, Motor delay
266581	Flat occiput, Sparse scalp hair, Low-set ears, Bilateral ptosis, Broad lateral eyebrow, Short philtrum, Abnormality of the nose, Abnormality of the lip, Infantile muscular hypotonia, Wide intermamillary distance, Deep palmar creases, Deep plantar creases, Abnormality of the heart valves, Overlapping fingers, Neonatal respiratory distress, Global developmental delay, Short thumb, Congenital laryngeal stridor, Asymmetry of the thorax, Peripheral pulmonary artery stenosis, Bicuspid aortic valve, 11 pairs of ribs, Impaired ocular abduction
266931	Intellectual disability, Aplasia cutis congenita of midline scalp vertex, Low hanging columella, Downturned corners of mouth, Obesity
270667	Aganglionic megacolon, Microcephaly, Intellectual disability moderate, Low anterior hairline, Broad thumb, Synophrys, Ventricular septal defect, Pulmonic stenosis, Proportionate short stature
271037	Vesicoureteral reflux, Nystagmus, Moderate global developmental delay, Hypotelorism, Plagiocephaly,

	Broad forehead, Sacral dimple, Joint hypermobility, Low-set posteriorly rotated ears
271552	Severe global developmental delay, Sleep disturbance, Horizontal eyebrow, Deeply set eye, Prominent nose, Clinodactyly of the 5th finger
271631	Pulmonic stenosis, Intrauterine growth retardation, Gastroesophageal reflux, Drooling, Talipes equinovarus, Abnormality of coordination, Overfriendliness
271839	Rhizomelic short stature, Tetralogy of fallot, Arachnoid cyst, Global developmental delay, Periauricular skin pits, Clinodactyly of the 5th finger, Preauricular skin tag, Nevus flammeus, Hypertelorism, Epicanthus, Uplifted earlobe, Abnormality of the helix, Triangular mouth, Conductive hearing impairment
273401	Intrauterine growth retardation, Postnatal growth retardation, Broad forehead, Asymmetric growth, Global developmental delay, Small face
273472	Jaundice, Global developmental delay, Tall stature, Truncal obesity, Brachycephaly, Abnormality of skin pigmentation, Hypotelorism, Abnormal number of incisors, Joint hypermobility, Pes cavus, Specific learning disability
277020	Short stature, Microcephaly, Moderate global developmental delay, Delayed skeletal maturation
277316	Umbilical hernia, Mild global developmental delay, Protruding tongue, Uplifted earlobe, Drooling, Brachycephaly, Tall stature

Table 2-5 Phenotypes recorded in Decipher for each of the children with detected UPD events.

2.5 Discussion

In this chapter I described the development and implementation of UP*D*io, a new software tool to detect uniparental disomy from exome sequence data. UP*D*io has unique advantages compared with existing trio-based UPD detection programs for mitigating the effect of genotype errors and heterozygous deletions. First, genotype errors have the potential to over-segment UPD calling in SNPtrio and UPDtool, tools that detect runs or blocks of UPD, but have little effect on disrupting the per chromosome rate of informative genotypes, the metric used by UP*D*io. Second, SNPtrio and UPDtool are vulnerable to false isodisomy created by hemizygous regions in the proband, while UP*D*io has an integrated CNV filter to avoid common CNV and user-specified sample-specific CNV regions before the binomial test is applied. Since deletions generate genotypic signatures identical to isodisomy, this step is essential to prevent the unintentional ascription of deletions as UPD. UP*D*io enables users to remove these erroneous signatures from UPD analyses using data from a single platform, by providing sample-specific CNVs in BED[176] or VCF format. In addition, the statistical test applied in UP*D*io intrinsically adjusts for differences in platform genotyping density, which varies in orders of magnitude between exome data, SNP data, and whole-genome data. Also, only UP*D*io outputs a measure of statistical confidence, a p value that can be calibrated by the user to achieve the desired sensitivity and specificity. Only UP*D*io can read single-sample and multi-sample VCF files, the modern genotype file standard, and thus can be more easily assimilated as a module into existing pipelines. While UPDtool was the fastest method of the three tested, UP*D*io performs additional processing to cleanse poor-quality genotypes and avoid copy number regions; nevertheless, it completes UPD calling on high-density SNP trio data in under three minutes, and is the least memory intensive of the three methods for detecting UPD events. In fact, memory efficient iterator functions enabled UP*D*io to process a whole-genome trio using less memory than either of the competing programs used to process a SNP microarray trio.

The relative accuracy of the three trio-based UPD calling software was compared using each tool's default parameter settings on the same set of simulated data. Marked differences in the sensitivity and specificity of these three software tools were observed. The practical utility of SNPtrio is greatly hampered by its lack of specificity, whereas UPDtool exhibited very low sensitivity, was only capable of detecting the very

largest of simulated UPD events, and would miss most small UPD events. In contrast, using default parameters, UP*D*io was sensitive and specific for simulated UPD events at 1 Mb from SNP data and 10 Mb from exome data, with broadly equivalent sensitivity to SNPtrio. There are several factors that likely account for these dramatic differences in calling accuracy. Probably the most important factor is due to the need to finely calibrate SNPtrio and UPDtool, which use statistical approaches that are more vulnerable than is UPDio to platform differences in genotype density and genotype error rates. Unfortunately, unlike UP*D*io, SNPtrio and UPDtool do not offer a convenient user-adjustable threshold of statistical threshold, such a p value.

In this study, the sensitivity for detecting smaller UPD events was lower for trios in exome data primarily because the number of informative sites genotyped was approximately 10x fewer, although other factors, such as less even distribution and slightly higher genotyping error rate may have been contributory. The use of multi-sample VCF files in stage two of the analysis increased the number of assayed sites, by 50% on average, compared with the use of single-sample VCFs, which was likely in part to the recovery of rare variants in the proband, which had been excluded, in the first stage analysis. Nevertheless, the detection sensitivity measured by simulations was 100% for whole-chromosomal UPD events, and was sensitive for most simulated segmental events at the 1 Mb level in SNP data and the 10-Mb size for exome data. This size is clinically relevant as non-trio-based studies of UPD typically only investigate potential UPD when regions of homozygosity exceed 10 Mb[36].

Smaller UPD events, such as those affecting 1 Mb in size, are challenging to detect due to a paucity of informative genotypes. For example, SNP microarray data contain on average only 14 informative genotypes per megabase window. Still, with high-quality genotypes, the occurrence by chance of 14 contiguous UPD characteristic genotypes is a very unlikely event, and the previously developed contiguous runs of informative genotypes method may be marginally more sensitive than the proposed method at detecting events at this size. However, the contiguous runs method is also more likely to be sensitive to small runs of UPD-mimicking genotypes occurring by chance across the whole genome, lowering specificity. Moreover, smaller UPD events are less likely to be pathogenic and are much more likely to be mosaic[107], implying that alternative UPD detection approaches, based on BAF of proband genotypes, would be more appropriate for segmental UPD events.

I implemented UPD detection with UP*D*io on 1,057 unique trios in the first stage of analysis and UPD was detected in six probands. Using UP*D*io, all six UPD events were easily called from both platforms yielding highly significant p values in both SNP and exome data. Given this finding and the simulation results, this suggests that exome-based trio designs are appropriate to detect UPD, without the requirement to run SNP microarrays specifically for this purpose. In the second stage of analysis, 15 UPD events were detected among 3,263 children. Among all UPD events, eight were at least 75% heterodisomic, and would have likely escaped detection using a proband-only homozygosity approach for detection.

All segmental UPD cases were isodisomic, consistent with mitotic loss of one allele and reduplication of the remaining allele. The most common reported mechanism underlying UPD is trisomy rescue[122], which suggests that that meiotic non-disjunction is the most common generating mechanism of UPD. Meiotic non-disjunction most often occurs in maternal meiosis I[177]. The association of trisomy rescue and maternal non-disjunction predicts that the majority of heterodisomic and mixed UPD events should be maternal in origin; concordant with this prediction, 8 of 10 such events were maternally derived. Complete isodisomy can originate from a monosomy compensated for by reduplication, or by a trisomy rescue event of chromosomes that had not undergone recombination. In this study, 3 of 11 complete isodisomies were paternally derived and 8 of 11 were maternally derived. Given that meiotic non-disjunction is more common in females, the former may likely reflect monosomic eggs rescued by reduplication, while the latter may likely represent trisomic eggs with non-recombinant chromosomes which underwent trisomy rescue.

The rate of UPD abnormalities in the studied children was 0.5%, a statistically increased rate (p value of 10^{-19}), and represents a 20-fold enrichment compared to population prevalence estimates. There are several explanations that could cause the high rate seen in this study: 1) a high false-positive rate in UPD detection in DDD, 2) the estimation of UPD prevalence in the population is an underestimate and the DDD study has higher prevalence of benign UPD by chance alone, 3) some of the UPD events are disease causing. There is over-whelming statistical evidence of UPD in the six cases from two independent platforms, suggesting that 1) is not the explanation. To address the question of whether UPD prevalence in the population has been underestimated an empirical estimate the rate of UPD using SNP microarray data on unrelated individuals from the Wellcome Trust Case Control Consortium was

performed. There are limitations to this approach, mainly that it is indirect (only can identify UPD by observing single-chromosome large runs of homozygosity, not directly from the inheritance patterns of individual genotypes), and confounded by other causes of large runs of homozygosity, such as identity by descent, identity by state, or loss of heterozygosity. Notwithstanding these limitations, previous prevalence estimates about uniparental disomy in the human population are compatible with these observations. Therefore, the suggestion that some individuals with UPD in our study may have UPD-related disorders warrants further investigation.

I examined several sources of genetic variation to identify the basis of disease in children with detected UPD events. In 14 of 21 cases, the UPD chromosome provided the best source of candidate pathogenic variants. These included seven UPD events associated with imprinting syndromes. One UPD event was associated mechanistically with a pathogenic 10 Mb triplication. In at least one case, disease was best explained by the contribution of both a UPD event (causing the imprinting syndrome Temple Syndrome) and a mutation elsewhere on the chromosome (a compound heterozygous mutation causing deafness). Exome analysis provided a rich source of plausible candidate variants for a follow-up investigation, especially in isodisomic regions, as such regions convert to homozygosity an allele inherited from a carrier parent, a precarious genetic phenomenon prone to cause recessive diseases. For seven patients (8-14), the best candidates were located in isodisomic regions of UPD chromosomes. In two cases, strong candidate *de novo* mutations, not located on the UPD chromosome, were identified. Previous analysis has found that *de novo* SNV mutations are the most common mutations causing disease in undiagnosed DDs; therefore, it would not be surprising if mutations of this class were identified in some of the isodisomic UPD cases. Experimental follow-up is required to definitively implicate these novel variants with disease causation.

The ascertainment of patients in this study, whom are only recruited once clinical genetics services have failed to obtain a diagnosis, may bias against the discovery of UPD events that result in a well-recognized imprinting or recessive disorders for which routine diagnostic assays are available. Given the broad range of recessive and imprinted phenotypes associated with UPD, its detection should be a part of the genetic analysis for disease studies more broadly, as it is a small, but important piece of the puzzle of pathogenic genomic variation.

As sequencing technologies continue to increase the cost-effectiveness of genome-wide sequencing data, the ability to interrogate UPD will improve. The tool presented here efficiently scales as files are read line-by-line without storing large data hashes, thus making efficient use of memory. Although UPD detection is fundamentally limited to a resolution on the scale of tens of kilobases, defined by the density of informative genotype configurations in the parents. In addition, the availability of sequence data enables the exploration of sequenced-based methods as an orthogonal approach for the detection of mosaic UPD, and mosaic structural rearrangements, which, due to incomplete aneuploidy rescue and mitotic recombination, are closely associated. Chapter 4 presents the investigation of using exome and whole-genome sequence data for the detection of large mosaic abnormalities. But first, mosaic structural variation using SNP microarray is discussed in chapter 3.

3 MOSAIC STRUCTURAL VARIATION FROM SNP MICROARRAY

3.1 Publication Note

Most of the work described in this chapter was previously published earlier this year[178]. Unless explicitly stated otherwise, the analysis described herein is the work I performed myself, under the supervision of Matthew Hurles.

3.2 Introduction

Rearrangements of genomic structure, termed structural variation, consist of copy-number and copy-neutral events. Pathogenic structural variation is the cause of genomic disorders[179]. As discussed in chapter 2, constitutive copy-neutral UPD is enriched in children with DD and can be detected from trio genotypes but mosaicism distorts allele fraction, which confounds genotype prediction and hinders the detection of mosaic copy-*neutral* variation from predicted genotypes. In addition, mosaic copy-*number* variation is not typically detected using genotypes but results in deviation of allele fraction. SNP microarray data enable access to a quantitative measure of allele fraction, the b allele frequency, which, compared to categorical genotype data, defines with more granularity the mixture of alleles underlying mosaic structural abnormalities. This chapter discusses the use of SNP microarray data in identifying mosaic copy-number and copy-neutral abnormalities, primarily using deviation in allele fraction.

The detection sensitivity for mosaic abnormalities is a function of several parameters, some of which are intrinsic to the mosaic event – including event size, clonality, type (i.e. loss, gain, LOH); others which are technology dependent – including platform (e.g. karyotyping or microarray), number of molecular probes, signal to noise ratio of molecular probes; and others which are algorithmic (e.g. single-sample vs. trio-based tests).

Mosaicism can involve multicellular clonality for mutations of any size[180,181]. Reliable detection of small-scale mosaicism requires sequencing data of very high depth. Generating such data may be feasible to interrogate specific genes for mutations suspected in rare disease and cancer[182,183] but it is prohibitively expensive for genome-wide screening. In contrast, large-scale genomic variation can be detected using karyotyping and microarray analysis. In this study, I focussed on mosaic events of at least 2 Mb in size, a generally accepted threshold for large structural alterations[184], allowing a fair basis of comparison for the different chip designs I analysed, and concordant with a recent study that used a SNP microarray design and algorithmic protocol similar the platform used in the DDD study[50]. Henceforth in this chapter, the term *mosaicism* will refer to mosaic events of at least 2 Mb in size.

Mosaicism of low clonality is difficult to detect because there is a low proportion of abnormal cells, reducing the mosaic signal. While karyotyping is still widely used in many clinical centres, this approach is insensitive to sub-microscopic rearrangements and small supernumerary marker chromosomes[185], and is labour-intensive, since, for example, 30 cells must be counted to exclude 10% mosaicism with 95% confidence[26]. Compared to karyotyping, SNP microarray offers a higher-resolution, higher-throughput assay and has been proposed as a standard of care for clinical diagnostics in children with developmental disabilities[101]. The resolution of SNP microarray for mosaicism detection is influenced by probe density and the signal to noise ratio of the experiment and the type of mosaic abnormality.

The SNP platform generates quantitative measures of summed allelic intensity, the log R ratio (LRR), and of allele balance, the B-allele frequency (BAF). When genetic heterogeneity exists in assayed cell populations, the BAF deviates from expected diploid frequencies (B_{dev}) and algorithmic approaches translate B_{dev} into mosaic detections. These approaches generally calculate B_{dev}, then cluster B_{dev} values using a segmentation step, and then use a quality-control step to identify deviations that are significant. For example, in the analyses recently presented by Laurie *et al.*[130] and

Jacobs *et al.*[50], B_{dev} is calculated, segmentation is performed by CBS[186] or GADA[187], and quality control is performed by a automated curation (filtering of constitutive abnormalities based on the bivariate distribution of BAF and LRR in putative segments) and manual curation of remaining putative detections. A similar approach has been used to detect structural variation in tumour-normal admixture using ASCAT[188], a mosaic detection tool for tumours, which uses a tumour-normal sampling approach to identify informative mosaic loci, uses piecewise constant fitting[189] for segmentation, and then uses a grid search to identify the most likely tumour ploidy and clonality that fit the data. Mosaic Alteration Detection[49] (MAD), introduced in chapter 1 of this dissertation is the software tool that was used by Jacobs *et al.* as the primary engine for mosaic detection. As a review, MAD is a popular software tool that identifies segments as described above and then uses the average LRR value in each segment to classify segments into mosaic type: loss, gain, or loss of heterozygosity. The detection sensitivity for MAD on SNP microarrays with approximately 1 million probes for events at least 2 Mb in size has been estimated to be limited to loss or LOH events in about 10%-90% of cells and gain events in about 20%-80% of cells[49,50].

The B_{dev} calculation is based on the absolute value of the difference of BAF from expected allele fraction. However, detection power can be improved if phased genotype data are available, since it can then be shown that BAF consistently deviates towards one parental haplotype, which is less likely to occur by chance alone. Phasing can be imputed based on reference haplotypes when dense (high resolution SNP microarray) genotyping data are available. For example, a haplotype-aware upgrade of ASCAT (the 'Battenberg' algorithm) was recently reported[190], and J-LOH, an HMM-based approach also for tumour-normal SNP data, was recently published[191]. When proband-parent trio data are available, proband genotypes can be phased directly, an approach avoiding imputation error, and yielding higher quality haplotype prediction. triPOD[51] is a trio-based mosaic detection tool that leverages parental genotype data to phase child genotypes, and has been shown to have increased sensitivity, compared to MAD, for detecting events below approximately 10% clonality, but this trio-based method requires parent genotype data, which are not always available.

Recent investigation using MAD in 60,000 adults who lacked rare genetic diseases showed a positive correlation between mosaic frequency and sample age, with frequency of mosaic events rising after the age of 45[50]. In children with DD, the

frequency of LOH mosaicism was estimated at 0.26%[35], while the frequency of CNV mosaicism, based on an average of three studies, was estimated at 0.56%[192-194]. Combining these rates yields a frequency of 0.82%. Conlin *et al.* detected a higher rate, 1.1%[36] (Table 3-1). One plausible explanation for this higher rate is that one third (8 of 23) of the events detected in the Conlin *et al.* study were XX/X0 mosaics, the cause of Turner syndrome[195], a disease causing short stature and amenorrhoea, phenotypes which may not be appreciated until children reach adolescence. Such children are unlikely to have been enrolled in the other studies or DDD study, which typically assess children with more severe diseases and congenital abnormalities.

	Platform	Variation type	No. of Probes	Tissue	No. of Samples	No. of Mosaics	Frequency (%)
Bruno[35]	Illumina HumanCytoSNP-12	LOH	220k	blood, skin biopsy, saliva	5,000	13	0.26
Ballif[192]	SignatureChip CGH	CNV	969 BACs	blood	3,600	18	0.5
Cheung[193]	CGH	CNV	853 BACs	blood	2,585	18	0.5
Pham[194]	BCM V8 OLIGO (aCGH)	CNV	180k	blood	10,362	57	0.55
Conlin[36]	IlluminaQuad610 (SNP)	LOH, CNV	620k	blood, fibroblasts	2,019	23 (1 chimera)	1.1

Table 3-1 Example. Clinical diagnostic microarray studies investigating mosaicism in children with congenital or developmental abnormalities. SNP: Single nucleotide polymorphism. (aCGH) Array comparative genomic hybridisation; (BACs) Bacterial artificial chromosomes

In comparison to studies of clinically ascertained children with DD, the prevalence of mosaicism among children without DD is less well established, although evidence suggests that the frequency is extremely low[50,130]. In the cohort studies analysed by Laurie *et al.*, no mosaicism was detected in any of 1,600 individuals aged 10–19 years old. While 13 mosaic events were found among 6,810 children aged 0-4, a frequency of 0.19%, this may reflect ascertainment bias, as the youngest stratum of children in this study included children from a cohort study of oral clefts, a potential manifestation of pathogenic mosaicism. Thus, the frequency of mosaicism in children without DD remained an open question.

In this study, to quantify the burden of pathogenic structural mosaicism in children with developmental disorders, I determined the frequency of structural mosaicism in thousands of children with and without developmental disorders, using

both single-sample (MAD), and trio-based (triPOD) detection of structural mosaicism from SNP microarray data. Both clinical review of the specific variants and a statistical analysis of enrichment of structural mosaicism in cases indicated that the majority of the mosaic events detected in probands were pathogenic.

3.3 Materials & Methods

3.3.1 Description of studies

SNP microarray data from four studies were used in this analysis.

The first study was DDD, designed to study children with undiagnosed DD. SNP microarray data were available for 3,669 samples, which included 1,303 probands and most of their parents. Of the 3,669 total, 3,419 (93%) were derived from saliva and the remainder from blood, and of the 1,303 probands, 1,057 (81%) were derived from saliva and the remainder from blood. A clinical geneticist prepared a detailed family history, documented complications during the pre-natal, peri-natal, and neonatal periods, assessed development milestones, recorded phenotypic features in Human Phenotype Ontology format (HPO format), and uploaded clinical photographs with parental consent[3].

The second study was the Scottish Family Health Study (SFHS), designed to study the genetics of complex traits. Like DDD, this is a trio study, but the main subjects are young adults who lacked delays in development. This study was included in this experiment as a control study. SNP microarray data were produced primarily from blood (84.5% of samples) and the remainder from saliva[196].

Both the DDD and SFHS cohorts were processed on the same custom Illumina® SNP genotyping chip, a design combining 733,059 HumanOmniExpress-12v1_A-b37 positions and 94,840 additional selected positions. DNA was sourced from saliva using Oragene® OG-500 (parent) or OG-575 (child) collection tubes (DNA Genotek Inc.). The Sanger Genomics core performed genotyping using Illuminus [148], and recorded the results in PLINK format[149]. I converted these data to VCF format[141] using plinkseq version 0.08. Probe-level quality control measures selected polymorphic, well-covered positions that were absent from copy number regions of at least 1% frequency (as calculated from a composite of multiple CNV studies)[150,151]. This resulted in 679,891 assayed positions (Table 3-2). Samples were not excluded on outlier levels of BAFs or LRRs since large (especially genome-wide) mosaicism will skew these measures and I wanted to prevent unintentional filtering of real mosaicism.

The third and fourth studies included for analysis were two prospective, longitudinal, birth cohort studies: TEDS and ALSPAC. The child participants from Avon Longitudinal Study of Parents and Children (ALSPAC), a cohort called “Children of the 90s”, consisted of approximately 15,000 children. Illumina SNP microarray data

were available for 8,970 unique samples. BAF and LRR metrics were derived by Tom Gaunt and Hashem Shihab from the ALSPAC group using raw data and published guidelines[38]. For 5,667 samples, DNA was sourced from cell line material, 3,290 from blood or tissue, and 13 had unknown origin. The SNP genotyping chip assayed 478,184 sites on autosomes and chromosome X aligning to GRCh37 and absent from copy number regions of at least 1% frequency (Table 3-2). I excluded samples as controls if the child had phenotypes suggesting developmental problems; the exclusion criteria were: child has ever had developmental delay (sa032a): 'Yes'; parent worries over development (kd075): greater than zero. The ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary: Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

The Twins Early Development Study (TEDS) includes approximately 13,000 unrelated twin pairs from England and Wales. A main aim of the study is the investigation of genes and environment on cognitive and behavioural development in children. SNP genotype data were derived from buccal swab sampling using Affymetrix 6® chips. This genotyping chip assayed 695,017 sites on autosomes and chromosome X aligning to GRCh19 and absent from common copy number regions (Table 3-2). Samples were excluded from selection as controls if the child had phenotypes suggesting perinatal or developmental problems at four years were noted: Perinatal outlier overall exclusion 'YES', medical exclusion 'YES', talking problem (dhtalk1) 'YES', or above 90th centile for total behaviour problems (dbhbeht1 and dsdbeht1).

DDD & SFHS SNP Probe Quality Control	
#Positions	Filtering Step
810110	all designed positions
793968	removing non-SNV or non {A,T,C,G} positions
695516	removing maf < 0.01, hwe > 0.001, missingness > 0.1
679891	removing positions in common CNV regions
ALSPAC SNP Probe Quality Control	

# Positions	Filtering Step
610259	provided QC polymorphic hg18 positions
500527	Passed ALSPAC QC
488199	Mapping to GRCh37
478164	Outside common CNVs
TEDS SNP Probe Quality Control	
# Positions	Filtering Step
723257	provided QC polymorphic NCBI36 positions
710992	Mapping to GRCh37
695017	Outside common CNVs

Table 3-2 SNP Probe Selection

3.3.2 Mosaic event detection

I used MAD and triPOD to detect structural mosaicism from probands and proband-trios. The advantage of triPOD is increased sensitivity compared with MAD for detecting events of low clonality, however triPOD additionally requires parental genotype data, which are not available in all studies.

I ran MAD using the following default parameter values: aAlpha = 0.8, T = 9, and MinSegLen = 75. Because the published version of MAD processes samples in series and the score of this analysis required implementation on several thousand samples, I modified the MAD code to more easily process samples in parallel. These modifications did not alter the statistical approach used by MAD. I ran triPOD using default settings (alpha = 0.1, nc_thresh = 0.03) but changed 'genome build' to 'hg19'.

3.3.3 Methods of evaluating of clinical significance

I evaluated the clinical significance of copy-number and copy-neutral mosaic events differently.

For mosaic copy-number events, I assessed whether online genomic disorder databases, DECIPHER[104] and OMIM[10], reported CNVs overlapping in location and consistent in direction (losses or gains) with the mosaic copy number detections. If a genomic disorder was identified, I assessed whether the child's phenotypes were

concordant with the genomic disorder, and if so considered the mosaic CNV likely pathogenic.

For mosaic copy-neutral (aUPD) events, I investigated whether these events caused imprinting syndromes or recessive diseases. To evaluate the first possibility, I assessed whether the abnormality was present on a chromosome associated with imprinting syndromes, based on the frequently updated Liehr UPD online database[132]. LOH-mediated recessive disease occurs when LOH in mosaic tissue results in homozygosity of a pathogenic allele. To detect candidate pathogenic alleles underlying recessive disease I interrogated the exome data for rare (below 0.5% MAF) functional and loss-of-function variants in the LOH interval. To ensure that the candidate allele was homozygous in the mosaic tissue, I only included for analysis variants for which the allele fraction of the rare allele was greater than 0.5, i.e. skewed toward homozygous non-reference. With the collaboration of clinical geneticist Dr. Helen Firth, I assessed whether detected candidate variants were pathogenic based on her clinical expertise and my literature review.

3.3.4 Exome sequencing

Exome sequencing was performed by the Sanger sequencing core and DDD informatics team, as fully described elsewhere[6]. In brief, the exome capture design was Agilent® SureSelect v.3 50-Mb baits and augmented with 5 Mb of custom regulatory sequences. Sequencing was performed using Illumina® HiSeq 2000 platform to greater than 50x mean coverage using paired-end 75-bp read-length sequence reads. Alignment to the genome reference GRCh37, version hs37d5 (a version of the human reference genome used by the 1000G Project[146] that includes decoy sequences aimed to improve the fidelity of single nucleotide polymorphism detection), used the Burrows-Wheeler Algorithm[57] version 0.5.9. Quality control filters (genotype quality below 30.0, homopolymer runs above 5, variant quality by depth below 5.0, read depth below 4 or above 1200, strand bias above 10.0) were applied. Genotype data were stored in VCF files.

3.4 Results

The main analysis goal was the assessment of mosaic burden in children with DD compared to children without DD. This analysis involved the execution of MAD and triPOD in a case-control setting.

Initial attempts running MAD and triPOD yielded thousands of putative detections. Inspection of a subset of these 'calls' demonstrated that the vast majority were false-positives. I identified systematic classes of detection-error, and, as described in more detail below, I evaluated different approaches to best account for these failure modes, finally selecting a strategy based on the number of peaks in the BAF distribution and the percentage of genotypes that were homozygous, to reduce the number of putative detections for manual curation.

There were two case-control analyses performed using SNP microarray data. First, I ran MAD on child cases in the Deciphering Developmental Disorders study (DDD, N=1,303)[1] and on controls derived from two UK birth cohort studies: the Avon Longitudinal Study of Parents and Children (ALSPAC, N=2,168) [197] and the Twins Early Development Study (TEDS, N=3,588)[198]. The second case-control analysis used trio data, in the hope of including lower-clonality mosaicism; here the trio analysis was performed using the triPOD method on DDD trios and on a control group from the Scottish Family Health Study, a study of young adult healthy controls and their parents (SFHS, N=478)[196].

3.4.1 Filtering Strategies for MAD output from DDD & SFHS samples

Initial testing of MAD on all 5,103 DDD and SFHS samples produced 2,299 putative mosaic detections, orders of magnitude higher than expected. Manual inspection quickly identified recurrent sources of error (listed in order of descending observation frequency): (1) incorrect classification of long tracts of constitutive homozygosity as mosaic (Figure 3-1); (2) over-segmentation of single contiguous regions (Figure 3-2) (3) unimodal skews of heterozygous BAFs (Figure 3-3); (4) incorrect classification of constitutive copy number events, mainly duplications, as mosaic.

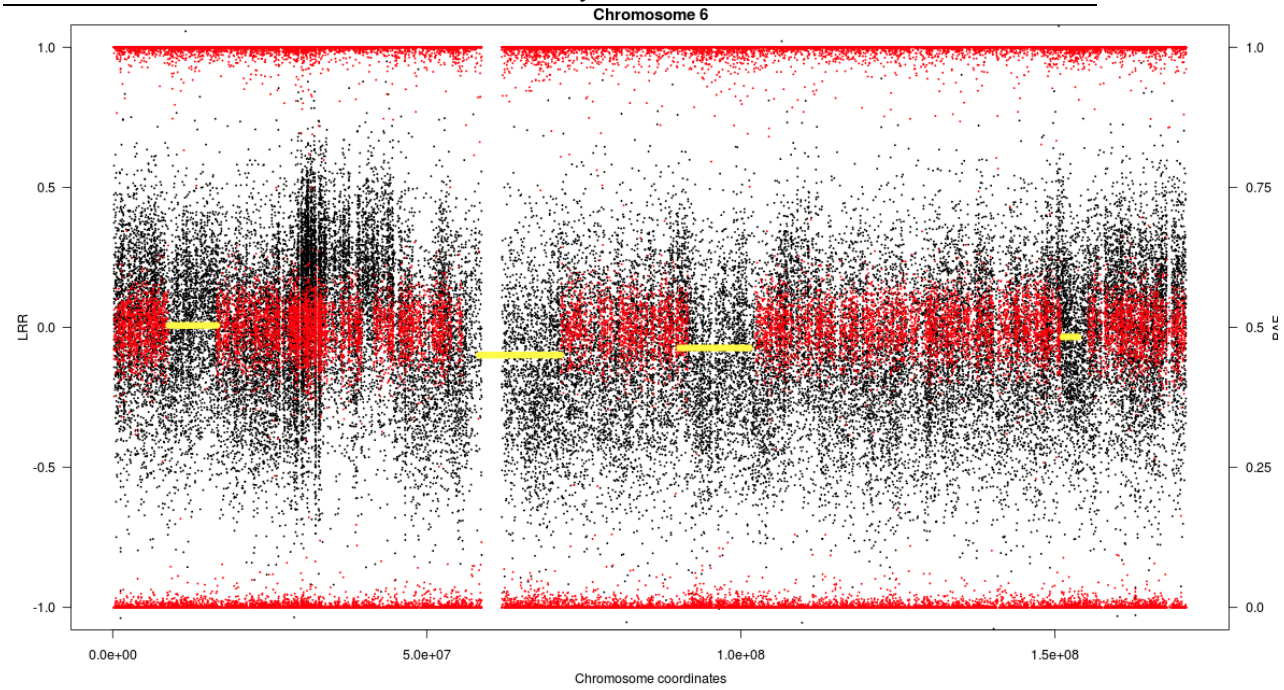

Figure 3-1 Four tracks of constitutive homozgygosity classified (incorrectly) as mosaic.

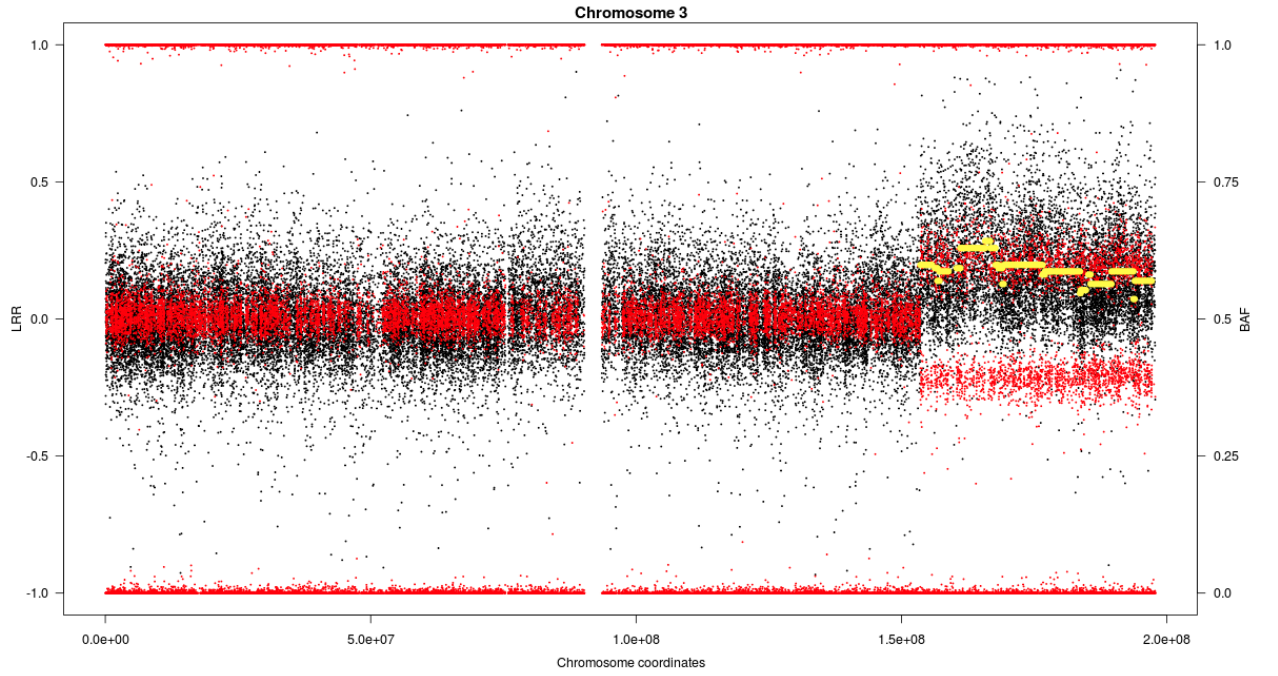

Figure 3-2 An example of over-segmentation. The single mosaic duplication is broken into many smaller duplications.

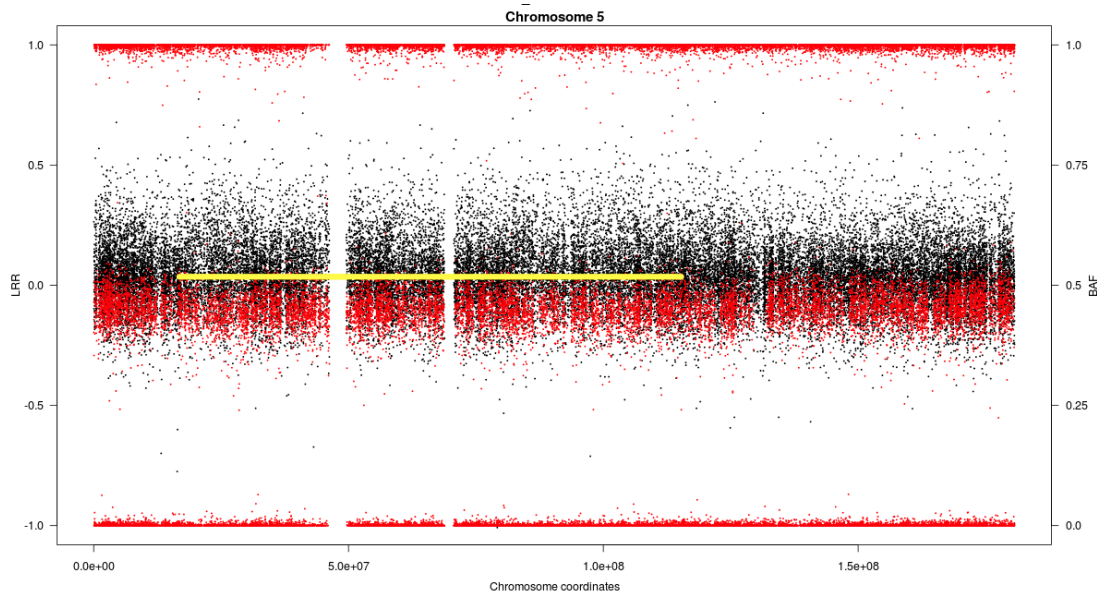

Figure 3-3 An example of unimodal skew, in this case, BAFs systematically depressed slightly below 0.5. This results in an increase in B_{dev}, which then results in a false mosaic detection.

3.4.1.1 Managing over-segmentation

Of these sources of error, it was most straightforward to manage over-segmentation. This is an artefact characterised by imperfect delineation of event boundaries and is a common pitfall for segmentation algorithms. To reduce over-segmentation I merged nearby (within 1 Mb) putative detection sub-segments representing the same event type (loss, gain, or loss of heterozygosity). The LRR and B_{dev} values for the final merged segment were calculated using a weighted-average (based on the number of probes in segments) of the LRR and B_{dev} values among the sub-segments. Segments beyond 2 Mb in size after merging were retained for analysis.

3.4.1.2 Managing constitutive homozygosity & unimodal BAF deflection

Tracks of constitutive homozygosity are relatively frequently observed in the DDD study as families often have familial relatedness[3], which results in large blocks of inherited homozygosity (identity by descent). Due to imperfect measurement of BAF, some homozygous genotypes have BAF values different from 0 or 1. This results in non-zero B_{dev}, although rarely sufficiently displaced to result in heterozygous genotypes. Thus, I devised a strategy to manage constitutive homozygosity based on the ratio of heterozygous to homozygous genotypes in the putative detection.

Secondly, real mosaic events have heterozygous genotypes with bilateral departures from 0.5, but I found that one recurrent error mode was characterised as putative detections with unilateral (usually downward) deflection from 0.5 from an unknown cause. To distinguish unilateral and bilateral BAF deflections, I evaluated

several peak-finding software tools on a training set of positive and negative events but found superior performance (data not shown) using a simple, heuristic strategy using the R density function, based on the difference in height of the tallest peak of the BAF density function to the next-tallest height. Segments with one prominent single peak reflected unimodal distributions, while density functions with at least one additional large peak was characterised as bimodal.

Real mosaic events should have high proportions of heterozygous genotypes and an obvious bimodal distribution, whilst constitutive homozygosity events are likely to have low proportions of heterozygous genotypes, and segments with unilateral BAF deflections are likely to appear unimodal. Therefore, I suspected that segments underlying these three possibilities should segregate well in a bivariate plot of het:hom ratio and peak:next-peak ratio (Figure 3-4).

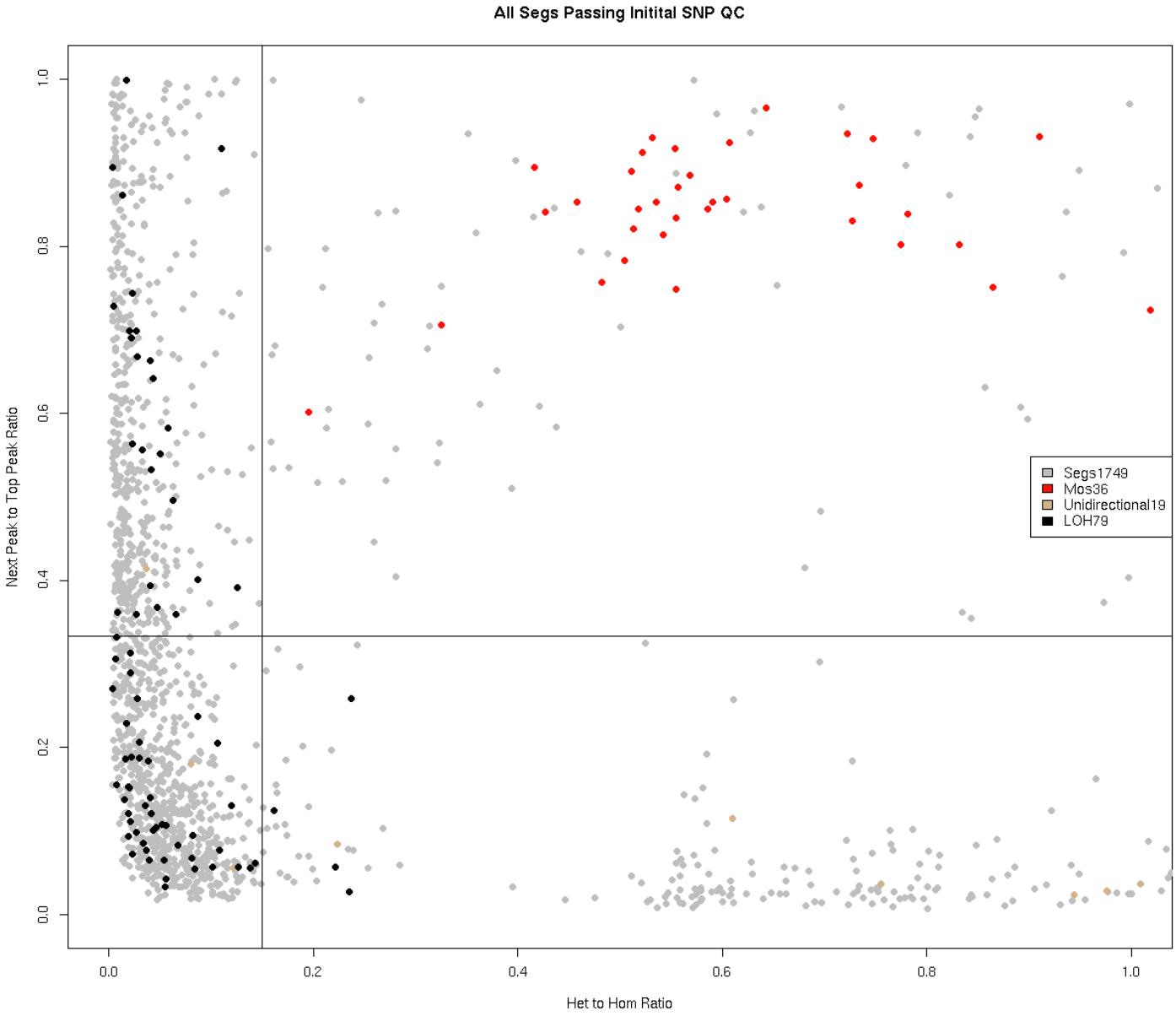

Figure 3-4 Filtering unimodal BAF deflections and constitutive homozygosity using the het:hom ratio and peak:next-peak ratio. grey dots: putative detections, yellow dots: unimodal deflections, black dots: constitutive LOH, red dots: suspected real events

I plotted the location of segments I had classified as constitutive LOH or unimodal during initial manual review, and found that, according to expectation, the constitutive LOH events fell on the left side of the graph, and the unimodal segments fell on the bottom-right. I calibrated thresholds for het:hom ratio and peak:next-peak ratio based on the distribution of segments belonging to the constitutive homozygosity cluster and unimodal cluster and manually inspected all putative detections in the upper-right quadrant. Among the putative detections in the upper-right quadrant I found 36 putative detections (red dots) that appeared to represent real mosaic events, and false-segments representing stochastic fluctuations in the data. Of the 36 putative events, some were found to be constitutive duplications (next section) and others required further merging to consolidate sub-segments into final mosaic detections.

In addition to the filtering strategies listed above, I also manually reviewed all putative segments on chromosome X to prevent exclusion of segments in males with aberrant BAF characteristics due to mosaicism in the context of hemizygosity.

3.4.1.3 Managing constitutive CNVs

Ten putative mosaic detections among DDD and SFHS samples had a large magnitude of upward deviation of LRRs and wide separation of BAFs. Jacobs *et al.*[50] identified a similar signature in their study and concluded that such events represented constitutive CNVs detected as mosaic. Two of these ten events were found in probands and parental data were available that showed the same CNV present in at least one parent, substantiating the constitutive nature of these two proband events and suggesing that the remaining eight were also likely constitutive.

To further assess whether these remaining events were constitutive, I gathered known constitutive duplications in the DDD study and calibrated thresholds of LRR and BAFs based on the distribution cluster of these constitutive events. The list of known constitutive duplications came from Dr. Tomas Fitzgerald who used trio data to identify as inherited (and thus constitutive) 1,813 CNVs in the DDD study. I manually curated this list to a high-quality set of 148 CNVs at least 200 kb in size and plotted the B_{dev} and LRR for each CNV. I observed that all ten suspicious duplications overlapped with the cluster of inherited duplications; thus were all very likely constitutive, and I removed

these from further analysis. The curated mosaic and constitutive events for DDD and SFHS are discussed in greater detail and plotted below (Figure 3-5).

3.4.1.4 Inclusion of aberrant standard deviation of BAFs rescues one mosaic event

A commonly employed QC criterion used in GWAS studies is exclusion of samples on the basis of high average standard deviation of heterozygous BAFs. However, to avoid unintentional exclusion of mosaicism, I did not employ this filter. As a result, I found eight samples with a consistent multi-band skew of BAFs across all chromosomes, a signature of contamination, and removed these from analysis. However, this strategy also retained one sample with a high BAF standard deviation of 0.06, which reflected a real mosaic structural event (see patient ID259709 in section 3.4.6).

3.4.1.5 Filtering strategies for TEDS and ALSPAC

The MAD results for the TEDS and ALSPAC cohort were merged and filtered as above, and events of 2 Mb size or greater in samples passing phenotypic exclusion criteria were included for analysis. There were 87 putative events at this size or greater; these included 7 events with large skews in LRRs and BAFs, 30 that reflected two sibling contamination events, and the remaining were due to spurious X chromosome deviations in males, and small peri-centromeric events. Four of seven events were deletion events, with BAFs not strictly at 0 and 1, but skewed inwards. These events had consistent levels of LRR and BAFs and clustered together, suggesting they were constitutive events, but skewed due to a noisy background. The remaining three of the seven were gains, and surprisingly, two of these three represented trisomy chromosome X. Extended phenotypic data of these two individuals, including school maths, reading and anxiety levels were scrutinised, but neither child was an outlier in any of these measurements, suggesting their trisomy X was benign or subclinical.

In ALSPAC, there were 347 putative mosaic events at least 2 Mb in size and I manually reviewed all of them. Of these, 47 appeared real, and filtering of constitutive duplications using the method described above identified four mosaic events.

The curated mosaic and constitutive segments from MAD analysis for all SNP-based cohorts are provided here (Figure 3-5).

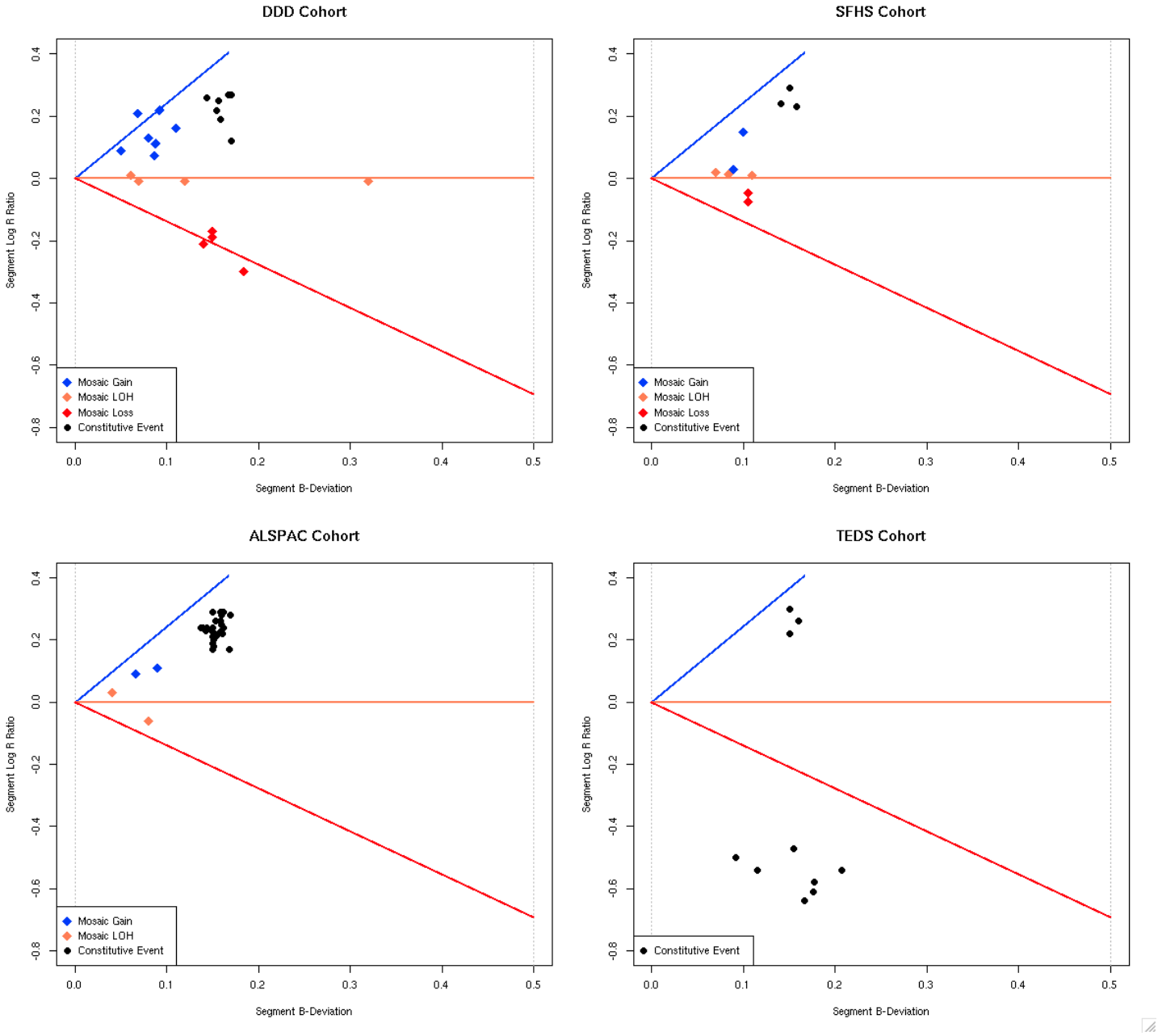

Figure 3-5 Characterisation of mosaic events and constitutive duplications in the DDD, SFHS, ALSPAC and TEDS studies.

3.4.2 Assessing the accuracy of filtering strategies

To assess the accuracy of this MAD-based workflow, I compared the frequency of mosaic events detected among the parents of the DDD and SFHS trio studies with established estimates of mosaicism frequency for individuals of these ages. The median age at sampling of DDD parents was 39 years old and of SFHS parents was 59 years old (Figure 3-6).

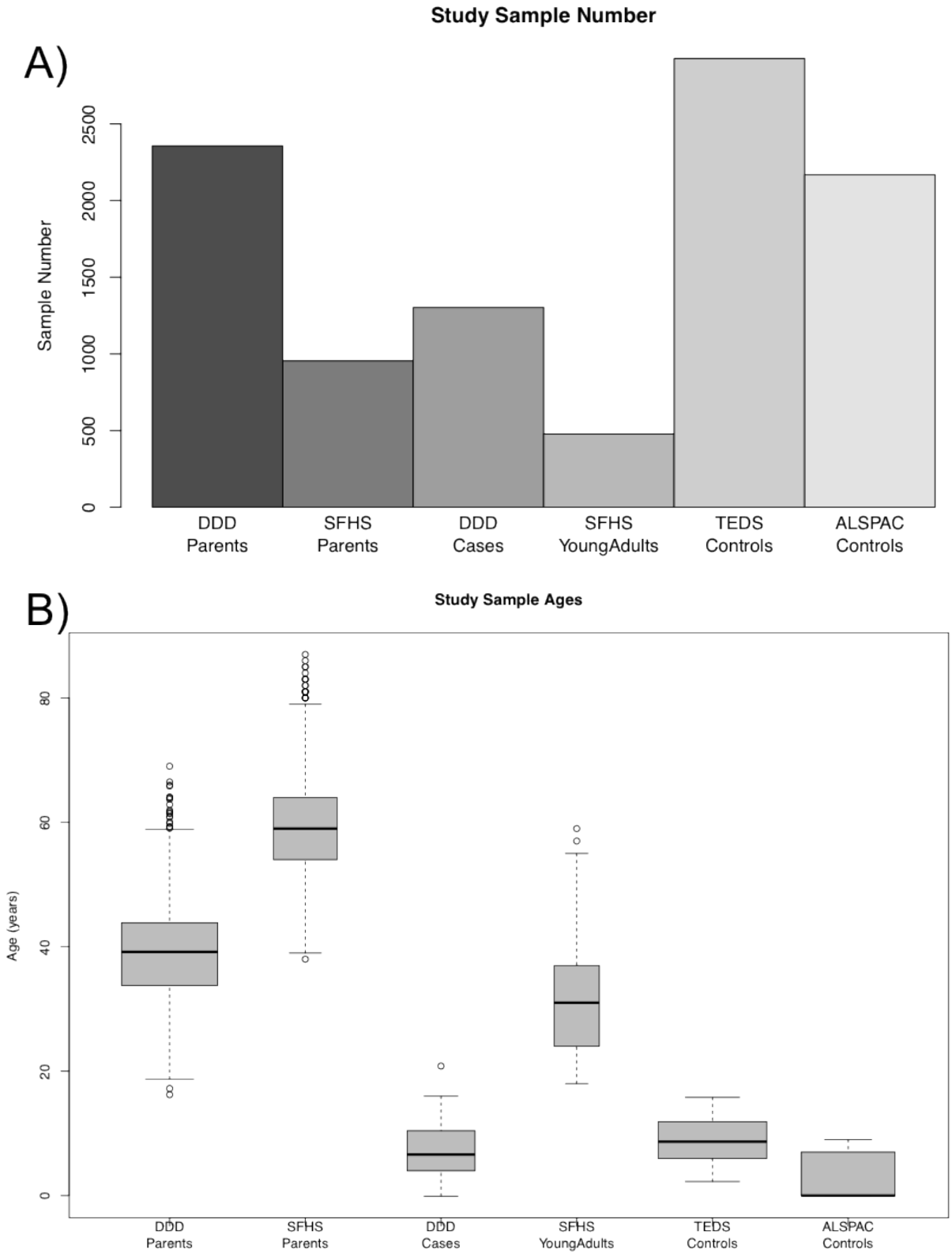

Figure 3-6 The (A) sample number and (B) ages corresponding to the analysed studies.

I identified 6 mosaic events among 955 parents of SHFS controls, a frequency of 0.6%, and 4 among 2,356 parents of DDD probands, a frequency of 0.1%, which are within the confidence interval estimates for these ages[50] (Figure 3-7). This suggested that the method, filtering strategy and manual curation used were consistent with expectations based on the published studies, and I next used this workflow to detect mosaicism in the child samples.

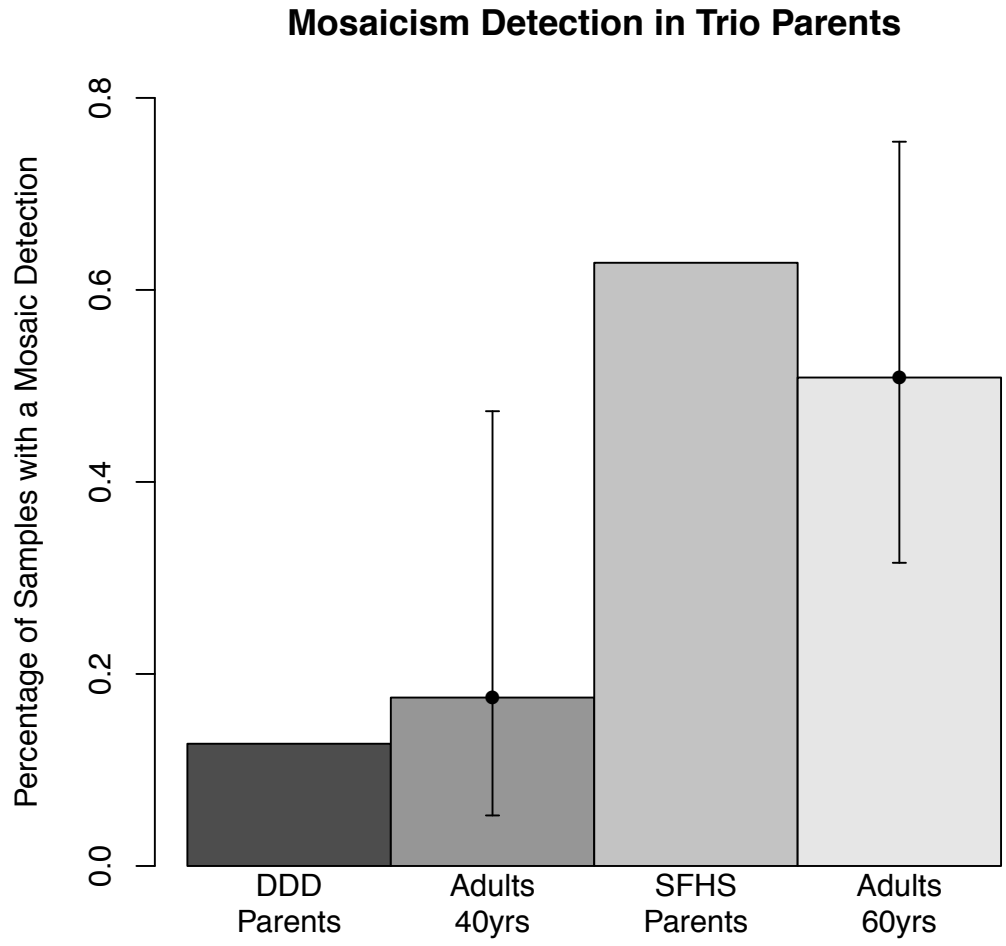

Figure 3-7 The frequency of mosaicism detected in the parents of the trio cohorts was within the confidence intervals of the frequency detected for samples of this age range.

3.4.3 Mosaicism Frequency in Cases & Controls using MAD

I assessed mosaicism frequency using MAD, described in this section, and using triPOD, described in section 3.4.4, and then I assessed the clinical consequences of detected mosaicism in section 3.4.6. These steps are summarised in Figure 3-8.

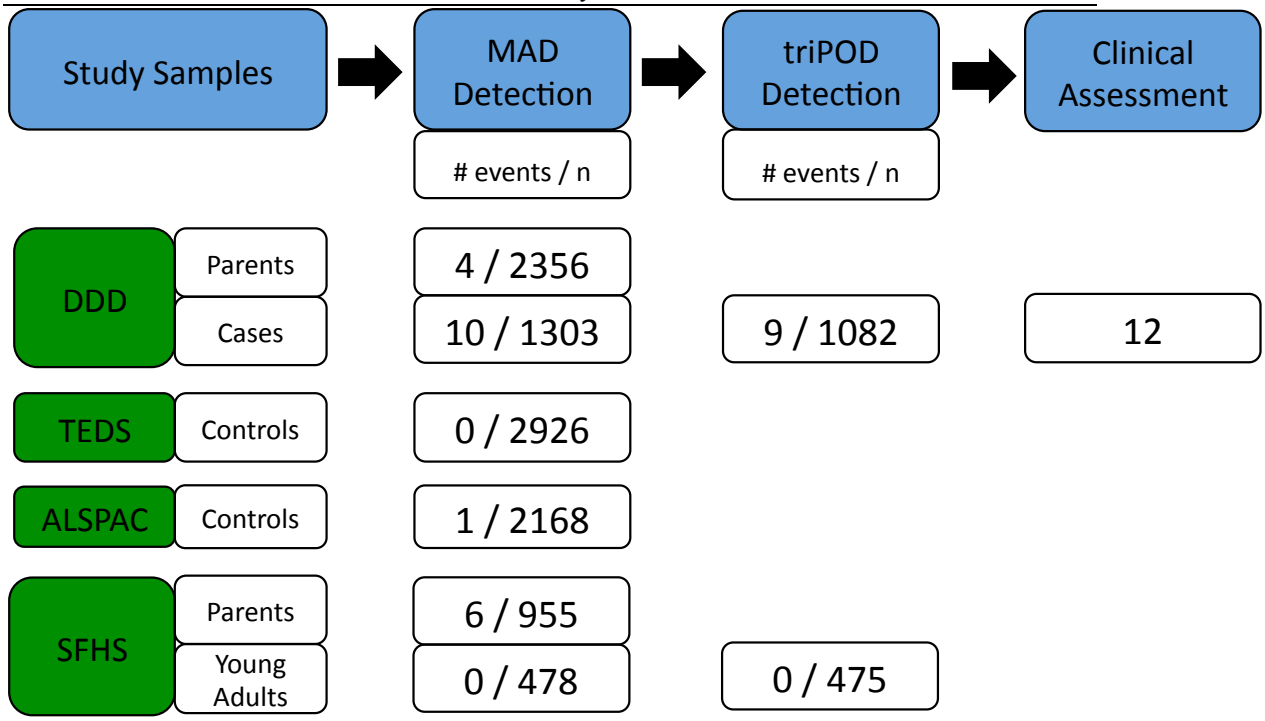

Figure 3-8 Overview. A MAD-based workflow was used to detect mosaicism. This workflow identified an enrichment of mosaicism in cases compared with controls, and triPOD detected two additional mosaic events not detected by MAD. Clinical assessment was performed on all 12 probands of the DDD study with mosaicism.

I ran MAD on children from the DDD study and used the filtering strategies listed above (section 3.4.1) to curate putative events. This resulted in the detection of 10 mosaic detections among 1,303 children analysed, a rate of 0.77% (Figure 3-9, A and B). The range of cellular fraction (clonality) of the detected abnormalities was 24% to 66%. Compared to the frequency of mosaicism derived by combining studies of LOH and CNV mosaicism, 0.82%, the frequency observed in this study was not significantly different (binomial test p value 1.0). A more conservative comparison, based on the frequency observed among children ascertained for genetic testing in Conlin *et al.*, 1.1%, also yielded no significant difference (Fisher exact test p value 0.37).

With respect to distribution of mosaicism across tissue, all 10 of the detections were among the 1,057 samples derived from saliva, while no mosaicism was detected among the 247 samples derived from blood. The tissue-specific frequency difference was not significant (binomial test p value 0.096) but there was little power to detect a difference given the rarity of mosaic events.

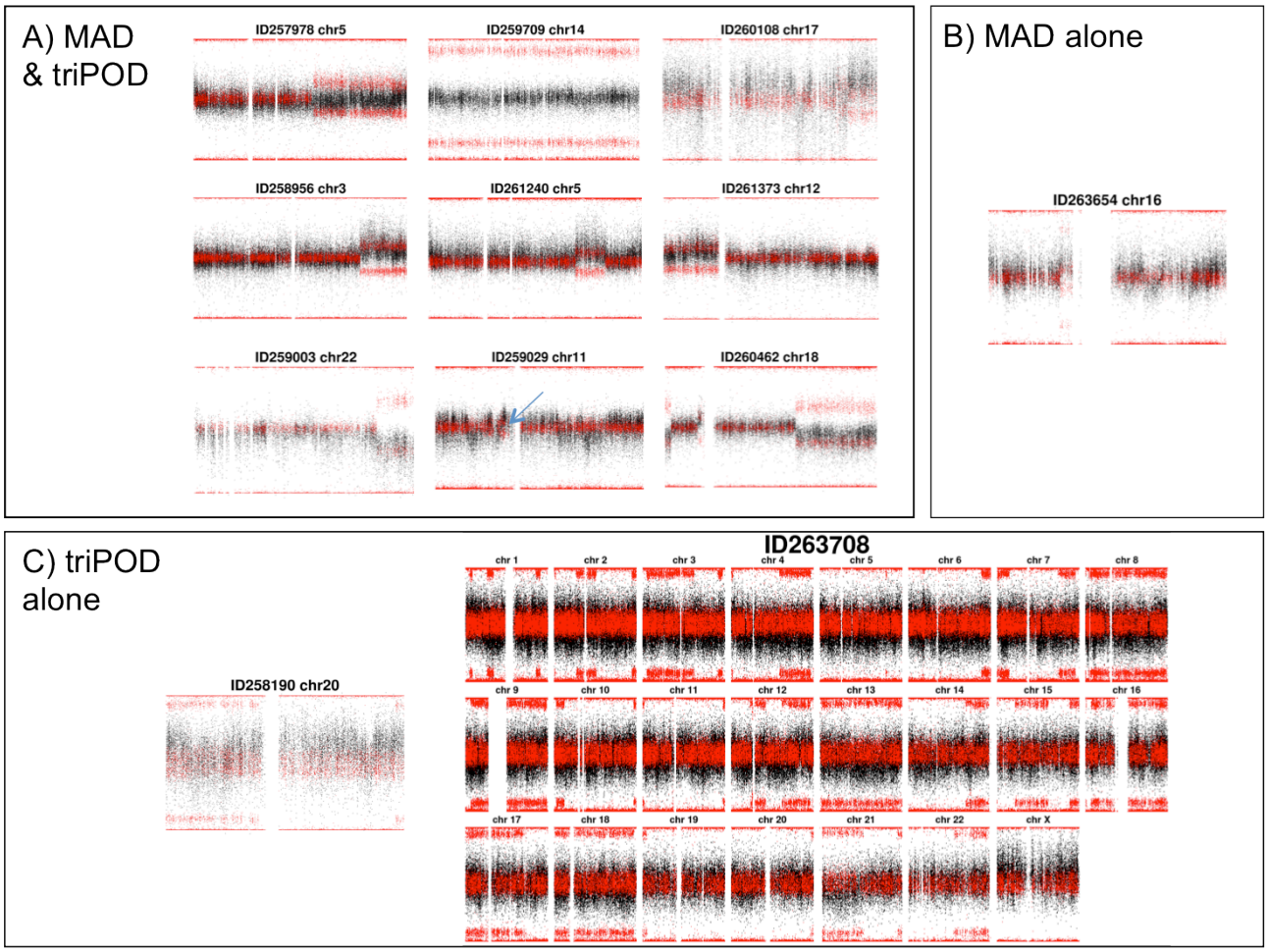

Figure 3-9 All proband detections: The detections made by (A) MAD & triPOD, (B) by MAD alone and (C) by triPOD alone.

I ran MAD on TEDS and ALSPAC to include frequency comparison to these children lacking DD. There were 3,588 children in the TEDS cohort with genotype data from blood-derived DNA available. Analysis was performed on 2,926 samples for which phenotypic data were available and samples were not medically excluded nor had developmental problems. There were zero mosaic events retained after accounting for seven constitutive duplications. There were 8,970 children in ALSPAC with genotype data available from DNA derived from blood or cell-lines. An initial attempt at detecting mosaicism in data from both DNA sources detected more mosaicism in samples derived from cell-lines (two-sided Fisher's exact test p value 5e-5), suggesting the presence of cell-line induced chromosomal rearrangements[199,200], which would overestimate *in vivo* mosaicism. To assess frequency in children accurately, I analysed the 3,290 DNA samples sourced from blood or saliva (but not cell-lines). Of 2,538 children with phenotypic data available, 2,168 (85%) lacked developmental disorders or

major developmental problems. One sample contained a mosaic LOH, representing a frequency of 0.05%.

I also investigated a collection of 478 individuals from the Scottish Family Health Service (SFHS). These were samples without DD recruited in early adulthood, median age 31. There were zero mosaic events remaining after automated filtering and manual curation of 28 possible mosaic events.

Compared to the fraction of mosaic detections among all child control samples (2 in 5,345), the frequency of mosaicism in DDD probands (10 in 1,303) was highly statistically significant (odds ratio 20.66, one-sided Fisher's exact test p value 3.627e-6). A meta-analysis additionally incorporating 7,119 samples from two previous studies[35,36] strongly supports a statistical enrichment of mosaicism in children with developmental disorders (p value 9.919e-11).

.

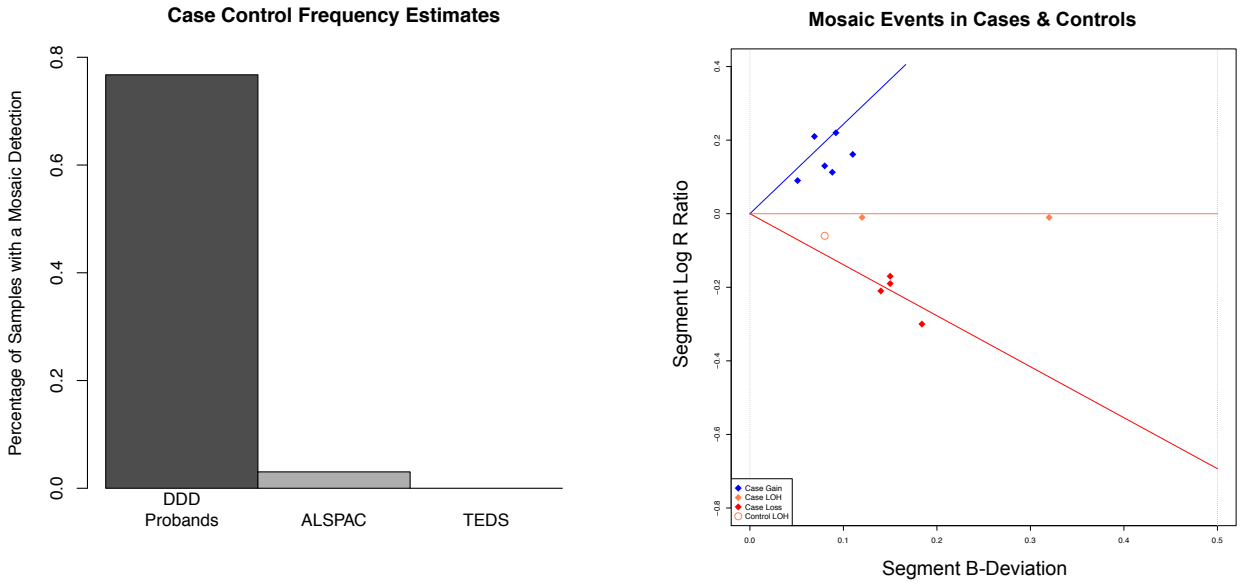

Figure 3-10 (A) The percentage of samples with mosaic events in the case and control cohorts. (B) A depiction of each mosaic event, where the line segments represent the ideal location of mosaicism for gains (blue), LOH (orange) and losses (red).

3.4.4 Additional detections using triPOD

triPOD leverages haplotype information in trio data to yield improved sensitivity to detect lower-clonality mosaic events compared with MAD[51]. I implemented this tool on DDD trio data to improve detection of mosaic events of lower clonality.

Complete trio genotypes were available for 1,082 of 1,303 (83%) probands, and these were processed with triPOD. There were a vast number (4,920) of putative detections, of which 148 were at least 5 Mb and 876 were at least 2 Mb. All putative detections at least 5 Mb were manually reviewed. I also reviewed 200 randomly selected events at least 2 Mb or greater, which identified two error modes: no deflection in BAFs (spurious), or CNV present in parent (inherited). Due to the large number of detections, and the rationale to use triPOD mainly for the detection of low clonality events, computational filtering was implemented to select segments at least 2 Mb and having a median BAFs below 0.70 (as segments with very higher BAFs appeared to reflect constitutive events). Several hundred events with BAF values of "NA" or 0.50 (no BAF shift) were observed, which on the basis of no visually apparent mosaicism appeared spurious, so a 0.51 minimum threshold cut-off was used. triPOD identified 11 events with highly skewed BAFs and LRRs that were suggestive of inherited CNVs; 10 of 11 CNVs were also present in a parent, substantiating the constitutive nature of the event, and the remaining event clustered with the inherited events, so it too was considered likely constitutive.

Detections at the 2 Mb size or greater identified 7 of the 10 mosaic events that had been detected in single-sample analysis by MAD. Two of the three remaining events lacked complete trio data so they could not be analysed by triPOD. The third remaining undetected event was a mosaic duplication characterised by an additional haplotype not present in the diploid cell line (Figure 3-9 part C); this third event had a lower clonality (26%), lower than all but one of the abnormalities detected by MAD.

Two events were identified among the 148 putative events greater than 5 Mb detected by triPOD that were each reviewed manually. One event appeared to have a chromosome-wide elevation of LRR and a BAF pattern reflecting meiotic crossover, perhaps resulting from incomplete trisomy rescue.

The second event was extraordinary for a genome-wide pattern of large segments of consistently aberrant BAF interspersed with segments of normal BAF. These segments of aberrant BAF were present on most chromosomes in three or fewer large segments per chromosome. The clonality of this abnormality was approximately 17%, the lowest of all detected abnormalities. I investigated the parental origin of the aberrant BAF segments by plotting the proband BAFs within these segments separately for each configuration of parental genotypes. The sites with aberrant BAF were only observed where the father was heterozygous, suggesting that the aberrant BAF was due

to the presence of both paternal chromosomes. In addition, the BAF at obligate heterozygous sites in the proband (parents homozygous for different alleles) was always skewed toward a greater contribution from the inherited paternal allele, suggesting a second paternal haplotype, while only a single maternal haplotype (Figure 3-11 Interrogating possible haplotype combinations to determine the alleles present and their origin in the chimeric sample.

These observations are potentially compatible with a triploid cell line, however, karyotypic analysis failed to identify any triploid cells. An alternative explanation is "androgenetic / bipaternal mosaicism or chimerism"[201,202], which has been hypothesised to occur from one or two zygotes (Figure 3-12)[201]. The homozygous BAF skews had BAF deviations consistent with approximately 15% clonality, which is a smaller cellular burden than any event detected by MAD.

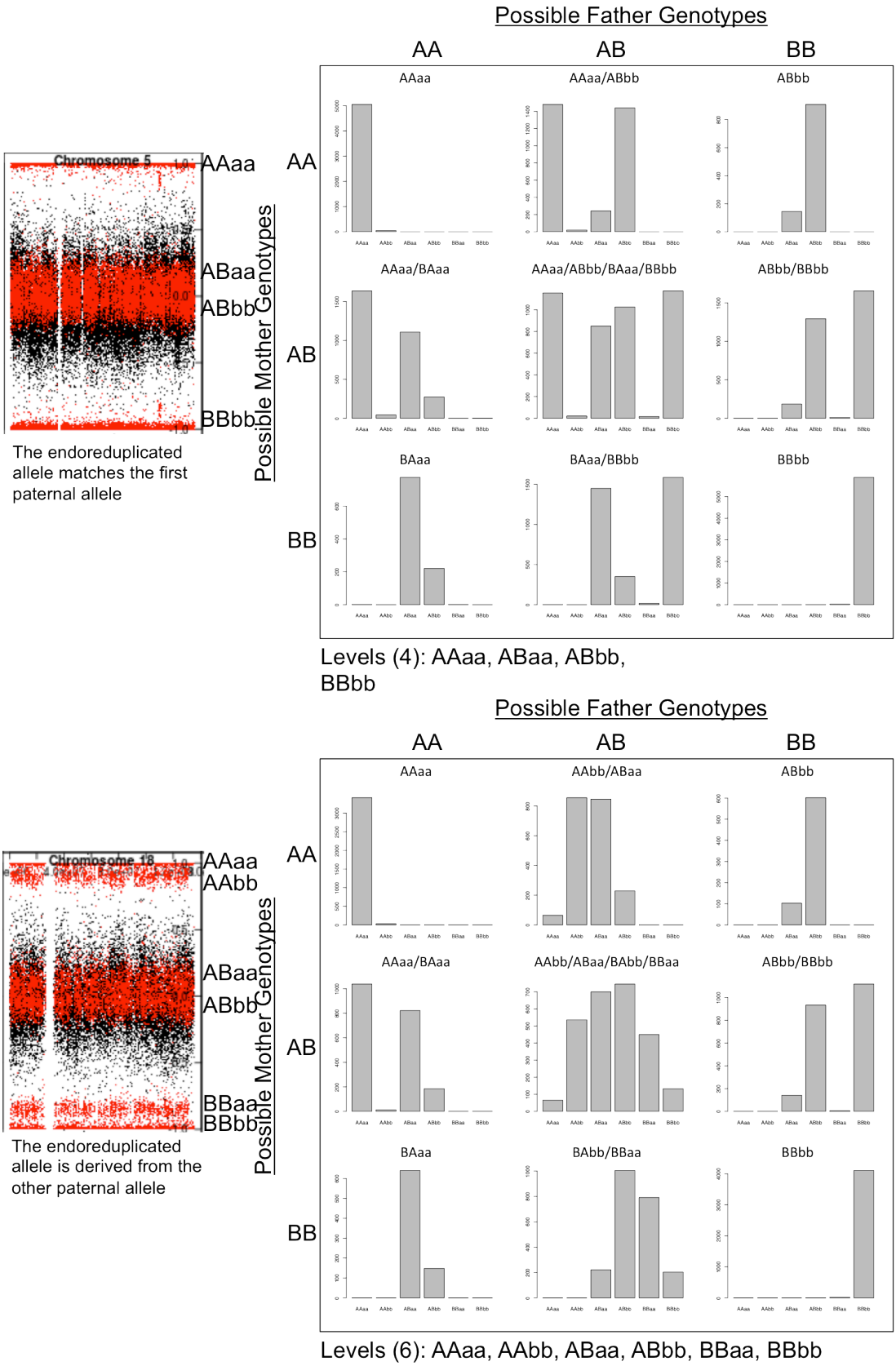

Figure 3-11 Interrogating possible haplotype combinations to determine the alleles present and their origin in the chimeric sample.

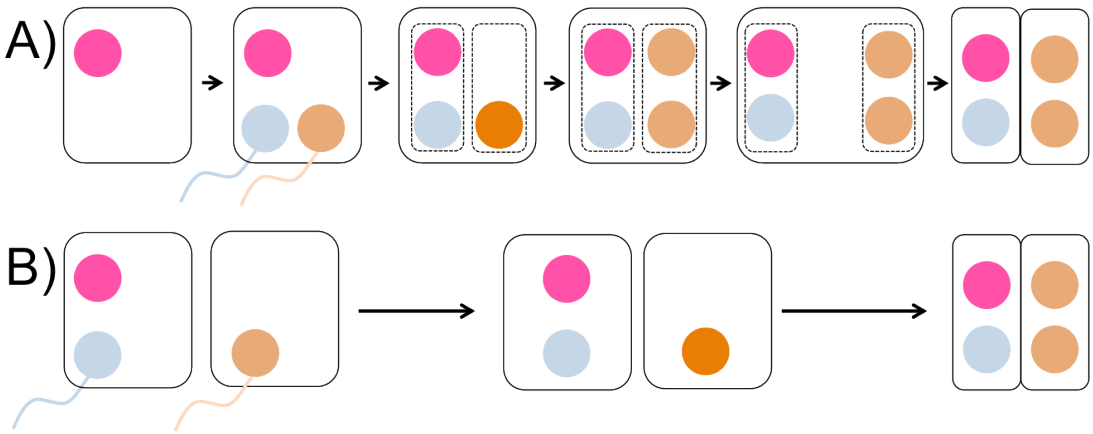

Figure 3-12 An illustration of two possibilities hypothesised by Robinson *et al.*[201] underlying androgenetic /bipaternal mosaicism or chimerism. In A) a one zygte mechanism, an ovum is fertilised by two sperm (dispermy), while in B) a two zygote mechanism, a fertilised zygote fuses with an endoreduplicated sperm cell-line.

triPOD was also applied to detect structural mosaicism in the 475 SFHS control trios. There were 26 putative events, of which 3 were constitutive and 23 were spurious, all but two in a narrow peri-centromeric region of chromosome 11; therefore there were zero mosaic detections uncovered.

3.4.5 Validation experiments to explore tissue distribution

Combining the results of MAD and triPOD, there were twelve children with mosaic abnormalities. Working with clinical centres and the DDD lab team, I attempted to validate each mosaic event in at least one tissue by aCGH or FISH and was able to determine whether the nine CNV events were distributed in both or either of epithelium-derived (saliva or buccal) and mesoderm-derived (blood) tissue. Of the nine children with CNV events, seven exhibited tissue-limited mosaicism. In all seven cases, the mosaicism was observed in epithelium-derived but not in blood, while two were observed in both tissues.

3.4.6 Clinical Interpretation of Probands with Mosaicism

Phenotypic data for the perinatal period for each proband were collected by clinical geneticists, who assessed developmental milestones and recorded phenotypes at time of recruitment using a standardised nomenclature called the Human Phenotype Ontology[135].

Mosaicism was detected in twelve individuals with developmental disorders (Table 3-3).

		birth records			measurements at time of recruitment					mosaic abnormality							validation				
sample	sex	gestation (weeks)	birth weight (kg)	required NICU (days)	age	height (cm)	weight (kg)	OFC (cm)	ID	type	chr	start (GRCh37)	end (GRCh37)	size (Mb)	B-Dev	clonality	aCGH results		FISH results		tissue limited?
																	blood	saliva	blood	saliva	
260462	F	37	2.6 (35)	no	5 yr	89 (3)	10.86 (1)	45.5 (1)	GDD	loss	18	650816	2804129	2.2	0.14	0.44	no deviation	downward	not detected	56% (buccal)	Yes:E
										gain	18	13422042	15265500	1.8	0.1	0.5					
										loss	18	48362664	78015180	29.7	0.1	0.46					
261240	F	37	1.9 (25)	7	16 yr	152 (7)	52 (48)	53 (7)	moderate	gain	5	123828524	145717285	21.9	0.08	0.38	not done	upward	double ring	not done	No
258956	F	38	2.6 (17)	10	4 wk	73.5 (26)	7.58 (1)	43.8 (1)	moderate	gain	3	153567441	197148984	43.6	0.11	0.56	no deviation	upward	failed QC	not done	Yes: E
261373	F	38	2.0 (1)	no	4 yr	96 (7)	14 (10)	50 (17)	moderate	gain	12	193818	38453531	38.3	0.09	0.44	no deviation	upward	not done	12% tetrasomy (buccal)	Yes: E
11	M	32	2.2 (90)	19	7 yr	100 (14)	14 (6)	47 (1)	GDD	gain	16	27183151	31888684	4.7	0.07	0.33	no deviation	not done	not detected	50% (buccal)	Yes: E
259003	M	40	4.6 (98)	no	3 yr	NA	15 (59)	51 (33)	GDD	loss	22	47182944	51666786	4.5	0.184	0.54	downward	downward	43%	failed QC	No
260108	F	40	3.6 (80)	?	19 wk	60 (1)	5.1 (1)	38 (1)	GDD	gain	17	66922993	81006629	14.1	0.092	0.451	no deviation	upward	failed QC	failed QC	Yes: E
263708	F	38	2.8 (27)	yes, ? days	16 yr	157 (14)	59 (67)	56 (75)	moderate	GWpUPD	all	n/a	n/a	N/A	0.0477	0.174	no deviation	no deviation	not detected	results pending	NA
258190	M	38	5.9 (99)	7	6 yr	113 (7)	22.8 (60)	55 (cm)	GDD	gain	20	1	63025520	63	0.0578	0.261	no deviation	not done	not detected	30% (buccal)	yes: E
259709	M	34	2.9 (98)	31	10 yr	132 (64)	28 (67)	?	moderate	loh	14	20432664	107287663	86.9	0.33	0.66	no deviation	not done	N/A	N/A	NA
257978	F	40	4.2 (95)	no	15 yr	?	?	50 (4)	severe	loh	5	101118483	180710763	79.6	0.12	0.24	no deviation	not done	N/A	N/A	NA
259029	F	40	3.3 (41)	no	5 yr	109 (77)	18 (60)	50 (11)	moderate	gain	11	42322518	45512054	3.2	0.051	0.227	no deviation	results pending	results pending	results pending	yes:E (SNP, saliva)

Table 3-3 Mosaic events detected among 1,303 DDD probands. (NICU) Neonatal Intensive Care Unit. (GWpUPD) Genome-wide paternal Uniparental Disomy. (LOH) loss of heterozygosity. (ID) Intellectual Disability. (GDD) Global Developmental Delay. (OFC) Occipital Frontal (head) Circumference; (E) epithelium. Numbers in parentheses in the 'birth weight', 'height', 'weight' and 'OFC' reflect population centiles given child age and sex.

Each mosaic event was assessed for overlap with regions previously implicated in specific genomic disorders, and if so, whether the patient phenotypes were concordant with the manifestations of these genomic syndromes. To identify a relationship between the mosaic copy-number events found in probands to online databases of pathogenic CNVs required the assumptions that: 1) pathogenicity is due to disruption of overlapped regions, not due to disruption of long-range regulatory elements; and 2) constitutive CNVs that are pathogenic produce phenotypes which are similar in character, if perhaps larger in magnitude, than the corresponding CNV in mosaic state. Mosaic UPD mutations can be pathogenic by multiple mechanisms, such as imprinting syndromes, by disrupting differentially methylated regions[203] or by manifesting recessive diseases, by converting a single inherited deleterious allele to homozygosity. To investigate these possibilities, I assessed whether the UPD event is implicated in an imprinting syndrome, the paternal origin of the mosaic allele, and whether homozygous alleles in mosaic tissue may be implicated in recessive disorders.

Patient ID260462 had global developmental delay, intermittent horizontal nystagumus with alternating abnormal head position and bilateral, symmetric large optic nerves. Magnetic resonance imaging of the brain showed cortical atrophy, generalised delay in myelination, moderate sized left middle cranial fossa, arachnoid cyst and deficiency of the rostrum of corpus callosum and atrophic splenium. Copy number analysis by karyotype and aCGH, genetic testing for Pitt-Hopkins, Fragile X syndrome, *MECP2* gene test, spinal muscular atrophy, and Angelman syndrome were all normal. Upon recruitment to the DDD study, aCGH was performed on blood and saliva by the DDD laboratory and no large (>500kb) CNVs were reported by the DDD informatics team. Mosaic analysis on SNP microarray data from a salivary sample identified three mosaic events on chromosome 18, two deletions and one duplication in approximately 50% of cells. Results from triPOD showed that the deletions resulted from loss of the maternal allele, while the duplication was of the paternal allele (Figure 3-13).

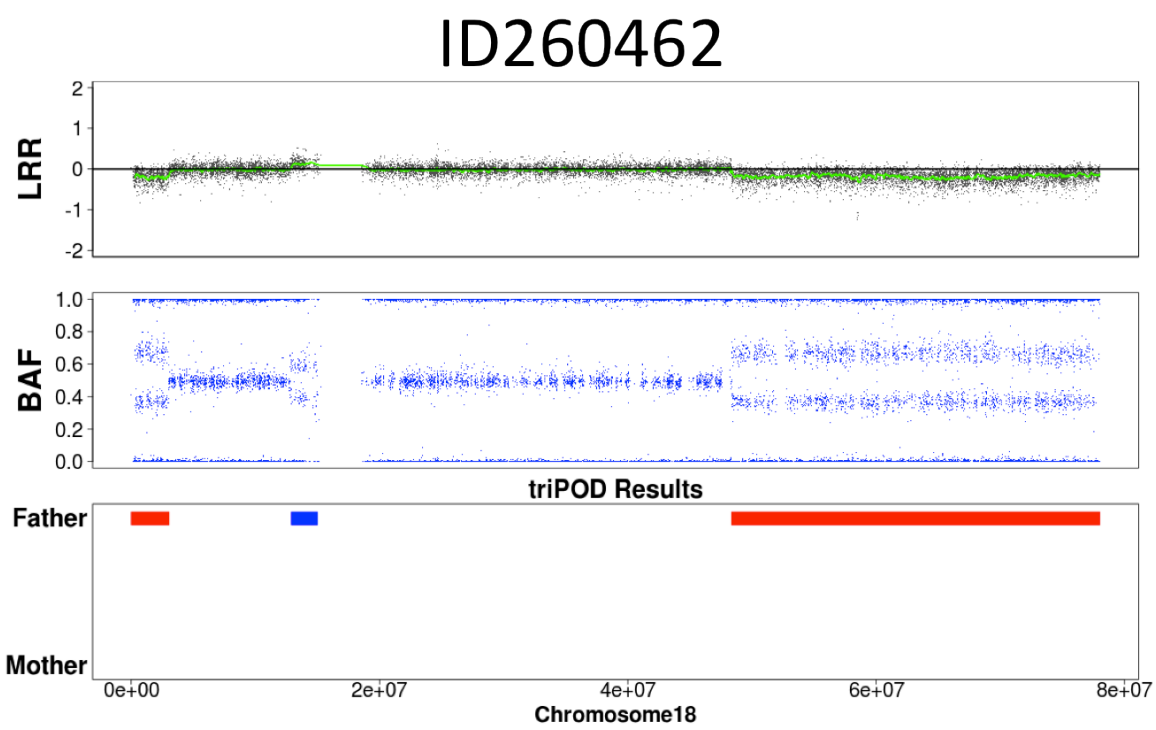

Figure 3-13 triPOD shows that the deletions and duplications arose from different alleles.

Fluorescent *in situ* hybridisation (FISH) analysis, performed by the local cytogenetics department on cells from a buccal sample, confirmed these events in 56 of 100 inspected cells. Retrospective scrutiny by the local cytogenetics department of the salivary CGH array identified deviations in aCGH probes but insufficient to be detected by the standard copy number detection pipeline. No deviation in blood aCGH probes was noted, suggesting the mosaicism was not present in all tissue types, and providing a likely explanation as why genetic testing, performed on blood, was negative. The mosaic deletion on chromosome 18 contains the gene *TCF4*, mutations in which cause Pitt-Hopkins syndrome[204], a diagnosis previously considered in this child. The SV was considered definitely pathogenic and the diagnosis was conveyed from the clinical geneticist to the family.

Female patient ID261240 required seven days in neonatal intensive care, and two weeks with nasogastric feeding. She had developmental delay, seizures, and short stature (154 cm, 3rd centile at 16 years). Before enrolment into DDD, clinical karyotyping was performed by the local centre on blood and showed a marker chromosome originating from chromosome 5; local inspection by aCGH did not detect any CNVs and the marker chromosome was classified as a balanced rearrangement. Local genetic testing for Fragile X syndrome was normal. At Sanger, mosaicism

analysis was performed on a saliva sample and identified a 22 Mb duplication on chromosome 5, present in approximately 40% of assayed salivary cells. Review of the interphase by the clinical cytogenetics team of karyotypic data noted that the suspected marker chromosome contained a double-ring chromosome. Retrospective manual review by the local cytogenetics team of the array CGH data on saliva identified stretches of raised LRR probes. Therefore, this event was classified as present in both blood and saliva. Duplications in this region, 5q23.2 to 5q32, have been previously implicated in seizure disorders (p.252)[205] and shared phenotypes and short stature are seen in a different patient with a overlapping duplication in the DECIPHER database (ID255372). Therefore, this mosaic aberration was considered likely pathogenic.

Female patient ID258956 had a number of congenital abnormalities, including a sacral meningocele, polydactyly, bilateral talipes, atrial and ventricular septal defects, pulmonary stenosis, EEG epileptiform activity, facial asymmetry, hirsuitism, hypomelanosis of Ito. At birth, she required neonatal intensive care for apnea and nasogastric feeding for 10 days. Clinical aCGH (Agilent 8 x 60K oligoarray) testing performed on blood by the local cytogenetics team was normal. Mosaicism analysis on saliva identified a 44 Mb duplication on chromosome 3q in approximately 55% of assayed cells. The DDD aCGH results from blood and saliva showed upward deviation in the data from assayed saliva tissue, only. Thus, it is likely this event is tissue limited. Duplications of 3q are associated with joint contractures, talipes, feeding difficulties, hirsuitism, and heart defects, including ASD and VSD[206]. There are several patients also present in the DECIPHER database who have duplications overlapping this large duplication in the child, including 280551, with hirsuitism, feeding difficulties, and global developmental delay; 283584, with sacral dimple, low set ears; and 1561, with frontal bossing, sacral dimple. Several examples of duplications of 3q have meningocele (p.145)[205]. Given the consistency of phenotypes with the proband and these patients, the mosaic mutation was considered likely pathogenic.

Female patient ID261373 had intrauterine growth retardation with a birth weight of 2.0 kg (1st centile). She had moderate developmental delay, severe speech delay, a high-arched palate and prognathism. An array on blood lymphocytes was performed at the local hospital and identified no abnormalities. Our SNP mosaicism analysis on saliva identified a gain of 12p in an estimated 44% of assayed cells, suggesting tissue-specific mosaicism as the cause. The event was detected also by confirmatory aCGH from saliva, and interphase FISH on buccal DNA of 100 cells

identified a triplication of 12p in 12% of cells. Triplications of 12p (tetrasomy 12p) are the cause of the clinical syndrome known as Pallister-Killian mosaic syndrome[207], which is consistent with many of her phenotypic features. The variant was considered definitely pathogenic and the diagnosis was conveyed from the clinical geneticist to the family.

Patient ID263654 required 19 days of neonatal intensive care to manage respiratory distress, jaundice and hypoglycemia. His speech and language were delayed and an MRI identified inferior vermis hypoplasia. Fragile X testing performed locally was normal. At Sanger, aCGH was performed by the DDD laboratory on blood and was normal. SNP mosaicism analysis identified a 4 Mb duplication in approximately 33% of salivary cells. The BAF pattern of the duplication was consistent with a meiotic origin of the duplication in the trisomic cell line. FISH was performed on blood and buccal tissues by the local cytogeneticist, and the event was detected in buccal tissue only, in 25 of 50 examined cells. As only interphase FISH was available for buccal tissue, positional information for the additional allele was not possible. The implicated region overlaps most of 16p11.2, a cytogenetic region in which duplications are well known to cause disruption of speech and language development [208] and this event was considered likely pathogenic.

Patient ID259003 had global developmental delay, no speech, and generalized hypotonia. Clinical aCGH (6K BAC array) and testing for Angelman syndrome were performed at the local hospital and were normal. At Sanger, SNP mosaic analysis on salivary cells identified a 5 Mb deletion in 54% of cells at chromosome 22q, from 22q13.31 to 22qter. Array CGH results showed a slight negative deviation in both blood and saliva probe data but not detected by the aCGH algorithm. FISH on blood lymphocytes performed by the local cytogenetics department identified the event in 43 of 100 of blood cells. This region overlaps with the well-characterised 22q13 Deletion syndrome, also known as Phelan-McDermid syndrome, which has as its main characteristics global developmental delay, absent or severely delayed speech and hypotonia; these manifestations are consistent with child phenotypes[209] and the mosaic event was considered definitely pathogenic.

Patient ID260108 had truncus arteriosis, hypertelorism, and feeding difficulties at birth. She demonstrated global developmental delay and required nasogastric feeding. An MRI performed at the local hospital was abnormal and showed possible arterial

shunting. Clinical testing performed locally for mutations in *SALL1*, *SALL4*, *CHD7,* and for Prader-Willi syndrome were normal. At Sanger, aCGH data in blood showed no abnormalities. SNP mosaic analysis identified a 14 Mb duplication on chr17 in approximately 45% of assayed saliva cells, confirmed by aCGH on saliva (6K BAC array). This mutation appears to be tissue-limited. FISH validation was not possible. Mosaic trisomies of chromosome 17 are associated with substantial heart defects, including truncus arteriosus and Tetralogy of Fallot, as well as speech delay[210], consistent with phenotypes in the proband, and considered likely pathogenic.

Patient ID263708 required neonatal intensive care with nasogastric feeding. At delivery, the placenta was hypertrophic, and numerous hemangiomata were noted. She had macroglosia, macrocephaly, and hepatic hemangiomata; as well as episodic hypoglycaemia, oligodontia, esotropia, and gynecomastia. The patient had pigmentary mosaicism following Blashko's lines. Clinical karyotype performed locally was normal. Beckwith-Wiedemann syndrome was suspected but clinical testing performed locally was negative. At Sanger, analysis of SNP microarray data for mosaicism identified genome-wide skews of BAFs, believed to reflect a cell-line with unipaternal disomy (Figure 3-9). Some ten or so examples of genome-wide unipaternal disomy have now been reported, with different underlying mechanisms[201]. The dominant manifestation of unipaternal disomic mosaicism is Beckwith-Wiedemann disorder, which is consistent with the majority of the phenotypes in this case. In addition, since Beckwith-Wiedemann is associated with increased tumour risk, this diagnosis can help increase surveillance of tumour development through increased screening[211]. Given the overlap of phenotypes known in genome-wide paternal UPD and the child's phenotypes, the variant was considered likely pathogenic.

Patient ID258190 required seven days neonatal intensive care due to hypoglycaemia and macrosomia (birth weight and head circumference > 99th centile). Congenital muscular torticollis, partial cryptorchidism, and vertebral abnormalities (joint fusions in cervical spine) were noted. He had global developmental delay, and autism. At Sanger, aCGH assay was performed by the DDD informatics team on blood and was negative and mosaic SNP analysis on saliva using MAD was negative. Analysis using triPOD on saliva detected a low level trisomy on chromosome 20. FISH confirmed trisomy in 30% of cells from buccal sampling but absent in cells from lymphocytes, suggesting the mutation is likely tissue limited. Mosaic trisomy 20 syndrome includes head tilt, developmental delay, autistic features, spinal and genital

abnormalities[212], all phenotypes consistent with those observed in this patient; therefore, the mosaic event was considered likely pathogenic.

Patient ID259709 required neonatal intensive care for 31 days with enteral feeding. Developmental milestones were delayed: sitting independently was achieved at 23 months and walking independently began at 3 years. At recruitment, recorded phenotypes included joint laxity, hyper-extensible skin, anterior 'beaking' of lumbar vertebrae and delayed speech and language development. Our analysis of SNP microarray data identified a chromosome-wide loss of heterozygosity (acquired UPD) on chromosome 14 in approximately 65% of assayed salivary tissue. Informative parental genotypes overlapping the mosaic region identified that the UPD resulted from a mosaic loss of the maternal allele (Figure 3-14).

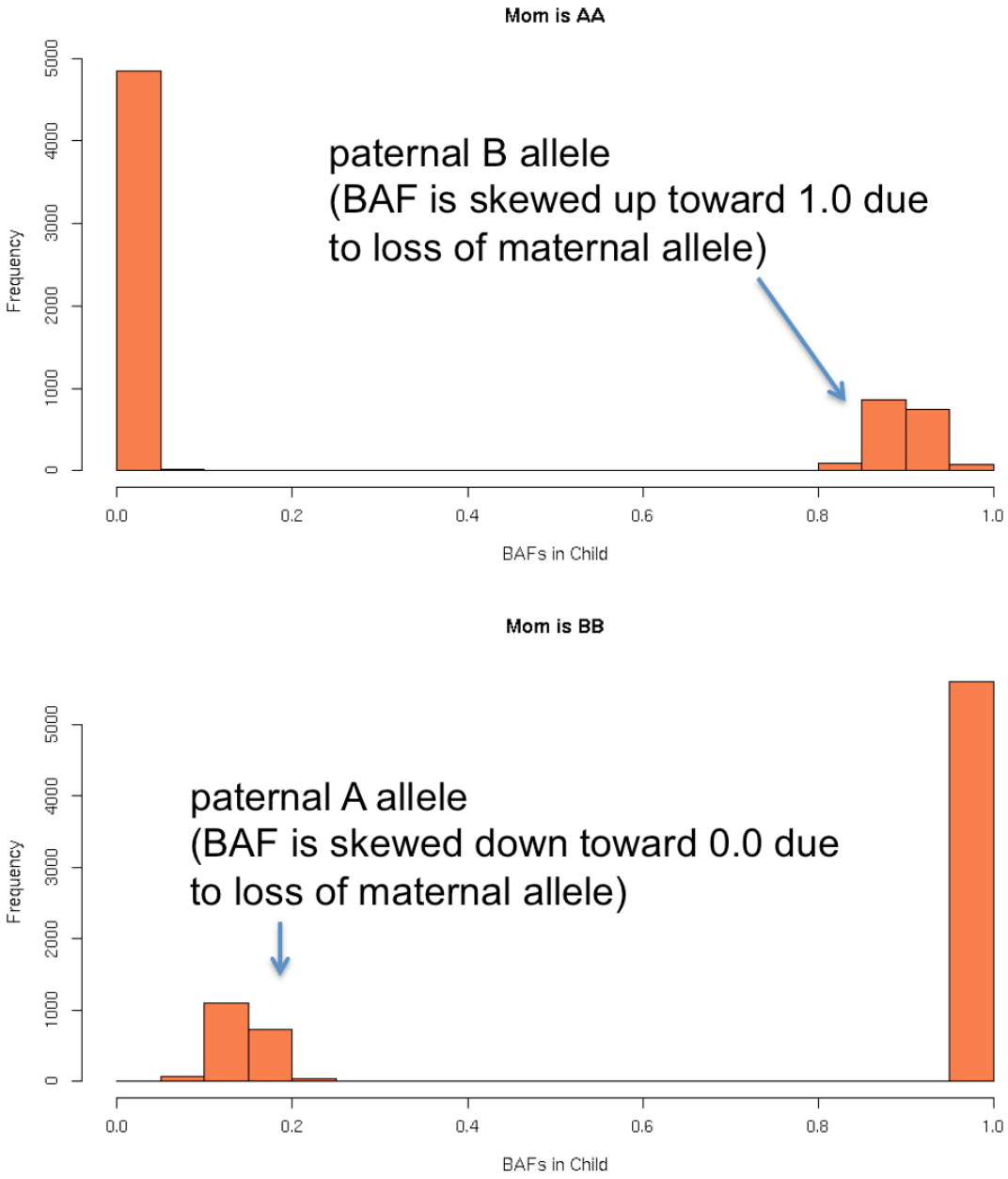

Figure 3-14 aUPD due to loss of maternal allele.

UPD may be pathogenic by causing imprinting disorders or by inheritance of a deleterious variant, present from a carrier parent, to homozygosity. Constitutive UPD 14 maternal is known to cause Temple syndrome, for which feeding difficulties at birth, joint laxity and developmental delay are present[158]. These features are consistent with the child's phenotypes and considered likely pathogenic.

Patient ID257978 had thoracolumbar scoliosis, seizures, somnolence and abnormality of neuronal migration. She demonstrated profound intellectual disability and achieved no developmental milestones. Clinical karyotyping and telomeric MLPA performed locally were normal. At Sanger, SNP mosaicism analysis identified an 80 Mb loss-of-heterozygosity (acquired UPD) region on chromosome 5 in 24% of assayed

salivary cells. Conversion to homozygosity of a deleterious variant in the UPD was suspected to underlie the pathogenicity. Of seven such variants, the most interesting candidate was a missense variant in *N4BP3*, a gene recently reported to be required for normal neuronal axonal branching[213]. The sequencing reads of this variant were inspected to test whether the deleterious allele was skewed toward homozygosity and it was observed that of the sequencing reads overlapping this variant position, 46 supported the alternate alleles, while only 28 supported the reference allele, suggesting that the alternate allele is homozygous in the mosaic cell line. Nevertheless, this gene has not previously been implicated in developmental disorders; therefore, a definitive relationship between this variant and the phenotype in the child was difficult to assess, and the variant as considered of uncertain pathogenicity.

Patient ID259029 was born at 40 weeks gestation with a birth weight of 3.3 kg (41st centile). The child has dysmorphic facies including severe hypertelorism and local clinical testing for craniofrontonasal dysplasia was negative. At Sanger aCGH performed by the DDD laboratory and informatics team on saliva was not obviously abnormal. Mosaic analysis detected a low-clonality (23%) 3 Mb mosaic event on chromosome 11, with a small elevation of LRR (0.09). Intellectual disability and hypertelorism are shared phenotypes with patient 255428 in the DECIPHER database with an overlapping duplication. This region contains *ALX4*, a gene implicated in skull ossification defects, which may be consistent with hypertelorism[214]. However, this region has not been consistently identified with other specific phenotypic features in the child and therefore the variant was considered of uncertain pathogenicity.

3.5 Discussion

The main aim of this experiment was to investigate whether children with developmental disorders have a significant burden of mosaic structural abnormalities relative to age-matched controls. A ~40-fold enrichment of mosaicism in cases compared to controls was observed. Using single-sample and trio-based approaches, 0.9% of DDD probands were found to have large-scale mosaicism. The substantial burden in cases suggests that many of these events were pathogenic. The phenotypes in each child were assessed for consistency with the known consequences of the underlying mosaic mutations and clinical evaluation assessed that 10 of 12 were highly likely to be pathogenic.

One component of this study explored the relative performance of single-sample vs. trio-based mosaic detection methods. Both methods discovered a majority of the total detections and neither software tool was clearly advantageous compared to the other. triPOD identified two events of lower-clonality not found by MAD. While MAD has diminished sensitivity to lower clonality events, it does not require complete trio data, a resource not always available; in this analysis, two real mosaic events detected by MAD lacked complete trio data and were not analysed by triPOD. Also, one third-haplotype gain was not found by triPOD and the false positive rate of triPOD was higher than MAD. These findings suggest that employing either tool can identify the majority of mosaic events but that maximal sensitivity can be gained by leveraging the complementary strategies of both tools if trio data are available.

Assessing the pathogenicity of mosaic copy-number and copy-neutral events requires several assumptions, primarily, that events present in mosaic form cause phenotypes similar in character, if perhaps less severe, than events present in constitutive form. The majority of events detected were copy-number variable mosaicism, which is consistent with previous studies, such as Conlin *et al*.[36]. However, in contrast to that study of mosaic aneuploidy, much lower levels of gonosomal aneuploidy were observed (0 in 1,303, compared with 9 of 2,019), and only a single event affected the whole chromosome. This may be due to differences in ascertainment, as nearly 80% of DDD probands were pre-screened by clinical aCGH testing performed locally, which would have high sensitivity to detect chromosome-size CNVs present in a majority of cells. In addition, gonosomal aneuploidy results in distinctive phenotypes, which are likely to trigger specific genetic investigations; this may compound the bias against recruiting such patients to a research study focusing on undiagnosed patients.

For these reasons, the observed estimate of mosaic frequency in children with undiagnosed disorders is likely an underestimate of frequency among all children with DD.

Mosaic copy-number events were typically not detected by standard aCGH analysis. The detection of mosaicism requires two conditions: the event must be present in the assayed tissue, and the detection tool must be sufficiently sensitive to identify minimal skews in intensity or allele fraction. No large mosaic copy-number events were identified in healthy controls, supporting prior evidence that large copy-number events are highly pathogenic. On the other hand, one LOH-type event, a category of mutation imperceptible by aCGH, was detected in healthy controls. While constitutive LOH has been identified in 1%-1.5% of children with developmental disorders[37,137], a significant burden compared to the population-level rate (1 in 3,500), the cases studied here did not have a statistically significant enrichment of LOH mosaicism (p greater than 0.05). It remains to be seen whether with increased sample sizes, a burden may become apparent, especially with respect to chromosomes sensitive to imprinting disorders.

The filtering strategy used to identify structural mosaic events was tuned to identify mosaicism 2 Mb or larger, a size threshold that allowed fair comparison across data sets given the variability in SNP density. Intuitively, larger events are more likely to be associated with pathogenicity and empirical observation demonstrates that larger constitutive CNVs are rarely found in healthy children[102]. More powerful genetic assays, such as high-depth whole-genome sequencing will enable a higher-resolution comparison of mosaic events at smaller sizes and allow improved detection of pathogenic mosaicism[215].

The strategy of using inherited duplications to characterise BAF and LRR properties of constitutive duplications for exclusion of putative detections with similar BAF and LRR profiles may have inadvertently filtered some mosaic duplications of very high-clonality. Since the TEDS dataset had SNP microarray data with a higher noise level compared with DDD, this effect may have been more pronounced in the TEDS analysis, which could potentially result in an underestimate of mosaicism in this control group. Nevertheless, the data quality from TEDS was sufficient to detect the size and clonality of mosaic events that were detected in the other cohorts.

The SNP microarray data in the DDD study were mostly derived from salivary DNA extraction. While salivary sampling is non-invasive and represents a mixture of

two tissue types (epiderm via buccal tissue epithelium, and mesoderm via lymphocytes)[216] saliva-derived DNA may have limited sensitivity to low-clonality events confined to a single tissue type. Because ALSPAC and TEDS data were derived from only one tissue type (blood) and the distribution of mosaic events may differ across tissue types, it is possible that our frequency comparison of mosaicism between cases and controls may have been partially confounded by hidden stratification, and indeed some mosaic abnormalities (such as the 12p tetrasomy leading to Pallister Killian syndrome) are rarely detected in blood. Indeed, the observation that the majority of mosaicism detected in DDD was present in epithelial-derived but not mesoderm-derived tissue calls for a future analysis of saliva from healthy children. In addition, this may provide some evidence that mosaicism underlying DD need not propagate into all germ layers to result in syndromic dysfunction. However, our assessment of tissue distribution was limited, as endoderm-derived tissue was not available, and factors that hinder the extrapolation of germ-layer distribution from assayed tissue distribution, such as purifying selection against deleterious mosaicism and sampling error, may have played a role. The subject of tissue distribution is revisited in greater detail in chapters 4 and 5.

Detection of mosaicism in probands and subsequent genetic diagnosis offers reassurances to parents that a subsequent child is not at increased risk of developing the same mutation. Nevertheless, the majority of children with previously undiagnosed genetic disorders still receive no genetic diagnosis after extensive interrogation, including aCGH, exome and SNP-based analyses. Improved detection of all forms of mosaicism is needed, including smaller mosaic abnormalities, such as indels and point mutations. This will require further reductions in sequencing cost and the development of accurate sequence-based mosaicism detection algorithms.

Chapter 4 of this dissertation addresses the development and implementation of a new software tool that analyses targeted and whole-genome sequencing data to detect structural mosaicism.

4 MOSAIC STRUCTURAL VARIATION FROM TARGETED AND WHOLE-GENOME SEQUENCING

4.1 Publication Note

Most of the work described in this chapter has been described in a manuscript and is now under editorial review. Unless explicitly stated otherwise, the analysis described herein is the work I performed myself, under the supervision of Matthew Hurles.

4.2 Introduction

Chapter 3 discussed the detection of structural mosaicism in children with DD using SNP microarray data. The metrics and methods used to detect mosaicism from SNP microarray data influenced the mechanics of the sequencing-based tool I developed and describe in this chapter.

Modern SNP microarray technology is well suited for detecting mosaicism because probe density is high (often above 1 million sites per genome) and probes generate allele ratio data with high signal to noise ratio. SNP microarray platforms generate two metrics useful for detecting mosaicism: 1) b allele frequency (BAF): the fraction of the alleles at a locus representing the less-common allele and 2) log R ratio

(LRR): a measure of copy-number, based on the log ratio of signal intensity compared to a reference. These metrics are perturbed differently depending on the nature of the structural abnormality: whereas copy-neutral (loss of heterozygosity; LOH) mosaicism results in a deviation of BAF alone, copy-number (gain or loss) mosaicism additionally alters the LRR. Absolute deviation from the BAF expected for constitutive genotypes (e.g. the expected BAF for a heterozygous genotype is 0.5), called B-deviation (B_{dev}), occurs in mosaic regions when the locus has a mixture of genotypes from wild-type and mosaic tissue. Several software tools (Partek® Genomics Suite, Illumina® cnvPartition, BAFsegmentation[217], and Mosaic Alteration Detection (MAD)[49]) harness this deviation as a signal of mosaicism. As reviewed in chapter 3, the MAD algorithm is open source and has been recently used in several large SNP microarray-based projects[50,218,219]; it identifies mosaic segments using aberrations in $\mathbf{B_{dev}}$ and then labels aberrant segments as copy-loss, copy-gain, or copy-neutral events based on the alteration of the LRR from baseline, a deviation referred to here as copy-deviation, or $\mathbf{C_{dev}}$.

Most DDs are caused by rare, small (SNV and indel) variants that are rarely assayed on microarrays[137]. Therefore, to achieve more comprehensive assessment of pathogenic mutations, rare disease studies rely heavily on targeted sequencing of the protein-coding regions (‘exons’) of the genome, an approach called whole-exome sequencing (WES)[220]. Indeed, sequencing of the whole genome (WGS) offers several advantages compared to WES, including greater breadth of the genome and more consistent coverage of exons[221]. Due to high cost, WGS is currently used in a minority of rare disease studies, but it will likely become more popular as costs decrease.

In addition to small-scale variation, forms of large-scale structural variation, including copy-number[222] and copy-neutral variation (uniparental disomy (UPD))[105], are also important causes of DD. CNV burden analysis of nearly 16,000 children with DD[102] demonstrated that nearly all CNVs greater than 2 Mb are likely pathogenic (odds ratios for CNVs of 1.5 Mb and 3 Mb were 20 and 50, respectively), and that, for a given size, deletion events are more often pathogenic than duplication events. UPD has been estimated to occur in about 1 in 3,500 healthy individuals[121], but is enriched in children with DD[137], and may result in highly penetrant imprinting disorders, recessive diseases, or may be associated with chromosomal mosaicism[125]. Low-clonality mosaicism is difficult to observe by karyotyping, as inspection of at least 10 cells is required to exclude 26% mosaicism with 95% confidence[26], and is also difficult to observe in

microarray, as the detection sensitivity of mosaic duplications by SNP microarray with about 1 million probes for events of at least 2 Mb in size is limited to events of at least 20% clonality[49]. The median average clonality in recent SNP-based studies of DD for mosaic aneuploidy was 40%[36], and for mosaic structural variation (2 Mb and greater) was 44%[178]. Among children investigated with clinical diagnostic testing, the frequency of autosomal mosaic copy-neutral events was 0.24% (12 in 5,000)[35] and the frequency of autosomal mosaic copy-number events was 0.35% (36 in 10,362)[194]. Combining these frequencies yields a combined frequency of 0.59% of mosaic structural variation in children with DD.

The detection of large-scale mutations from WES data is challenging because the input data typically represent a sparse sampling of the genome, as the targeted regions typically cover only about 2% of the genome[221], and sequence read-depth at exons is biased by enrichment efficiency and other factors[223]. Despite these limitations, exome-based software tools have been successfully engineered to detect large-scale *constitutive* mutations, including copy-number variation[62,224-227] and copy-neutral variation (bcftools roh (in preparation) and UP*D*io[137]). These tools are insensitive to *mosaic* abnormalities, however, because they typically rely on single metrics, such as copy-number change (rather than copy-number *and* allele-fraction), or on genotype, which is not well assessed in mosaic state. Specialised methods have been developed for the analysis of cancer exomes where tumour and normal tissue can be isolated[228,229] or, in the context of a parent-foetus trio, for foetal DNA in maternal plasma[75]. However, a method to detect copy-number and copy-neutral mosaicism from an individual's exome (or genome) is lacking, but if available, could further extend the range of sequence-based analyses.

I developed MrMosaic, a method that detects structural mosaicism using joint analysis of B_{dev} and C_{dev} in targeted or whole-genome sequencing data. Simulations demonstrated superior performance of MrMosaic compared to the MAD algorithm. Using MrMosaic, I analysed WE data from 4,911 children with developmental disorders and identified 11 structural mosaic events in 9 individuals, 6 of whom exhibited tissue-specific mosaicism.

4.3 Materials & Methods

4.3.1 MrMosaic

I worked with Alejandro Sifrim, Ph.D., a post-doctoral researcher in Matt Hurles' group, to create MrMosaic. Alejandro introduced me to the tricube distance as a decay function and the use of the Fisher's Omnibus method to combine p values from statistical tests. The other statistical steps in the algorithm were developed in collaboration with Drs. Sifrim and Hurles. I integrated multi-threaded support to provide faster implementation on a multi-core CPU, developed 'wrapper' functions to facilitate implementation in a 'pipeline' environment, executed MrMosaic on DDD data and analysed and interpreted the results.

The algorithm consisted of several steps: statistical testing, segmentation, filtering, and results visualisation. 'BAF' is used below as shorthand for 'non-reference proportion'.

The input data for MrMosaic consist of genomic loci with measured B_{dev} values, C_{dev} values, and genotypes, stored in a tab-delimited file.

The loci selected for inclusion in the input data were di-allelic, single-nucleotide, polymorphic (1% - 99% MAFs among European individuals in the UK10K[230] project), autosomal positions. For exome analysis, only loci overlapping targeted regions of the exome design were used. At these loci, B_{dev} and C_{dev} values were calculated as described in the following two paragraphs.

B_{dev} values were generated using the following method: the identity of the alleles at each locus was extracted using fast_pileup function in the perl module Bio::DB::Sam (https://github.com/GMOD/GBrowse-Adaptors/tree/master/Bio-SamTools), using high-quality reads (removal criteria: below base quality Q10, below mapping quality Q10, improper pairs, soft- or hard-clipped reads) and BAF was calculated as the number of reference bases divided by the number of reference bases and non-reference bases. Heterozygous sites were defined as loci with a BAF between 0.06 and 0.94, inclusive, instead of defining heterozygous sites based on a genotype caller, as this static threshold range is more lenient of sites with small numbers of alternate reads, and I wanted to be sensitive to detect low clonality mosaicism. The B_{dev} was calculated at heterozygous sites as the absolute difference between the BAF and 0.5. Only loci with sufficient read coverage (at least 7 reads) were used for analysis.

C_{dev} values were generated using the following method: the average read-depth for each target region was counted, the $\log_2$ ratio for that target region was calculated by comparing its read-depth to a reference read-depth, where the reference value was defined as the median read-depth among the distribution of read-depths at that target region from dozens of highly correlated samples. This $\log_2$ ratio was normalised based on several covariates pertaining to each target region (covariates included were: GC-content, hybridisation melting temperature, delta free energy), a process used in an exome-based CNV detection algorithm called Convex[6]. Lastly, I generated the C_{dev} value using the Aberration Detection Algorithm v2 (ADM2) method by Agilent® (p.496 of http://www.chem.agilent.com/library/usermanuals/public/g3800-90042_cgh_interactive.pdf), which produces a value from the normalised $\log_2$ ratio that is error-weighted to reflect higher confidence in regions with more depth.

The statistical testing step of the MrMosaic algorithm began by data smoothing, using a rolling median (width of 5) across heterozygous and homozygous sites, so as to utilize the depth information in homozygous sites to reduce variance. From this point forward, only heterozygote sites were considered, as mosaic abnormalities do not affect B_{dev} of homozygous loci. Statistical testing assesses whether a given locus is significantly deviated from the B_{dev} and C_{dev} means given the null hypothesis of no chromosomal abnormality. At every heterozygote site I computed two Mann Whitney U tests, one for B_{dev} and one for C_{dev}, testing the alternative hypothesis that the distribution of the metric in the neighborhood of the chosen site was greater (has a higher median rank) than the distribution of the background. I used 10,000 randomly selected sites, from all autosomes excluding the current chromosome, as the background population. In order to account for non-uniform spacing of the data points when generating the neighbourhood metric I applied a distance-weighted resampling scheme, to down-weight more distant points from the chosen site. The tricube distance, inspired by Loess smoothing, was chosen as a decay function for the resampling weights and considered data points up to 0.5 Mb upstream and downstream of the given position. An equal number of data points was then sampled around the chosen site and from the background (n=100) and the Mann-Whitney U test was performed. Finally, I combined the p values of the two statistical tests (one for B_{dev} and C_{dev}) for every position using Fisher's Omnibus method.

The segmentation step operated on the combined p value generated above. Segmentation was performed using the GADA[42] algorithm, using the parameters values as follows: SBL step: maxit of 1e7; Backward Elimination step: T value of 10 and MinSegLen value of 15. This step generated contiguous segments of putative chromosomal abnormalities. Segments in close proximity (within 1Mb) that showed the same signal direction (loss, gain, LOH) were merged during post-processing to reduce over-segmentation.

The filtering step was required to enrich the segments generated above for those that were likely reflective of true mosaicism. Whilst testing MrMosaic in exome simulation analyses, I observed that true-positive detections (those overlapping simulated events) tended to be larger (had greater number of probes) and had stronger evidence of deviation (had higher GADA amplification values) than putative segments that did not overlap with simulated regions (i.e. false-positive, spurious calls) (Figure 4-1).

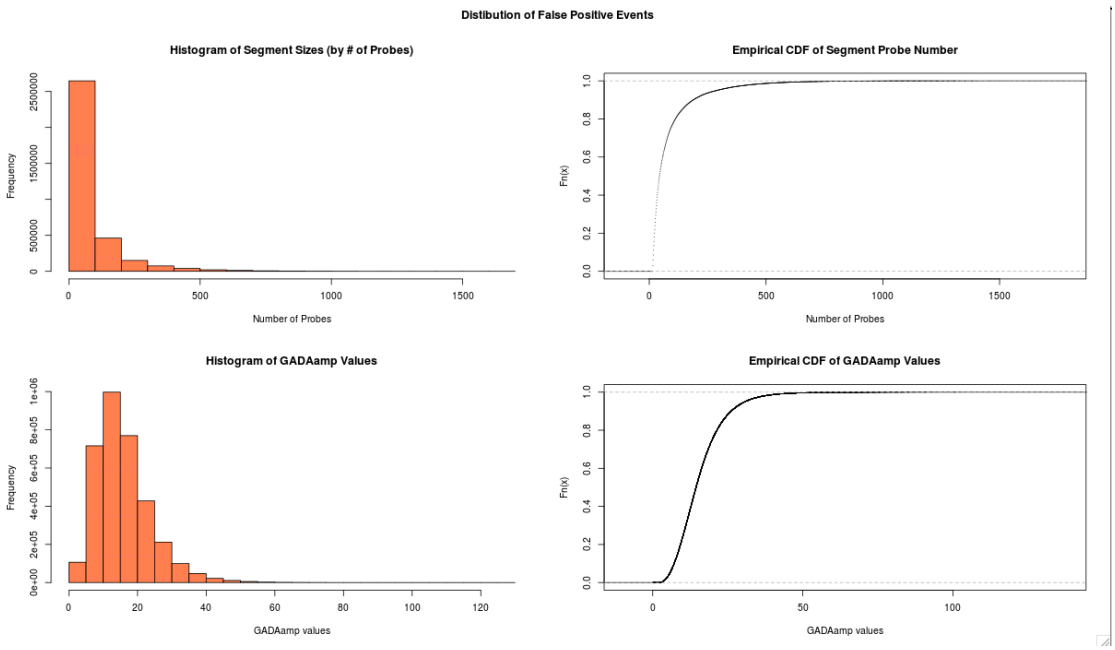

Figure 4-1 Distribution of size and signal-strength of false positives. The histograms of probe-number and GADAamp values both show long tails, with the majority of putative events being smaller and weak. The cumulative distribution functions from the data (right column) showed that events with greater than about 100 probes or about 25 GADAamp were very rare in the false positive events; true events (shown in the next figure) had far larger and have stronger signals.

I integrated these two observations into a single scoring metric calculated from the empirical cumulative distribution functions for 'number of probes' and 'GADA amplification value' of false-positive segments, and assessed the composite probability

that a given segment comes from these distributions, such that: Mscore = abs($-\log_2(x)$ + $-\log_2(y)$) where x and y refer to these empirical cumulative distribution functions. Thus, the Mscore is a quality-control metric derived by combining the size and signal-strength of detections. I then used the Mscore to filter out those events most likely to represent false positives. I selected events with an Mscore of 8 or greater for analysis because I observed that this appeared to provide a good balance between sensitivity and specificity Figure 4-2 and Figure 4-3).

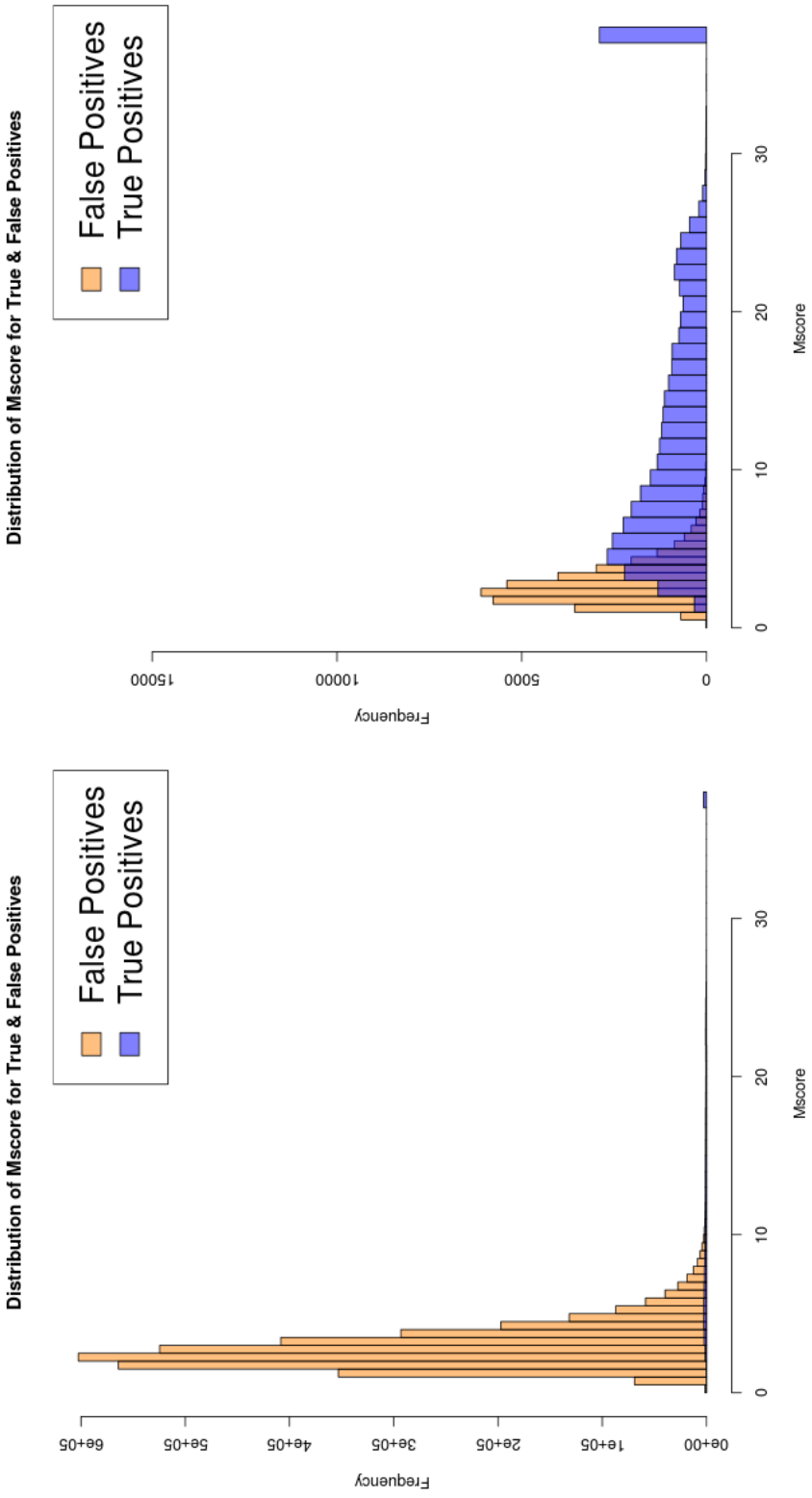

Figure 4-2 Comparing Mscores of true positives and false positives. The Mscore distributions for all simulated false positive events (first graph) and for a random subselection of false positive events equal to the number of true positive events (second graph) demonstrated that the true positive events in general have higher Mscores. The accumulation of true positive events at ~40 was an artefact of assigning a maximum cut-off to an R "-Inf" value.

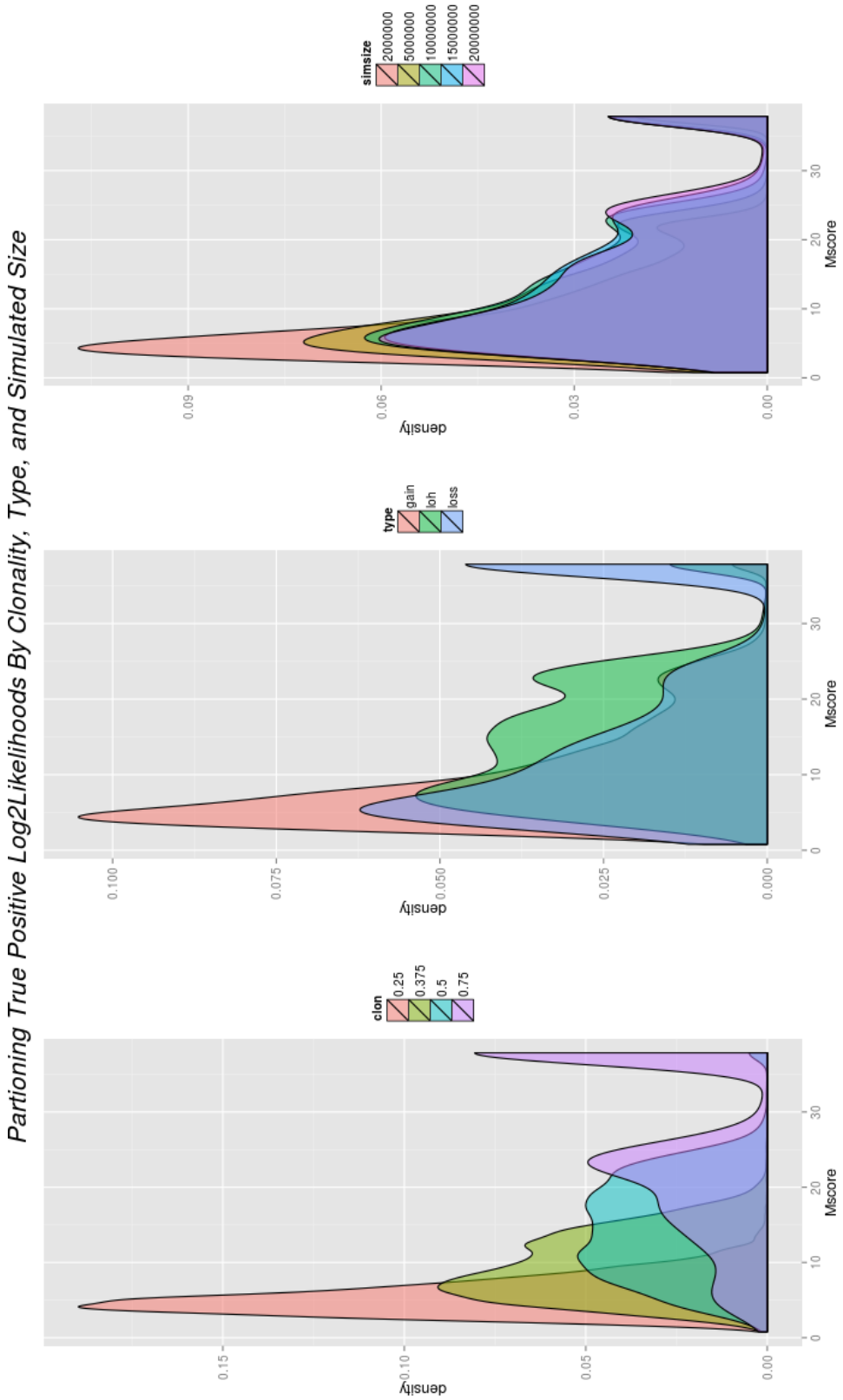

Figure 4-3 Stratifying Mscore by simulation clonality, type, and size. I stratified the true positive events by Mscore to better define the relationship between Mscore thresholds and simulated mosaic events. The mosaic events with the lowest Mscore were those at the lowest clonality (left side of left graph).

The visualisation step generates a detection table and detection plots. The detection table consists of mosaic abnormalities detected and contains the following data: chromosome, start_position, end_position, log2ratio_of_segment, bdev_of_segment, clonality, type, number_of_probes, GADA_amplification, p_val_nprobes, p_val_GADA_amplification, Mscore. Event clonality was calculated by assessing the type of mosaic event based on LRR and converting the B_{dev} value to clonality based on the type of event (Table 4-1). The detection plots are showing the loci and BAF and C_{dev} data for each chromosome in which a mosaic abnormality is detected, as well as a genome-wide lattice plot using the data for all chromosomes.

Simulation metrics	**Normal**	**Loss**	**Gain**	**LOH**
LRR	0	$log2(\frac{2-m}{2})$	$log2(\frac{2+m}{2})$	0
Simulated Read Depth (SDP)	$\lambda_i = \widetilde{DP_i} \cdot S$ $SDP_i \sim Poiss(\lambda_i)$	$\lambda_i = \widetilde{DP_i}\left(\frac{2-m}{2}\right)S$ $SDP_i \sim Poiss(\lambda_i)$	$\lambda_i = \widetilde{DP_i}\left(\frac{2+m}{2}\right)S$ $SDP_i \sim Poiss(\lambda_i)$	$\lambda_i = \widetilde{DP_i} \cdot S$ $SDP_i \sim Poiss(\lambda_i)$
B-allele frequency (B_{dev})	$p = 0.5$ $B_{dev,i} \sim Binom(SDP_i, p_i)$	$p = 0.5 \pm \frac{m}{2(2-m)}$ $B_{dev,i} \sim Binom(SDP_i, p_i)$	$p = 0.5 \pm \frac{m}{2(2+m)}$ $B_{dev,i} \sim Binom(SDP_i, p_i)$	$p = 0.5 \pm \frac{m}{2}$ $B_{dev,i} \sim Binom(SDP_i, p_i)$

Table 4-1 Functions to Prepare Simulations. m: Clonality as in proportion of cells with abnormality; $\widetilde{DP_i}$: Median read depth (after quality filtering) at position *I; S*: Scaling factor so that $Target\ Average\ Read\ Depth = 75.2 \times S$; SDP_i : Simulated Read Depth at position *I;* p: Proportion of reads with alternative allele at position *i*

MrMosaic is primarily written in the R language, available as an open-source tool at https://github.com/findingdan/MrMosaic. The algorithm can be used in multi-threaded mode to facilitate whole genome analysis. Analysis of a single whole-exome using a single thread was completed in 15 minutes when tested using a single core of an Intel Xeon 2.67Ghz processor and 500 Mb of RAM. Whole-genome analysis using 24 cores required 30 Gb of RAM and 7 hours. Whole-genome analysis can be substantially shortened if the number of sliding windows is reduced or the window size is increased.

4.3.2 Simulating Mosaicism

I devised a series of simulation experiments to assess MrMosaic performance for various events, across type (LOH, gains, losses), clonalities, sequencing depths, platforms (whole-exome (WE) and whole-genome (WG)) and to compare performance

to the MAD method. I compared performance to a modified version of MAD I adapted to enable more flexible execution in a parallel-computing environment, but identical with respect to statistical methods.

The simulation method consisted of these steps: (1) loci selection, (2) calculating depth at these loci, (3) parameter space and number of trials, (4) adjusting read depth in simulated regions, (5) calculating final real depth, (6) selecting sites based on minimum depth, (7) calculating relative copy-number, (8) assigning genotypes, (9) calculating the BAF for each site, (10) calculating performance. Steps 1-3 differed between the WE and WG simulations and are described first below. The remaining steps 4-10 were executed consistently for WE and WG simulations.

For WE simulations, loci selection (1) was based on di-allelic single nucleotide polymorphic positions (between 1% and 99% UK10K[230] European minor allele frequency) in the V3 version of the target-region design (Agilent® Human All Exon V3+). To calculate depth at these loci (2), at each locus *i,* baseline sequence read depth ($\widetilde{DP_i}$) for these sites was defined as the median of the read depth distribution among 100 parental exomes for each site, considering only high-quality reads (mapQ at least 10, baseQ at least 10, properly mapped read-pairs), where parental exomes had a mean average sequencing output of 67x (calculated where x was the number of QC-passed & mapped reads without read-duplicates * 75 bp read length / 96 Mb targeted bp). The parameter space (3) consisted of the following: target average sequencing coverage (in fold coverage) {50, 75, 100}, event clonality $m \in \{0.25, 0.375, 0.5, 0.75\}$, type {loss, gain, LOH}, and size {2e6, 5e6, 1e7, 2e7}. Two hundred trials (4) were conducted per parameter combination for a total of 36,000 simulations.

For WG simulations, the loci selection (1) was based on di-allelic single nucleotide polymorphic (1% - 99% European MAFs from 1000G[146] May-2013 release) autosomal positions. To calculate expected depth at these loci (2), I calculated a scaling factor for each locus based on the median read depth of the first two median absolute deviations of the distribution of coverage for that site seen across 2,500 low-coverage samples in the 1000Genomes[146] project. A site-specific scaling factor was calculated as the deviation of each site's read depth from the average read depth across all polymorphic positions. Simulation depth was defined at each site as the desired simulation coverage multiplied by site-specific scaling factor. The parameter space (3) consisted of two experiments: 1) average genome coverage of 25x, event clonality $m \in$

{0.25, 0.375, 0.5, 0.75}, type {loss, gain, LOH}, and size (Mb) {1e5, 2e6, 5e6}; and 2) 5 Mb 50% clonality event captured at average genome coverages (in x) {30, 40, 50, 60} for the three mosaic types {loss, gain, LOH}. One hundred trials (4) were conducted per WG simulation.

The remaining simulation steps 4-10 described below were performed consistently for WE and WG simulations. For each simulation a single mosaic event was introduced into each simulation trial. The adjustment of read-depth in simulated regions (4) was performed using a scaling factor based on the type and clonality of the simulated event, m, while sites not overlapping copy-number simulated events would not undergo this scaling step. To calculate the final simulated read depth (5) for each site i (SDP_i), I sampled from a Poisson distribution with λ_i equal to the scaled read depth (Table 4-1). Only positions with a final read depth (6) of at least 7 reads were included for analysis. Relative copy-number (7) was defined as $\log_2$ of the ratio of the final read depth to the baseline read depth.

The simulation of genotypes (8) (AA, AB, or BB) at each position *i* was determined based on the site's minor allele frequency, which was used in a multinomial function with probabilities corresponding to Hardy Weinberg-assumed genotype proportions (p^2, 2pq, q^2). To calculating the BAF for each heterozygote at site *i* (9), I adjusted the expected heterozygote proportion of 0.5 with respect to the chosen event type and clonality, and sampled from a binomial distribution given this adjusted proportion and the simulated read depth at *i*. BAFs for homozygote reference (AA) and non-reference (BB) sites were chosen by sampling from a binomial distribution with p=0.01 or p=0.99 respectively and the simulated read depth at *i*.

MrMosaic and MAD were applied on the simulated WE and WG samples generated by the above procedure and performance was measured using precision-recall metrics (10). A 'success' in a trial was considered a detection overlapping the simulated mosaic event. Precision was calculated as the number of successes divided by the number of detections. Recall was defined as the proportion of trials with a success.

4.3.3 Description of Samples & Sequencing

The samples used in this analysis derived from the DDD study. DNA was extracted from blood and saliva by local clinical teams and was processed at the Wellcome Trust Sanger Institute. The array CGH and exome sequencing were performed by the Sanger Institute array and sequencing cores. There were 4,926 DNA samples analysed in this

study from 4,911 children, as some children were analysed using both blood and saliva. The majority, 3,260 of 4,926 (66%) of the DNA samples were extracted from saliva.

Exome sequencing was performed by the Sanger Institute sequencing core as fully described elsewhere[137]. In brief, DNA was enriched using a Agilent® exome kit, based on the Agilent Sanger Exome V3 or V5 backbone and augmented with 5 Mb of additional custom content (Agilent Human All Exon V3+/ V5+, ELID # C0338371). An 'extended target region' workspace was defined by padding the 5' and 3' termini of each target region by 100-bp yielding a total analyzed genome size of approximately 90 Mb. Sequencing was performed by the sequencing core using the Illumina® HiSeq 2500 platform with a target of at least 50x mean coverage using paired-end sequence reads of 75-bp read-length. Measured exome coverage ranged from 14x to 155x with a mean of 69x (Figure 4-4).

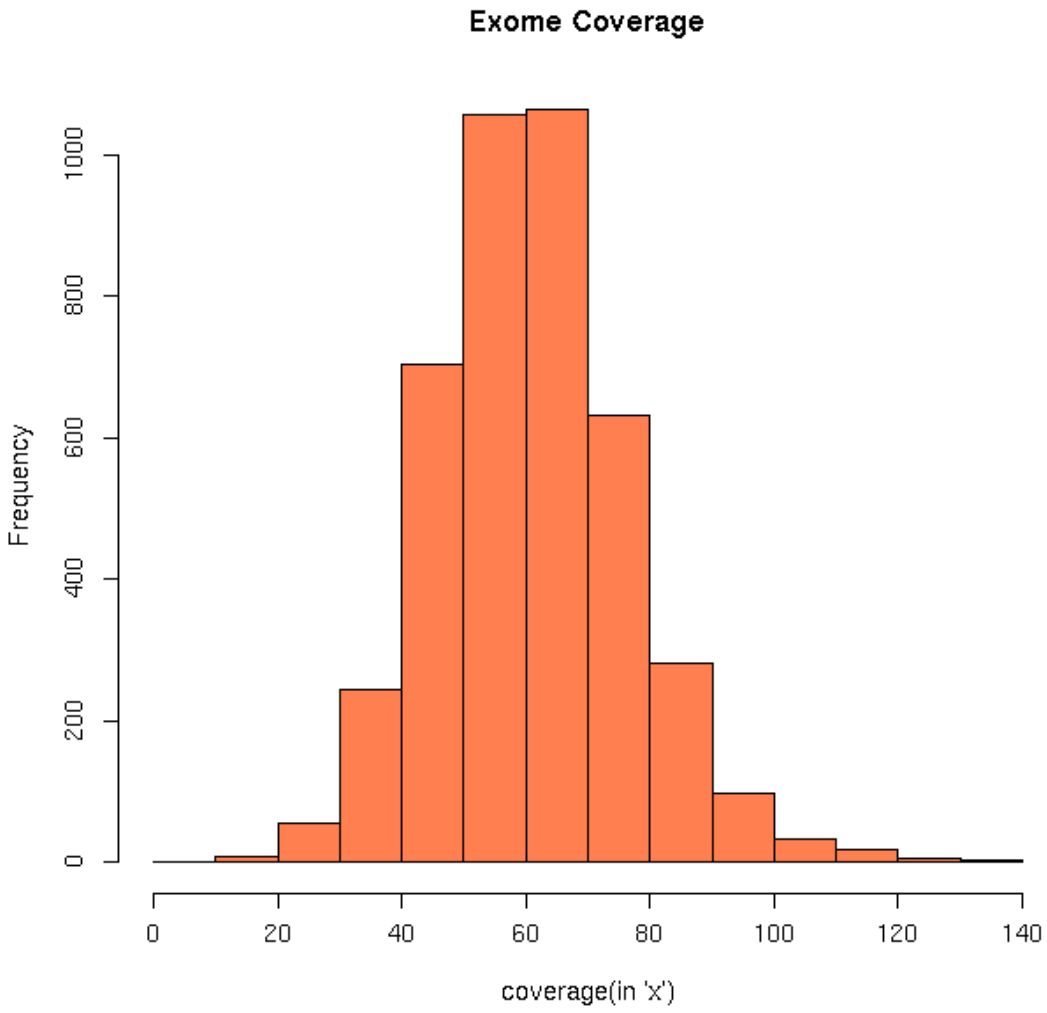

Figure 4-4 The distribution of average coverage of exomes used in this study.

Alignment to the reference genome GRCh37-hs37d5 was performed by the Human Genetics informatics team using bwa[57] version 0.5.9 and saved in BAM-format[58] files.

Additionally, I processed two exome samples *post hoc* from saliva after SNP genotyping chip analysis showed mosaicism was present in saliva but absent in blood. These two exome samples and the exome sample with suspected revertant mosaicism were processed separately from the exome experiment described in the previous paragraph. For these three exomes, the Agilent Sanger Exome V5 target kit was used, and sequence depth ranged from 387x - 455x coverage (reads = {465,522,627, 483,098,826, 549,766,632} * 75bp read-length / 90e6 target-region-size). The sample with suspected underlying mosaic reversion had 549,224,891 QC-passed & mapped reads, and 57,165,328 duplicates, and therefore had a mapped read coverage of 410x ((549,224,891-57,165,328)) * 75 / 90e6).

For the sample for which whole genome sequencing data were generated, sequencing was performed by the Sanger Institute sequencing core using an Illumina® X-Ten sequencing machine. Library fragments of 450-bp insert-size were used and paired-end 151-bp read-length sequence reads were generated. Alignment to the reference genome GRCh37-hs37d5 was performed by the Human Genetics informatics team using bwa mem[57] version 0.7.12. I calculated average coverage using samtools flagstat as the number of QC-passed mapped-reads without duplicates using 151 bp read-lengths in a 3Gb genome: (616,151,282 –124,325,581) * 151 / 3e9 = 24.8x. Rearrangement analysis was carried out using Breakdancer[231] v1.0.

4.3.4 Additional filtering implemented in addition to Mscore quality score

Some events with very high Mscores appeared to represent real, but constitutive, abnormalities. I identified two failure modes: constitutive duplications and homozygosity by descent (HBD). Constitutive duplications genuinely produce strong B_{dev} signals in MrMosaic, but also constitutive deletions and large regions of homozygosity (ROH) may potentially produce putative detections if individual probes have mapping artefacts that resulted in spurious signals. I used bcftools roh (developed by Vagheesh Narasimhan, manuscript in preparation) to identify and filter HBD regions and flagged as suspicious events with greater that 25% reciprocal overlap with CNVs detected through constitutive copy-number detection. In addition, I observed several recurrent putative detections, especially prevalent in pericentromeric and acrocentric regions that appeared spurious on the basis of inconsistencies between BAF and LRR,

and I filtered such events by filtering putative mosaic events seen in more than 2.5% of samples.

4.3.5 SNP genotyping chip validation

The Sanger Institute genotyping core used Illumina® HumanOmniExpress-24 Beadchips (713,014 markers) for SNP genotyping, Illumina® GenomeStudio to generate log R ratio and BAF metrics, and Illumina® Gencall software to calculate genotypes. I performed structural mosaic detection using MAD[49]. Initial mosaic events were merged if events were within 1 Mb, and were the same type (loss, gain, or LOH) of mosaic event. Results were plotted using custom R code.

4.4 Results

I developed a new computational method, MrMosaic, to detect structural mosaic abnormalities from high-throughput sequence data (Methods). In summary, this method identifies chromosomal segments with clustered deviations in allelic proportion and copy number, relative to randomly selected sites on other chromosomes from the same data. Initially, measures of deviation of allelic proportion (B_{dev}) and copy number (C_{dev}) are computed from the WE/WG data at well-covered known polymorphic SNVs. Whereas B_{dev} is only assessed at heterozygous sites, C_{dev} integrates information from flanking non-heterozygous sites to reduce noise. The statistical significance of the observed B_{dev} and C_{dev} are assessed separately, using non-parametric testing, and the resultant p values are subsequently combined and then segmented using the GADA algorithm[42]. I devised a confidence score, the Mscore, to curate putative detections of mosaic segments by integrating metrics that discriminate between true positive and false positive mosaic detections (Figure 4-5).

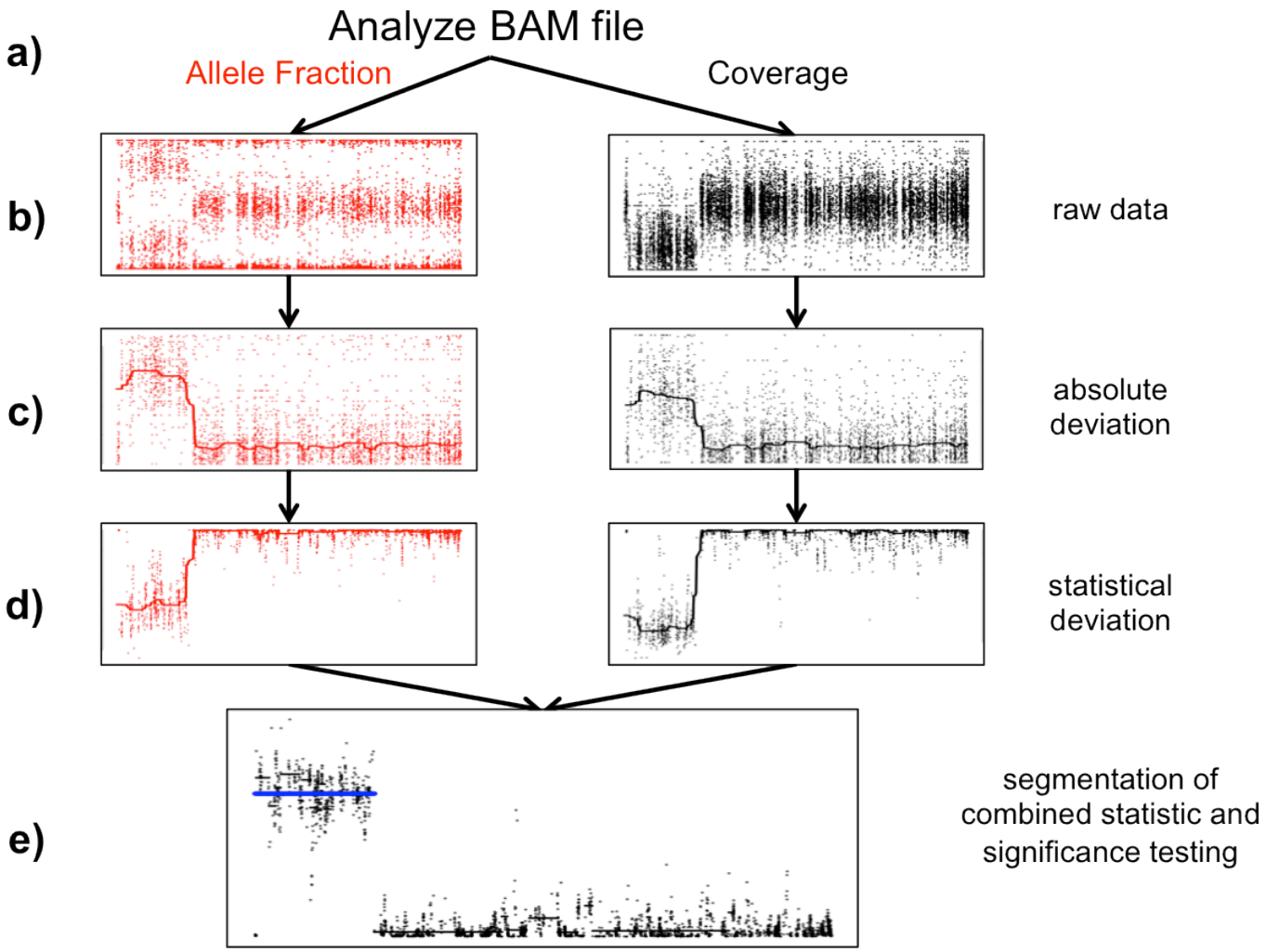

Figure 4-5 Detecting structural mosaicism using MrMosaic. a) Exome data are stored in a BAM file from which allele fraction (left column) and coverage (right column) are measured at polymorphic positions within or near target regions. b) A simulated mosaic deletion is depicted and the raw data, consisting of BAFs and normalized coverage are plotted for a simulated mosaic deletion. c) Absolute deviation of BAF (B_{dev}) and normalized coverage (C_{dev}) at heterozygous sites are analyzed. d) Mann Whitney U Tests are performed separately for B_{dev} and C_{dev}, comparing the signal detected in sliding windows in this chromosome, compared with a randomly selected chromosome for background. The test statistics are depicted on the log scale. e) The p values of the Mann Whitney U Tests are combined and segmented (black lines). Segments passing the Mscore significance threshold are plotted in blue.

4.4.1 Simulations

I performed simulations (Methods) to explore the performance of MrMosaic for three different classes of structural mosaicism: gains, losses and LOH, in several contexts.

The performance results across mosaicism of different *sizes*, *clonalities* and sequencing *coverage* are summarised in Figure 4-6 or both WE and WG data.

Across all measured categories, mosaic duplications were more difficult to identify than deletion or LOH events, especially at lower (25%) clonality (Figure 4-6).

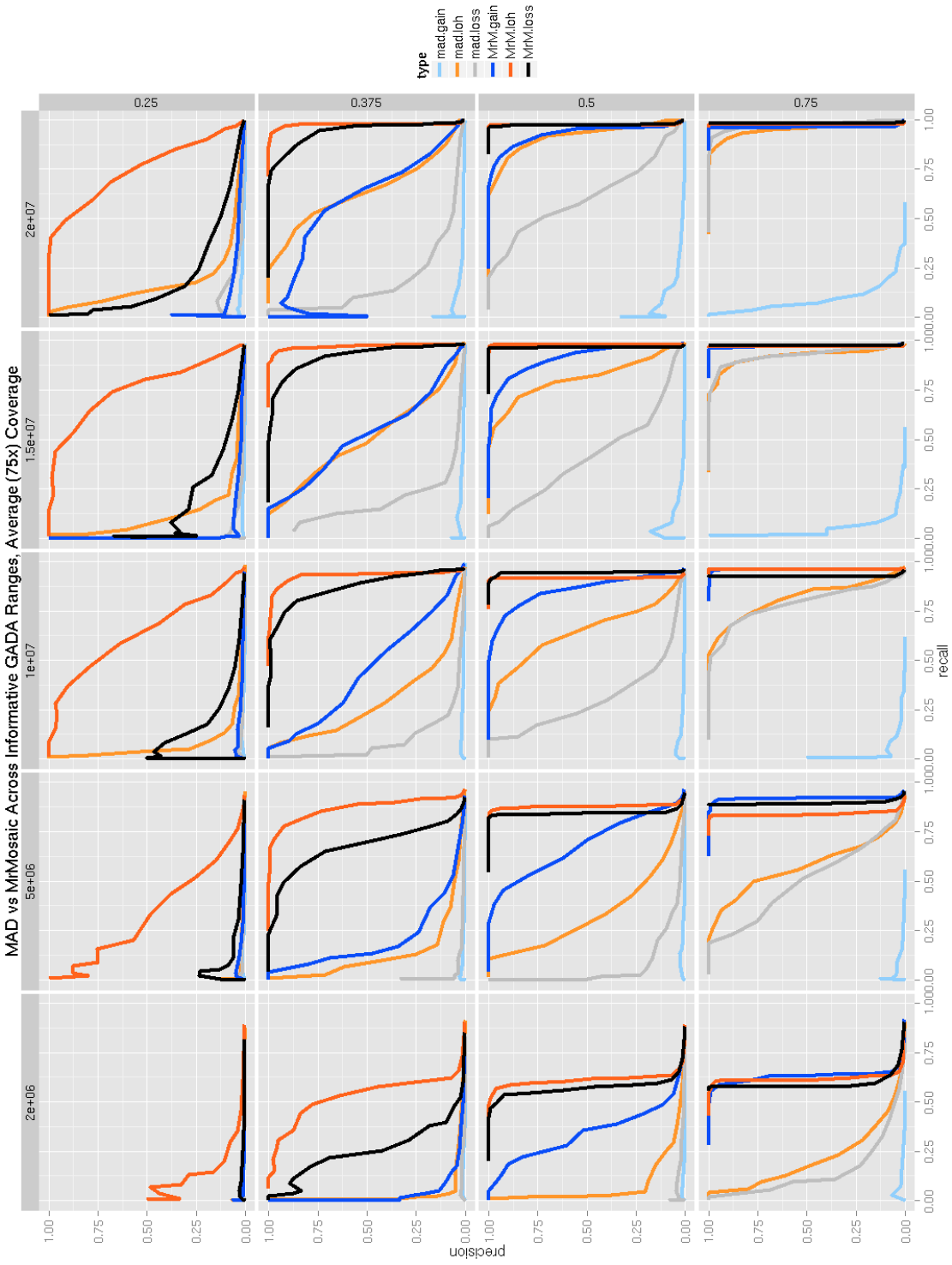

Figure 4-6 WE performance of MAD and MrMosaic algorithms. In this grid of precision-recall graphs, the performance of MAD and MrMosaic is compared at 75x average coverage for a range

of sizes (columns), clonalities (rows), and for the three types of mosaic abnormalities (colors) run with either MAD or MrMosaic (shades). Performance of both algorithms improves with increasing simulated event size (due to more assayed informative points) and at higher clonalities (due to a stronger deflection of non-reference proportion (B_{dev}) and coverage (C_{dev})). MrMosaic performs favorably compared to MAD in all measured categories. This effect is especially apparent for mosaic gains, which is the type of mosaicism that generates the smallest deviations in B_{dev}; unlike MrMosaic, which analyses B_{dev} and C_{dev}, MAD analyses B_{dev} alone.

The most likely explanation for this relative weakness is that duplications result in the smallest deviation of B_{dev}, compared with deletion and LOH events and that the C_{dev} signal does not overcome sampling noise at low clonality. Figure 4-7 shows the relationship between clonality and C_{dev} and B_{dev} for the three classes of mosaicism.

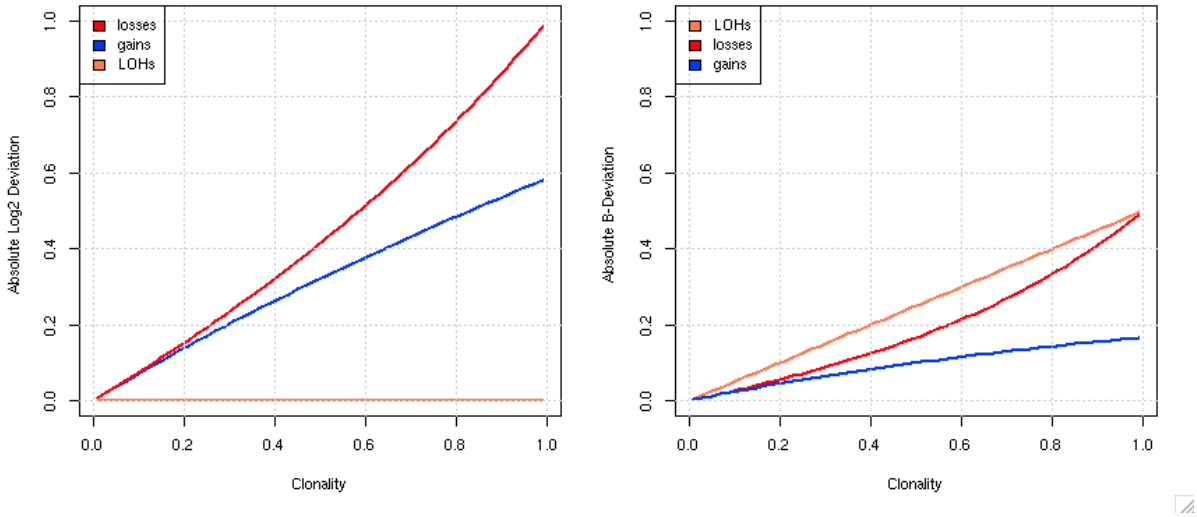

Figure 4-7 Relationship between Clonality and Metrics. The relationship between clonality and measured metrics (C_{dev} and B_{dev}) indicates that while LOH events result in no deviation of C_{dev}, gains have the smallest deflection of B_{dev}, compared to other events of a given clonality.

To further explore the effect of including C_{dev} in addition to B_{dev}, I investigated the performance of MrMosaic using B_{dev} alone compared with joint analysis of B_{dev} and C_{dev}. This analysis showed that incorporation of C_{dev} substantially improved detection of copy-number events above lower clonality, while only a marginally decreased performance of LOH detection (Figure 4-8), consistent with the intuition that C_{dev} yields a valuable net signal when clonality is above the C_{dev} noise floor.

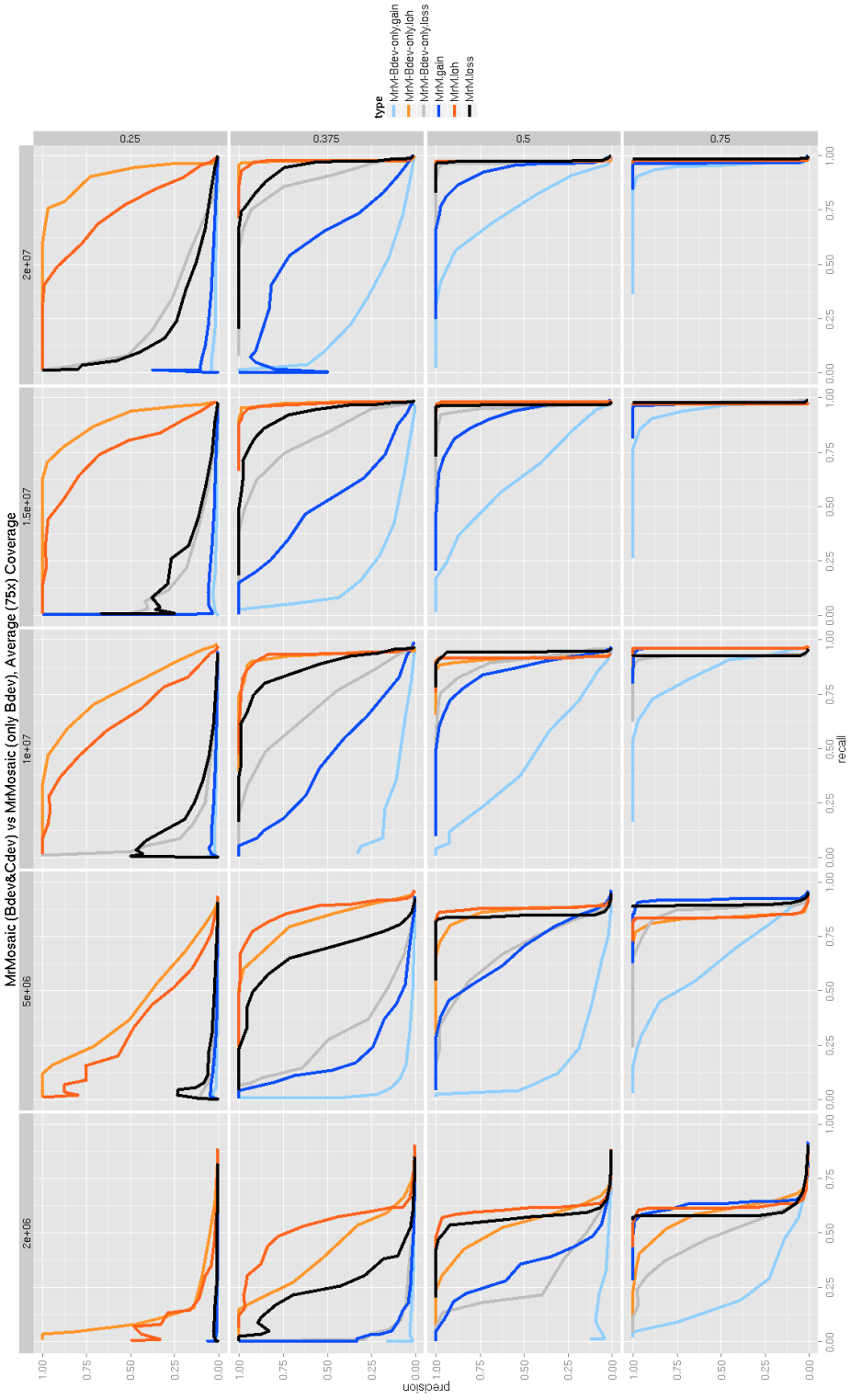

Figure 4-8 WE MrMosaic, Cdev & Bdev vs Bdev-alone. MrMosaic combines the statistical deviation from differences in coverage (C_{dev}) and non-reference proportion (B_{dev}) while the MAD approach uses B_{dev} alone. I ran MrMosaic in standard joint-mode and also using B_{dev} alone. The results demonstrate improved detection when considering joint calling, especially for copy number events above 0.25 clonality. LOH-type mosaicism does not affect copy number (C_{dev}), so considering C_{dev} adds no additional information and has the potential to add noise to the calculation, which may

explain the slightly lower performance of LOH calling in the low-clonality (0.25), large (20 Mb) category.

Simulations showed detection performance increased with larger event *size* (Figure 4-9). WE simulation analysis demonstrated high area under the precision-recall curve (AUC) for all events at least 10 Mb in size and at least 50% in clonality; and, for deletion and loss of heterozygosity (LOH) events at least 5 Mb in size. MrMosaic performed favourably compared to MAD in all measured categories. For WG data simulations demonstrated an AUC of about 0.9 for 100 kb LOH and loss events, and greater than 0.95 for all megabase-size events. WG analyses interrogated nearly 50-fold more sites than exome data (Table 4-2). In the WE simulations, the number of informative sites increased with increasing coverage, a finding driven primarily from an increasing number of sites passing the minimal depth threshold. Whilst the number of sites assayed did not differ in WG simulations, because sequencing coverage is more uniform and at the levels of coverage simulated here (20x minimum), sites always had sufficient coverage. Incidentally, the number of informative sites actually decreased very slightly in the WG simulations at higher coverage, with more sites classified as homozygous (non-informative) because of sampling artefacts, but this effect was small, and far outweighed by the benefit of assaying far greater number of sites compared to WE simulations.

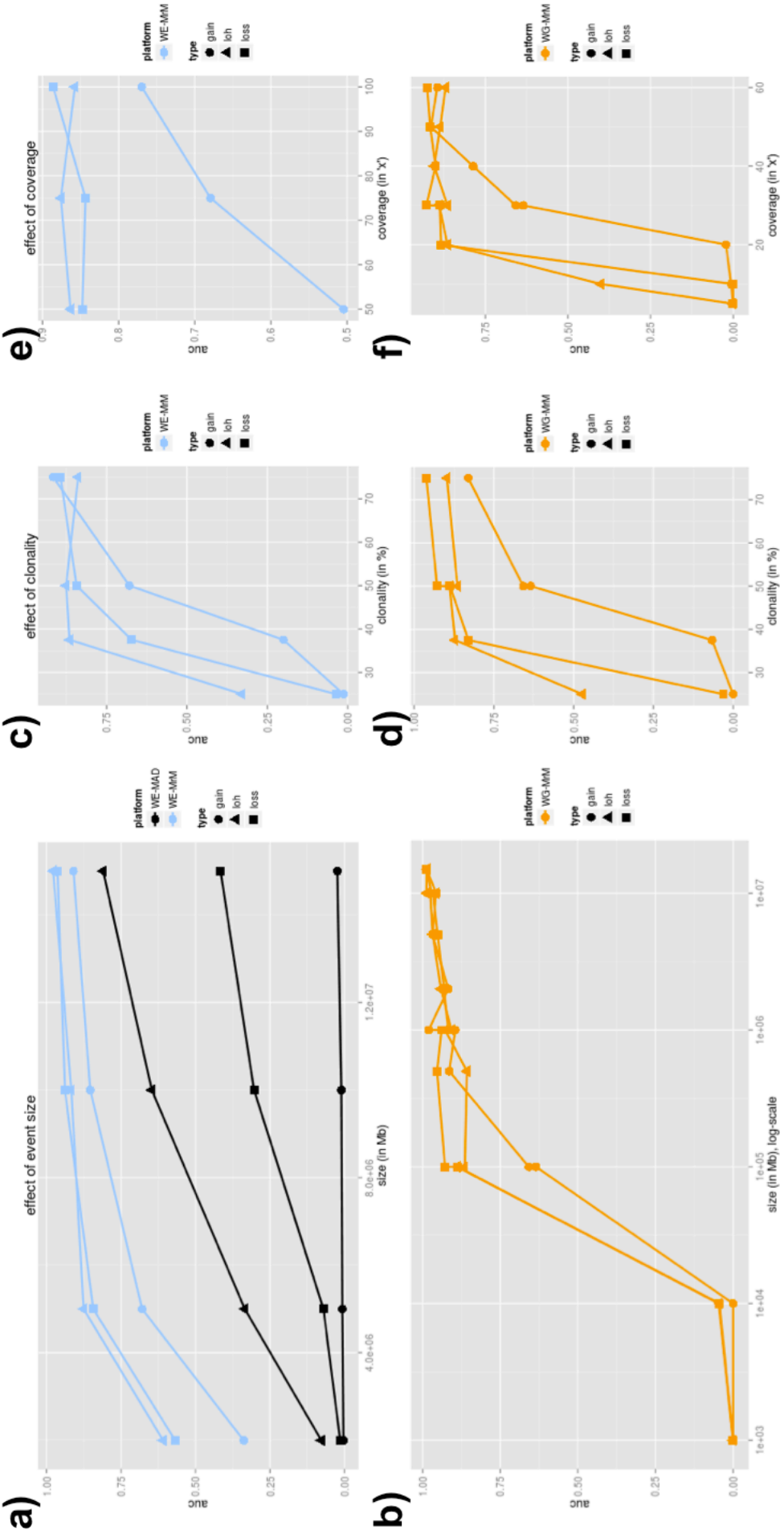
a)
effect of event size
auc
size (in Mb)
platform
WE-MAD
WE-MM
type
gain
loh
loss
b)
size (in Mb), log-scale
platform
WG-MM
c)
effect of clonality
clonality (in %)
d)
e)
effect of coverage
coverage (in 'x')
f)

Figure 4-9 Simulation performance summarised by AUC. I measured the average precision (area under the precision recall curve) for MrMosaic implemented on whole-genome (WG) simulations, and MrMosaic & MAD implemented on whole-exome (WE) simulations. The depth, size, and coverage measured for WG and WE simulations were selected to accentuate informative differences in performance. The first column of figures measures AUC across size. Simulated events of 50% clonality were studied for WG (a) and WE (b) simulations. Whereas for WE simulations, simulated exome depth was 75x depth, for WG simulations it was 30x depth. MrMosaic on whole-genome data (WG-MrM) outperformed MrMosaic on exome data (WE-MrM), which outperformed MAD on exome data (WE-MAD). The second column of figures measures AUC across clonality. Whereas for WE (c) simulations the simulated size and coverage was 5 Mb & 75x, for WG (d) simulations it was 100 kb & 30x. The third column of figures measures AUC across average coverage. Simulated events of 50% were studied for both WE (E) and WG (F) simulations. Whereas for WE simulations, simulated event size was 5 Mb, for WG simulations it was 100 kb.

Depth (in x)	Platform	Mean # Assayed Positions	Mean # Informative Positions	Median Distance between Informative Positions	Mean sampling variance
20	WG	7858070	2014409	1503	0.130282
30	WG	7866967	1949467	1554	0.129219
40	WG	7867003	1932357	1568	0.128347
50	WG	7867003	1924407	1574	0.128340
50	WE	163521	39382	59719	0.12264
75	WE	181053	43131	54581	0.12247
100	WE	191104	45233	52046	0.12213

Table 4-2 Number of assayed positions in WE and WG simulations. This table lists the mean number of assayed positions, the number of informative (heterozygous) sites, the average distance between informative sites and the mean sampling variance for each simulated coverage. Average distance between was calculated using sites on the p arm of chr1. All averages were calculated using 50 simulated samples per depth. There was a positive correlation between increasing depth and number of assayed sites, with a more pronounced effect in WE compared with WG. The interprobe distance is higher in the exome compared with the genome. This is due to having fewer sites and more variable distance between sites in WE compared with WG. The variance of the b allele frequency for heterozygous sites decreases with increasing sampling depth.

Detection performance in simulations increased between 25% and 75% *clonality* (Figure 4-9). The WE and WG clonality performance results were measured at 5 Mb and 100 kb sizes, respectively, as events at these sizes were most sensitive to changes in clonality. Previous studies of children with DD have reported a median mosaicism of approximately 40% clonality and at the event sizes studied detection performance is strong at this level of clonality. As clonality increases, the mosaicism is present in a greater proportion of cells, resulting in a greater signal to detect.

Simulation performance increases with respect to sequencing *coverage* (Figure 4-9). The WE and WG performance with respect to sequencing coverage were assessed for events of 50% clonality, using 5 Mb events for the WE simulations, and 100 kb events for the WG simulations. WE simulations demonstrated a marginal improvement of detection performance across a range of coverage from 50-100x, which was notable for mid-clonality gains (Figure 4-10).

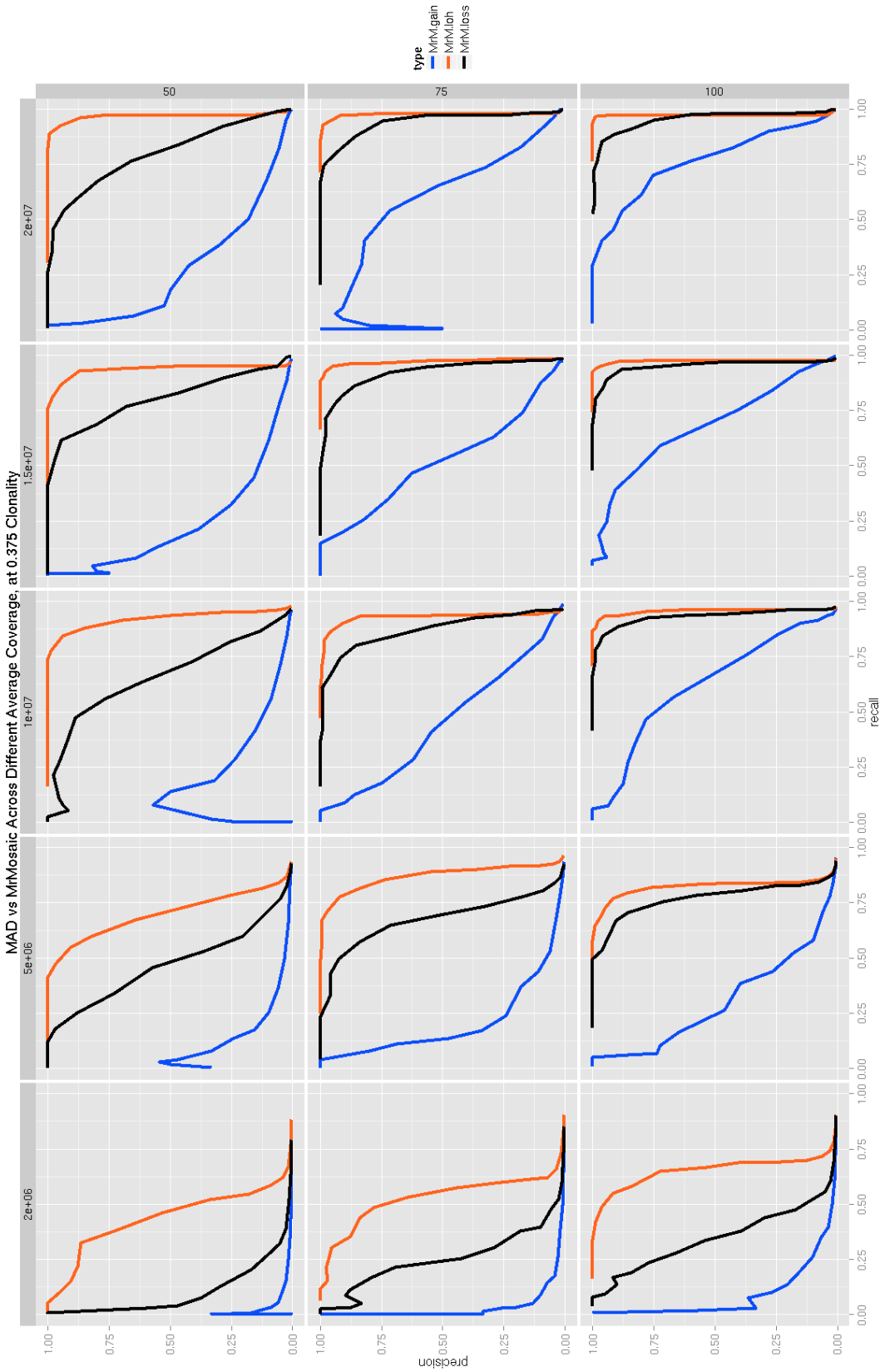

Figure 4-10 WE performance of MrMosaic across 50-100x. I generated simulated exomes of 50x, 75x, and 100x depths and measured MrMosaic detection performance across coverage. Detection

was measured at events of 50% clonality. Simulated event size and coverage (in 'x') are denoted in column and row headers, respectively. Increasing coverage is positively correlated with higher performance. This is likely due to a greater number of events passing minimum depth threshold (more signals) and a more precise estimate of non-reference discrepancy (better signal:noise ratio).

Previous work has suggested that 75x average coverage in WE data enables high resolution constitutive copy-number analysis[8] and these coverage simulations demonstrated that this exome coverage is also sufficient for the detection of mosaic structural abnormalities.

Increasing coverage has an effect on the number of assayed sites (number of signals) if some simulated sites fail to meet the minimum depth criterion, and has an effect on sampling variance ('noise') (see Figure 4-14 below). In WE data, both of these characteristics operate, whilst WG data have a much more even coverage distribution (it is not vulnerable to the enrichment biases of WE data) and increased simulation performance at higher coverage is likely primarily driven by decreased sampling noise.

In the WG results, AUC rose dramatically between 15x and 20x coverage for LOH and loss events and between 25x and 30x for gains. AUC was above about 0.9 for LOH and loss events at 30x depth, the standard sequencing depth generated by Illumina® X-Ten™ sequencing system. Nearly all structural mosaic events of 100 kb and 50% clonality were detected (Figure 4-11).

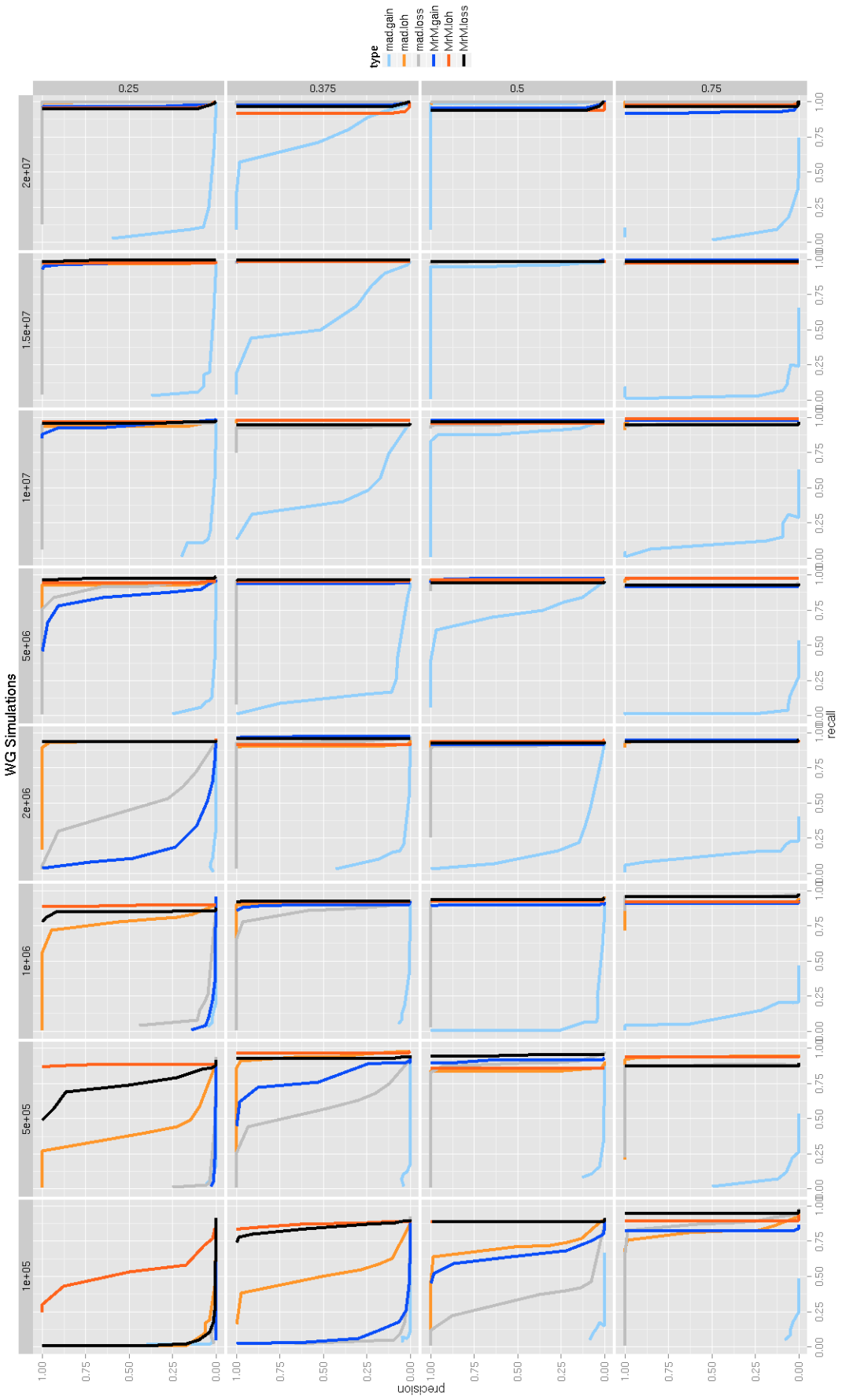
type
mad.gain
mad.loh
mad.loss
MrM.gain
MrM.loh
MrM.loss
0.25
0.375
0.5
0.75
WG Simulations
2e+07
1.5e+07
1e+07
5e+06
2e+06
1e+06
5e+05
1e+05
recall
precision
1.00
0.75
0.50
0.25
0.00

Figure 4-11 WG performance of MrMosaic and MAD. The performance of MAD and MrMosaic is compared at 30x WG average coverage for a range of sizes, clonalities, and for the three types of mosaic abnormalities simulations. The performance of MrMosaic detection is extremely high (high recall, high precision) at the same size ranges (2 Mb to 20 Mb) tested in exome simulations. In addition, detection performance is high at small-sized (100,000 bp) medium-clonality (0.5) events.

Average coverage of 20x was sufficient to detect nearly all 50% clonality deletion and LOH events at 100 kb. Detection performance of gains improved at 30x and 40x (Figure 4-12).

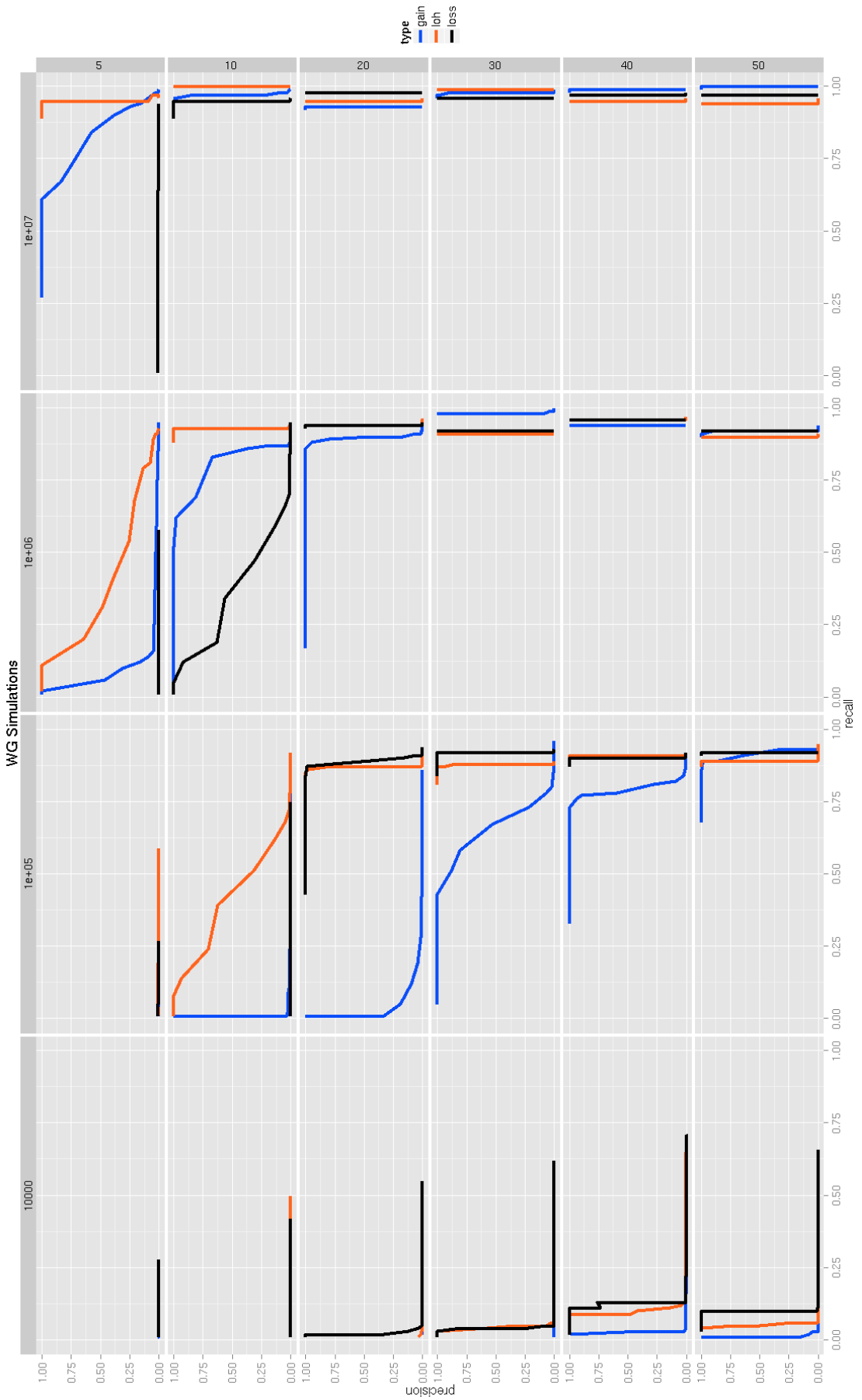
type
gain
loh
loss
5
10
20
30
40
50
1e+07
1e+06
1e+05
10000
WG Simulations
recall
precision
1.00
0.75
0.50
0.25
0.00

Figure 4-12 WG MrMosaic performance across 5-50x. I generated simulated genomes of 5x-50x depths and measured MrMosaic detection performance across coverage. Performance was measured of simulated events of 50% clonality. Simulated event size and coverage (in X) are denoted in column and row headers, respectively. Increasing coverage is positively correlated with higher performance. Events at 1Mb were detected easily at standard X-Ten coverage (30x) (http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html).

4.4.2 Detections in Exome Data

DNA for WES data were derived from saliva (66%) or blood sampling (34%), for 4,911 children with undiagnosed DDs. Analysis for structural mosaicism identified 11 mosaic abnormalities among 9 individuals, a frequency of 0.18%. The detections consisted of five losses (median size: 13 Mb, median clonality: 46%), four gains (median size: 25 Mb, median clonality: 55%), and two LOHs (median size: 50 Mb, median clonality: 26%) (Figure 4-13, Table 4-6 at end of chapter).

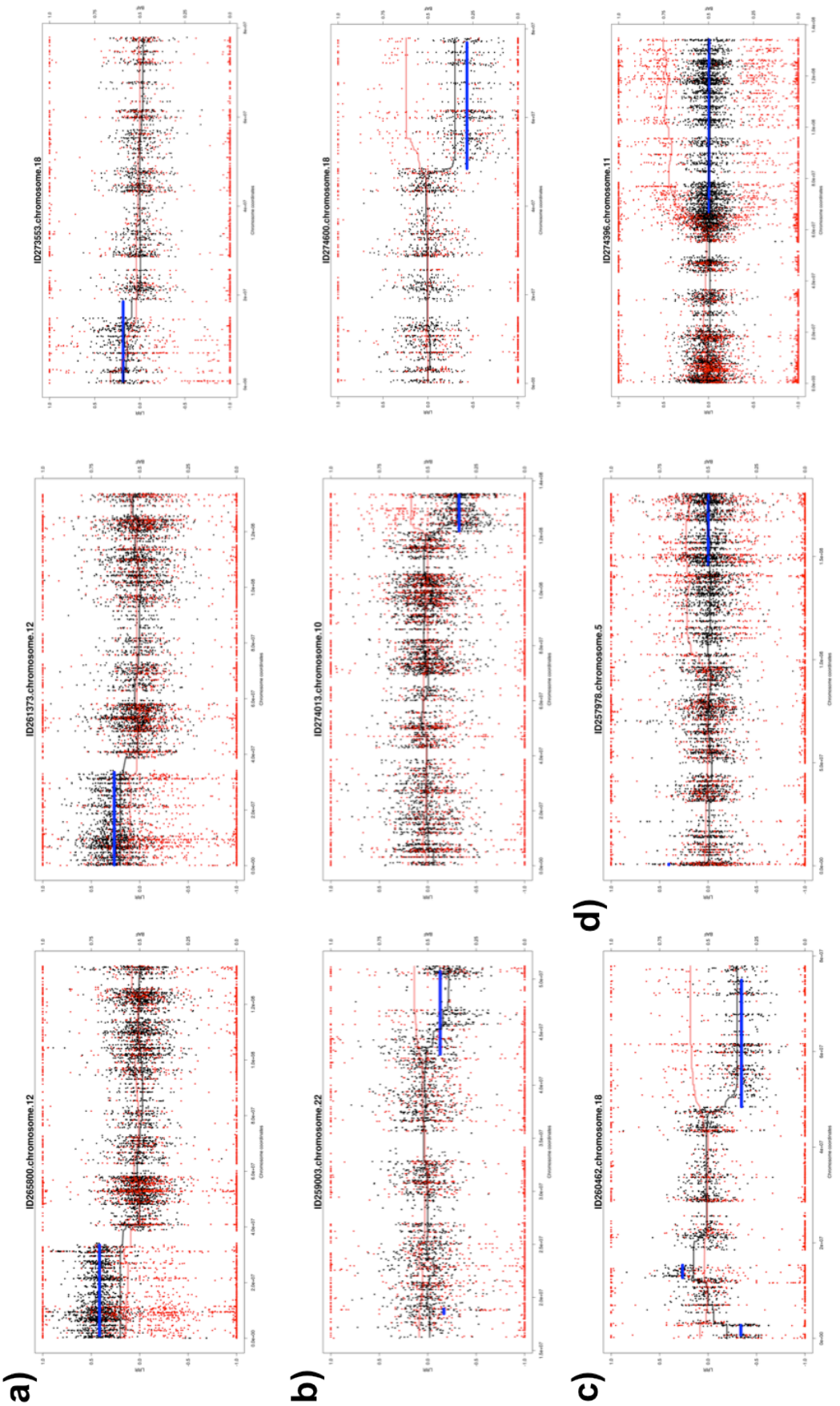
a)
b)
c)
d)
ID265800.chromosome.12
ID259003.chromosome.22
ID260462.chromosome.18
ID257978.chromosome.5
ID261373.chromosome.12
ID274013.chromosome.10
ID274600.chromosome.18
ID274396.chromosome.11
ID273553.chromosome.18

Figure 4-13 Structural mosaicism detected by MrMosaic from exome data in nine DDD samples, grouped into four categories. Black and red dots represent copy-number and allele fraction, respectively. C_{dev} and B_{dev} are plotted in black and red trend lines. The blue line represents statistically significant segmented detections passing a threshold. a) mosaic gains; b) mosaic losses; c) mixed copy-number; d) loss-of-heterozygosity events

In chapter 3, I presented analysis results for a subset (1,226 of 4,911) of these samples which had been analysed using SNP microarray[178] and among the samples in this subset, the SNP microarray approach had identified 10 events (in 8 samples), whilst exome analysis performed here yielded 8 events (in 6 samples). Of the two (missed) events not detected by exome but detected by SNP microarray, one of these events was a 4 Mb duplication below 25% clonality. The other missed event was an LOH event with low sequencing depth (33x, one of the lowest of our study, Figure 4-4). Low depth results in lower statistical significance of deviations in allelic proportion and copy number and higher sampling variance. Variance was much higher in WE samples with lower coverage (Figure 4-14).

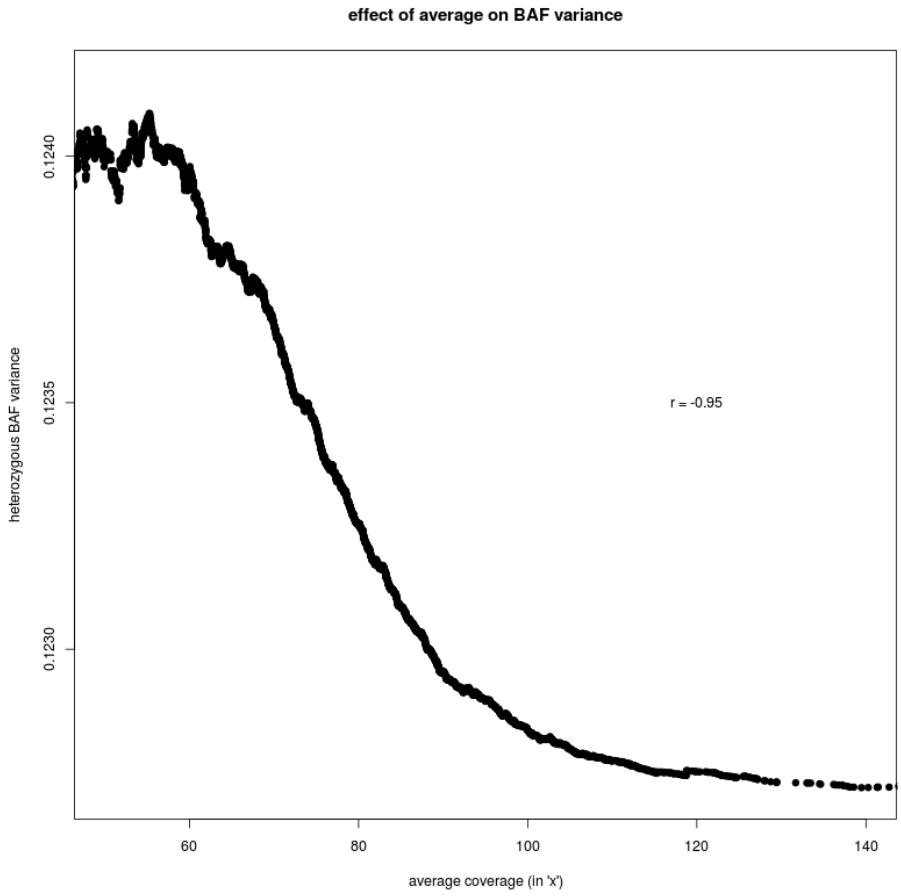

Figure 4-14 Observed BAF variance at heterozygous sites in WE data across samples with different sequencing depth.

Given the high clonality (about 75%) of this missed LOH event, it may have been detected using constitutive (genotype-based) UPD analysis (although, as paternal data were not available for this sample, it was not analysed by trio-based UPD[137] detection).

The frequency of mosaicism detected in this study, 0.18%, is lower and significantly different ($p < 10^{-4}$, binomial test) from the 0.59% estimate of structural mosaicism frequency calculated above (in §§§§§21Section 4.2). One likely explanation for the discrepancy in these frequencies is ascertainment bias, as 11 of the 36 events underlying the copy number frequency estimate were mosaic trisomies and children with trisomy are likely to have been diagnosed by clinical karyotype or microarray and not enrolled into the DDD study. Another component of this discrepancy may be due to decreased sensitivity, as mosaicism smaller than 2 Mb is challenging to detect by exome and 9 of the 36 events underlying the 0.59% frequency estimate were smaller than 2 Mb. The rate of mosaic events detected in the first 1,226 samples, 0.41%, is higher than the rate detected in the remaining 3,685 samples, 0.24%. This may suggest that the detection of mosaicism in real data is less sensitive than I estimated from simulations, or that clinical ascertainment has changed over the course of the project, which may be due in part to the increasing use of microarray over karyotyping by clinical centres in the last few years.

Validation data were generated using SNP microarrays for each of the 11 mosaic abnormalities assaying both blood and saliva derived DNA for individual. In these data I detected all abnormalities in at least one tissue (Table 4-6). Notably, six of the seven mosaic copy-number mutations detected by MrMosaic in exome data had been undetected by both clinical and high-resolution aCGH investigation of the same tissue, despite most events being at least 5 Mb in size and exhibiting 50% clonality (Table 4-3).

ID	tissue	chr	aCGH_appearance	clonality_by_SNP	detected_in_aCGH?
265800	Blood	12	no_deviation	absent	na
265800	Saliva	12	no_data	0.68	na
261373	Saliva	12	no_data	0.45	na
261373	Blood	12	no_deviation	absent	na
273553	Blood	18	no_deviation	absent	na
273553	Saliva	18	no_data	0.6	na
259003	Saliva	22	deviation_but_no_call	0.54	no
259003	Blood	22	deviation_but_no_call	0.34	no
274013	Blood	10	no_deviation	absent	na
274013	Saliva	10	no_data	0.44	na
274600	Saliva	18	no_data	0.49	na
274600	Blood	18	no_deviation	absent	na
260462	Saliva	18	deviation_no_call	0.5	all-three-missed
260462	Blood	18	no_deviation	absent	na
258956	Blood	3	failed_QC	absent	na
258956	Saliva	3	partially_detected	0.94	yes
261240	Blood	5	no_data	absent	na
261240	Saliva	5	partially_detected	0.39	partially_seen_escaped_review

Table 4-3 Validation results of all structural mosaic events in blood and saliva. Most mosaic copy number events escape detection by aCGH.

Examination of the raw aCGH data in one case (Figure 4-15) showed that only small fragments of one of the events were detected but these called segments were individually much smaller than the actual event and escaped review.

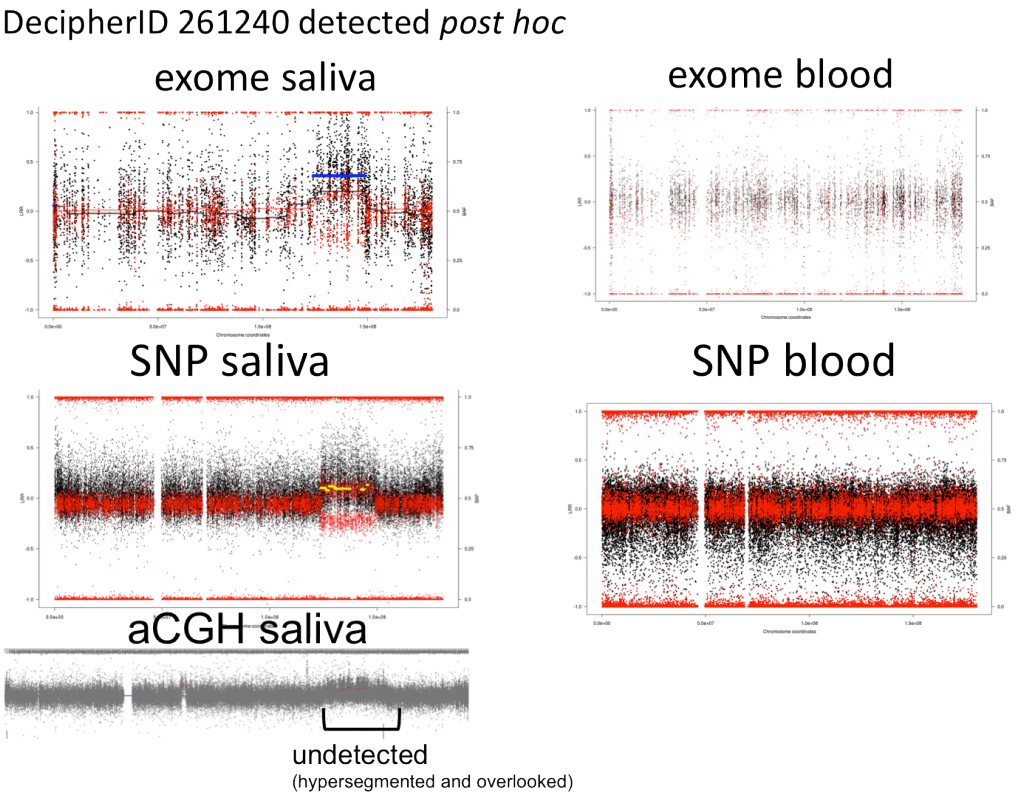

Figure 4-15 Detection of 261240 was *post hoc* in that originally, DNA from blood was analysed and no event was detected, although SNP microarray data data which had been previously analysed identified an abnormality in saliva, suggesting that either the event was missed by exome in blood, or that the mosaic event is not present in blood. I generated SNP microarray data for blood, which showed no evidence for the mosaic event in blood. And, I generated exome data from saliva, and MrMosaic detected the mosaic abnormality, with an Mscore of 12. Note that array CGH of saliva identified small segments of elevation but none was sufficiently large to pass size filtering.

Both of the mosaic events initially observed in blood-derived DNA were also observed in saliva, however, only one out of the eight events observed in saliva-derived DNA was also detected in blood (Table 4-6). There were 2 abnormalities detected from 1,036 blood samples and 9 detected from 3,260 saliva samples, a non-significant proportional difference ($p > 0.05$, Fisher's exact test). One of the mosaic events detected in both blood and saliva was an LOH-type event, remarkable for having a gradient of increasing clonality toward the telomere (Figure 4-16 and Figure 4-17).

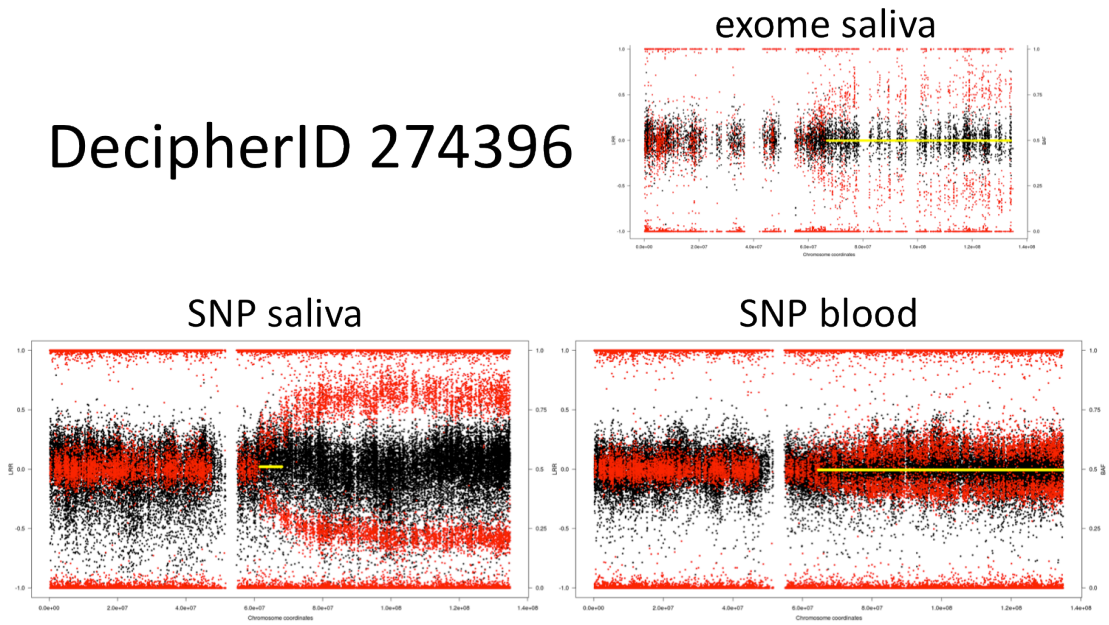

Figure 4-16 SNP Validation of 274396. A gradient of clonality present on chromosome 11, extending to the 3' end of the chromosome.

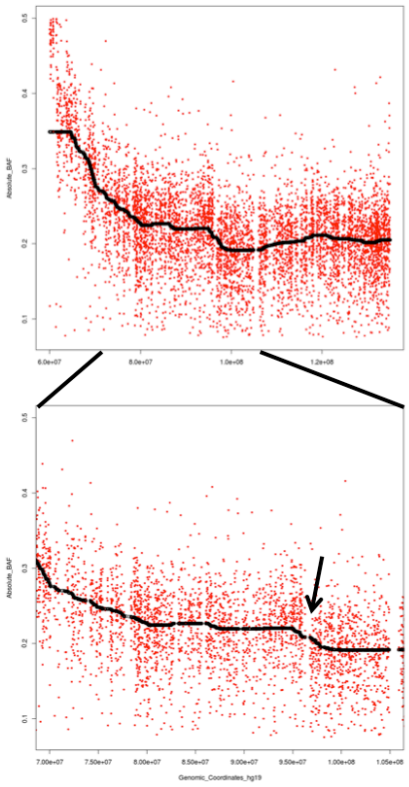

Figure 4-17 Investigating the mosaic reversion event. I examined SNP microarray data to help localise the cause of the suspected reversion. These plot displays heterozygous BAFs (BAFs above 0.5 are reflected below the 0.5 line) from SNP microarray data on the 3' end of chromosome 11, with a median trend line included. The bottom plot is a zoomed-in version of the top plot. Just 5' to the 100 Mb position there is a sudden increase in mosaic clonality (arrow), followed by a plateau of

clonality toward the 3' end. I investigated the rare (below 1%) variants present in the region from 90 Mb – 105 Mb.

This gradient of increasing clonality along the chromosome is compatible with incomplete LOH-mediated mosaic reversion. Reversion is the somatic recovery of a functional allele. The genotype data present here are consistent with distinct cell populations carrying partially overlapping independent LOH events (Figure 4-18), a mechanism reported elsewhere recently[232].

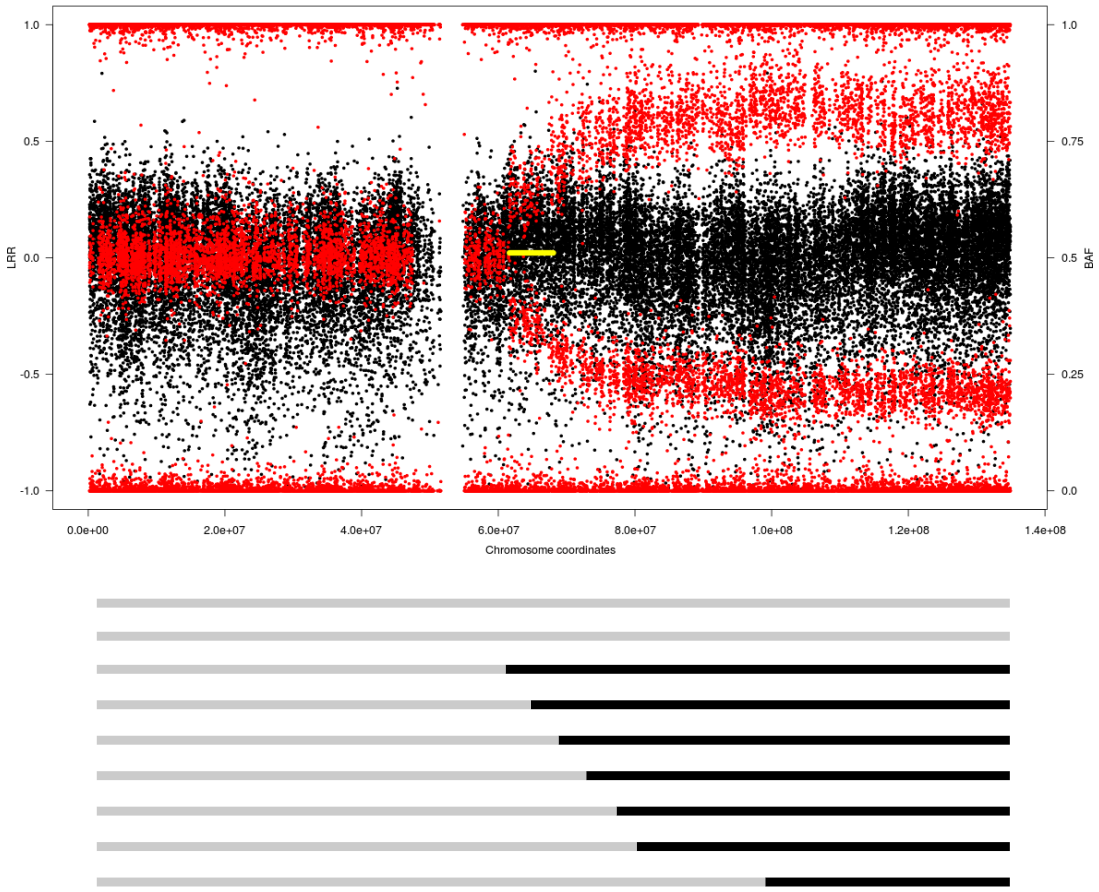

Figure 4-18 The revertant mosaic event detected in this study, and below, a schematic depicting the hypothesised mechanism, with black lines representing segments of LOH in independent revertant clones, while the gray represent wild-type. This reversion is 'incomplete' in the sense that, at least at the time of sampling, some clones still contain the wild-type allele.

I scrutinised the genomic interval in the most proximal (5') portion of this LOH segment (just distal to the arrow in Figure 4-17), suspected to contain a pathogenic allele and present the variants in the following table (Table 4-4).

chr	pos	ref	alt	af	gene	ddg2p?	consequence
11	92087959	G	A	0.005931	FAT3	no	missense_variant
11	93170909	T	TCC	none	CCDC67	no	3_prime_UTR_variant
11	94039561	G	A	0.008177	IZUMO1R	no	intron_variant
11	94564757	G	A	0.000276	AMOTL1	no	intron_variant
11	94696714	T	C	0.000366	CWC15	no	intron_variant
11	95569170	T	G	0.007078	CEP57	yes	intron_variant
11	100665791	C	T	0.000414	ARHGAP42_no	intron_variant	11

Table 4-4 Rare variants in the most proximal region of the smallest LOH region.

Nevertheless, despite generation and analysis of high-depth (~400x) WES data for this sample, and the identification of several strong candidate genes, including *CEP57* (the cause of mosaic aneuploidy syndrome[233]) in the reversion-localised region, no plausibly pathogenic *de novo* or rare (below 1% minor allele frequency) coding sequence variants were identified. Another possibility is that the suspected mutation responsible for driving the reversion may be absent from the exonic regions, i.e. is a regulatory mutation, or be a class of mutation not well detected in exome data. Deep sequencing of this entire genomic region may be warranted for further study.

4.4.3 Empirical evaluation of detection of mosaicism from WGS data

I selected one sample with three mosaic abnormalities detected on a single chromosome to demonstrate MrMosaic performance on whole-genome sequence data and to investigate the structure of the mosaic rearrangement. MrMosaic easily detected these multi-megabase mosaic events, found with very high Mscores of 36, 117, and 32. The presence of three mosaic events of similar clonality on the same chromosome is suggestive of a complex chromosomal rearrangement. I analysed the read -pair WGS data using Breakdancer[231], which identified read-pairs mapping across the centromere and evidence of a breakpoint spanning from the q-arm deletion to the centromere. Ring chromosomes are associated with bi-terminal deletions[234] and inverted duplications[235] and I suspected that the underlying abnormality in this child is a ring chromosome, although the cellular material required to generate the cytogenetic data to test this hypothesis was not available for study (Figure 4-19).

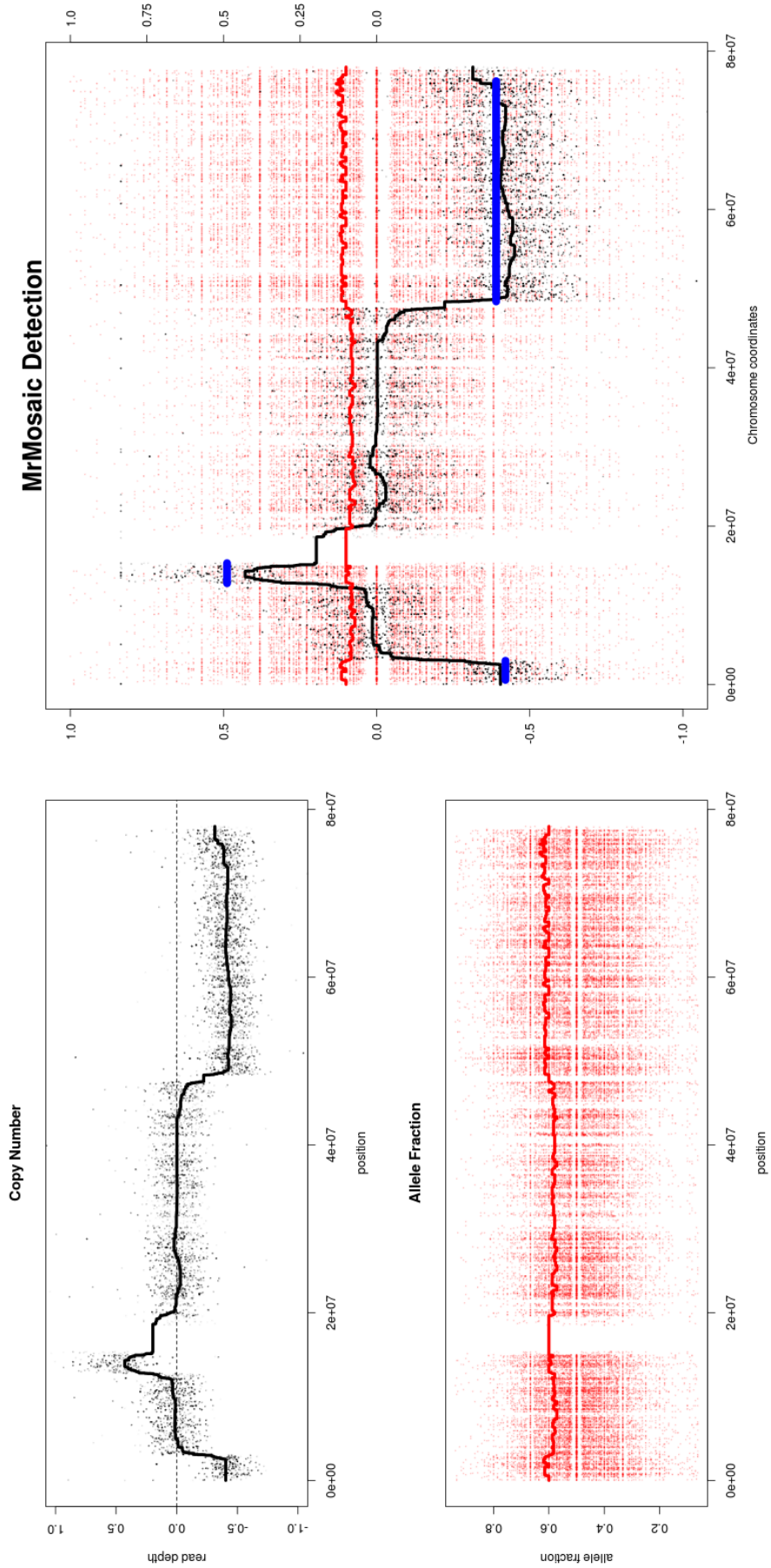
MrMosaic Detection
Chromosome coordinates
Copy Number
read depth
position
Allele Fraction
allele fraction
position

Figure 4-19 WGS analysis of Decipher 260462. Measurement of copy number (left, top) was generated using CNVnator[236], using bins of 10k reads and normalizing by GC content. The allele fraction plot (left, bottom) shows slight more variance in BAFs at the termini of the chromosomes. MrMosaic detection (Tgada of 20, minSegLen of 30) identified the three mosaic abnormalities (blue lines).

The BAF signal is 'noiser' here than in the exome analysis because measurement of BAF is sensitive to sampling variance, which is related to read coverage, and coverage is much lower in the WGS (25x) compared to the WES data (75x).

4.5 Clinical assessment

I investigated the clinical impact of the detected mosaic mutations to determine whether each was diagnostic, that is, providing the likely explanation of the child's phenotype (Table 4-7). In chapter 3, I presented the clinical evaluation of four (Decipher IDs: 261373, 259003, 260462, and 257978) of the nine mutations presented here and the clinicians and I assessed that in three of the four children the mutations were definitely pathogenic and considered diagnostic of the child's disease (three multi-megabase mosaic CNVs causing genomic disorders) whilst one child (257978) with a mosaic LOH mutation, had absence of neuronal migration, seizures, somnolence, scoliosis, but no loss of function variants or functional variants in known DD genes in the LOH region, and the mosaic LOH was considered of uncertain pathogenicity. I investigated the phenotypic profile of the remaining five patients and present the results from that analysis here; the clinicians and I assessed that the mosaic mutation is the likely explanation for disease in each of these children. I summarise the diagnostic results in the following table (Table 4-5) and discuss each patient in detail below.

mosaic structural variation from targeted and whole-genome sequencing

DecipherID	Diagnosis
265800	Pallister Killian syndrome
273553	18p mosaic tetrasomy
274013	distal 10q deletion syndrome
274600	Pitt Hopkins syndrome
274396	mosaic reversion of unknown de novo mutation

Table 4-5 Diagnoses resulting from mosaic abnormalities

Female patient 265800 had feeding problems, hypotonia, moderate developmental delay, severe speech delay, joint laxity, macroglossia, meningocele, delayed closure of the anterior fontanelle with short stature (2^{nd} centile). An array CGH was performed on blood lymphocytes but no copy number events were detected. Additionally, testing for mucopolysaccharidosis, *SMARCA2*, Fragile X, and FISH for 17p11.2 were negative. The exome analysis on saliva detected a gain of 12p. Mosaic tetrasomy 12p is the genetic basis of Pallister Killian syndrome[207], a well known cause of developmental delay. Simultaneous skin biopsy confirmed mosaicism for isochromosome 12p, considered definitely pathogenic. The child's clinical features are consistent with Pallister-Killian syndrome and the diagnosis was conferred to the family.

Male patient 273553 has moderate developmental delay, proportionate short stature, mild dysmorphism, significant behavior problems, undescended testes, strabismus, hypermetropia, joint laxity, indistinct speech, palatal insufficiency and communication difficulties. He had surgical correction of a patent ductus arteriosis. Multiple clinical array CGH investigations were performed on blood and all were negative. Exome analysis of saliva detected a mosaic abnormality of 18p, and the abnormality was validated using SNP analysis of saliva (clinical aCGH of the saliva is pending). The variant was considered definitely pathogenic. The gain in chromosome 18 appears to have two extra haplotypes, which may be consistent with a mosaic trisomy condition. Tetrasomy 18p is a recognized genomic disorder, responsible for causing a variety of clinical symptoms. The mosaic form, mosaic tetrasomy 18 presents with milder phenotypes[237]. In this case, the phenotypes present in the child were

considered likely to be due to this mosaic chromosomal abnormality and the diagnosis was conferred to the family.

Male patient 274013 required 35 days of neonatal medical intensive care for feeding difficulties. The child developed with severely restricted growth (1st centile, below -3.5 standard deviations of height, weight, and head circumference) and developmental delay, characterized by severe expressive language disorder and dyspraxia. The child had an abnormal facial shape, abnormal facial musculature, joint stiffness, brachydactyly, short stature, and was mildly dysmorphic. Testing was performed for acroosteolysis and was negative. Clinical array CGH performed in blood was negative. Exome analysis of saliva detected a 13 Mb mosaic deletion affecting the nearly all of 10q26 (10q26.12-10qter). Deletions of 10q26 are responsible for a variety of phenotypes, most commonly pre- and post-natal growth restriction, mental retardation, and abnormal facial facies (broad 'beak-like' nose)[238]. This mosaic abnormality was considered definitely pathongenic, diagnostic of the child's disease, and returned to the family.

Female patient 274600 had severe global developmental delay, with absent speech at 5 years of age, severe and progressive microcephaly (below -3.5 standard deviations), muscular hypotonia, hypotelorism, brachycephaly, narrow palate, apneas as a baby, abnormal extensor posturing, beaked nose, bow-shaped upper lip, broad terminal phalanges, and lack of intracranial myelination. Pitt Hopkins was suspected but clinical testing for mutations in the *TCF4* gene, the cause of Pitt Hopkins[204] were normal. Additionally, tests for mutations in *UBE3A*, and for abnormalities in 15q methylation were performed and were normal. Exome analysis of saliva detected a 28 Mb mosaic deletion in 18q, overlapping the *TCF4* gene, considered definitely pathogenic. The child's phenotypes are suggestive of Pitt Hopkins disorder and the diagnosis was conferred to the family.

Male patient 274396 had mild global developmental delay with severe growth restriction, including substantial microcephaly (below 7 standard deviations), restricted height (below -3.5 standard deviations) and restricted weight (below -5 standard deviations). The child had several abnormalities including progressive hypo- and hyper-pigmentation of the skin especially in the axilla, groin and neck. Skin wrinkling on dorsum of the hands, sparse & fine hair and a wide mouth were also noted. Dyskeratosis congenita was suspected, premature chromosome condensation testing was performed

and showed no abnormalities. This is the child discussed earlier with the suspected revertant mosaic mutation.

In summary, combining the results for the nine children with mosaic abnormalities, seven of nine mosaic events were considered definitely pathogenic on the basis of being multi-megabase CNVs that overlap known genomic-disorder regions. The reversion mosaic event was considered indicative of a likely pathogenic mutation as the presence of multiple overlapping mosaic clones suggests strong and on-going negative selection against a deleterious allele. One LOH event was of uncertain pathogenicity as no rare loss-of-function or functional variants were detected.

4.6 Discussion

Structural mosaic abnormalities are multi-megabase, post-zygotic mutations and are well recognised in developmental disorders[36,178]. This work introduces a novel method to detect these mutations from next generation sequencing data.

In an extensive simulation study, I observed adequate power to detect abnormalities in WES and WGS data across a large, clinically relevant range of size and clonality in different types of mosaic structural variation. I compared this method to the popular array-based mosaic detection method, MAD, and showed a substantial boost in performance, which derives primarily from the joint analysis of allelic proportion and copy-number deviations. Simulation results suggested that exome sequencing data can be used to identify many of the known clinical mosaic duplication syndromes involving chromosome-arm events, such as 12p and 18p mosaic tetrasomy as MrMosaic easily detected events of this size.

I hoped to use MrMosaic to uncover pathogenic structural mosaicism as an explanation for disease for children with undiagnosed DD. Applying this method to a set of 4,911 exomes from children with undiagnosed developmental disorders, I identified nine individuals with structural mosaicism and the majority of these mutations were considered pathogenic. In this WES-based analysis I recovered 8 of 10 abnormalities previously detected in a subset of 1,226 samples previously analysed with SNP genotyping chip data. One of the missed abnormalities was likely undetected because the exome data were of low depth, which increases the variance of measured B_{dev} and C_{dev}. Most of the detected mosaic copy number abnormalities had escaped detection by previous aCGH analysis. This demonstrates that detection of mosaic events requires assay of tissue containing the abnormality and tailored methods with sufficient sensitivity for mosaicism.

In one sample I observed a gradient of mosaicism, a phenomenon likely associated with mosaic reversion of a *de novo* mutation inducing genome instability. Analysis of the mosaic LOH region with high-depth exome data identified a strong candidate gene and investigation for the suspected *de novo* mutation is on-going. Whole genome sequencing data were generated for one individual with three mosaic abnormalities on the same chromosome. Analysis of these data recapitulated the mosaic events and analysis of read pair analysis identified a pericentromeric inversion and supported the hypothesis of an underlying complex chromosomal rearrangement, likely a ring chromosome.

Whole genome analysis had superior performance compared to exome analysis, which was likely due to a combination of advantages of whole-genome data, including higher density of assayed sites (by nearly 50 fold) and more consistent coverage across sites, compared to exome coverage, which is subject to exome bait hybridisation biases. Nevertheless, even detection from whole genome data is difficult at low depth. Compared to whole genome data, the exome data had higher average coverage (75x to 25x) for sites within targeted regions compared to the whole genome data and whilst simulation results showed increasing performance with higher depth sequence data, this effect was outweighed by the greater density of sites in whole genome data.

Although the general performance of the method is adequate in many clinically relevant cases, some classes of event proved more difficult to detect. For example, low clonality mosaic gains generate the smallest deviation in B_{dev} and C_{dev} compared to other types of events, explaining their comparatively poor detection sensitivity in simulations, and the failure to detect one mosaic duplication found using SNP data but not in exome data. More lenient detection thresholds may be preferred to increase detection sensitivity if clinical suspicion of mosaic duplication exists. Increasing the clonality of mosaicism by the biopsy of affected tissue, as is performed when pigmentary mosaicism provides evidence of underlying mosaicism, should also theoretically improve detection. Given the size and clonality of the two missed events and the simulation results from whole genome sequencing, both events would likely have been detected had they been analysed using higher depth exome sequencing or whole genome sequencing, which are likely to become more common in the future.

The majority of the mosaic events I observed were in saliva-derived DNA but not in blood-derived DNA. The samples with these abnormalities were recruited into our study because they remained undiagnosed after assessment by clinical laboratories of blood-derived DNA failed to detect the mosaic abnormalities detected in saliva. DNA derived from saliva has a mixed origin, mainly lymphocytes (derived from mesoderm) and epithelium (derived from epiderm)[216]; therefore the events detected in saliva, but not blood, are believed to reflect epithelial mosaicism. There are two possible explanations for the disparity in tissue distribution we observed: first, that the epithelium-derived mutational events occurred late, i.e. after the differentiation of lymphocytes and epithelial cells, or second, that these events occurred early, i.e. prior to

the split between lymphocytes and epithelial cells with subsequent removal from blood cell lineages by purifying selection. Several lines of evidence suggest the second explanation is more likely: 1) existing precedent, as the second phenomenon has been directly observed in Pallister-Killian syndrome, where the percentage of abnormal cells decreases with age in blood but not fibroblasts[239], and tissue-limited mosaicism has been observed in mosaic tetrasomies of chromosomes 5p, 8p, 9p and 18p[240]; 2) the clonality of events observed in both blood and saliva is not greater than the clonality of events in only saliva, which would be expected if events seen across tissue arose earlier in development; 3) both observed LOH events are shared between tissues but only 1 of 9 CNV events are shared between tissues, perhaps suggesting increased pathogenicity of CNV events compared to copy-neutral events, thus more likely to be negatively selected in blood. Given these considerations underlying the disparity in tissue-type, and the observation that the majority of observed abnormalities were detected in saliva but not blood, it is possible that, compared to the sampling of saliva, the sampling of blood could lead to a substantial loss of power, possibly less than 50% power, to detect pathogenic mosaic events, resulting in missed diagnoses.

Additional work is required to investigate for which developmental disorders tissue-limited mosaicism is common. Another intriguing question regarding tissue distribution is the relationship between clonality and pathogenicity. While mosaicism limited to a small number of cells is unlikely to cause developmental disorders, it is conceivable that low-level mosaicism present in a vulnerable tissue, such as white matter neurons, may have clinical consequences. More work is needed to address this question, including more extensive analysis of the tissue distribution of mosaicism, for example, by analysing diverse tissues sampled from all three germ layers, and assays with improved resolution, allowing single or oligo-cell sequencing. The availability of more sensitive detection methods will improve the detection of a larger fraction of events limited to a single tissue.

Next generation sequencing, in the form of exome and genome sequencing, can be harnessed to detect a wide range of mutations, including, as presented here, mosaic structural abnormalities. Given that sequencing costs continue to decline and the multifaceted detection capabilities of exome data, it may be that exome sequencing will supersede microarray technology as a first-line test for developmental disorders. Widespread incorporation of high-depth exome and whole-genome sequencing will

revolutionise our understanding of the extent of mosaicism in the body and better define the relationship of mosaicism and disease.

In the next chapter, I will review the main findings of this dissertation, discuss its limitations, suggest future improvements, and predict the relevance of UPD, mosaicism, and sequencing in the future of genomics.

Exome Detections									SNP Validation	
DecipherID	chr	type	start (GRCh37)	end (GRCh37)	bdev	l2r	tissue	clonality	clonality in saliva	clonality in blood
265800	12	gain	988894	33535510	0.201	0.140	saliva	1.34	0.68[@]	absent
261373	12	gain	283642	33535289	0.131	0.262	saliva	0.72	0.45[@]	absent
273553	18	gain	670541	18534702	0.186	0.185	saliva	1.18	0.6[@]	absent
259003	22	loss	42912136	50717129	0.131	-0.129	blood	0.42	0.54	0.34
274013	10	loss	121717932	134916366	0.159	-0.324	saliva	0.48	0.44	absent
274600	18	loss	48458662	76870586	0.190	-0.434	saliva	0.55	0.49	absent
260462	18	loss	662103	2740714	0.171	-0.339	saliva	0.51	0.46	absent
260462*	18	gain	12702610	15323214	0.118	0.263	saliva	0.41	0.5	absent
260462	18	loss	48466843	74962645	0.153	-0.3455	saliva	0.47	0.45	absent
257978	5	LOH	146077526	179731635	0.167	-0.0020	blood	0.33	0.24	0.26
274396	11	LOH	66834252	134126612	0.255	-0.0047	saliva	0.51	0.28	0.17

Table 4-6 Detections by exome and validation by SNP microarray

The 11 mosaic abnormalities detected in the 9 samples with exome data were validated using SNP microarray chips. All exome detections were validated in at least one tissue. In the majority of cases (8 of 11), the mutation was detected in only one of two assayed tissues, and in all such cases, the mutation was detected in saliva but not in blood.

Clonality was calculated from B_{dev} using Table 4-1 and ranged from 17% to 68%. This calculation is based on the assumption that the mosaic event is an alteration of a single allele. However, this calculated clonality is an overestimate for one of the events which was

found (by previous FISH analysis[178]) to be a mosaic tetrasomy, and two others were are suspected to also be rearrangements of multiple alleles (another gain of chromosome 12p and one gain of chromosome 18p, thought to reflect mosaic tetrasomy 18). [@]adjusted tetrasomy clonality. *located in peri-centromeric region and detected during *post hoc* analysis.

Decipher ID	Phenotypes
257978	Intellectual disability profound, Seizures, Somnolence, Thoracolumbar scoliosis, Gastroesophageal reflux, Abnormality of neuronal migration
259003	Generalized hypotonia, Global developmental delay
260462	Microcephaly, Muscular hypotonia, Short philtrum, Upslanted palpebral fissure
261373	Moderate global developmental delay
265800	Global developmental delay, Meningocele, Delayed closure of the anterior fontanelle, Macroglossia, Sparse scalp hair, Ligamentous laxity, Delayed speech and language development, Coarse facial features
273553	Global developmental delay, Joint laxity, Hypermetropia, Strabismus
274013	Severe expressive language delay, Global developmental delay, Abnormal facial shape, Brachydactyly syndrome, Thick hair, Coarse facial features, Abnormality of facial musculature, Joint stiffness
274396	Congenital hypothyroidism, Congenital microcephaly, Moderately short stature, Mild global developmental delay, Premature anterior fontanel closure, Fine hair, Sparse scalp hair, Long palpebral fissure, Wide mouth, Short broad hands, Excessive wrinkling of palmar skin, Excessive skin wrinkling on dorsum of hands and fingers, Strabismus, Generalized hypopigmentation of hair, Progressive hyperpigmentation, Mixed hypo- and hyperpigmentation of the skin, Axillary and groin hyperpigmentation and hypopigmentation
274600	Microcephaly, Progressive microcephaly, Severe global developmental delay, Abnormal posturing, Brachycephaly, Epicanthus, Muscular hypotonia, Narrow palate, Hypotelorism, Broad distal phalanx of finger

Table 4-7 Phenotypes listed in Decipher for children with identified structural mosaicism.

5 DISCUSSION

5.1 Summary of Findings

Understanding the genetic causes of DD is a priority of contemporary medical research. Modern rare disease studies rely heavily on exome sequencing, yet prior to the research described in this dissertation, software tools to detect uniparental disomy or structural mosaicism from sequencing data were lacking. This limitation led to the development of UP*D*io and MrMosaic, software tools which have extended the diagnostic reach of sequencing data and have been made freely available. Simulation studies have shown that these tools can detect the large-scale abnormalities identified by karyotyping or microarray in standard clinical testing. Implementation on nearly 5,000 children with undiagnosed diseases has shown that UPD and structural mosaicism are enriched in children with developmental disorders compared with healthy children. The estimated odds ratios compared to apparently healthy population controls suggested that most of the detected abnormalities are likely to be pathogenic. Assessment of the clinical impact of the detected events identified several disease-causing mechanisms, including UPD-associated imprinting and recessive diseases, and genomic disorders associated with large mosaic deletions and duplications. Some pathogenic mechanisms were unexpected and opened new research opportunities, such as UPD associated with triplication and mosaic reversion. The results of the analyses presented here have enabled genetic diagnoses for about 25 children, ending for them and their families, their quest for diagnosis.

5.2 Implications

The new methods described in this dissertation detected abnormalities and enabled diagnoses in approximately 1% and 0.5%, respectively, of the probands enrolled in DDD. The implication of this finding is that UPD and mosaic structural variation are small but important parts of genetic diagnosis in rare disease studies.

Heterodisomy is difficult to detect without genome-wide trio data and no large trio dataset had existed prior to the DDD study. Therefore, some of the outstanding questions in the field related to heterodisomy, such as the prevalence and diagnostic rate of heterodisomy in children with DD, could now be answered. For instance, of the 21 UPD events detected among 4,320 samples, 8 (38%) were entirely heterodisomic and likely to have escaped detection by non-trio-based screening. The implication of this finding is that trio-based methods increase UPD detection by about 50%. About half of the all-heterodisomy UPD chromosomes appear to be diagnostic, suggesting that trio-based analysis increases UPD diagnostic yield by 25%. The 0.49% UPD detection rate (21 of 4,320 samples) is, given assessment of both isodisomy and heterodisomy in this large trio study, and not withstanding the ascertainment bias of children selected for DDD recruitment, the best estimate of UPD frequency in children with DD to date.

Investigation of structural mosaicism identified a disparity in the tissue-distribution of mosaicism since in 8 of 11 cases, mosaicism was not observed in blood but was observed in saliva (likely from buccal epithelium). This tissue-difference may reflect greater negative selection against pathogenic mosaicism in lymphocytes, as suggested in Pallister-Killian syndrome[240]. An alternative possibility is differential rate of generation, but this is less likely, as studies of cadavers have shown that non-pathologic somatic CNVs are commonly found in many tissue types[241,242]. The tissue disparity observed in this study lends support for the assessment of saliva in disease studies, as, other factors equal, this tissue yielded greater numbers of mosaic diagnoses. There are several additional arguments supporting the collection of DNA from saliva rather than blood for high-throughput analysis, including that it is less invasive, less expensive[243], easier to store and ship[244], and genotyped equally well as blood[243]. Arguments against the use of saliva may include the absence of biomarkers present in blood that may also be of interest[245], lower DNA yield compared to blood[243], increased contamination of foreign (i.e. bacterial) DNA[246], or that higher rates of mosaicism in saliva may make it theoretically more challenging to assess genotype. However, for the purpose of high-throughput genetic analysis in studies of rare disease, DNA extraction

is the primary concern over biomarkers and mosaicism, and can increase diagnostic yield. Therefore, saliva sampling may become more popular for future research studies, and diagnostic testing.

An implication of high-throughput assays, such as WES and WGS, in connection with variant detection software, such as the algorithmic techniques developed in this work, is that the discovery of genomic variation has outpaced its interpretation. In the near term, the interpretation gap is likely to widen as WGS provides the resolution to detect smaller structural variants, whose significance will be unknown, and may add diagnostic anxiety[247]. This pressure highlights the importance of collaborative efforts, such as DECIPHER, and continued aggregation of genomic variation across centres to facilitate pathological assessment of structural mosaicism and UPD.

The most common trisomy in pregnancy is trisomy 16[248], and the most common UPD-generating mechanism is trisomy rescue[124]; but UPD 16 is observed less often than UPD of chromosomes 15, 11, 7, and 14 (descending order of observed frequency)[124]. Ascertainment bias almost certainly plays a role in this discrepancy, as these higher-frequency UPD chromosomes are involved in imprinting disorders, and are observed following scrutiny from characteristic phenotypes in children. While UPD 16 is controversially implicated in imprinting disorders, it is known that constitutive 16 trisomy is lethal, and that trisomy rescue is often incomplete, resulting in mosaic trisomy; perhaps lower levels of UPD16 reflects the fact that trisomy rescue is often incomplete and children with mosaic 16 rarely survive.

5.3 Limitations

5.3.1 Estimates of prevalence

Only about one third of the full DDD sample set was available for the work presented in this dissertation. Therefore, the assessment of UPD and mosaicism frequency is less precise than will be possible when the study is complete. Nevertheless, UPD frequency in the first-stage 1,000 trios was not significantly different from either the second-stage 3,000 trios or from estimates of UPD frequency in other DD studies; these pieces of evidence suggest limited benefit of acquiring additional samples for the purpose of improving the genome-wide estimate of UPD frequency in DD children. There was a non-significant lower frequency of mosaicism from nearly 4,000 additional children

beyond the first analysed 1,000 trios so it is conceivable that collecting greater number of samples will lower our frequency estimate of structural mosaicism. The trio set available in DDD enabled frequency estimates of heterodisomy, but the lack of trio data in the WTCCC dataset hindered heterodisomy frequency estimates in that dataset and relied on extrapolation from the identification of UPD with mixed heterodisomic and isodisomic regions.

The DDD population is not representative of all children with DD but reflects a pre-screened population as recruitment is generally only offered to children for which prior investigation of genetic abormalities failed to yield diagnostic abnormalities. Since many UPD and mosaic structural variants lead to phenotypically evident, syndromic manifestations, some children with such abnormalities and DD may be excluded from recruitment. Therefore, DDD likely has an ascertainment bias that lowers the estimate of UPD and mosaicism compared to the full population of children with DD. Children in DDD are unlikely to have large high-clonality mosaic events, unless perhaps, if such mosaicism is limited to tissue not analysed. Thus, it likely that the frequency estimates made in this work of UPD and structural mosaicism are underestimates compared to children in the general DD population.

DDD is primarily an exome-driven study. Exome read-coverage varies substantially across the genome by design, to maximize limited sequence resources for the genomic locations most likely to disrupt genes. However, whilst such exonic read-coverage enrichment is desired for identifying genic point mutations, it is not necessarily optimal for the detection of large-scale abnormalities. Abnormalities may be harder to detect in genes with widely spaced exons or genes with fewer exons, although, this limitation is mitigated by the target size of event detection (2 Mb and greater). Indeed, analysis for mosaicism of approximately one thousand samples by SNP and exome platforms showed that exome analysis missed two of ten events detected by the SNP platform. Thus, it is likely that exome-based calculation of frequency would produce a slight underestimate because of platform differences.

5.3.2 Algorithmic

Uniparental disomy describes two homologous alleles originating from the same parent and reflects an inheritance aberration. UP*D*io detects abnormal inheritance as an enrichment of uniparental trio genotype configurations on a single chromosome and data for proband and both parents are required to assess inheritance. There are two

failure modes that disrupt UPD detection: 1) missing genotypes and 2) missing parental samples.

Extending the method to account for the first failure mode is fairly straightforward. This approach could work by phasing parental haplotypes and then imputing the genotypes that have failed genotyping. On a practical level, this would likely make little difference for UPD detection because the genotyping error rate is low and UPD events are sufficiently large to be detected even in the context of missing genotypes.

However, for DDD probands now not analysed for UPD because full trio data are not available, the development of a proband single-parent UPD software tool should be possible. The approach might first phase the child's haplotypes and the known parent's haplotypes, and then determine which known parental haplotype the child has inherited. Based on the child's genotypes and the available haplotypes in the population, the other parent's haplotypes could be assessed. Each of the child's haplotypes should derive from a different parent and a discrepancy could reflect UPD or inheritance by descent, the latter distinguished by occurrence on multiple chromosomes.

MrMosaic uses a backbone of autosomal polymorphic di-allelic point mutations from which heterozygous sites are extracted for B_{dev} and C_{dev} calculations. There are three ways to improve the number of assayed sites: first, the number of assayed sites could be increased by adding to this backbone rare and private polymorphisms in each patient; second, the C_{dev} information from non-heterozygous (i.e. homozygous) sites can still be used in detecting deviation in copy number, even though the B_{dev} is not informative; third, gonosomal sites can be included.

MrMosaic has not been tested on the gonosomes but this extension should be possible. Mosaicism of chromosome X will detect the genetic aneuploidies associated with mosaic Klinefelter Syndrome and Turner Syndrome, diseases identified with high frequency in the Conlin *et al*[36] study. Implementing MrMosaic on gonosomes requires an ADM score generated on a sex-specific pool of samples. Mosaicism of the chromosome Y may be less useful, as the XYY karyotype in itself does not result in abnormal phenotypes[249], although mosaicism involving Y may signal other pathogenic events, such as complex aneuploidy involving multiple chromosomes, or chimerism.

Interpreting the output of MrMosaic is fairly labour-intensive because at the Mscore cut-off (8) chosen to be sensitive to mosaic events of 2 Mb and despite filtering based on event detection frequency and exclusion of peri-centromeric regions, approximately one putative detection is made per sample. In this large experiment presented of 4,911 probands, manual curation of 4,643 putative detections was undertaken, which required approximately 12 hours. The full data set will involve approximately three times the number of samples. The number of putative detections for review can be reduced by increasing the Mscore threshold, but is likely to lower the sensitivity of detecting smaller events.

5.3.3 Number of diagnoses

In about half of the cases for which a UPD or mosaic structural event was detected, a direct association between that event and the child's pathology could not be determined. UPD has a prevalence in the general population of about 1 in 3,500 and should therefore appear at least once among the nearly 5,000 studied children here in a benign form. However, given the enrichment of UPD and mosaicism in children with DD compared to generally healthy children, it is reasonable to suspect that the majority of the detected events are pathogenic, although diagnosis has only yet been possible for about half of those with detected abnormalities.

The diagnostic workup differs for UPD events compared with large mosaic abnormalities. For UPD events, the main pathological mechanisms are imprinting disorders, recessive diseases, and incomplete trisomy. The detection of UPD events on imprinting chromosomes in children with manifestations of known imprinting disorders provides definitive diagnosis. The majority of UPD events detected in this study did not lie on chromosomes vulnerable to imprinting, nor were they implicated in incomplete trisomy rescue. Instead, many resulted in regions of isodisomy, which can result in conversion to homozygosity of a deleterious allele inherited from a carrier parent. Assigning pathology to such homozygous variants is challenging and requires at least three broad categories of evidence: the variant causes disruption in the gene, pathology results when the gene is disrupted, and that this pathology matches the phenotypes in the child. This is fairly straightforward when the identified homozygous variant is predicted to be loss-of-function (such as a nonsense mutation), loss-of-function mutations in that gene have been closely associated in a specific disease, and the child's phenotypes match the manifestations of that disease. Knowledge gaps in gene function

and disease-gene associations hinder pathogenic analysis and require further investment in gene function.

The diagnostic workup for structural mosaicism is similar to the assessment of structural variation as a cause of genomic disorders and relies heavily on disease databases. Genetic diagnosis is fairly straightforward if the copy-number event in the child has been observed in other children who share the same phenotypes as the proband. Partially clouding diagnostic assignment in mosaic structural abnormalities is the effect of clonality on physiological disruption; this requires the assumption that an abnormality in mosaic state causes phenotypes similar in quality (but perhaps less severe) than the corresponding constitutive state. The assessment of mosaic UPD is slightly more complicated because incomplete aneuploidy often coexists with imprinting or recessive defects.

UPD and mosaicism are only detected in about 1% of children in the DDD study, and even after comprehensive assessment of constitutive copy-number analysis and other genetic abnormalities detected in the exome, genetic diagnosis still lacks for the majority (69%) of children in DDD. Improvements in understanding of gene function and variant ascertainment are essential and will hopefully lead to substantial reductions in the number of undiagnosed children.

5.4 Future work

Given the limitations above and the increasing trend for larger datasets, there are exciting opportunities for improved methods, which invariably will expand our understanding of DD.

Future trends may benefit from increasing integration of datasets and algorithms. With respect to integration of data, many of the analyses presented in this dissertation have made direct comparisons of the use, suitability, and performance of SNP vs. exome array. However, studies often use multiple platforms to assay genetic variation given unique advantages offered by each platform. In DDD, SNP, exome and aCGH data were generated for thousands of probands. Therefore, it is reasonable to consider the development of a tool that can integrate data gathered by multiple platforms. For example, mosaic analysis using SNP and exome platforms could increase the number of sites by including both common and rare variation, inside and outside of coding regions. Trio data facilitate the possibility of a haplotype-aware version of

MrMosaic, which is challenging given the sparse distribution of exome data, but should be possible for WGS analyses.

With respect to integration of algorithms, UP*D*io and MrMosaic were designed to detect constitutive UPD and structural mosaicism but it may be possible to integrate these two functions into one software tool as subroutines or "plug-ins" that function in a larger part of pipeline. Next-generation sequencing technology provides a substrate for simultaneously assaying a wealth of genomic variation, including structural variation, uniparental disomy, and mosaicism. In addition, there are likely statistical methods that can be learnt from transcriptomics, as this field must deconvolute signals of expression or transcript-assembly from heterogenous collections of tissue-types. Joint analysis of mosaicism and disruptions in expression could yield fascinating insight.

One of the limitations of MrMosaic is the number of putative detections that require manual review and future work could better automate the filtering strategy. A hurdle in such an approach is the lack of a strong positive-control training set, relative to the negative-control dataset. It may suffice to create the positive-control dataset using simulations, and then real mosaic events could be incorporated dynamically as they are discovered. Approximate Bayesian Computation is a Bayesian statistical technique that can be used in the absence of a known underlying likelihood model but when the sampling distributions of parameters are available; this approach may be useful for this automated filtering application as simulation analyses can generate the sampling distributions needed for multiple parameters (number of probes, strength of signal, event frequency, distance to centromere) underlying putative detections.

Regions of heterodisomy on non-imprinted chromosomes without evidence of mosaic aneuploidy are not predicted to be damaging. Despite this, eight examples of such heterodisomic chromosomes were found in this dataset. This invites speculation that many of these heterodisomic events may be pathogenic, perhaps by mechanisms already known, such as hidden trisomy-rescue, or by entirely new mechanisms. Maybe UPD is incompletely penetrant for some chromosomes, or results in highly variable phenotypes, as suspected for chromosome 16. Experiments that investigate the effect of heterodisomy on expression may yield interesting insights.

Decreasing sequencing costs have enabled acceleration in DNA sequencing data availability. Whilst whole-genome sequence data is still expensive to generate and were not available for analysis, such data are likely to be available in future studies of children with DD. Such data will enable unprecedented discovery of smaller mosaicism.

The somatic point mutation rate is approximately 0.3e-9 per site per cell division[250]; therefore mosaicism arises *de novo* with nearly every cell division. Despite this ubiquity, mosaicism is elusive, only detected when present in at least approximately 3,000 cells (based on: standard microarray input requirements require 200 ng (about 30,000 'genomes-worth' of DNA assuming 6 pg per cell) and mosaicism minimal detection threshold is 10% clonality). Future work will benefit from the use of single-cell sequencing or high-depth sequencing to detect mosaicism of lower levels of clonality tissue-specific mosaicism. Intuition suggests that mosaic abnormalities may often result in an intermediate phenotype (i.e. are less severe) than constitutive abnormalities and that mosaic events with greater tissue involvement are more pathogenic. These assumptions are difficult to assess empirically because tissue-sampling resolution is poor, often limited to blood or saliva. Study of mosaic trisomy 21 has found that mosaicism was more frequent in epithelial-derived tissue compared to lymphocytes and that phenotypic severity is linked to mosaic clonality in a tissue-specific manner[251]. These findings highlight the importance of developing a greater understanding of the distribution of mosaicism for diagnostics (identifying the mutation) and prognostics (interpreting its severity and outcome).

Analysis of one structural mosaic abnormality predicted that the most likely generative mechanism was LOH-mediated mosaic reversion, a mechanism previously reported[252]. Recently, chromothripsis has been implicated as an additional reversion mechanism[253] and it is reasonable to hypothesise that additional reversion mechanisms may be uncovered. It is speculative but interesting to consider that reversion may be fairly common; the disconnect between the theoretically-predicted commonality of mosaicism and the poor ascertainment of such events lends credence to this possibility. Several questions for reversion remain for future study: How common is reversion? Are most reversion events triggered by genomic instability? Are reversion events 'in response' to an underlying physiological disruption or an indication that stochastic genomic instability is commonplace? Do other reversion mechanisms, such as single codon deletions, exist? Do reversion clones have a common ancestor? Is the age-related dissipation of epidermal neoplasms (skin moles) immunologic or genetic (reversion)? Nature uses LOH and chromothripsis as reversion mechanisms; can man harness these mutational events therapeutically?

5.5 And then...

Forecasting the future of genomics is a useful exercise for planning but can be challenging. James Crow stated about prediction, "for the near future, I can follow the principle...that tomorrow's weather is best predicted by today's...for a somewhat longer future we can extend current trends. But for the long-term future, we can only guess"[254].

5.5.1 Achieving a higher fidelity genome

There is tremendous societal investment in genomics with an estimated 796 billion US dollars investigated in genomics between 1988 and 2010[255]. Such investment has empowered technological innovation, leading to a 100-fold decrease in sequencing costs within the period between 1991 and 2001[256], and an accelerated 1000-fold decrease between 2008 and 2014[257]. Yet, the cost of sequencing a human genome by WGS today is still expensive, more than $1,000[257], which also does not account for ancillary costs, such as data storage and interpretation[258]. Illumina® "has essentially monopolized the high-throughput sequencing market"[259], controlling 75% of the general genomics market share and 90% of high-throughput sequencing. It is reasonable to predict that continuing investment in genomics will spur industry competition, which will continue to drive down sequencing costs. Additional sequencing methods, such as those that measure changes in electrical current[260] or pH[261] avoid the overhead of optics, are extremely fast, and seem likely to rise in popularity. Inevitably, sequencing costs and technological advances will produce a portable, inexpensive, fast, high-fidelity whole-genome & whole-epigenome sequencing tool, perhaps within 15 years.

The technical implications of this new sequencing era will be profound: 1) long read-length sequencing will enable *de novo* assembly as the primary form of genome reconstitution; 2) reduction of mapping artefacts and sequencing errors will identify genomic variation with greater confidence and will reduce the computational complexity of assembly; 3) high-confidence genotyping will lead to more efficient storage[262], as less intermediate data need to be stored; improved knowledge of population haplotypes will enable an even more compressed haplotype-reference version of storage; re-sequencing a sample will be sufficiently inexpensive if long-term storage is not possible.

5.5.2 Having achieved a higher fidelity genome

The development of third generation (long-read single-molecule) sequencing[56] will especially have important consequences on the assessment of structural variation. Long

read-lengths will greatly facilitate the detection of structural variation via *de novo* reconstruction of the genome[263]. The resulting genome-wide frequency-map of structural variation will provide an empirical catalogue of all haploinsufficient genes and greatly reduce the number of CNVs of unknown clinical significance. More broadly, as sequencing becomes routine, catalogues of all forms of genomic variation will begin to saturate with all possible combinations of non-lethal mutations; this will identify which gene knock-outs are tolerated[142] and improved allele frequency data will facilitate interpretation of mutations in children with DD.

In contrast to constitutive structural variation, the detection of *mosaic* structural variation may prove challenging for some time to come because of sampling difficulties. The detection of mosaicism requires increasing read- and tissue- sampling, but low error rates may reduce the impetus to sequence the genome to high-depth, and accessing multiple tissue types is invasive and therefore not likely to become commonplace. High-depth sequencing is likely to be a continued priority of the cancer genetics community and may yield important insights of distribution of mosaicism throughout the body. Perhaps, sequencing can one day be performed non-invasively, as seen with *in vivo* magnetic resonance spectroscopy[264] for metabolomics, which would profoundly improve the ease of tissue sampling.

Large collections of WGS data are likely to come from healthcare settings, and eventually from domestic and municipal sources. In the Cold Spring Harbor Laboratory Biology of Genomes conference in 2013, Dr. Mike Snyder presented research (a lecture entitled "Integrative personal omics profiling for monitoring healthy and disease states") demonstrating that the distribution of his microflora fluctuated in a consistent and characteristic pattern each time he had 'a cold'. Toilet sensors, in the form of 'smart plumbing', may provide a method to detect early infections (microbiome sequencing) and cancer (detection of new mutations previously characterised as cancer driver mutations). Analysis of sewage microbiota can demonstrate the viruses circulating in the community and inform on community diet[265] (some viruses are endemic to certain types of plants only, for example). Analogous to telemetry used in the clinical setting to identify arrthymias remotely, it may be in the public interest to screen municipal sewers to identify epidemics, for example.

The majority of detected genetic variation today has unknown biological significance. Yet, complex disease studies operate with the assumption that a great

number of variants exhibit low-level effects on phenotype. Higher resolution phenotyping is needed to better understand low-effect variants with better granularity. Currently phenotyping is largely restricted to external traits and standardised human terms[266] but phenotyping is likely to become increasingly molecular, quantitative, and comprehensive ('phenomics'). Computational interpretation of facial dysmorphology is beginning to overtake human performance[267] and the integrated analysis of deep phenotyping data, such as transcriptomics and metabolomics, is likely to exacerbate this gap. The detection of UPD events may one day more appropriately be detected directly, using disruptions in epigenetics and alterations in expression, than indirectly by genotype. It also may be the case that detection of altered transcription or metabolic products will trigger the investigation of low-clonality mosaicism in children with DD.

Further ahead, widespread use of genomics and phenomics perhaps may mean that computational representation of each person's genome and phenome is recorded. Family studies could be performed quickly, entirely using stored data. Social media may allow contact with others who are most genetically similar (yielding interesting implications in genealogy, such as tracing ancestry or finding relatives), or metabolically similar, perhaps finding those who share similar disease states.

5.5.3 Challenges further ahead

Despite the battle cry of exuberant contemporary research papers[268], *determining* the genetic cause of Mendelian disease is not the same as *solving* Mendelian disease. Recent advances have treated some metabolic deficiencies using enzyme replacement and gene therapy[269], and others suggest that reversion of phenotype in children with Rett syndrome and Down syndrome may indeed be possible[270,271]. Nevertheless, a cure for the vast majority of DD has not been found.

Some treatments for DD may require intervention during early embryonic life. Non-invasive prenatal testing (NIPT) is now widely used in the United States, with 90% of pre-natal genetic counsellors having integrated NIPT into their clinical practice[272]. Currently NIPT is limited to detection of foetal aneuploidy and large structural variation but advances in genomics will inevitably lead to the incorporation of whole-genome sequencing in NIPT and the detection of pathogenic variation.

Many of the challenges in medical genetics ahead will be ethical. Intervention on human embryos has already generated substantial ethical debate, with respect to selective abortion[273,274], the right to access a child's genome[275], and whether gene

editing of human embryos[276] should be allowed[277]. It seems inevitable that genomic editing will be eventually welcomed, even by pro-life activists, as a method to cure a child's disease, in a way that preserves the child's life. The privacy implications of databasing and reporting of personal genomics are certain to become contentious but likely to become adopted given the potential impact on medicine and health.

Challenging questions ahead relate to analysis, thorough space and time, of transient and tissue-dynamic components of genomic activity, such as transcriptomics, metabolomics, and 3-dimensional chromatin architecture. The new concept that the 3D layout of the genome is informative[278] is exciting and throws dirt over the grave to the concept that non-exonic genomic regions are 'junk'[279] (although I sympathise with the somewhat unpopular view that much of the genome probably has little biological function[280], despite the widely publicised claim to the contrary[281]). Notwithstanding technical limitations to High-C technology[282], the field now appreciates that intergenic regions hold regulatory value[283] and the way chromatin is spaced is important[284]. It should be possible to quantify how important each DNA base is in terms of the spacing and positioning of regulatory elements beside their targets, a 'white-space' metric of the genome. For aneuploidy, in addition to disruption of gene dose, what proportion of pathogenesis is contributed by the disruption of long-range interactions and regulatory spacing?

DNA, like the heavens, once had complexity seemingly beyond reach. A breakthrough in cosmology research, the construction of a three-dimensional map of our local galactic neighbourhood, has just been completed[285]. Efforts to create a 3D map of the genome may benefit from a cross-disciplinary collaboration involving the mapping techniques of astronomers, the expertise of physicists in electrostatic interactions, and the biological experience held by genomicists. Eventually such maps of our genome will be available and if fortune grants me the opportunity, I would be eager to explore them.

6 REFERENCES

1. Firth, H.V. & Wright, C.F. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* **53**, 702-3 (2011).

2. Markiewicz, K. & Pachalska, M. Diagnosis of severe developmental disorders in children under three years of age. *Med Sci Monit* **13**, CR89-99 (2007).

3. Wright, C.F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* (2014).

4. Rehm, H.L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* **15**, 733-47 (2013).

5. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880-7 (2014).

6. DDD_Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).

7. van Karnebeek, C.D. *et al.* Etiology of mental retardation in children referred to a tertiary care center: a prospective study. *Am J Ment Retard* **110**, 253-67 (2005).

8. study, D.D.D. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* (2014).

9. Boycott, K.M., Vanstone, M.R., Bulman, D.E. & MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14**, 681-91 (2013).

10. Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat* **32**, 564-7 (2011).

11. Driscoll, D.A. & Gross, S.J. Screening for fetal aneuploidy and neural tube defects. *Genet Med* **11**, 818-21 (2009).

12. Vorstman, J.A. & Ophoff, R.A. Genetic causes of developmental disorders. *Curr Opin Neurol* **26**, 128-36 (2013).

13. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).

14. Liehr, T. Uniparental disomy - clinical consequences due to imprinting and activation of recessive genes. *Mol Cytogenet* **7**, I21 (2014).

15. Freeman, J.L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61 (2006).

16. Poduri, A., Evrony, G.D., Cai, X. & Walsh, C.A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758 (2013).

17. von Winiwarter, H. Etudes sur la spermatogenese humaine. *Arch. de Biol.* **XXVII**, p.91 (1912).

18. Barr, M.L. & Bertram, E.G. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* **163**, 676 (1949).

19. Hsu, T.C. Mammalian chromosomes in vitro. IV. Some human neoplasms. *J Natl Cancer Inst* **14**, 905-33 (1954).

20. Eagle, H. Nutrition needs of mammalian cells in tissue culture. *Science* **122**, 501-14 (1955).

21. Levan, A. Chromosome studies on some human tumors and tissues of normal origin, grown in vivo and in vitro at the Sloan-Kettering Institute. *Cancer* **9**, 648-63 (1956).

22. Giemsa, G. Eine Vereinfachung und Vervollkommnungmeiner Methylenblau Eosin Farbemethode zur Erzielung der Romanowsky Nochtschen Chromatinfarbung. *Centralblatt fur Bakteriologie* **32**, 307-313 (1904).

23. Seabright, M. A rapid banding technique for human chromosomes. *Lancet* **2**, 971-2 (1971).

24. Yunis, J.J., Sawyer, J.R. & Ball, D.W. Characterization of banding patterns of metaphase-prophase G-banded chromosomes and their use in gene mapping. *Cytogenet Cell Genet* **22**, 679-83 (1978).

25. Mattei, M.G., Mattei, J.F., Ayme, S. & Giraud, F. X-autosome translocations: cytogenetic characteristics and their consequences. *Hum Genet* **61**, 295-309 (1982).

26. Hook, E.B. Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use. *Am J Hum Genet* **29**, 94-7 (1977).

27. Mayall, B.H. *et al.* The DNA-based human karyotype. *Cytometry* **5**, 376-85 (1984).

28. Kwok, P.Y. & Gu, Z. Single nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today* **5**, 538-43 (1999).

29. Pardue, M.L. & Gall, J.G. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A* **64**, 600-4 (1969).

30. Langer, P.R., Waldrop, A.A. & Ward, D.C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc Natl Acad Sci U S A* **78**, 6633-7 (1981).

31. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21 (1992).

32. Oostlander, A.E., Meijer, G.A. & Ylstra, B. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet* **66**, 488-95 (2004).

33. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-82 (1998).

34. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**, 4181-93 (2009).

35. Bruno, D.L. *et al.* Pathogenic aberrations revealed exclusively by single nucleotide polymorphism (SNP) genotyping data in 5000 samples tested by molecular karyotyping. *J Med Genet* **48**, 831-9 (2011).

36. Conlin, L.K. *et al.* Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum Mol Genet* **19**, 1263-75 (2010).

37. Wiszniewska, J. *et al.* Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing. *Eur J Hum Genet* **22**, 79-87 (2014).

38. Peiffer, D.A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16**, 1136-48 (2006).

39. Steemers, F.J. *et al.* Whole-genome genotyping with the single-base extension assay. *Nat Methods* **3**, 31-3 (2006).

40. Hsu, L. *et al.* Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211-26 (2005).

41. Huang, T., Wu, B., Lizardi, P. & Zhao, H. Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21**, 3811-7 (2005).

42. Pique-Regi, R. *et al.* Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309-18 (2008).

43. Broet, P. & Richardson, S. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911-8 (2006).

44. Marioni, J.C., Thorne, N.P. & Tavare, S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**, 1144-6 (2006).

45. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).

46. Hoque, M.O., Lee, C.C., Cairns, P., Schoenberg, M. & Sidransky, D. Genome-wide genetic characterization of bladder cancer: a comparison of high-density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis. *Cancer Res* **63**, 2216-22 (2003).

47. Altug-Teber, O. *et al.* A rapid microarray based whole genome analysis for detection of uniparental disomy. *Hum Mutat* **26**, 153-9 (2005).

48. Illumina. http://www.illumina.com/Documents/products/appnotes/appnote_cytogenetics.pdf. (2010).

49. Gonzalez, J.R. *et al.* A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics* **12**, 166 (2011).

50. Jacobs, K.B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* **44**, 651-8 (2012).

51. Baugher, J.D., Baugher, B.D., Shirley, M.D. & Pevsner, J. Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. *BMC Genomics* **14**, 367 (2013).

52. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).

53. Prober, J.M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336-41 (1987).

54. Karger, B.L. & Guttman, A. DNA sequencing by CE. *Electrophoresis* **30 Suppl 1**, S196-202 (2009).

55. Mardis, E.R. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)* **6**, 287-303 (2013).

56. Schadt, E.E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum Mol Genet* **19**, R227-40 (2010).

57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).

59. Phillips, K.A., Pletcher, M.J. & Ladabaum, U. Is the "$1000 Genome" really $1000? Understanding the full benefits and costs of genomic sequencing. *Technol Health Care* (2015).

60. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).

61. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502-11 (2013).

62. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91**, 597-607 (2012).

63. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).

64. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103 (2009).

65. Gorringe, K.L. & Campbell, I.G. High-resolution copy number arrays in cancer and the problem of normal genome copy number variation. *Genes Chromosomes Cancer* **47**, 933-8 (2008).

66. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-92 (2009).

67. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).

68. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).

69. Jiang, Y., Wang, Y. & Brudno, M. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**, 2576-83 (2012).

70. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-71 (2009).

71. Price, T.S. *et al.* SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **33**, 3455-64 (2005).

72. Ashoor, G., Poon, L., Syngelaki, A., Mosimann, B. & Nicolaides, K.H. Fetal fraction in maternal plasma cell-free DNA at 11-13 weeks' gestation: effect of maternal and fetal factors. *Fetal Diagn Ther* **31**, 237-43 (2012).

73. Chiu, R.W. *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105**, 20458-63 (2008).

74. Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L. & Quake, S.R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* **105**, 16266-71 (2008).

75. Rampasek, L., Arbabi, A. & Brudno, M. Probabilistic method for detecting copy number variation in a fetal genome using maternal plasma sequencing. *Bioinformatics* **30**, i212-8 (2014).

76. Jacobs, P.A. & Strong, J.A. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* **183**, 302-3 (1959).

77. Ford, C.E., Jones, K.W., Polani, P.E., De Almeida, J.C. & Briggs, J.H. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* **1**, 711-3 (1959).

78. Lejeune, J., Gautier, M. & Turpin, R. [Study of somatic chromosomes from 9 mongoloid children]. *C R Hebd Seances Acad Sci* **248**, 1721-2 (1959).

79. Patau, K., Smith, D.W., Therman, E., Inhorn, S.L. & Wagner, H.P. Multiple congenital anomaly caused by an extra autosome. *Lancet* **1**, 790-3 (1960).

80. Edwards, J.H., Harnden, D.G., Cameron, A.H., Crosse, V.M. & Wolff, O.H. A new trisomic syndrome. *Lancet* **1**, 787-90 (1960).

81. Carr, D.H. Genetic basis of abortion. *Annu Rev Genet* **5**, 65-80 (1971).

82. Lejeune, J. *et al.* [3 Cases of Partial Deletion of the Short Arm of a 5 Chromosome]. *C R Hebd Seances Acad Sci* **257**, 3098-102 (1963).

83. van Karnebeek, C.D., Jansweijer, M.C., Leenders, A.G., Offringa, M. & Hennekam, R.C. Diagnostic investigations in individuals with mental retardation: a systematic literature review of their usefulness. *Eur J Hum Genet* **13**, 6-25 (2005).

84. Hansteen, I.L., Varslot, K., Steen-Johnsen, J. & Langard, S. Cytogenetic screening of a new-born population. *Clin Genet* **21**, 309-14 (1982).

85. Maeda, T., Ohno, M., Matsunobu, A., Yoshihara, K. & Yabe, N. A cytogenetic survey of 14,835 consecutive liveborns. *Jinrui Idengaku Zasshi* **36**, 117-29 (1991).

86. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).

87. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).

88. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).

89. Pang, A.W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**, R52 (2010).

90. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).

91. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).

92. Korbel, J.O. *et al.* The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol* **18**, 366-74 (2008).

93. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**, 8-32 (2005).

94. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**, e49 (2005).

95. Cheung, S.W. *et al.* Development and validation of a CGH microarray for clinical cytogenetic diagnosis. *Genet Med* **7**, 422-32 (2005).

96. Shaffer, L.G. *et al.* Targeted genomic microarray analysis for identification of chromosome abnormalities in 1500 consecutive clinical cases. *J Pediatr* **149**, 98-102 (2006).

97. Rickman, L. *et al.* Prenatal detection of unbalanced chromosomal rearrangements by array CGH. *J Med Genet* **43**, 353-61 (2006).

98. Sahoo, T. *et al.* Prenatal diagnosis of chromosomal abnormalities using array-based comparative genomic hybridization. *Genet Med* **8**, 719-27 (2006).

99. Hochstenbach, R. *et al.* Array analysis and karyotyping: workflow consequences based on a retrospective study of 36,325 patients with idiopathic developmental delay in the Netherlands. *Eur J Med Genet* **52**, 161-9 (2009).

100. Reddy, U.M. *et al.* Karyotype versus microarray testing for genetic abnormalities after stillbirth. *N Engl J Med* **367**, 2185-93 (2012).

101. Miller, D.T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**, 749-64 (2010).

102. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-46 (2011).

103. Schinzel, A. Catalogue of unbalanced chromosome aberrations in man. (W. de Gruyter, Berlin; New York, 1984).

104. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524-33 (2009).

105. Yamazawa, K., Ogata, T. & Ferguson-Smith, A.C. Uniparental disomy and human disease: an overview. *Am J Med Genet C Semin Med Genet* **154C**, 329-34 (2010).

106. Zlotogora, J. Parents of children with autosomal recessive diseases are not always carriers of the respective mutant alleles. *Hum Genet* **114**, 521-6 (2004).

107. Kotzot, D. Complex and segmental uniparental disomy updated. *J Med Genet* **45**, 545-56 (2008).

108. Boue, A., Boue, J., Cure, S., Deluchat, C. & Perraudin, N. In vitro cultivation of cells from aneuploid human embryos. Initiation of cell lines and longevity of the cultures. *In Vitro* **11**, 409-13 (1975).

109. Engel, E. A new genetic concept: uniparental disomy and its potential effect, isodisomy. *Am J Med Genet* **6**, 137-43 (1980).

110. Dracopoli, N.C. & Fogh, J. Loss of heterozygosity in cultured human tumor cell lines. *J Natl Cancer Inst* **70**, 83-7 (1983).

111. Yokota, J., Wada, M., Shimosato, Y., Terada, M. & Sugimura, T. Loss of heterozygosity on chromosomes 3, 13, and 17 in small-cell carcinoma and on chromosome 3 in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* **84**, 9252-6 (1987).

112. Koufos, A. *et al.* Loss of heterozygosity in three embryonal tumours suggests a common pathogenetic mechanism. *Nature* **316**, 330-4 (1985).

113. Spence, J.E. *et al.* Uniparental disomy as a mechanism for human genetic disease. *Am J Hum Genet* **42**, 217-26 (1988).

114. Elder, F.F., Nichols, M.M., Hood, O.J. & Harrison, W.R., 3rd. Unbalanced translocation (15;17)(q13;13.3) with apparent Prader-Willi syndrome but without Miller-Dieker syndrome. *Am J Med Genet* **20**, 519-24 (1985).

115. Schinzel, A., Robinson, W.P., Bottani, A., Yagang, X. & Prader, A. Prader-Willi or Angelman syndrome in familial 15q11----q13 deletion of maternal origin? *Hum Genet* **88**, 361-2 (1992).

116. Vidaud, M. & Lavergne, J.M. [Prenatal diagnosis of hemophilia A and B]. *Rev Prat* **39**, 2689-96 (1989).

117. Engel, E. & DeLozier-Blanchet, C.D. Uniparental disomy, isodisomy, and imprinting: probable effects in man and strategies for their detection. *Am J Med Genet* **40**, 432-9 (1991).

118. Ledbetter, D.H. & Engel, E. Uniparental disomy in humans: development of an imprinting map and its implications for prenatal diagnosis. *Hum Mol Genet* **4 Spec No**, 1757-64 (1995).

119. Robinson, W.P. *et al.* Cytogenetic and age-dependent risk factors associated with uniparental disomy 15. *Prenat Diagn* **16**, 837-44 (1996).

120. Field, L.L., Tobias, R., Robinson, W.P., Paisey, R. & Bain, S. Maternal uniparental disomy of chromosome 1 with no apparent phenotypic effects. *Am J Hum Genet* **63**, 1216-20 (1998).

121. Robinson, W.P. Mechanisms leading to uniparental disomy and their clinical consequences. *Bioessays* **22**, 452-9 (2000).

122. Shaffer, L.G. *et al.* American College of Medical Genetics statement of diagnostic testing for uniparental disomy. *Genet Med* **3**, 206-11 (2001).

123. Liehr, T. & Unique. Uniparental disomy (UPD) in clinical genetics : a guide for clinicians and patients. (2014).

124. Liehr, T. Cytogenetic contribution to uniparental disomy (UPD). *Mol Cytogenet* **3**, 8 (2010).

125. Eggermann, T., Soellner, L., Buiting, K. & Kotzot, D. Mosaicism and uniparental disomy in prenatal diagnosis. *Trends Mol Med* **21**, 77-87 (2015).

126. Papenhausen, P. *et al.* UPD detection using homozygosity profiling with a SNP genotyping microarray. *Am J Med Genet A* **155A**, 757-68 (2011).

127. Ford, C.E., Polani, P.E., Briggs, J.H. & Bishop, P.M. A presumptive human XXY/XX mosaic. *Nature* **183**, 1030-2 (1959).

128. Hsu, L.Y. *et al.* Incidence and significance of chromosome mosaicism involving an autosomal structural abnormality diagnosed prenatally through amniocentesis: a collaborative study. *Prenat Diagn* **16**, 1-28 (1996).

129. Kotzot, D. Complex and segmental uniparental disomy (UPD): review and lessons from rare chromosomal complements. *J Med Genet* **38**, 497-507 (2001).

130. Laurie, C.C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* **44**, 642-50 (2012).

131. Rodriguez-Santiago, B. *et al.* Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am J Hum Genet* **87**, 129-38 (2010).

132. Liehr, T. http://www.fish.uniklinikum-jena.de/UPD.html. Vol. 2012 (2013).

133. South, S.T., Lee, C., Lamb, A.N., Higgins, A.W. & Kearney, H.M. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genet Med* **15**, 901-9 (2013).

134. El-Fishawy, P. & State, M.W. The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatr Clin North Am* **33**, 83-105 (2010).

135. Robinson, P.N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610-5 (2008).

136. Plon, S.E. *et al.* Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* **29**, 1282-91 (2008).

137. King, D.A. *et al.* A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res* **24**, 673-87 (2014).

138. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS One* **5**, e13996 (2010).

139. Ting, J.C. *et al.* Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. *Hum Mutat* **28**, 1225-35 (2007).

140. Schroeder, C. *et al.* UPDtool: a tool for detection of iso- and heterodisomy in parent-child trios using SNP microarrays. *Bioinformatics* **29**, 1562-4 (2013).

141. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).

142. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).

143. Evangelou, E., Trikalinos, T.A., Salanti, G. & Ioannidis, J.P. Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet* **2**, e123 (2006).

144. Li, J. *et al.* CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* **28**, 1307-13 (2012).

145. Love, M.I. *et al.* Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* **10**(2011).

146. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

147. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).

148. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741-6 (2007).

149. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

150. Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nat Genet* **40**, 1245-52 (2008).

151. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-12 (2010).

152. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).

153. Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* **10**, 985-7 (2013).

154. Eilers, P.H. & de Menezes, R.X. Quantile smoothing of array CGH data. *Bioinformatics* **21**, 1146-53 (2005).

155. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).

156. Li, L.H. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum Mutat* **27**, 1115-21 (2006).

157. Astle, W. & Balding, D.J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24**, 451-471 (2009).

158. Temple, I.K., Cockwell, A., Hassold, T., Pettay, D. & Jacobs, P. Maternal uniparental disomy for chromosome 14. *J Med Genet* **28**, 511-4 (1991).

159. Balciuniene, J. *et al.* Alpha-tectorin involvement in hearing disabilities: one gene--two phenotypes. *Hum Genet* **105**, 211-6 (1999).

160. Sagong, B., Park, H.J., Lee, K.Y. & Kim, U.K. Identification and functional characterization of novel compound heterozygotic mutations in the TECTA gene. *Gene* **492**, 239-43 (2012).

161. Moreno, J.C. *et al.* Inactivating mutations in the gene for thyroid oxidase 2 (THOX2) and congenital hypothyroidism. *N Engl J Med* **347**, 95-102 (2002).

162. Carvalho, C.M. *et al.* Absence of Heterozygosity Due to Template Switching during Replicative Rearrangements. *Am J Hum Genet* **96**, 555-64 (2015).

163. Horn, D., Schottmann, G. & Meinecke, P. Hyperphosphatasia with mental retardation, brachytelephalangy, and a distinct facial gestalt: Delineation of a recognizable syndrome. *Eur J Med Genet* **53**, 85-8 (2010).

164. Isidor, B., Pichon, O., Baron, S., David, A. & Le Caignec, C. Deletion of the CUL4B gene in a boy with mental retardation, minor facial anomalies, short stature, hypogonadism, and ataxia. *Am J Med Genet A* **152A**, 175-80 (2010).

165. Uldall, P., Alving, J., Hansen, L.K., Kibaek, M. & Buchholt, J. The misdiagnosis of epilepsy in children admitted to a tertiary epilepsy centre with paroxysmal events. *Arch Dis Child* **91**, 219-21 (2006).

166. Koch, M.C. *et al.* Evidence for genetic homogeneity in autosomal recessive generalised myotonia (Becker). *J Med Genet* **30**, 914-7 (1993).

167. Trip, J. *et al.* In tandem analysis of CLCN1 and SCN4A greatly enhances mutation detection in families with non-dystrophic myotonia. *Eur J Hum Genet* **16**, 921-9 (2008).

168. Yamatogi, Y. & Ohtahara, S. Early-infantile epileptic encephalopathy with suppression-bursts, Ohtahara syndrome; its overview referring to our 16 cases. *Brain Dev* **24**, 13-23 (2002).

169. Wolff, M., Casse-Perrot, C. & Dravet, C. Severe myoclonic epilepsy of infants (Dravet syndrome): natural history and neuropsychological findings. *Epilepsia* **47 Suppl 2**, 45-8 (2006).

170. Kearney, J.A. *et al.* A gain-of-function mutation in the sodium channel gene Scn2a results in seizures and behavioral abnormalities. *Neuroscience* **102**, 307-17 (2001).

171. Singh, N.A. *et al.* A role of SCN9A in human epilepsies, as a cause of febrile seizures and as a potential modifier of Dravet syndrome. *PLoS Genet* **5**, e1000649 (2009).

172. Caldovic, L. *et al.* Restoration of ureagenesis in N-acetylglutamate synthase deficiency by N-carbamylglutamate. *J Pediatr* **145**, 552-4 (2004).

173. Krakow, D. *et al.* Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis. *Nat Genet* **36**, 405-10 (2004).

174. Carmichael, H., Shen, Y., Nguyen, T., Hirschhorn, J. & Dauber, A. Whole exome sequencing in a patient with uniparental disomy of chromosome 2 and a complex phenotype. *Clin Genet* (2012).

175. Kotzot, D. & Utermann, G. Uniparental disomy (UPD) other than 15: phenotypes and bibliography updated. *Am J Med Genet A* **136**, 287-305 (2005).

176. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).

177. Koehler, K.E., Hawley, R.S., Sherman, S. & Hassold, T. Recombination and nondisjunction in humans and flies. *Hum Mol Genet* **5 Spec No**, 1495-504 (1996).

178. King, D.A. *et al.* Mosaic structural variation in children with developmental disorders. *Hum Mol Genet* (2015).

179. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**, 417-22 (1998).

180. Biesecker, L.G. & Spinner, N.B. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14**, 307-20 (2013).

181. Lupski, J.R. Genetics. Genome mosaicism--one human, multiple genomes. *Science* **341**, 358-9 (2013).

182. Lindhurst, M.J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* **365**, 611-9 (2011).

183. Behjati, S. *et al.* A Pathogenic Mosaic TP53 Mutation in Two Germ Layers Detected by Next Generation Sequencing. *PLoS One* **9**, e96531 (2014).

184. Machiela, M.J. & Chanock, S.J. Detectable clonal mosaicism in the human genome. *Semin Hematol* **50**, 348-59 (2013).

185. Robberecht, C., Fryns, J.P. & Vermeesch, J.R. Piecing together the problems in diagnosing low-level chromosomal mosaicism. *Genome Med* **2**, 47 (2010).

186. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).

187. Pique-Regi, R., Caceres, A. & Gonzalez, J.R. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics* **11**, 380 (2010).

188. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).

189. Baumbusch, L.O. *et al.* Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9**, 379 (2008).

190. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).

191. Xia, R., Vattathil, S. & Scheet, P. Identification of allelic imbalance with a statistical model for subtle genomic mosaicism. *PLoS Comput Biol* **10**, e1003765 (2014).

192. Ballif, B.C. *et al.* Detection of low-level mosaicism by array CGH in routine diagnostic specimens. *Am J Med Genet A* **140**, 2757-67 (2006).

193. Cheung, S.W. *et al.* Microarray-based CGH detects chromosomal mosaicism not revealed by conventional cytogenetics. *Am J Med Genet A* **143A**, 1679-86 (2007).

194. Pham, J. *et al.* Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10,362 consecutive cases. *Eur J Hum Genet* **22**, 969-78 (2014).

195. Zhong, Q. & Layman, L.C. Genetic considerations in the patient with Turner syndrome--45,X with or without mosaicism. *Fertil Steril* **98**, 775-9 (2012).

196. Smith, B.H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* **42**, 689-700 (2013).

197. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-27 (2013).

198. Haworth, C.M., Davis, O.S. & Plomin, R. Twins Early Development Study (TEDS): a genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Res Hum Genet* **16**, 117-25 (2013).

199. Liu, P. *et al.* Passage number is a major contributor to genomic structural variations in mouse iPSCs. *Stem Cells* **32**, 2657-67 (2014).

200. Narva, E. *et al.* High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat Biotechnol* **28**, 371-7 (2010).

201. Robinson, W.P. *et al.* Origin and outcome of pregnancies affected by androgenetic/biparental chimerism. *Hum Reprod* **22**, 1114-22 (2007).

202. Shin, S.Y., Yoo, H.W., Lee, B.H., Kim, K.S. & Seo, E.J. Identification of the mechanism underlying a human chimera by SNP array analysis. *Am J Med Genet A* **158A**, 2119-23 (2012).

203. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**, 21-32 (2001).

204. Zweier, C. *et al.* Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am J Hum Genet* **80**, 994-1001 (2007).

205. Schinzel, A. *Catalogue of unbalanced chromosome aberrations in man*, xx, 913 p. (W. de Gruyter, Berlin ; New York, 2001).

206. Steinbach, P. *et al.* The dup(3q) syndrome: report of eight cases and review of the literature. *Am J Med Genet* **10**, 159-77 (1981).

207. Reynolds, J.F. *et al.* Isochromosome 12p mosaicism (Pallister mosaic aneuploidy or Pallister-Killian syndrome): report of 11 cases. *Am J Med Genet* **27**, 257-74 (1987).

208. Shinawi, M. *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet* **47**, 332-41 (2010).

209. Phelan, M.C. *et al.* 22q13 deletion syndrome. *Am J Med Genet* **101**, 91-9 (2001).

210. Daber, R. *et al.* Mosaic trisomy 17: variable clinical and cytogenetic presentation. *Am J Med Genet A* **155A**, 2489-95 (2011).

211. Gogiel, M. *et al.* Genome-wide paternal uniparental disomy mosaicism in a woman with Beckwith-Wiedemann syndrome and ovarian steroid cell tumour. *Eur J Hum Genet* **21**, 788-91 (2013).

212. Willis, M.J., Bird, L.M., Dell'Aquilla, M. & Jones, M.C. Expanding the phenotype of mosaic trisomy 20. *Am J Med Genet A* **146**, 330-6 (2008).

213. Schmeisser, M.J. *et al.* The Nedd4-binding protein 3 (N4BP3) is crucial for axonal and dendritic branching in developing neurons. *Neural Dev* **8**, 18 (2013).

214. Mavrogiannis, L.A. *et al.* Haploinsufficiency of the human homeobox gene ALX4 causes skull ossification defects. *Nat Genet* **27**, 17-8 (2001).

215. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* (2014).

216. Endler, G., Greinix, H., Winkler, K., Mitterbauer, G. & Mannhalter, C. Genetic fingerprinting in mouthwashes of patients after allogeneic bone marrow transplantation. *Bone Marrow Transplant* **24**, 95-8 (1999).

217. Staaf, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol* **9**, R136 (2008).

218. Forsberg, L.A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet* **46**, 624-8 (2014).

219. Forsberg, L.A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* **90**, 217-28 (2012).

220. Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K. & Mardis, E.R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27-38 (2013).

221. Meynert, A.M., Ansari, M., FitzPatrick, D.R. & Taylor, M.S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**, 247 (2014).

222. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* **39**, S48-54 (2007).

223. Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747-54 (2012).

224. Magi, A. *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* **14**, R120 (2013).

225. Sathirapongsasuti, J.F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**, 2648-54 (2011).

226. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525-32 (2012).

227. Backenroth, D. *et al.* CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res* **42**, e97 (2014).

228. Lonigro, R.J. *et al.* Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. *Neoplasia* **13**, 1019-25 (2011).

229. Amarasinghe, K.C. *et al.* Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15**, 732 (2014).

230. Kaye, J. *et al.* Managing clinically significant findings in research: the UK10K example. *Eur J Hum Genet* **22**, 1100-4 (2014).

231. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-81 (2009).

232. Choate, K.A. *et al.* Frequent somatic reversion of KRT1 mutations in ichthyosis with confetti. *J Clin Invest* **125**, 1703-7 (2015).

233. Snape, K. *et al.* Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. *Nat Genet* **43**, 527-9 (2011).

234. Guilherme, R.S. *et al.* Mechanisms of ring chromosome formation, ring instability and clinical consequences. *BMC Med Genet* **12**, 171 (2011).

235. Knijnenburg, J. *et al.* Ring chromosome formation as a novel escape mechanism in patients with inverted duplication and terminal deletion. *Eur J Hum Genet* **15**, 548-55 (2007).

236. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-84 (2011).

237. Plaiasu, V., Ochiana, D., Motei, G. & Georgescu, A. A rare chromosomal disorder - isochromosome 18p syndrome. *Maedica (Buchar)* **6**, 132-6 (2011).

238. Wulfsberg, E.A., Weaver, R.P., Cunniff, C.M., Jones, M.C. & Jones, K.L. Chromosome 10qter deletion syndrome: a review and report of three new cases. *Am J Med Genet* **32**, 364-7 (1989).

239. Conlin, L.K. *et al.* Utility of SNP arrays in detecting, quantifying, and determining meiotic origin of tetrasomy 12p in blood from individuals with Pallister-Killian syndrome. *Am J Med Genet A* **158A**, 3046-53 (2012).

240. Choo, S., Teo, S.H., Tan, M., Yong, M.H. & Ho, L.Y. Tissue-limited mosaicism in Pallister-Killian syndrome -- a case in point. *J Perinatol* **22**, 420-3 (2002).

241. Piotrowski, A. *et al.* Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat* **29**, 1118-24 (2008).

242. O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E. & Snyder, M.P. Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci U S A* **109**, 18018-23 (2012).

243. Abraham, J.E. *et al.* Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics* **5**, 19 (2012).

244. Daksis, J.I. & Erikson, G.H. Heteropolymeric triplex-based genomic assay to detect pathogens or single-nucleotide polymorphisms in human genomic samples. *PLoS One* **2**, e305 (2007).

245. Ayoglu, B. *et al.* Affinity proteomics within rare diseases: a BIO-NMD study for blood biomarkers of muscular dystrophies. *EMBO Mol Med* **6**, 918-36 (2014).

246. Garcia-Closas, M. *et al.* Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiol Biomarkers Prev* **10**, 687-96 (2001).

247. Drost, M. *et al.* Genetic screens to identify pathogenic gene variants in the common cancer predisposition Lynch syndrome. *Proc Natl Acad Sci U S A* **110**, 9403-8 (2013).

248. Hassold, T., Merrill, M., Adkins, K., Freeman, S. & Sherman, S. Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *Am J Hum Genet* **57**, 867-74 (1995).

249. Gravholt, C.H. Chapter 44 - Sex-Chromosome Abnormalities. in *Emery and Rimoin's Principles and Practice of Medical Genetics (Sixth Edition)* (eds. Rimoin, D. & Korf, R.P.) 1-32 (Academic Press, Oxford, 2013).

250. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).

251. Papavassiliou, P. *et al.* The phenotype of persons having mosaicism for trisomy 21/Down syndrome reflects the percentage of trisomic cells present in different tissues. *Am J Med Genet A* **149A**, 573-83 (2009).

252. Choate, K.A. *et al.* Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in KRT10. *Science* **330**, 94-7 (2010).

253. McDermott, D.H. *et al.* Chromothriptic cure of WHIM syndrome. *Cell* **160**, 686-99 (2015).

254. Crow, J.F. Two centuries of genetics: a view from halftime. *Annu Rev Genomics Hum Genet* **1**, 21-40 (2000).

255. Tripp, S.G., M. Economic Impact of the Human Genome Project. (2011).

256. Meldrum, D.R. Tech.Sight. Sequencing genomes and beyond. *Science* **292**, 515-7 (2001).

257. Hayden, E.C. Technology: The $1,000 genome. *Nature* **507**, 294-5 (2014).

258. Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K. & Gerstein, M.B. The real cost of sequencing: higher than you think! *Genome Biol* **12**, 125 (2011).

259. Michaud, M. Illumina Is Dominting The Sequencing Market. (2015).

260. Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G. & Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci U S A* **106**, 7702-7 (2009).

261. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-8 (2009).

262. Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**, 734-40 (2011).

263. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat Biotechnol* **29**, 723-30 (2011).

264. Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R. & Griffin, J.L. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **40**, 387-426 (2011).

265. Bibby, K. & Peccia, J. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environ Sci Technol* **47**, 1945-51 (2013).

266. Robinson, P.N. & Mundlos, S. The human phenotype ontology. *Clin Genet* **77**, 525-34 (2010).

267. Ferry, Q. *et al.* Diagnostically relevant facial gestalt information from ordinary photos. *Elife* **3**, e02020 (2014).

268. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-55 (2011).

269. Dietz, H.C. New therapeutic approaches to mendelian disorders. *N Engl J Med* **363**, 852-63 (2010).

270. Garg, S.K. *et al.* Systemic delivery of MeCP2 rescues behavioral and cellular deficits in female mouse models of Rett syndrome. *J Neurosci* **33**, 13612-20 (2013).

271. Jiang, J. *et al.* Translating dosage compensation to trisomy 21. *Nature* **500**, 296-300 (2013).

272. Buchanan, A., Sachs, A., Toler, T. & Tsipis, J. NIPT: current utilization and implications for the future of prenatal genetic counseling. *Prenat Diagn* **34**, 850-7 (2014).

273. Hill, M. *et al.* Evaluation of non-invasive prenatal testing (NIPT) for aneuploidy in an NHS setting: a reliable accurate prenatal non-invasive diagnosis (RAPID) protocol. *BMC Pregnancy Childbirth* **14**, 229 (2014).

274. Gupta, K. Disability-selective abortion: denying human rights to make a "perfect world"? *Indian J Med Ethics* **10**, 70-1 (2013).

275. Yurkiewicz, I.R., Korf, B.R. & Lehmann, L.S. Prenatal whole-genome sequencing--is the quest to know a fetus's future ethical? *N Engl J Med* **370**, 195-7 (2014).

276. Liang, P. *et al.* CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell* **6**, 363-72 (2015).

277. Cyranoski, D. & Reardon, S. Embryo editing sparks epic debate. *Nature* **520**, 593-4 (2015).

278. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-4 (2013).

279. Ohno, S. So much "junk" DNA in our genome, in Evolution of Genetic Systems., (Gordon and Breach, New York, 1972).

280. Palazzo, A.F. & Gregory, T.R. The case for junk DNA. *PLoS Genet* **10**, e1004351 (2014).

281. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

282. van Berkum, N.L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* (2010).

283. Riethoven, J.J. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol* **674**, 33-42 (2010).

284. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403 (2013).

285. Ahn, C.P. *et al.* The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Sdss-Iii Apache Point Observatory Galactic Evolution Experiment. *Astrophysical Journal Supplement Series* **211**(2014).

}