3 MOSAIC STRUCTURAL VARIATION FROM SNP MICROARRAY

3.1 Publication Note

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

3.2 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3 Materials & Methods

3.3.1 Description of studies

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

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

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

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

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

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

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

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

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

Table 3-2 SNP Probe Selection

3.3.2 Mosaic event detection

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

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

3.3.3 Methods of evaluating of clinical significance

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

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

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

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

3.3.4 Exome sequencing

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

3.4 Results

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

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

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

3.4.1 Filtering Strategies for MAD output from DDD & SFHS samples

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

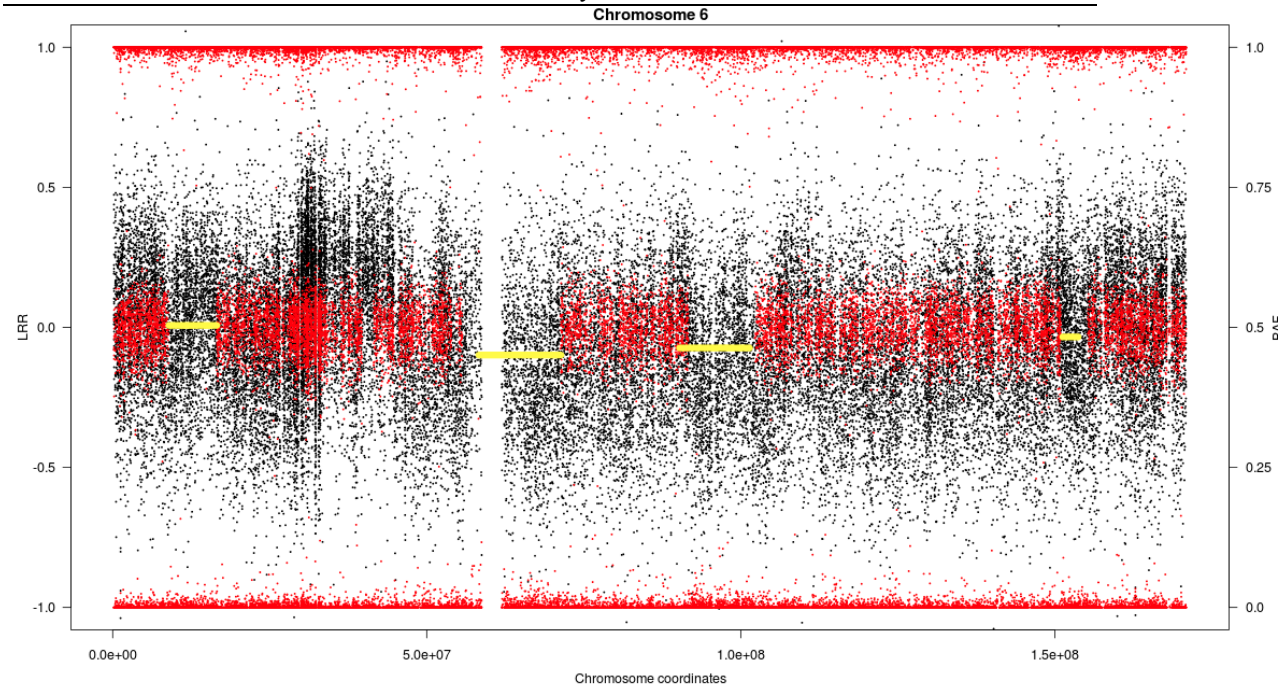

Figure 3-1 Four tracks of constitutive homozgygosity classified (incorrectly) as mosaic.

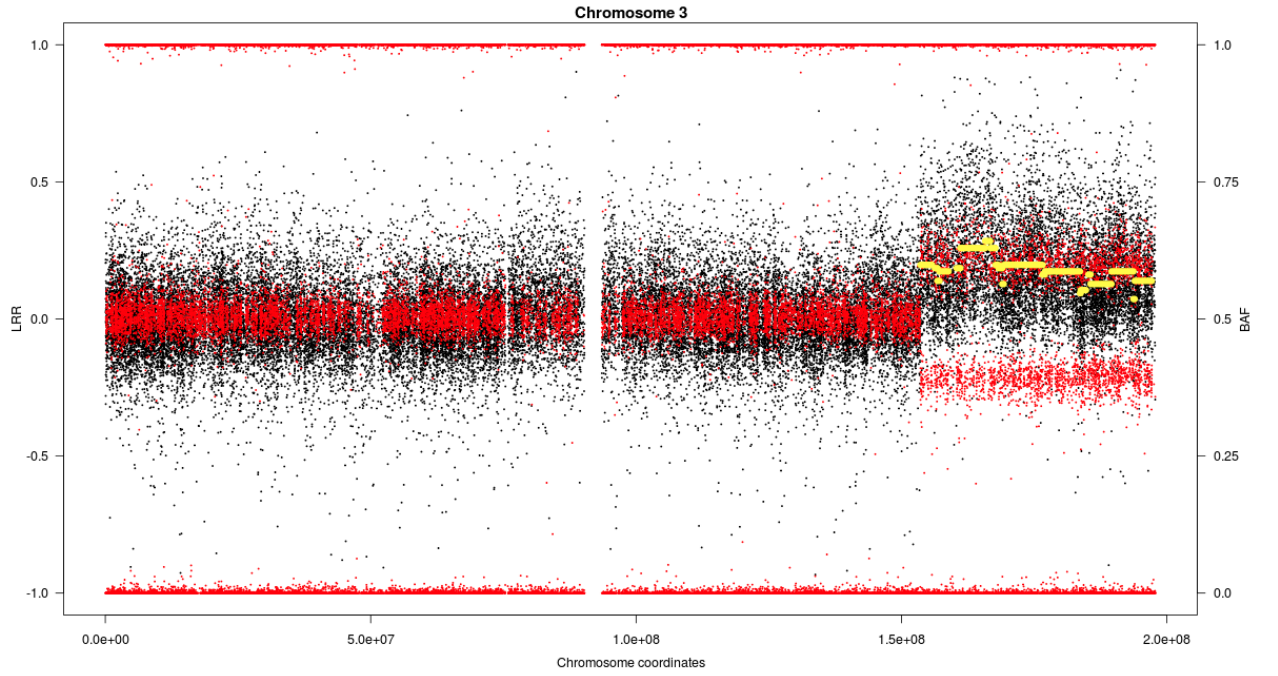

Figure 3-2 An example of over-segmentation. The single mosaic duplication is broken into many smaller duplications.

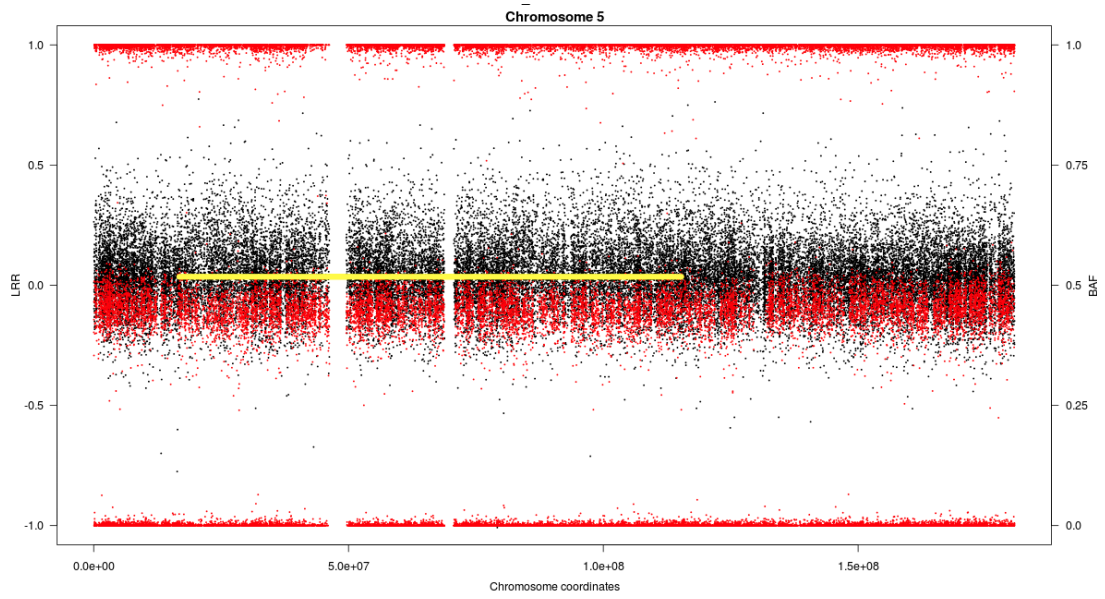

Figure 3-3 An example of unimodal skew, in this case, BAFs systematically depressed slightly below 0.5. This results in an increase in B_{dev}, which then results in a false mosaic detection.

3.4.1.1 Managing over-segmentation

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

3.4.1.2 Managing constitutive homozygosity & unimodal BAF deflection

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

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

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

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

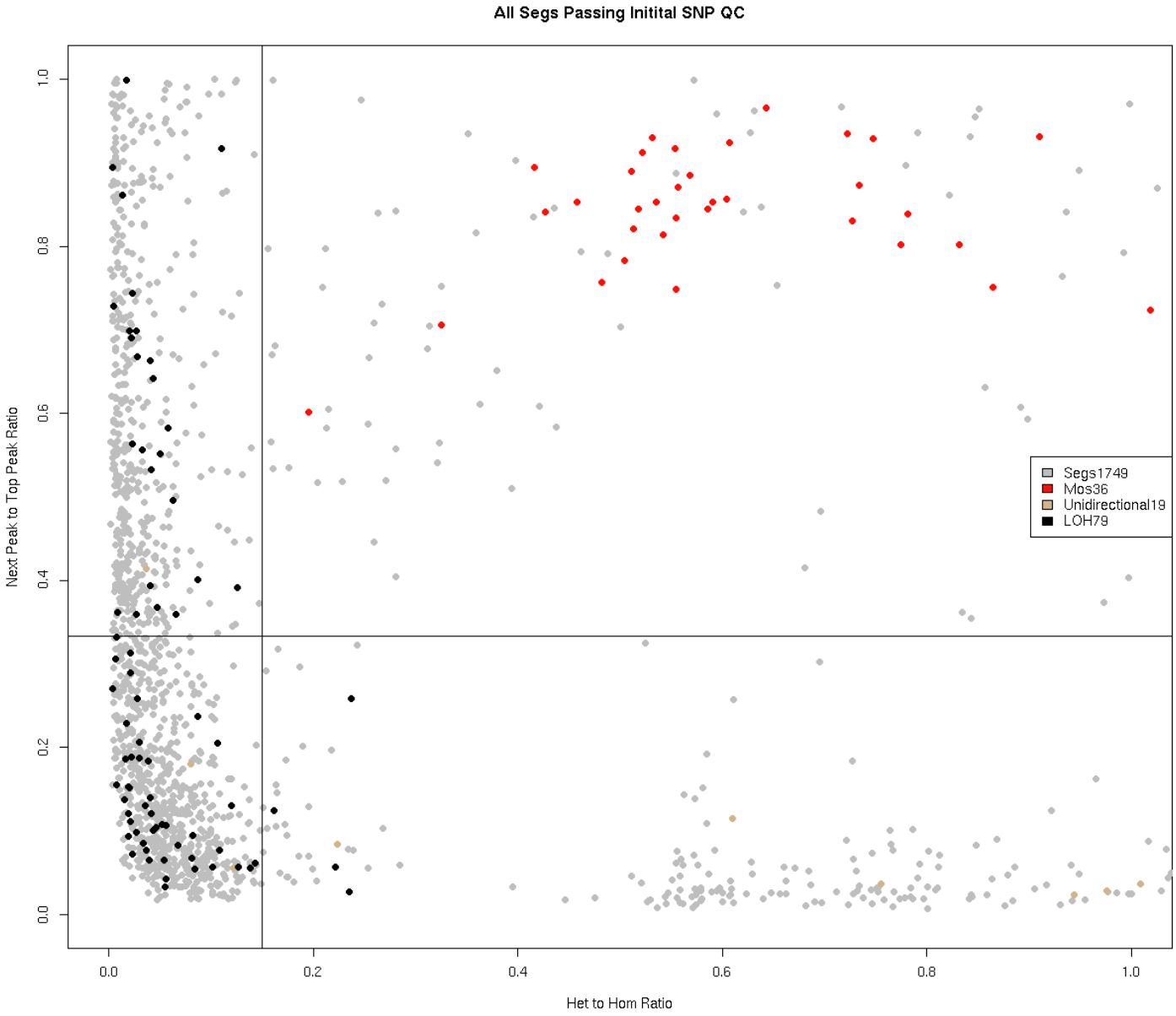

Figure 3-4 Filtering unimodal BAF deflections and constitutive homozygosity using the het:hom ratio and peak:next-peak ratio. grey dots: putative detections, yellow dots: unimodal deflections, black dots: constitutive LOH, red dots: suspected real events

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

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

3.4.1.3 Managing constitutive CNVs

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

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

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

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

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

3.4.1.5 Filtering strategies for TEDS and ALSPAC

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

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

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

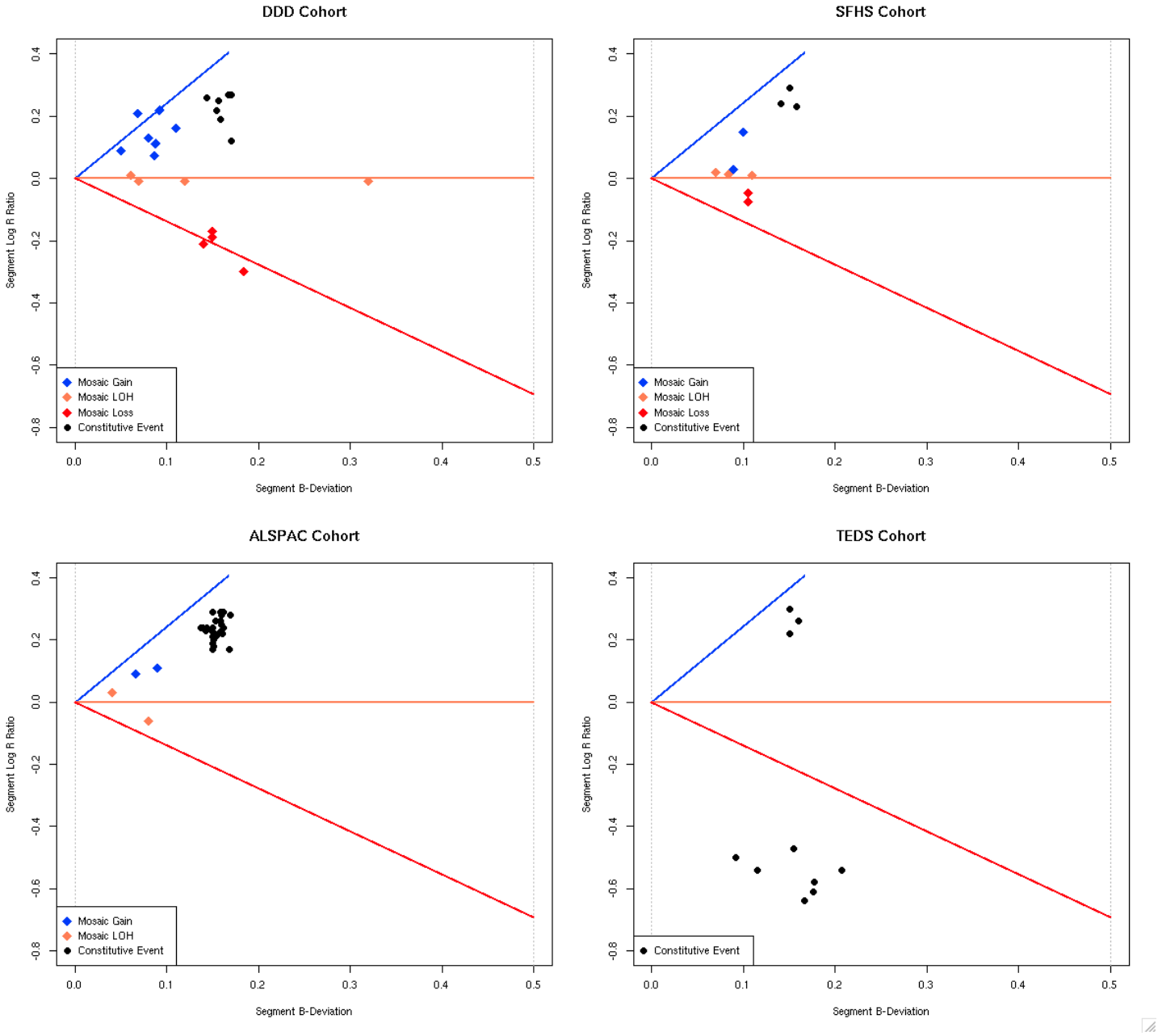

Figure 3-5 Characterisation of mosaic events and constitutive duplications in the DDD, SFHS, ALSPAC and TEDS studies.

3.4.2 Assessing the accuracy of filtering strategies

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

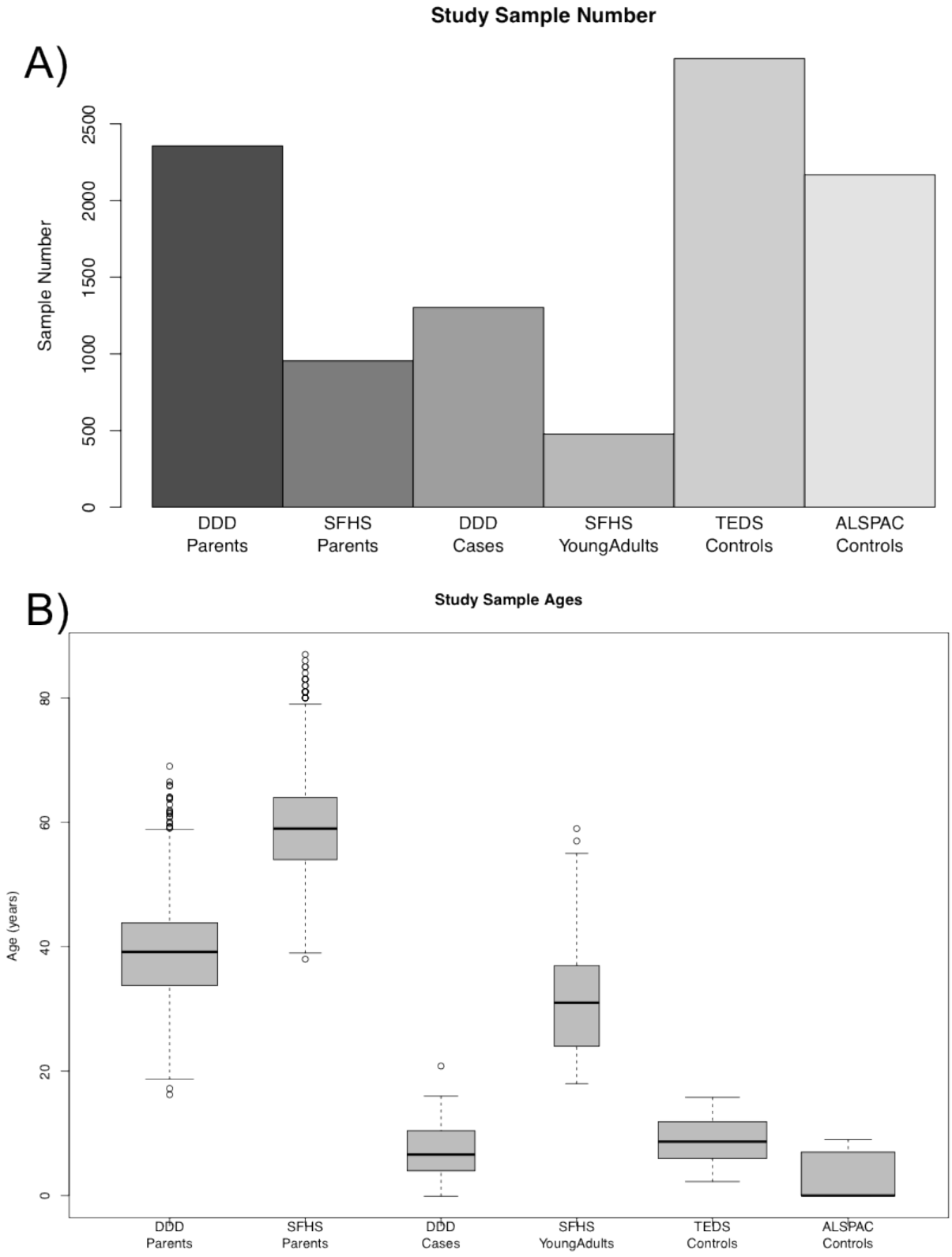

Figure 3-6 The (A) sample number and (B) ages corresponding to the analysed studies.

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

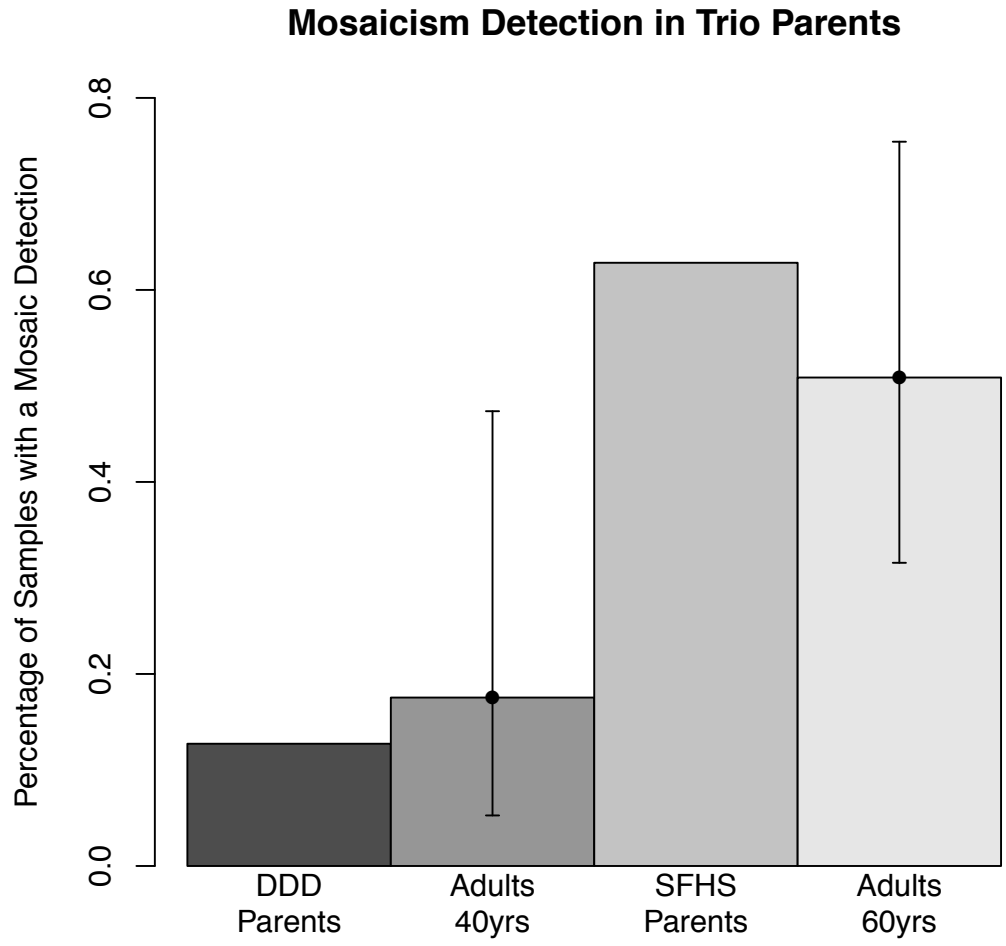

Figure 3-7 The frequency of mosaicism detected in the parents of the trio cohorts was within the confidence intervals of the frequency detected for samples of this age range.

3.4.3 Mosaicism Frequency in Cases & Controls using MAD

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

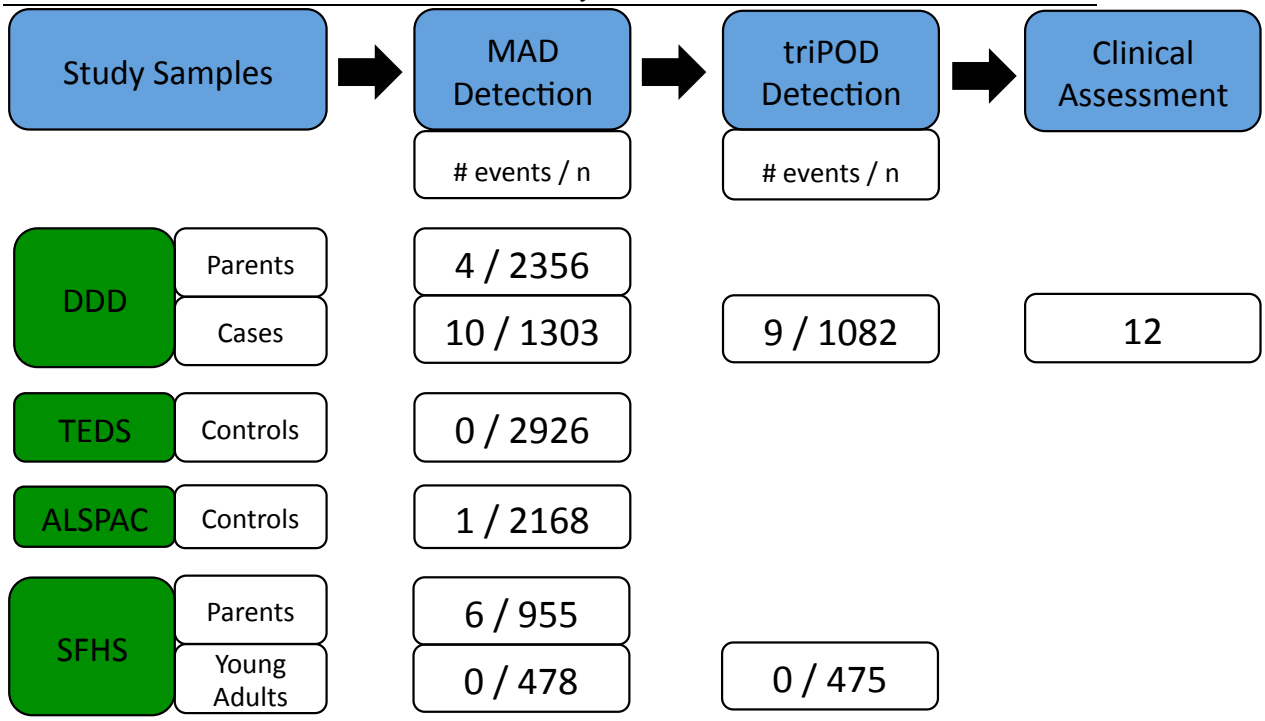

Figure 3-8 Overview. A MAD-based workflow was used to detect mosaicism. This workflow identified an enrichment of mosaicism in cases compared with controls, and triPOD detected two additional mosaic events not detected by MAD. Clinical assessment was performed on all 12 probands of the DDD study with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

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

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

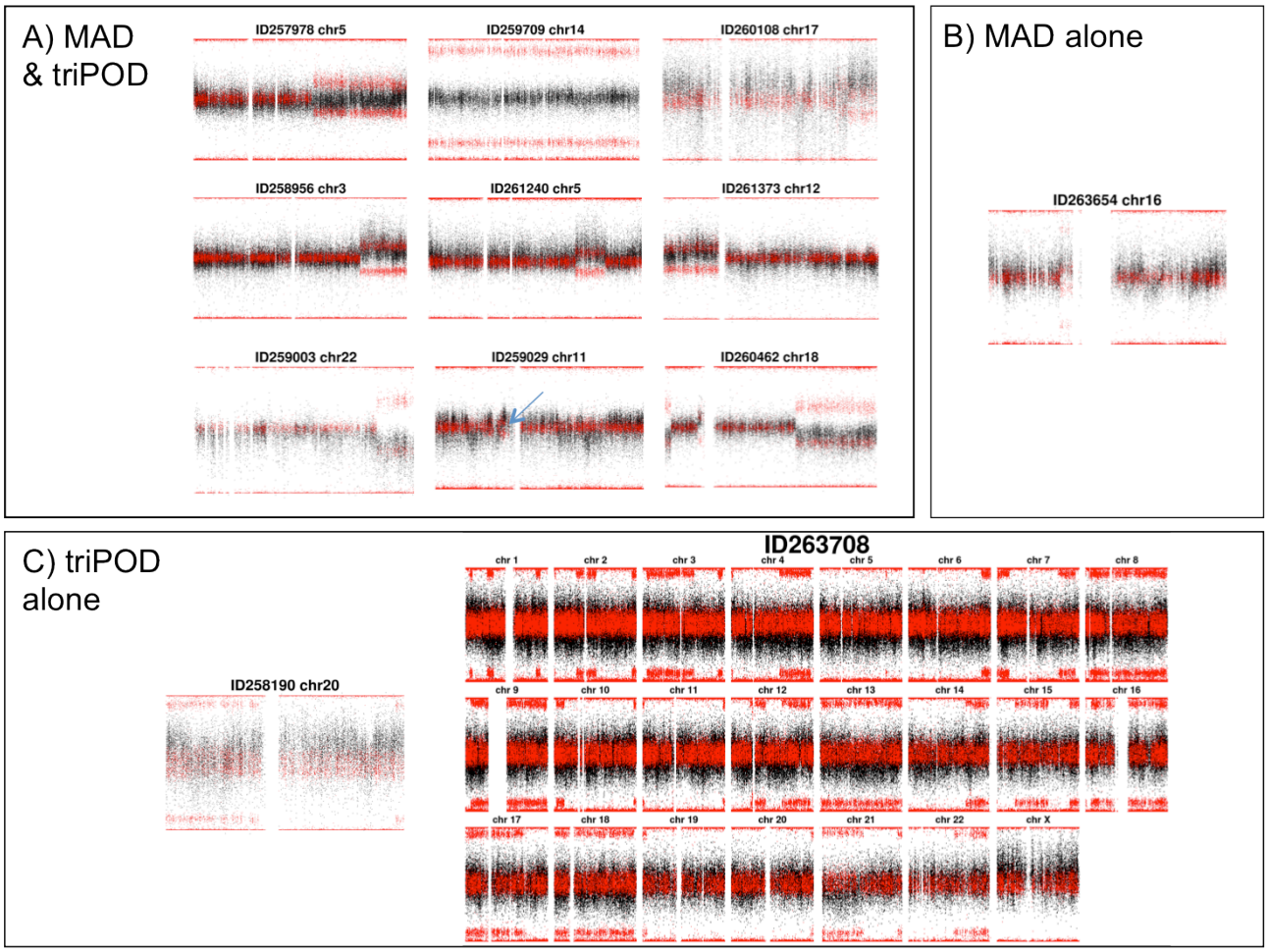

Figure 3-9 All proband detections: The detections made by (A) MAD & triPOD, (B) by MAD alone and (C) by triPOD alone.

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

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

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

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

.

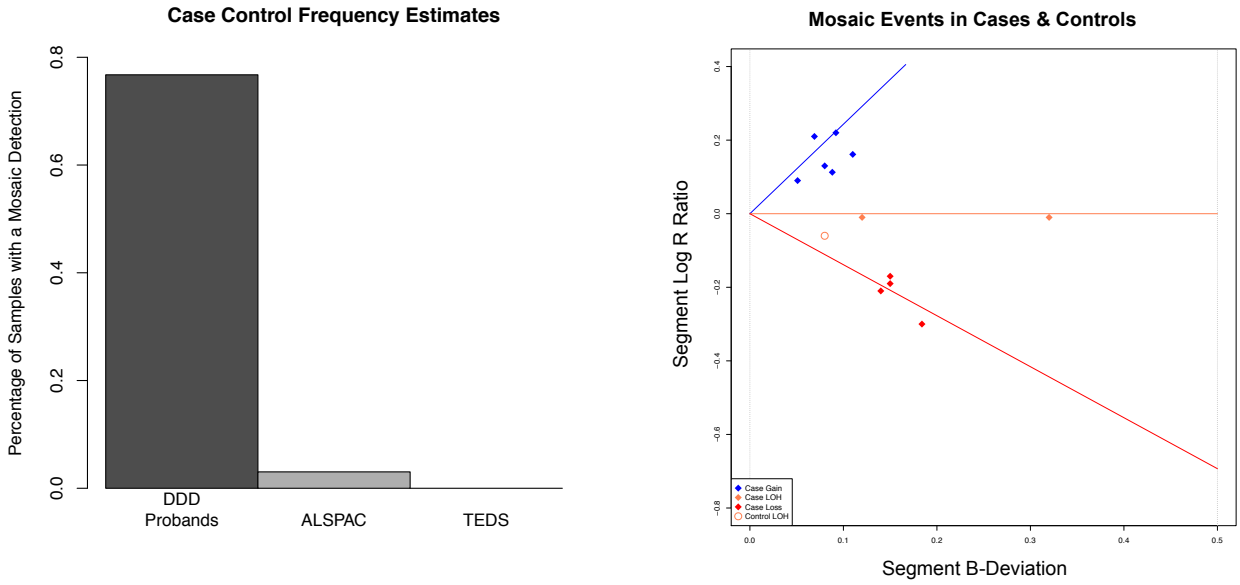

Figure 3-10 (A) The percentage of samples with mosaic events in the case and control cohorts. (B) A depiction of each mosaic event, where the line segments represent the ideal location of mosaicism for gains (blue), LOH (orange) and losses (red).

3.4.4 Additional detections using triPOD

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

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

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

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

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

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

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

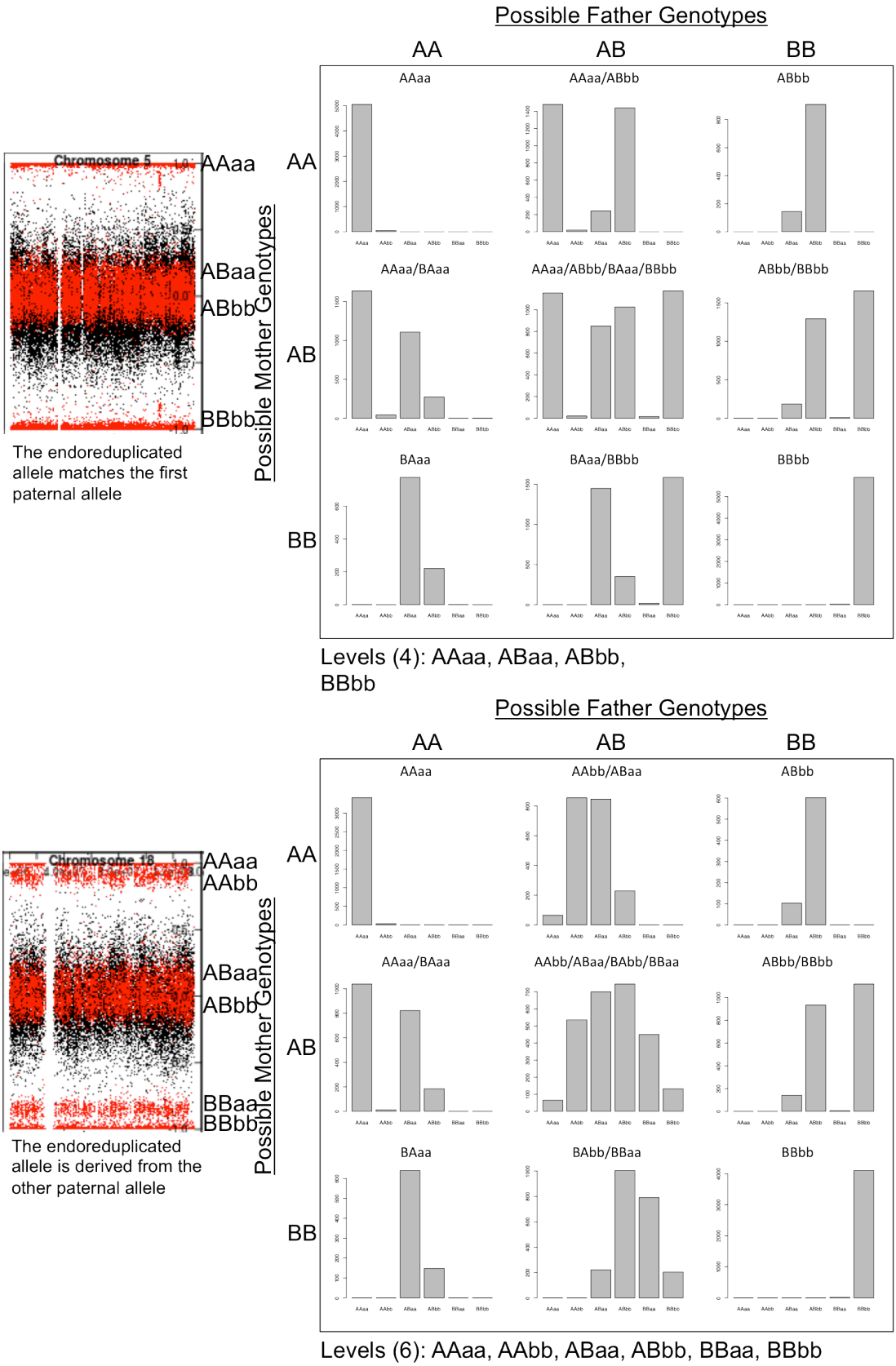

Figure 3-11 Interrogating possible haplotype combinations to determine the alleles present and their origin in the chimeric sample.

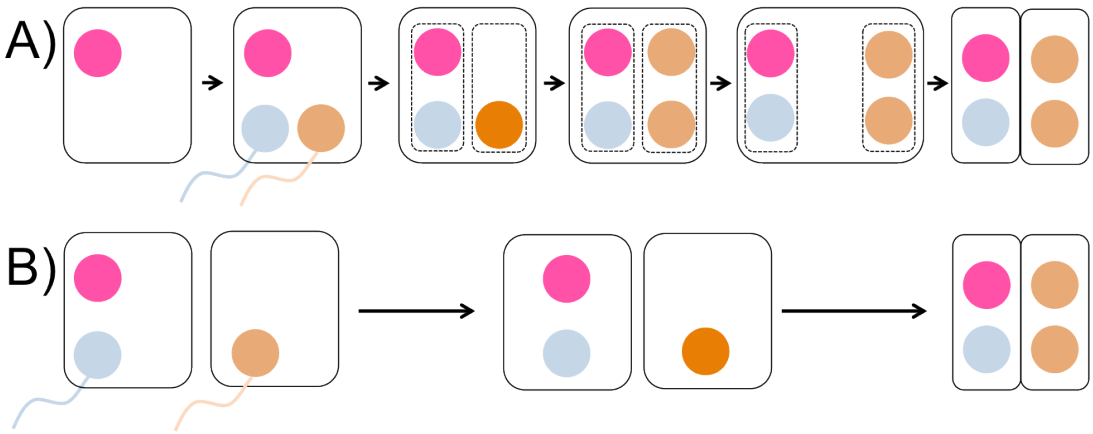

Figure 3-12 An illustration of two possibilities hypothesised by Robinson *et al.*[201] underlying androgenetic /bipaternal mosaicism or chimerism. In A) a one zygte mechanism, an ovum is fertilised by two sperm (dispermy), while in B) a two zygote mechanism, a fertilised zygote fuses with an endoreduplicated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

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

3.4.5 Validation experiments to explore tissue distribution

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

3.4.6 Clinical Interpretation of Probands with Mosaicism

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

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

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

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

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

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

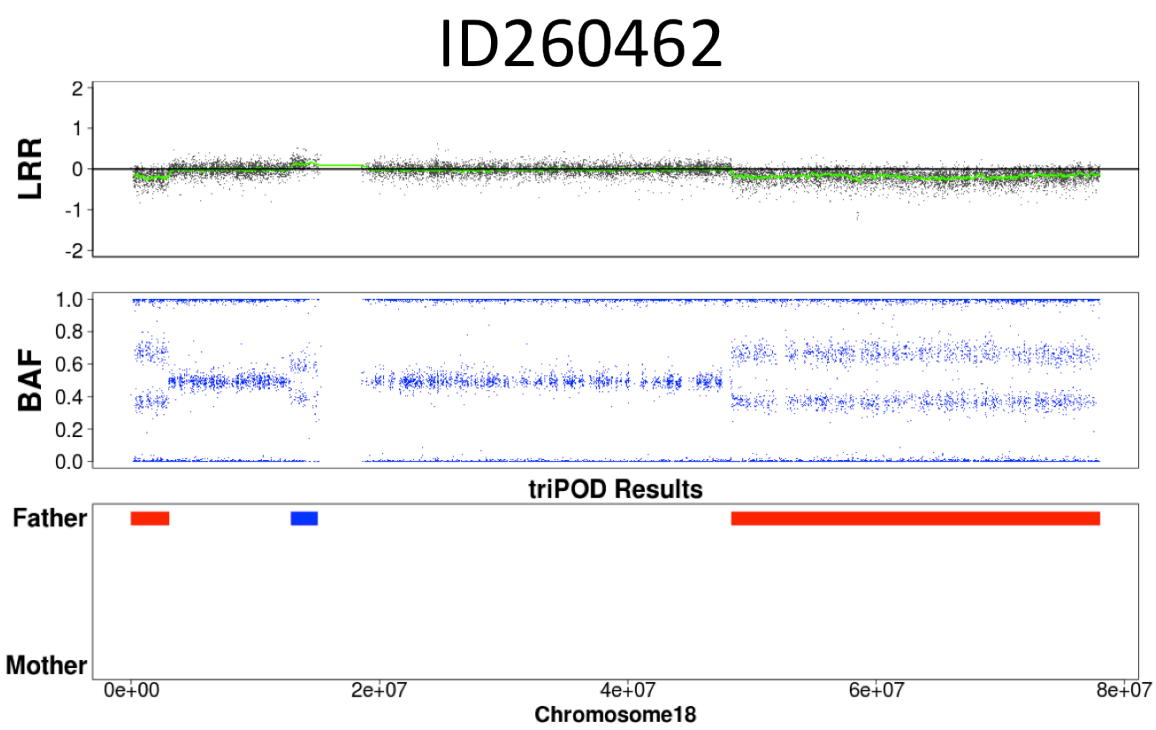

Figure 3-13 triPOD shows that the deletions and duplications arose from different alleles.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

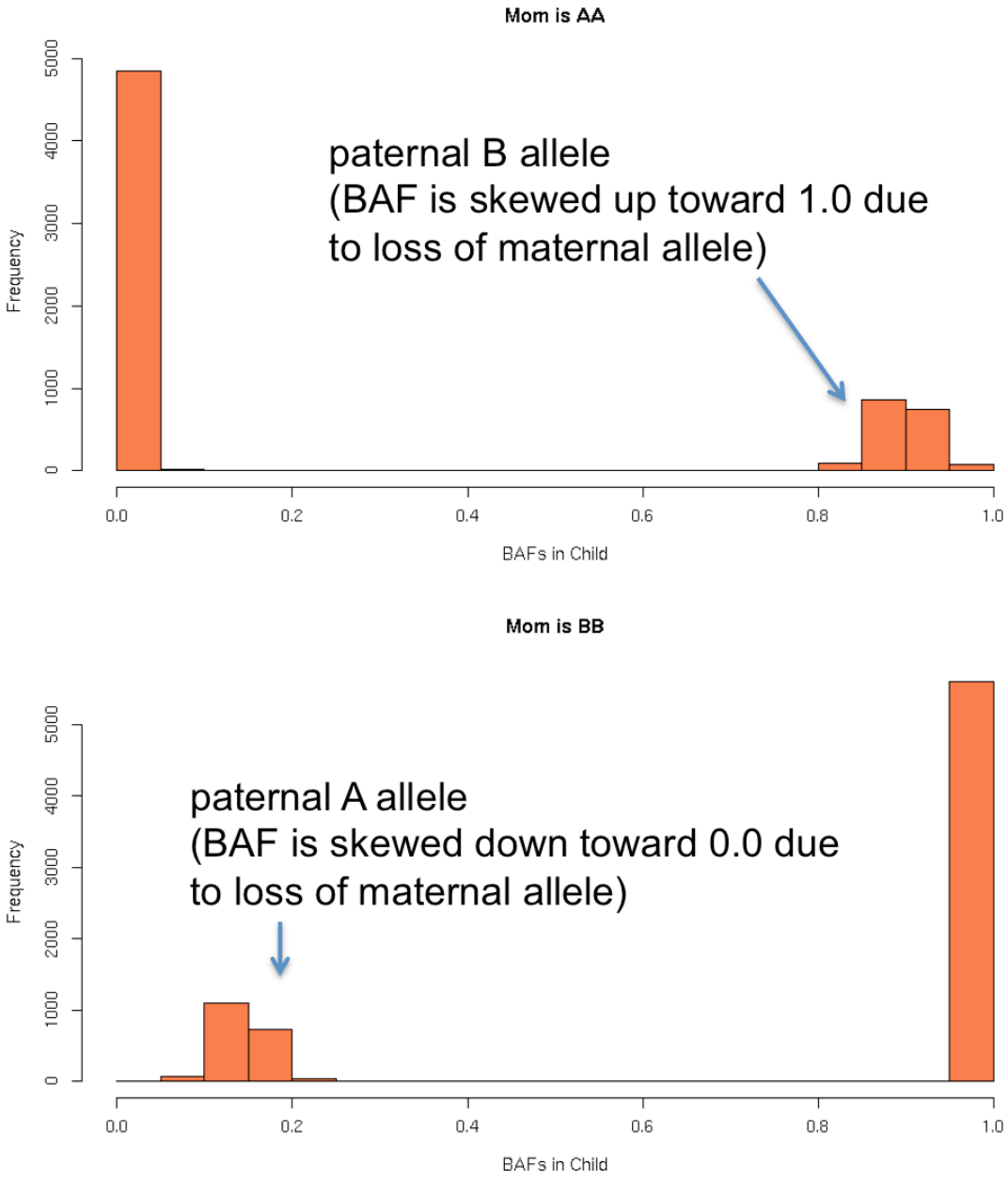

Figure 3-14 aUPD due to loss of maternal allele.

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

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

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

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

3.5 Discussion

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

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

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

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

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

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

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

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

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

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

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