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 Sections describing the second stage of analysis contain unpublished results.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.

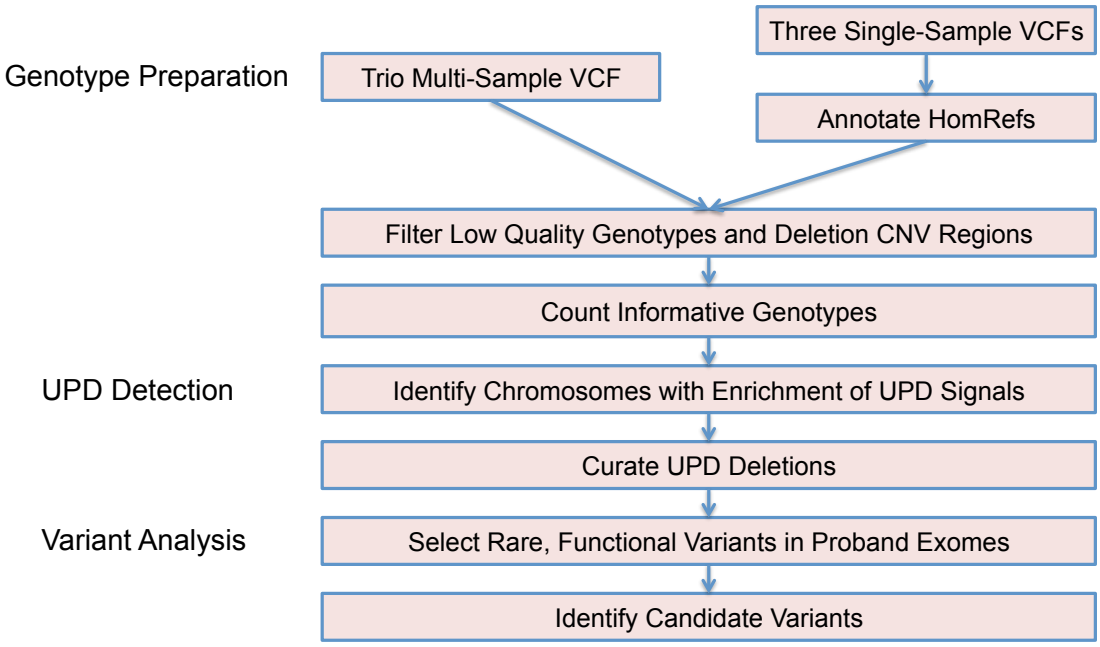

Figure 2-1 Study workflow. The study consisted of three main steps: data preparation, UPD detection, and candidate variant analysis. In the data preparation stage, informative genotypes were collected in all members of each trio. Either a multi-sample trio VCF or three single-sample VCFs can be used 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.

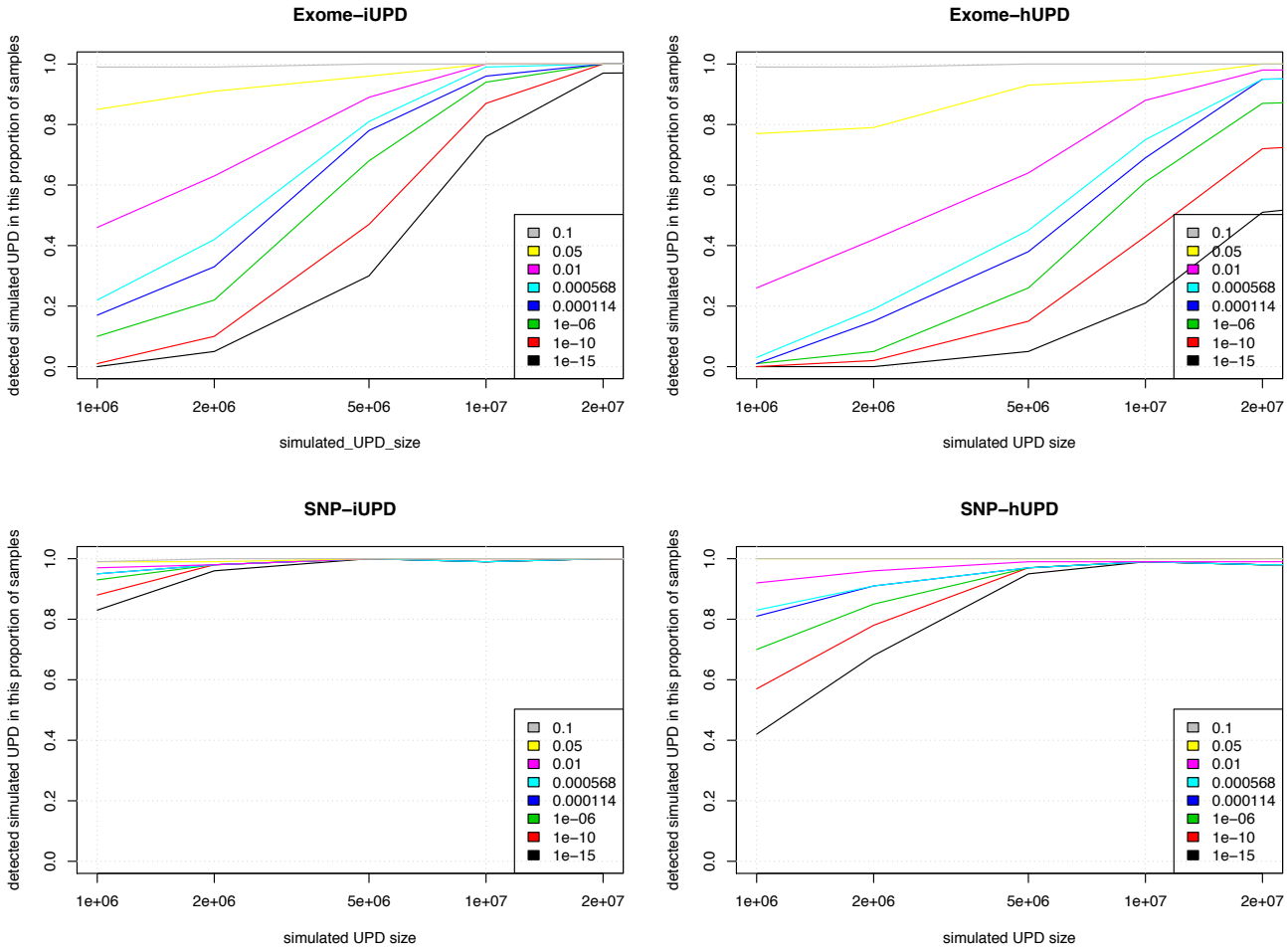

Figure 2-2 Sensitivity of UPD detection simulations. Simulations to assess sensitivity of UPD detections at different sizes, from different data sources. (iUPD) isodisomy; (hUPD) heterodisomy

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.

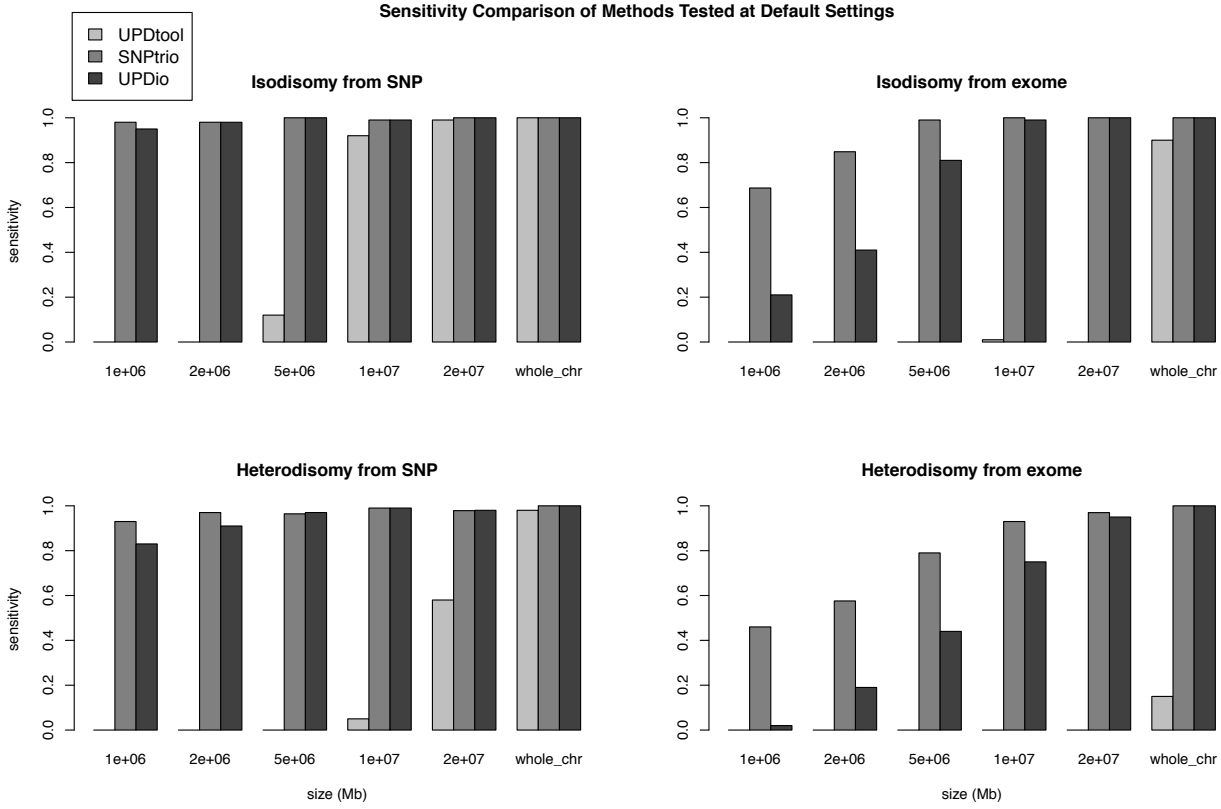

Figure 2-3 Sensitivity comparisons. Simulations were performed to measure the sensitivity of detecting introduced UPD events from SNP and exome data, ranging in size from 1 Mb to chromosomal.

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

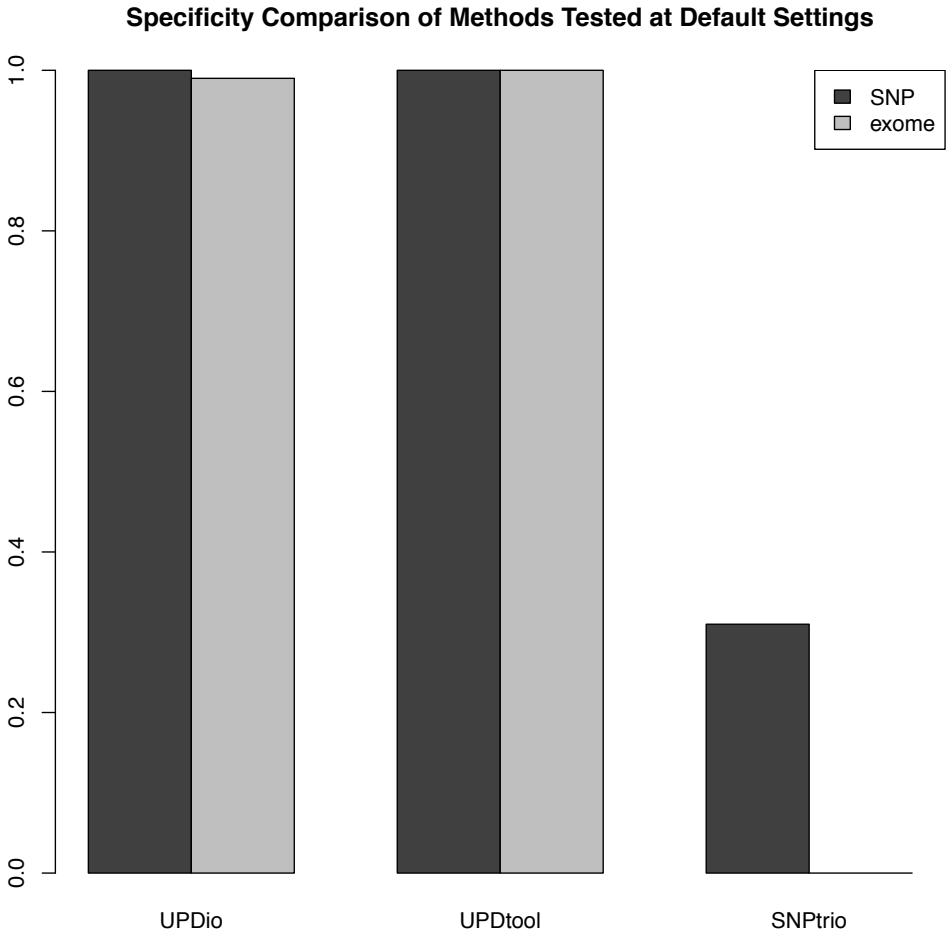

Figure 2-4 Specificity comparisons. Simulations on normal SNP and exome samples were compared to measure the proportion of samples without UPD detections.

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).

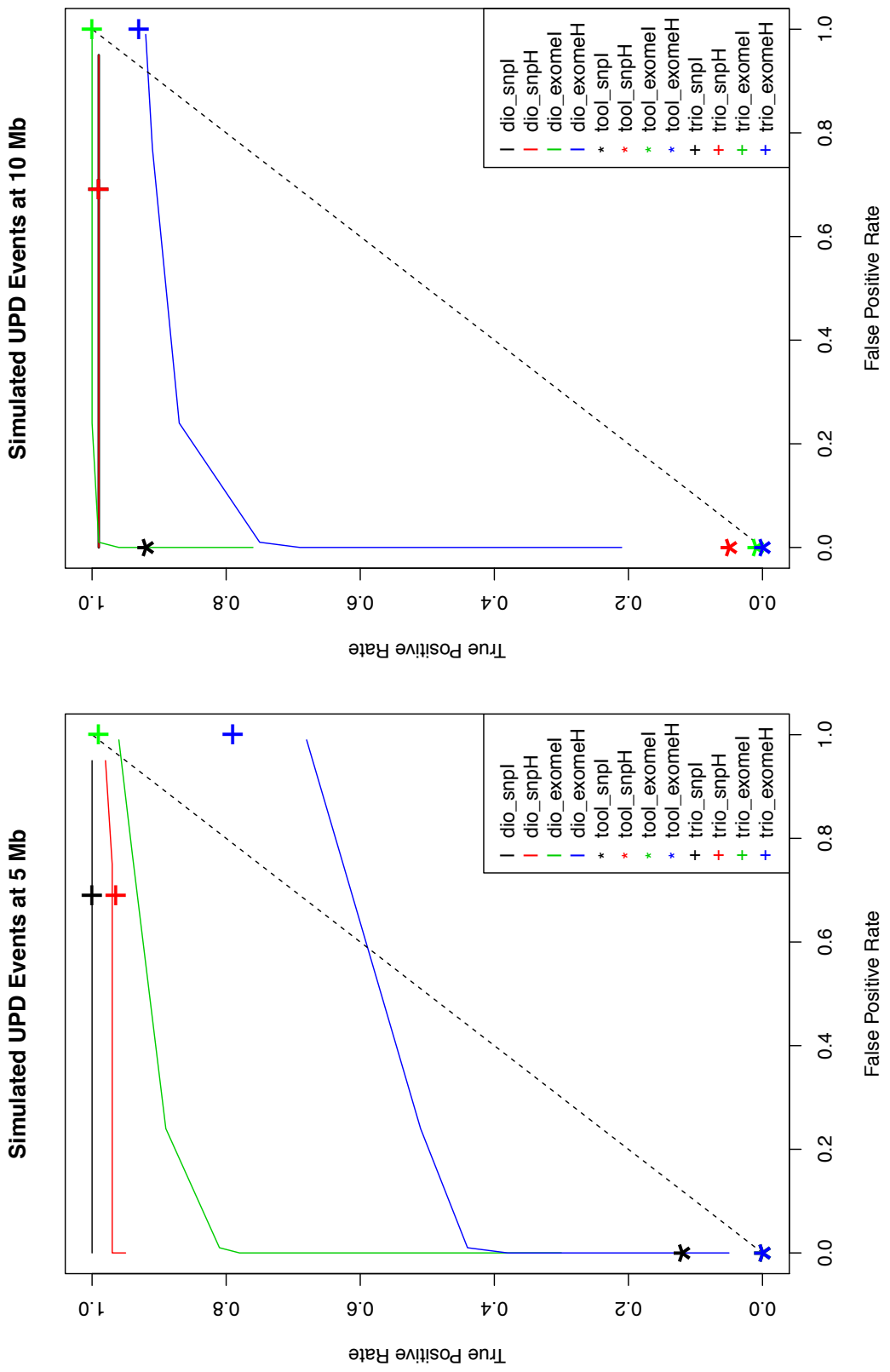

Figure 2-5 Receiver operator characteristic curve comparing UPD detection accuracy at different simulated UPD sizes. (dio) UP*D*io, (tool) UPDtool, (trio) SNPtrio.

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).

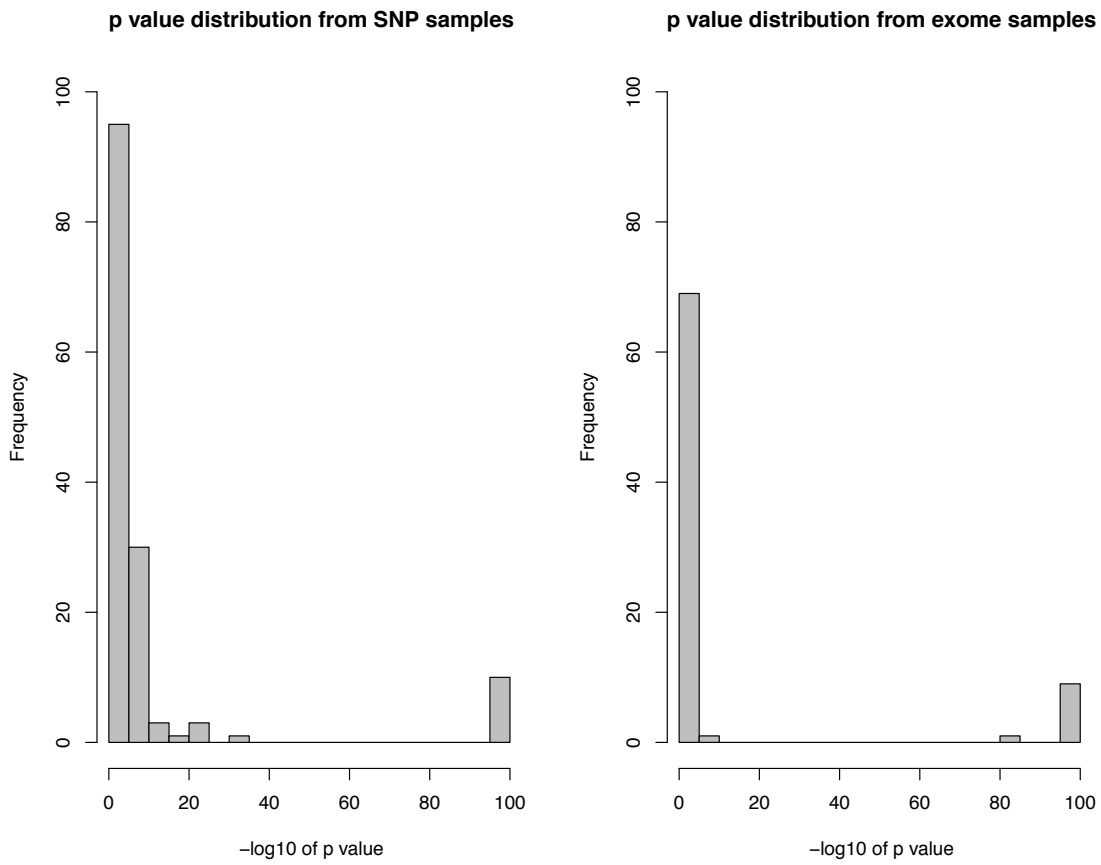

Figure 2-6 DDD UPD p value distributions. Distribution of the –log10 p values for UPD detections from different data sources, with or without CNV data. Presence of sample-specific CNV data increases the proportion of extremely significant events and decreases the proportion of events with p values less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

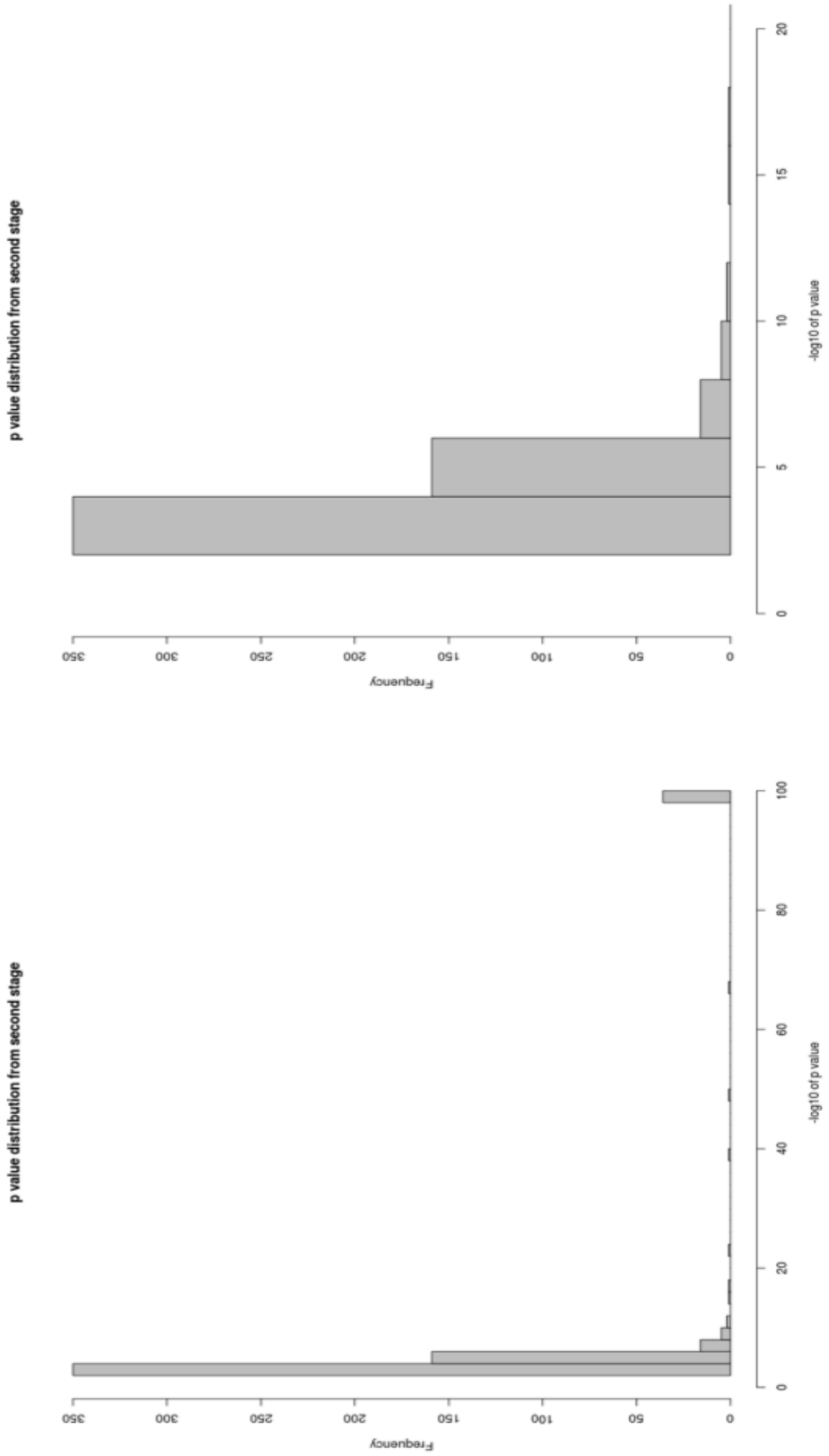

Figure 2-8 Distribution of –log10 p values for UPD detections in the second stage analysis. p value minimum truncated to 1e-100. The vast majority of candidate UPD calls are at low significance and cluster below 1e-10. The second graph depicts the events more significant than 1e-20.

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.