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).

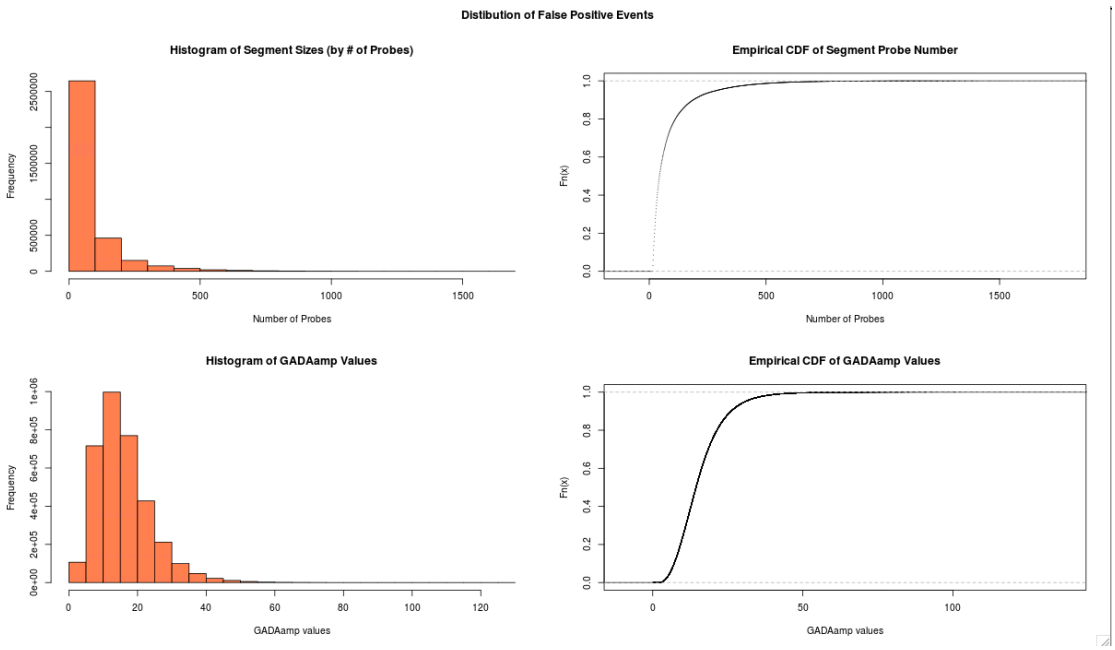

Figure 4-1 Distribution of size and signal-strength of false positives. The histograms of probe-number and GADAamp values both show long tails, with the majority of putative events being smaller and weak. The cumulative distribution functions from the data (right column) showed that events with greater 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).

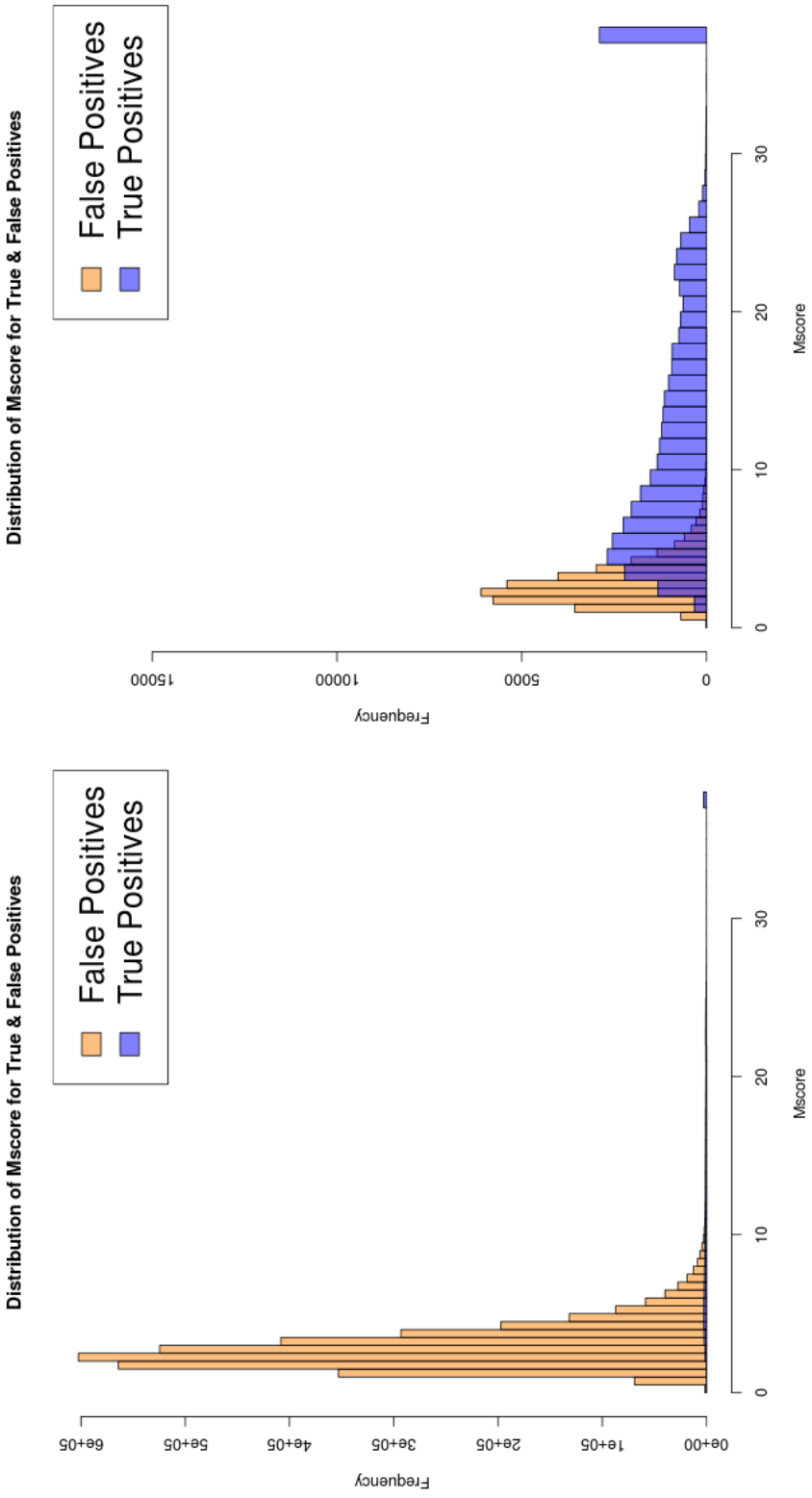

Figure 4-2 Comparing Mscores of true positives and false positives. The Mscore distributions for all simulated false positive events (first graph) and for a random subselection of false positive events equal to the number of true positive events (second graph) demonstrated that the true positive events 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.

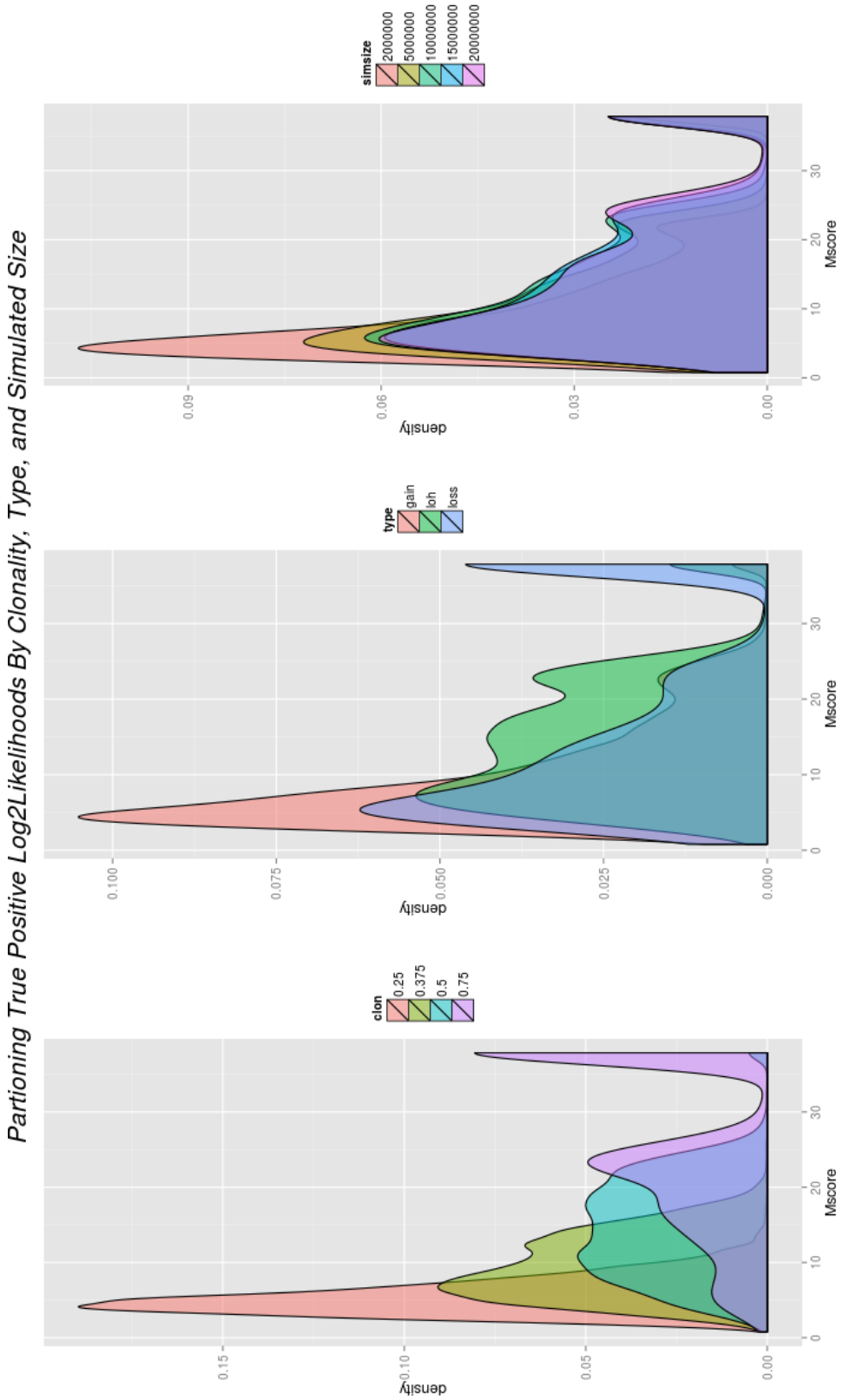

Figure 4-3 Stratifying Mscore by simulation clonality, type, and size. I stratified the true positive events by Mscore to better define the relationship between Mscore thresholds and simulated mosaic events. The mosaic events with the lowest Mscore were those at the lowest clonality (left side of left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

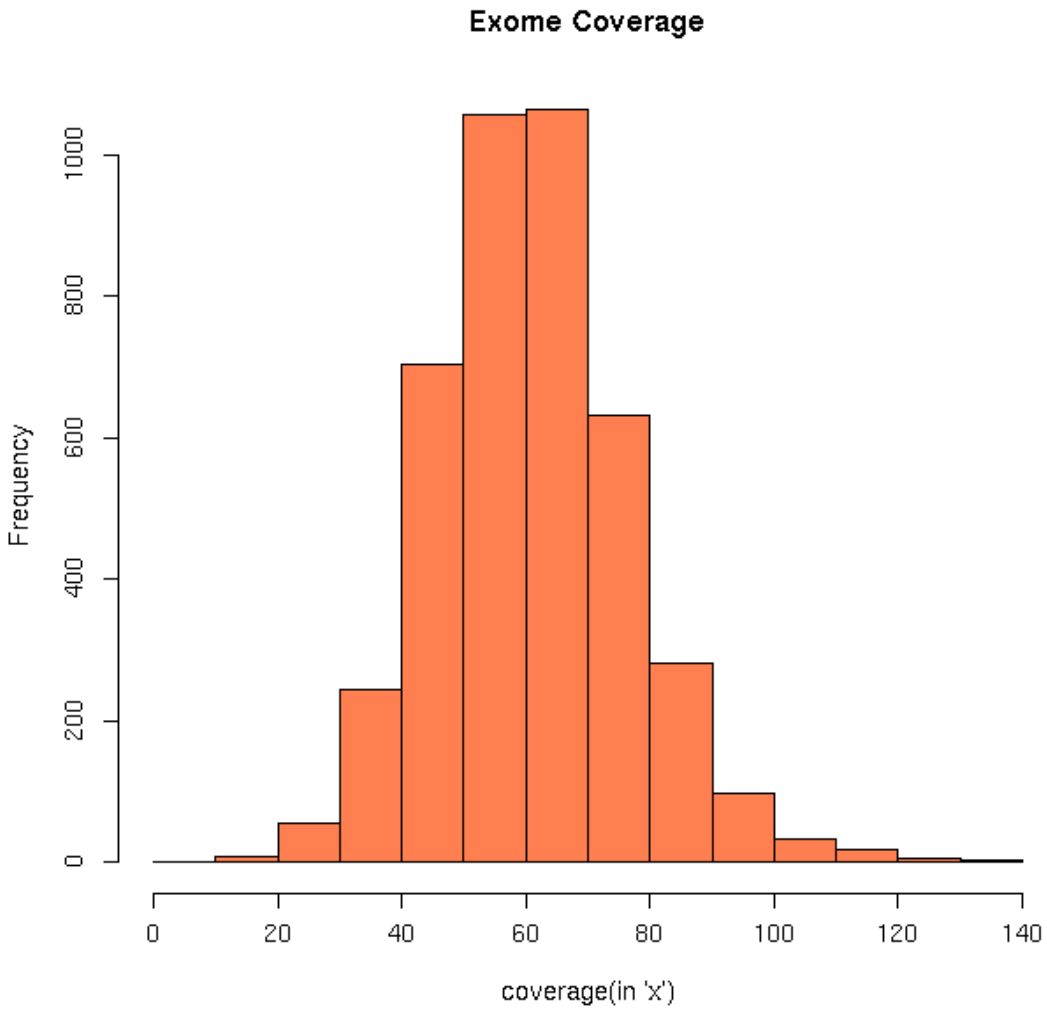

Figure 4-4 The distribution of average coverage of exomes used in this study.

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).

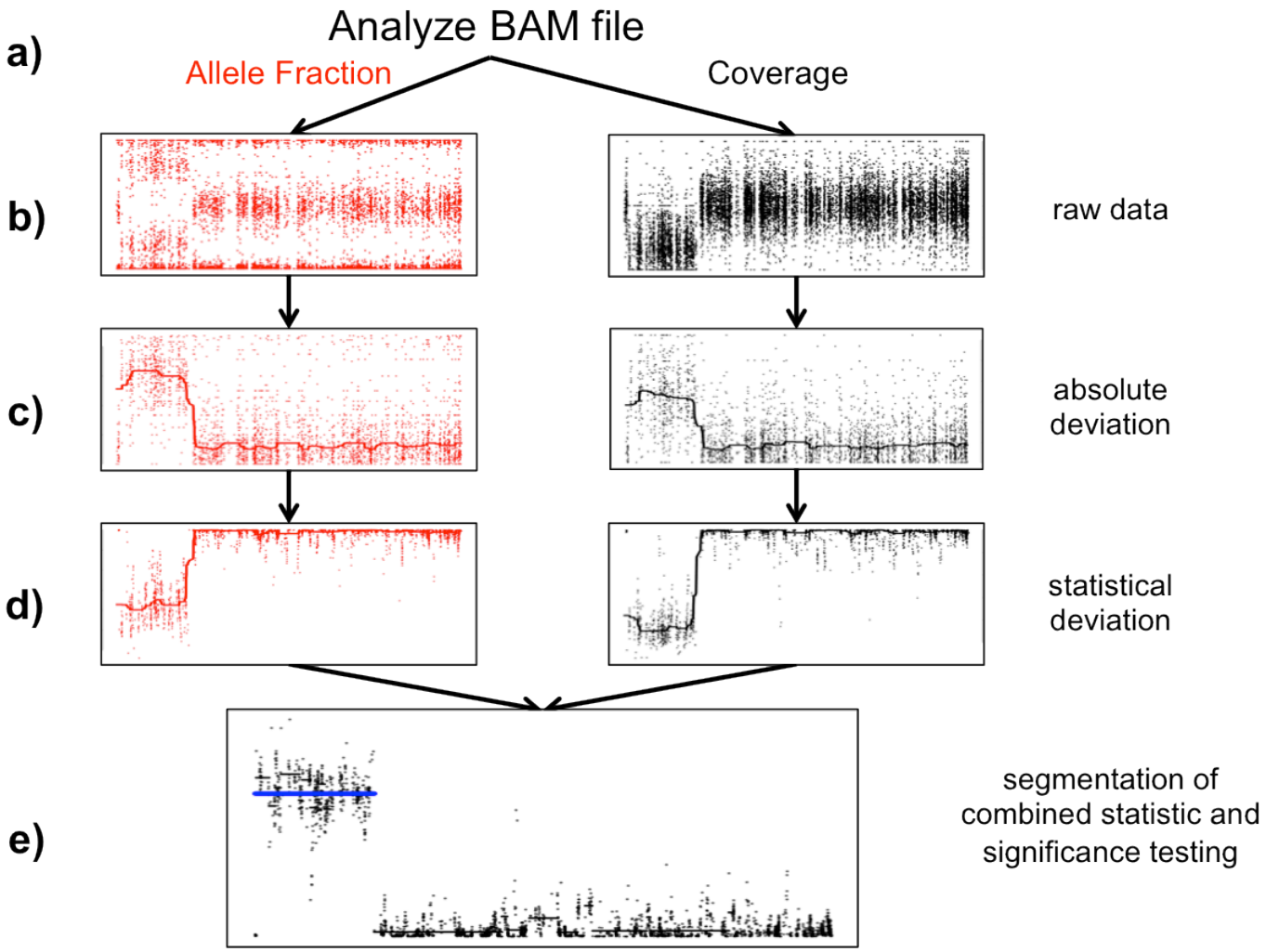

Figure 4-5 Detecting structural mosaicism using MrMosaic. a) Exome data are stored in a BAM file from which allele fraction (left column) and coverage (right column) are measured at polymorphic positions within or near target regions. b) A simulated mosaic deletion is depicted and the raw data, consisting 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).

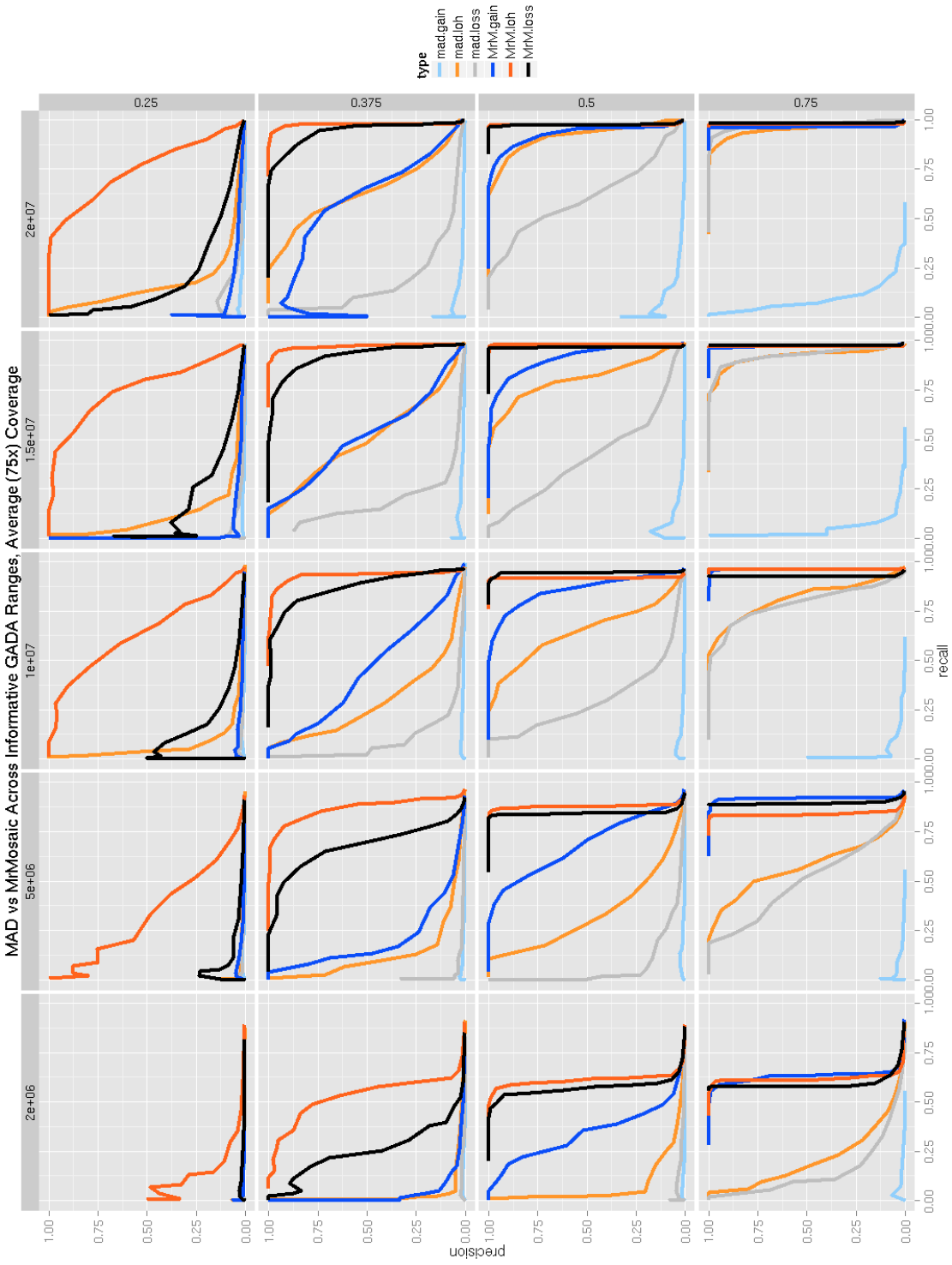

Figure 4-6 WE performance of MAD and MrMosaic algorithms. In this grid of precision-recall graphs, the performance of MAD and MrMosaic is compared at 75x average coverage for a range

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.

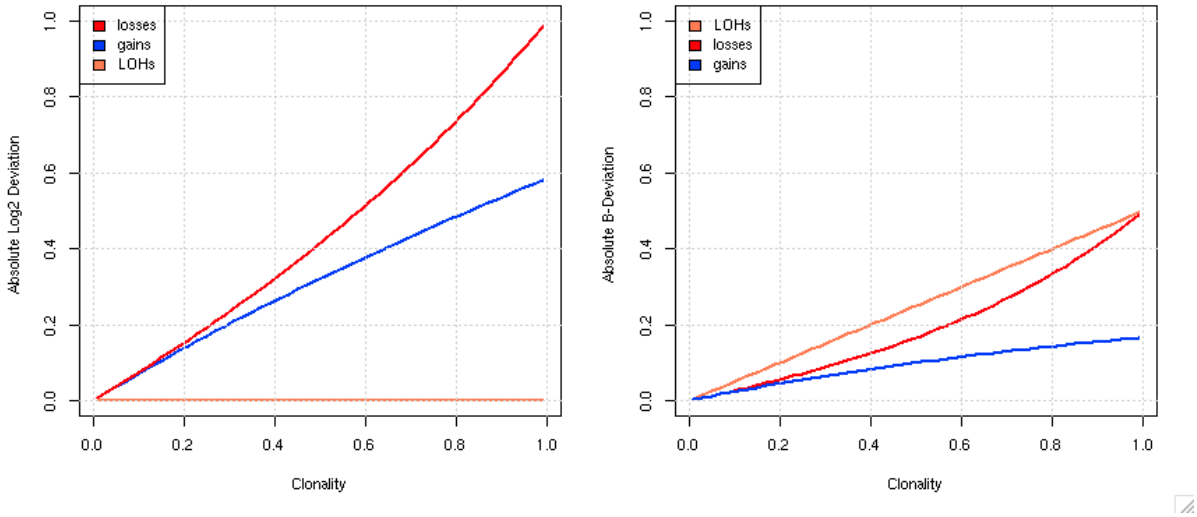

Figure 4-7 Relationship between Clonality and Metrics. The relationship between clonality and measured metrics (C_{dev} and B_{dev}) indicates that while LOH events result in no deviation of C_{dev}, gains have the smallest deflection of B_{dev}, compared to other events of a given clonality.

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.

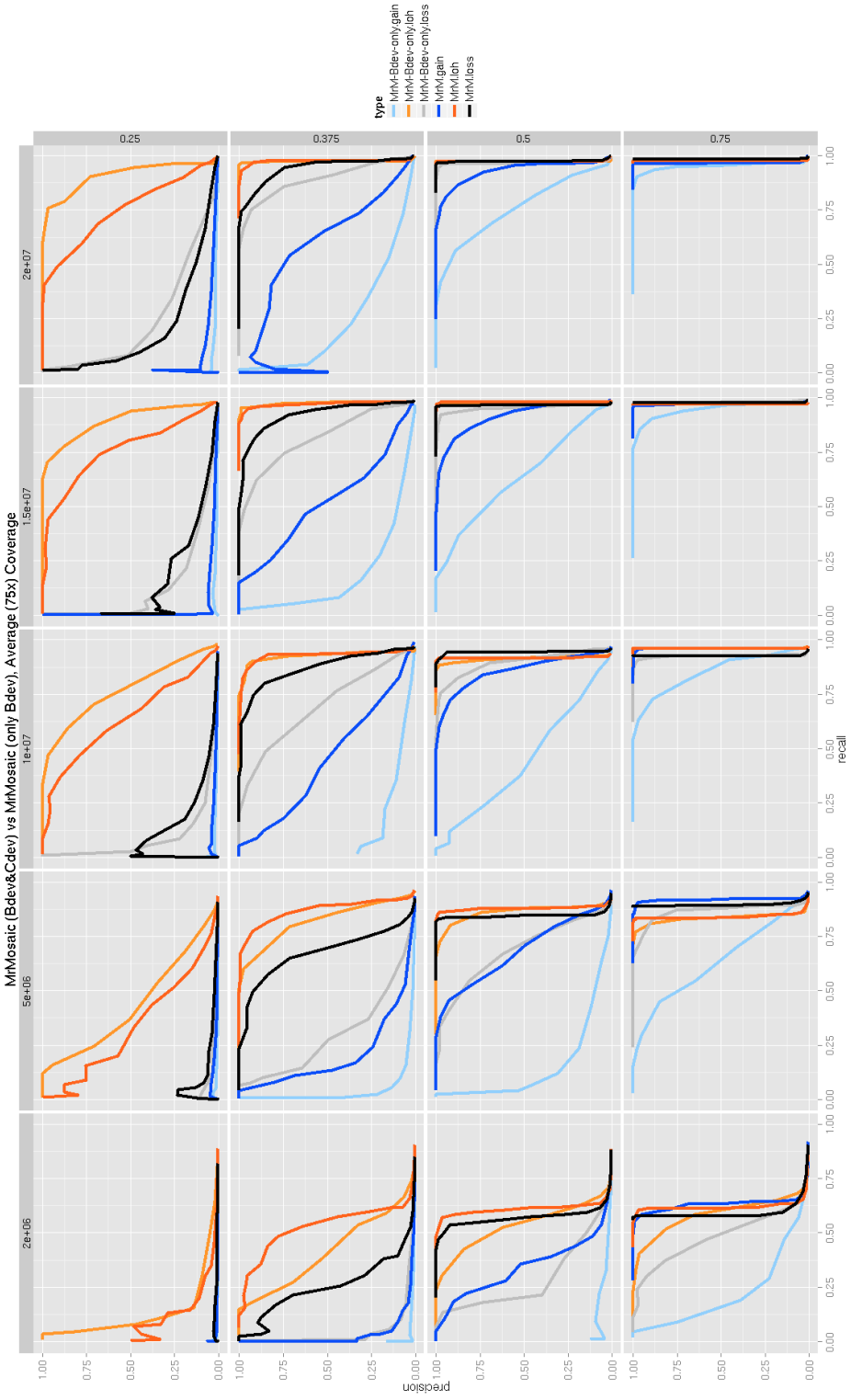

Figure 4-8 WE MrMosaic, Cdev & Bdev vs Bdev-alone. MrMosaic combines the statistical deviation from differences in coverage (C_{dev}) and non-reference proportion (B_{dev}) while the MAD approach uses B_{dev} alone. I ran MrMosaic in standard joint-mode and also using B_{dev} alone. The results demonstrate 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.

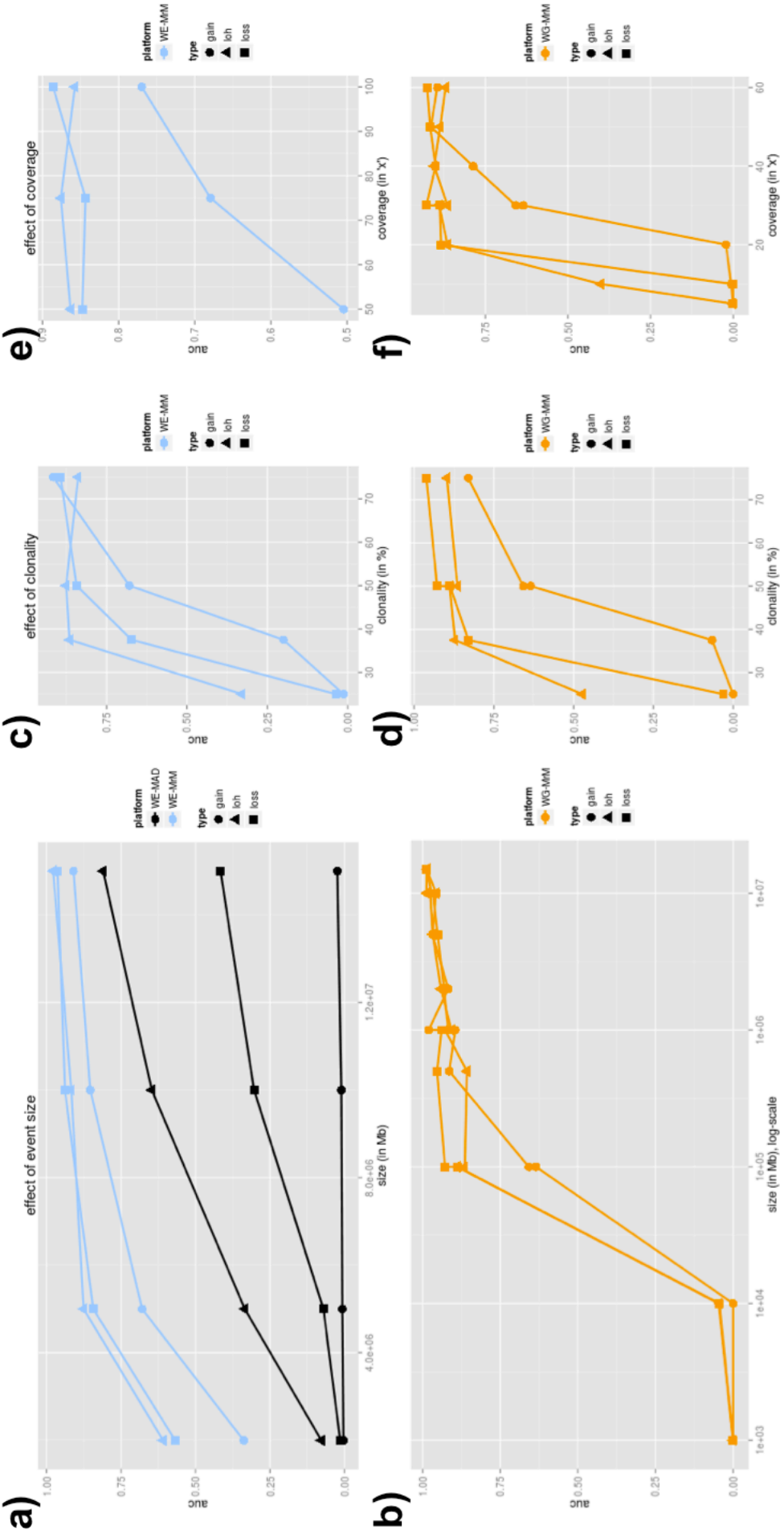
a)
effect of event size
auc
size (in Mb)
platform
WE-MAD
WE-MM
type
gain
loh
loss
b)
size (in Mb), log-scale
platform
WG-MM
c)
effect of clonality
clonality (in %)
d)
e)
effect of coverage
coverage (in 'x')
f)

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).

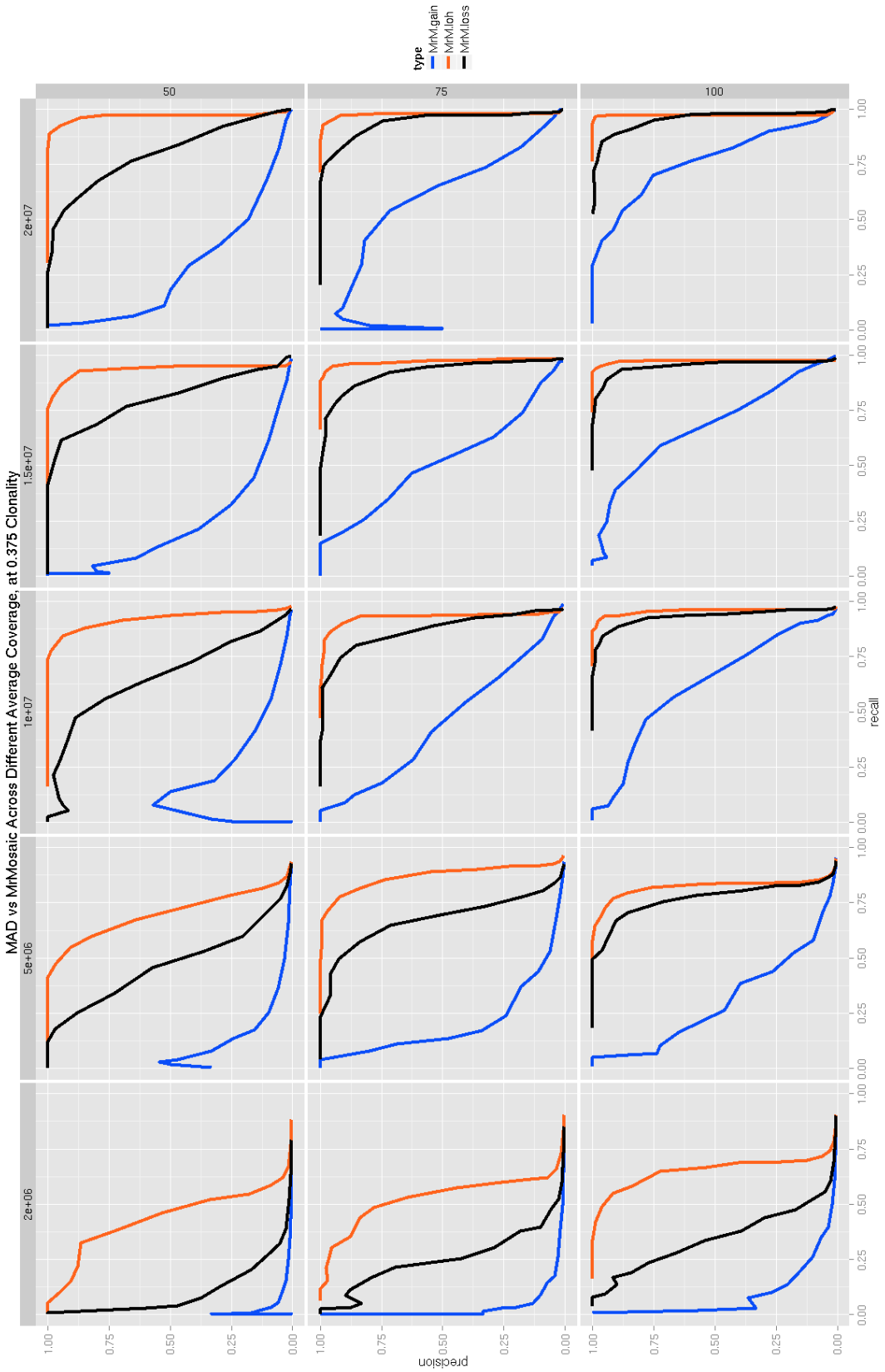

Figure 4-10 WE performance of MrMosaic across 50-100x. I generated simulated exomes of 50x, 75x, and 100x depths and measured MrMosaic detection performance across coverage. Detection

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).

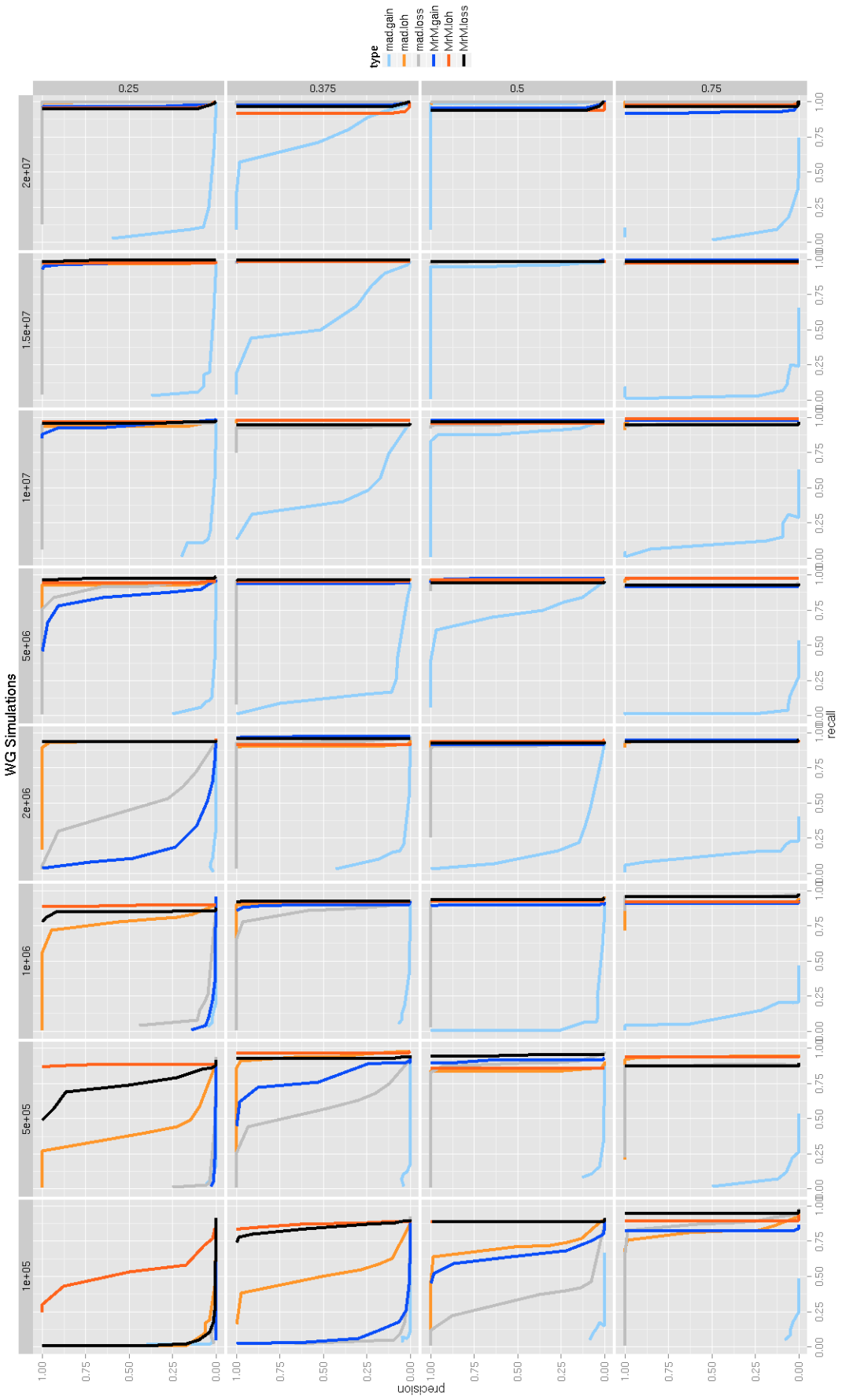
type
mad.gain
mad.loh
mad.loss
MrM.gain
MrM.loh
MrM.loss
0.25
0.375
0.5
0.75
WG Simulations
2e+07
1.5e+07
1e+07
5e+06
2e+06
1e+06
5e+05
1e+05
recall
precision
1.00
0.75
0.50
0.25
0.00

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).

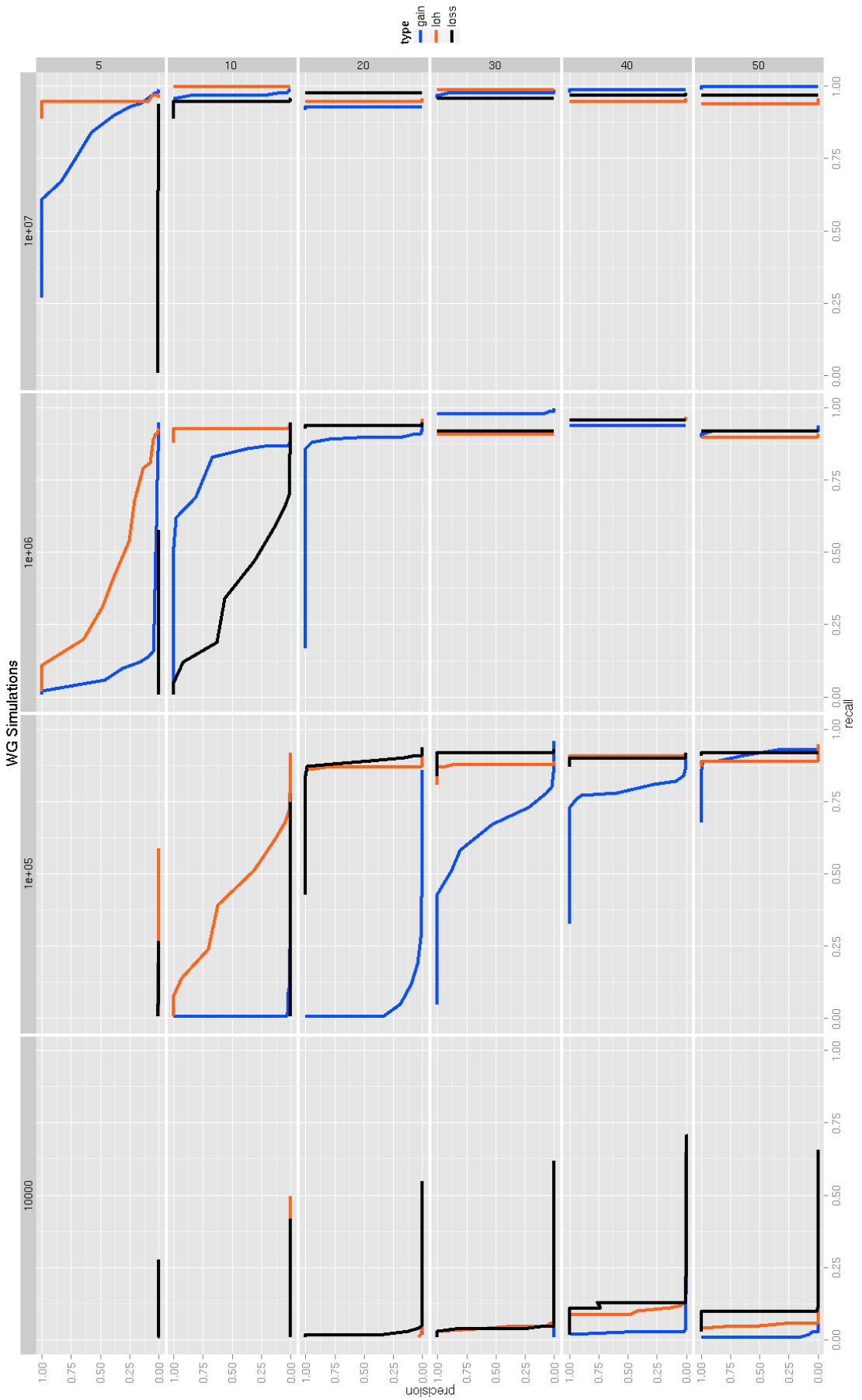
type
gain
loh
loss
5
10
20
30
40
50
1e+07
1e+06
1e+05
10000
WG Simulations
recall
precision
1.00
0.75
0.50
0.25
0.00

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).

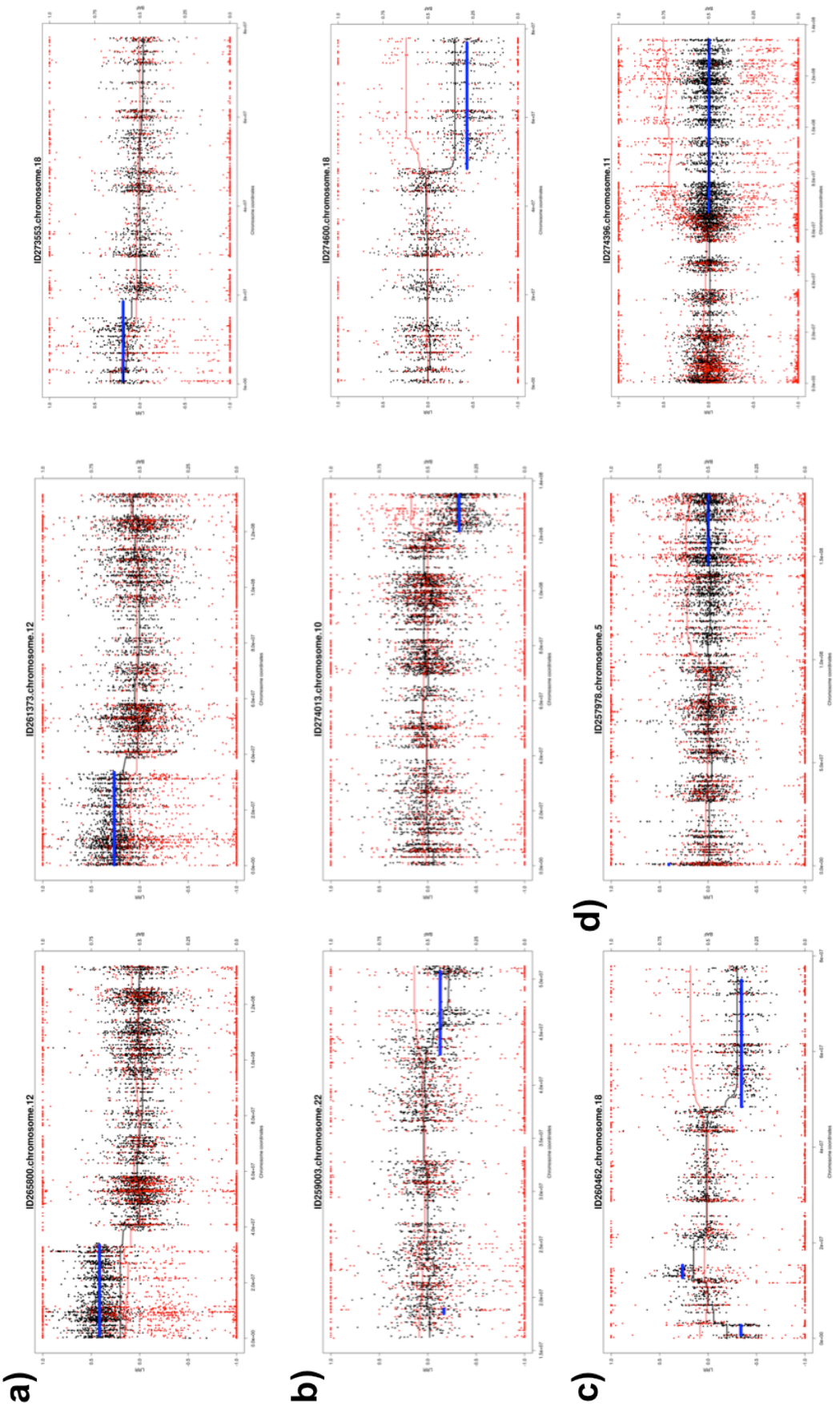
a)
b)
c)
d)
ID265800.chromosome.12
ID259003.chromosome.22
ID260462.chromosome.18
ID257978.chromosome.5
ID261373.chromosome.12
ID274013.chromosome.10
ID274600.chromosome.18
ID274396.chromosome.11
ID273553.chromosome.18

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).

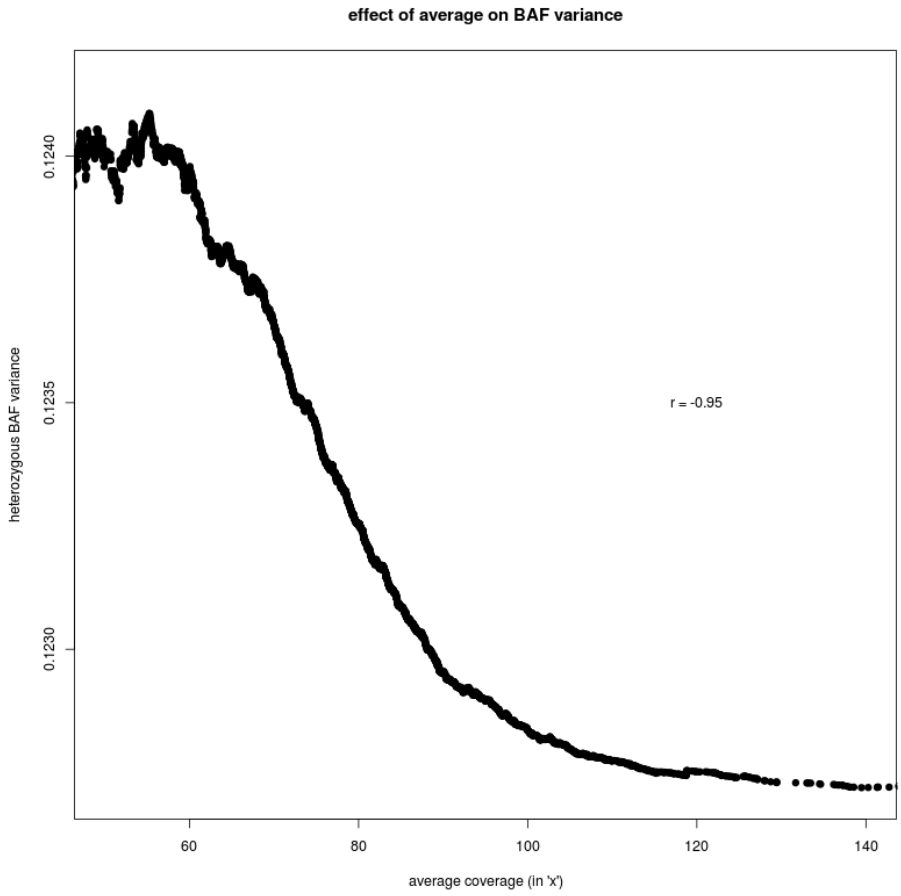

Figure 4-14 Observed BAF variance at heterozygous sites in WE data across samples with different sequencing depth.

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.

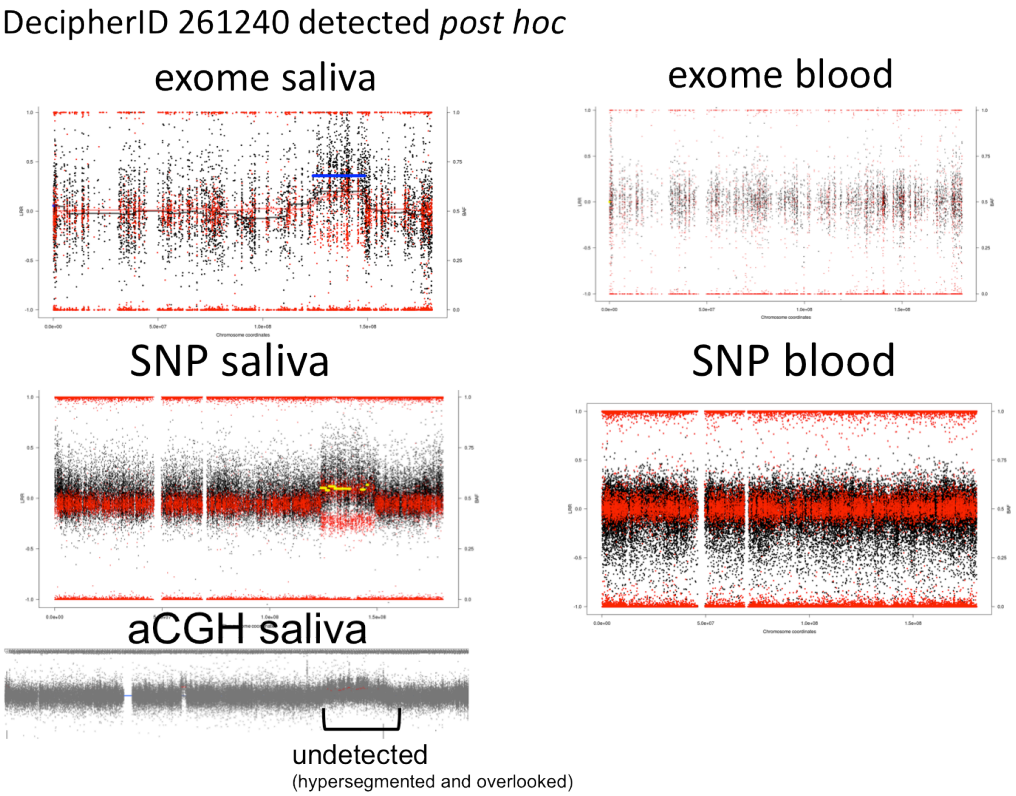

Figure 4-15 Detection of 261240 was *post hoc* in that originally, DNA from blood was analysed and no event was detected, although SNP microarray data data which had been previously analysed identified an abnormality in saliva, suggesting that either the event was missed by exome in blood, or that the 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).

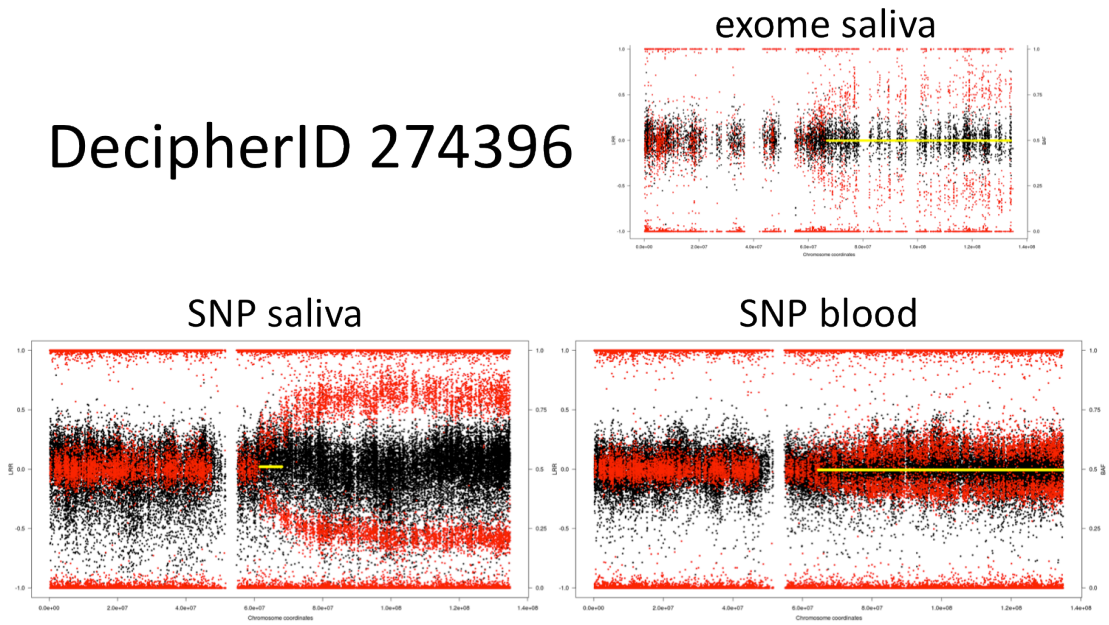

Figure 4-16 SNP Validation of 274396. A gradient of clonality present on chromosome 11, extending to the 3' end of the chromosome.

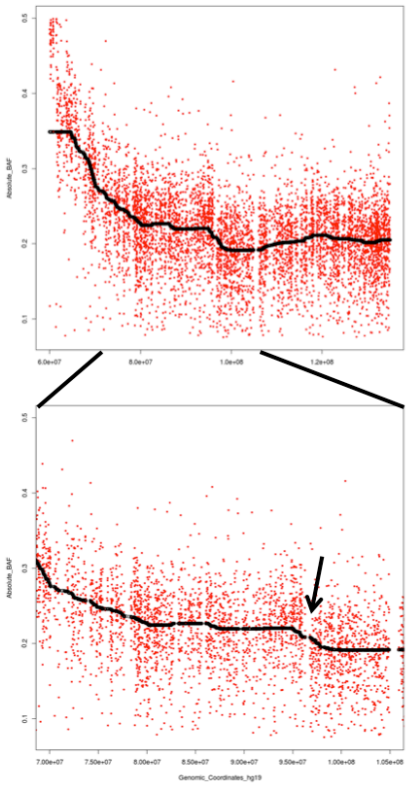

Figure 4-17 Investigating the mosaic reversion event. I examined SNP microarray data to help localise the cause of the suspected reversion. These plot displays heterozygous BAFs (BAFs above 0.5 are reflected below the 0.5 line) from SNP microarray data on the 3' end of chromosome 11, with a median trend 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].

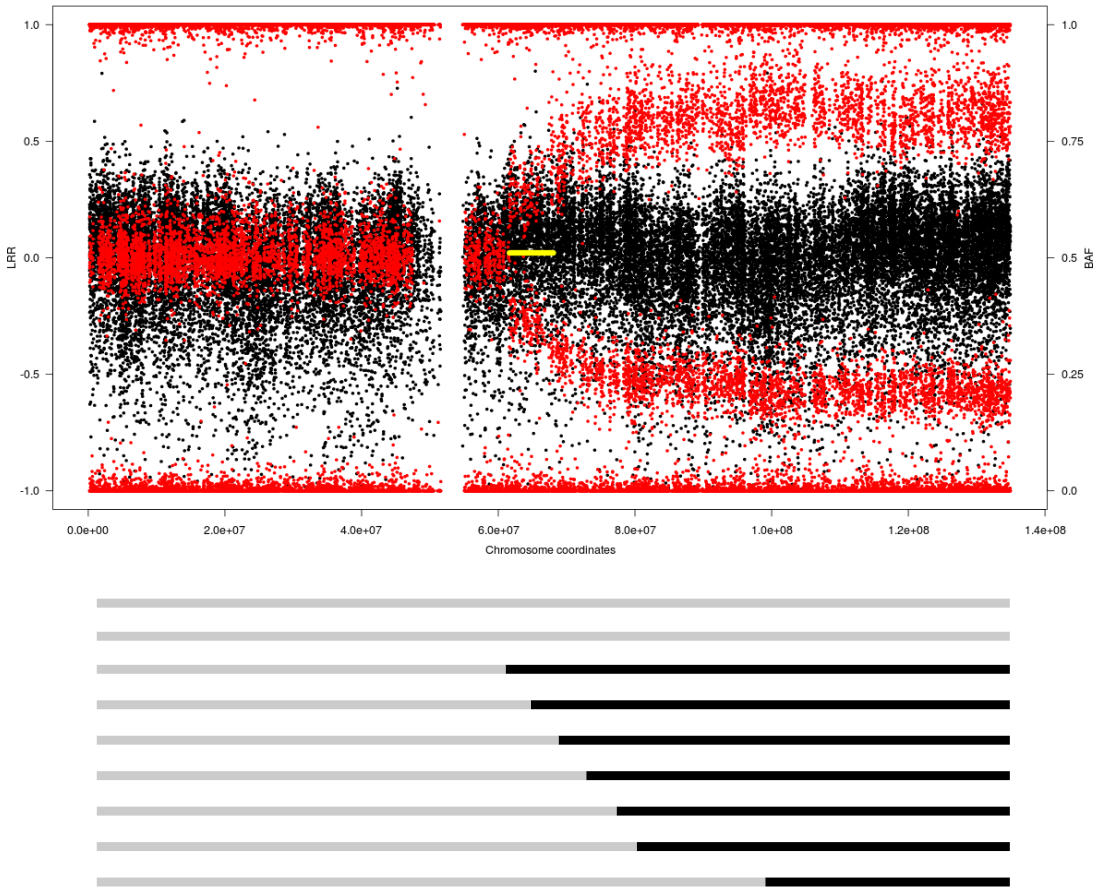

Figure 4-18 The revertant mosaic event detected in this study, and below, a schematic depicting the hypothesised mechanism, with black lines representing segments of LOH in independent revertant clones, while the gray represent wild-type. This reversion is 'incomplete' in the sense that, at least at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

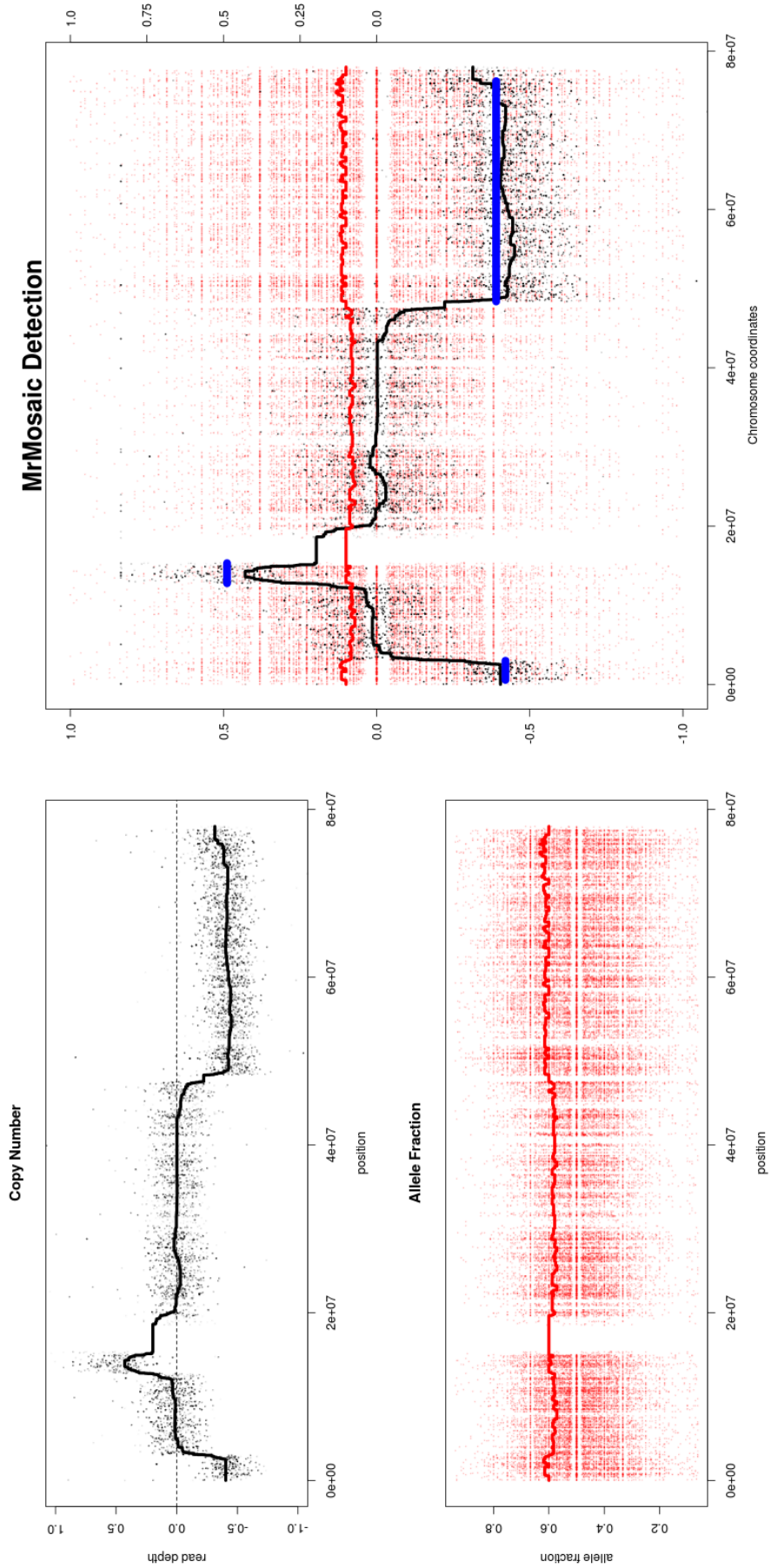
MrMosaic Detection
Chromosome coordinates
Copy Number
read depth
position
Allele Fraction
allele fraction
position

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.