Detection of structural mosaicism from targeted and whole-genome sequencing data

Daniel A King, Alejandro Sifrim, Tomas W. Fitzgerald, et al.

Genome Res. published online August 30, 2017
Access the most recent version at doi:10.1101/gr.212373.116

Detection of structural mosaicism from targeted and whole-genome sequencing data

Daniel A. King[1], Alejandro Sifrim[1], Tomas W. Fitzgerald[1], Raheleh Rahbari[1], Emma Hobson[2], Tessa Homfray[3], Sahar Mansour[3], Sarju G. Mehta[4], Mohammed Shehla[5], Susan E. Tomkins[6], Pradeep C. Vasudevan[7], Matthew E. Hurles[1], The Deciphering Developmental Disorders Study

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
[2] Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, United Kingdom
[3] Southwest Thames Regional Genetics Centre, St George's Healthcare NHS Trust, London, United Kingdom
[4] East Anglian Regional Genetics Service, Addenbrookes Hospital, Cambridge, United Kingdom
[5] South East Thames Regional Genetics Centre, Guys Hospital, London, United Kingdom
[6] Department of Clinical Genetics, St Michael's Hospital, Bristol, United Kingdom
[7] Leicester Royal Infirmary, Leicester, United Kingdom

Daniel A. King: dk6@sanger.ac.uk
Alejandro Sifrim: as33@sanger.ac.uk
Tomas W. Fitzgerald: tomas@ebi.ac.uk
Raheleh Rahbari: rr11@sanger.ac.uk
Emma.hobson: emma.hobson@nhs.net
Tessa Homfray: thomfray@sgul.ac.uk
Sahar Mansour: smansour@sgul.ac.uk
Sarju G Mehta: sarju.mehta@addenbrookes.nhs.uk
Mohammed Shehla: Shehla.Mohammed@gstt.nhs.uk
Susan E. Tomkins: Susan.Tomkins@UHBristol.nhs.uk
Pradeep C. Vasudevan: pradeep.vasudevan@uhl-tr.nhs.uk
Matthew E. Hurles: meh@sanger.ac.uk

ABSTRACT

Structural mosaic abnormalities are large post-zygotic mutations present in a subset of cells and have been implicated in developmental disorders and cancer. Such mutations have been conventionally assessed in clinical diagnostics using cytogenetic or microarray testing. Modern disease studies rely heavily on exome sequencing, yet an adequate method for the detection of structural mosaicism using targeted sequencing data is lacking. Here, we present a method, called MrMosaic, to detect structural mosaic abnormalities using deviations in allele fraction and read coverage from next generation sequencing data. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) simulations were used to calculate detection performance across a range of mosaic event sizes, types, clonalities, and sequencing depths. The tool was applied to 4,911 patients with undiagnosed developmental disorders, and 11 events among 9 patients were detected. For 8 of these 11 events, mosaicism was observed in saliva but not blood, suggesting that assaying blood alone would miss a large fraction, possibly more than 50%, of mosaic diagnostic chromosomal rearrangements.

INTRODUCTION

Genetic mutations that arise post-zygotically lead to genetic heterogeneity in an organism, a phenomenon called mosaicism. The detection of mosaic mutations that are small (single-base or indel) is still a great technical challenge, but can be achieved in specific experimental setups to restrict the number of candidate mutations (e.g. matched tumour-normal samples in cancer to discover somatic mutations) (Jaiswal et al. 2014; Genovese et al. 2014). However, multi-megabase (structural) mosaic rearrangements are now routinely detected using cytogenetics and microarray technology (Miller et al. 2010; Biesecker and Spinner 2013). Recent single nucleotide polymorphism (SNP) microarray-based studies have demonstrated that mosaic structural abnormalities are implicated in developmental disorders (Conlin et al. 2010; King et al. 2015), increase in incidence with age (Forsberg et al. 2012), and predispose to hematological malignancies in adults (Jacobs et al. 2012; Laurie et al. 2012).

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 assess two metrics useful for mosaicism detection: 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 affected 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 genotype-expected BAF (e.g. 0.5 for AB genotype), 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 (Staaf et al. 2008), and Mosaic Alteration Detection (MAD) (González et al. 2011a)) harness this deviation as a mosaic signal. MAD is open source and has been recently used in several large SNP-based mosaicism projects (Forsberg et al. 2012; Jacobs et al. 2012; Forsberg et al. 2014); it identifies mosaic segments using aberrations in 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 C_{dev}. Note that in contrast to loss of heterozygosity, other types of balanced structural variants, notably inversions and balanced translocations, do not typically disrupt BAF or LRR, cannot typically be detected using these methods, and are not addressed in this study.

Developmental disorders (DD) are often caused by rare, small (SNV and indel) mutations, genetic variation which is not easily captured using microarray (King et al. 2014). Therefore, to achieve a 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) (Koboldt et al. 2013). 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 (Meynert et al. 2014). However, WGS is not currently as widely used as WES for rare disease studies due to higher costs, so this work focuses primarily on exome-sequencing data.

In addition to small-scale variation, forms of large-scale 'structural variation', including copy-number (Lee et al. 2007) and copy-neutral variation (uniparental disomy (UPD)) (Yamazawa et al. 2010), are also important causes of DD. CNV burden analysis of nearly 16,000 children with DD (Cooper et al. 2011) 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 deletion events are more often penetrant than duplication events. UPD events are only present in about 1 in 3,500 healthy individuals (Robinson 2000), but are enriched in children with DD (King et al. 2014), and may result in highly penetrant imprinting disorders, recessive diseases, or may be associated with chromosomal mosaicism (Eggermann et al. 2015). Low-clonality mosaicism is difficult to observe in karyotyping, as inspection of at least 20 cells is required to exclude 14% mosaicism with 95% confidence (Hook 1977), and is also difficult to observe in microarray analysis, 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 approximately 10% clonality (González et al. 2011b; Laurie et al. 2012; Jacobs et al. 2012; Machiela et al. 2015). The median average *clonality* in recent SNP-based studies of DD for mosaic aneuploidy was 40% (Conlin et al. 2010), and for mosaic structural variation (2 Mb and greater), was 44% (King et al. 2015). With regard to *frequency* of mosaicism among children investigated with clinical diagnostic testing, the proportion of autosomal mosaic copy-neutral events was 0.24% (12 in 5,000) (Bruno et al. 2011) while the proportion of autosomal mosaic copy-number events was 0.35% (36 in 10,362) (Pham et al. 2014); summing both copy-neutral and copy-number proportions yields a combined frequency estimate among cases of 0.59% of mosaic structural variation.

The detection of large-scale genetic variation from WES data is challenging because input data are derived using sparse sampling of the genome, as targeted regions typically cover only about 2% of the genome (Meynert et al. 2014), and sequence read depth at exons is biased by enrichment efficiency and other factors (Plagnol et al. 2012). Despite these limitations, exome-based software tools have been successfully engineered to detect large-scale *constitutive* mutations, including copy-number variation (Magi et al. 2013; Sathirapongsasuti et al. 2011; Krumm et al. 2012; Backenroth et al. 2014; Fromer et al. 2012) and copy-neutral variation (BCFtools ROH (Narasimhan et al. 2016) and UPDio (King et al. 2014)). These tools are relatively insensitive to *mosaic* abnormalities (post-zygotic abnormalities, i.e. 'mutations'), 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. Specialized methods have been developed for the analysis of cancer exomes where tumor and normal tissue can be isolated (Lonigro et al. 2011; Amarasinghe et al. 2014) or, in the context of a parent-fetus trio, for fetal DNA in maternal plasma (Rampášek et al. 2014). 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 capacity of sequence-based analyses.

We developed MrMosaic, a method that detects structural mosaicism using joint analysis of B_{dev} and C_{dev} in targeted or whole-genome sequencing data (Figure 1). We used simulations to demonstrate the superior performance of MrMosaic compared to the MAD algorithm. We also applied MrMosaic to analyze WES data from 4,911 children with developmental disorders and identified 11 structural mosaic events in 9 individuals, 6 of whom exhibited tissue-specific mosaicism.

RESULTS

We developed a new computational method, MrMosaic, to detect structural mosaic abnormalities (copy number and loss-of-heterozyosity) from high-throughput sequence data (Methods). In summary, this method identifies chromosomal segments with elevated deviations in allelic proportion and copy number, relative to randomly selected sites on other chromosomes from the same data (Figure 1). Initially, measures of deviation of allelic proportion (B_{dev}) and copy number (C_{dev}) are calculated from the WES/WGS data at well-covered (at least 7 reads) known polymorphic SNVs. Whereas B_{dev} is only assessed at heterozygous sites, C_{dev} extracts and integrates read-depth 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 (Pique-Regi et al. 2008). We 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 (Methods).

Simulations

We performed simulations (Methods) to explore the performance of MrMosaic for three different classes of structural mosaicism: gains, losses and LOH, in several contexts. The variation in performance across mosaicism of different *sizes*, *clonalities* and sequencing *coverage* is summarised in Figure 2, for both WES and WGS data.

Across all measured categories, mosaic duplications were more difficult to identify than deletion or LOH events, especially at lower (25%) clonality (Supplemental Fig S1). We suspected that the most likely explanation for this lower sensitivity is that duplications result in the smallest deviation of B_{dev}, compared with deletion and LOH events (Supplemental Fig S2) and that the C_{dev} signal is masked by sampling noise at low clonality. To further explore the effect of including C_{dev} in addition to B_{dev}, we investigated the performance of MrMosaic using B_{dev} alone compared with joint analysis of B_{dev} and C_{dev}. This analysis showed substantially improved detection of copy-number events above lower clonality, while only a marginally decreased performance of LOH detection (Supplemental Fig S3), consistent with the intuition that C_{dev} yields a valuable net signal when clonality is above the C_{dev} noise floor.

Simulation performance increased with larger event *size* (Figure 2A). WES 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. Results for WGS 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. Larger events were assayed by more positions, and whole-genome simulations interrogated nearly 50-fold more sites than exome data (Supplemental Table S1).

Detection performance in simulations increased between 25% and 75% *clonality* (Figure 2B). The WES and WGS 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 (Supplemental Fig S4 and S5). Previous studies of children with DD have reported a median mosaicism of approximately 40% mosaicism and detection performance is strong for detecting mosaicism at this clonality at the studied sizes. As clonality increases, the mosaicism is present in a greater proportion of cells, resulting in a greater signal of detection.

Simulation performance increases with respect to sequencing *coverage* (Figure 2C). The WES and WGS performance with respect to sequencing coverage were assessed for events of 50% clonality, using 5 Mb events for the WES simulations, and 100 kb events for the WGS simulations. WES simulations demonstrated a marginal improvement of detection performance at higher coverage, which was notable for mid-clonality gains (Supplemental Fig S4). Previous work has suggested that 75× average coverage in WES data is sufficient for constitutive copy-number analysis (Fitzgerald et al. 2014) and these coverage simulations demonstrated that this exome coverage is also sufficient for the detection of mosaic structural abnormalities. In the WGS results, AUC rose dramatically between 15× and 20× for LOH and loss events and between 25× and 30× for gains. AUC was above about 0.9 for LOH and loss events at 30× depth, a standard sequencing depth used in WGS disease studies. Nearly all structural mosaic events of 100 kb and 50% clonality were detected (Supplemental Fig S5) and average coverage of 20× was sufficient to detect nearly all 50% clonality deletion and LOH events at 100 kb, while detection performance of gains improved at 30× and 40× (Supplemental Fig S6). This improved performance as coverage increases results primarily from sampling variance ('noise') decreasing (correlation r = -0.95; Supplemental Fig S7), with an additional minor contribution from more sites (more signals) passing the minimal depth threshold for consideration (Supplemental Table S1).

Detections in 4911 exome samples

We generated WES data for 4,911 children with undiagnosed developmental disorders. DNA was collected from either blood (n=1652), saliva (n=3246) or both (n=13), and sequenced to a median average coverage of 90×. 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 3, Table 1, Supplemental Fig S8-S18).

To improve our understanding of the accuracy of this sequencing-based method, we compared the results of the above analysis with the results of a prior experiment (King et al. 2015), which had analysed high-resolution SNP data of 1,303 DDD samples, among which 1,226 (of the 1,303) had both exome and SNP data available. Among these 1,226 for which the exome data could be compared with the gold-standard SNP data, detection using MrMosaic identified 8 events, while detection using SNP microarray data of probands identified 10 events. Of the two events not detected by exome but detected by SNP microarray, one of the missed events was a 4 Mb duplication below 25% clonality. The other missed event was an LOH event with low sequencing depth (33×, one of the lowest of our study - Supplemental Fig S19); low depth results in higher sampling variance and lower statistical significance of deviations in allelic proportion and copy number (Supplemental Fig S7). Given the high clonality (about 75%) of this 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 our trio-based UPD detection pipeline (King et al. 2014)).

Table 1: 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 variant was detected in only one of two assayed tissues, and in all such cases, the variant was detected in saliva but not in blood. Clonality was calculated from B_{dev} using Equation 2 (see Supplemental Table S7) 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 (King et al. 2015)) 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).

Exome Detections									SNP Validation	
DecipherID	chr	type	start (GRCh37)	end (GRCh37)	bdev	l2r	tissue	clonality	clonality saliva	clonality blood
265800	12	gain	988,894	33,535,510	0.201	0.140	saliva	1.34	0.68[@]	absent
261373	12	gain	283,642	33,535,289	0.131	0.262	saliva	0.72	0.45[@]	absent
273553	18	gain	670,541	18,534,702	0.186	0.185	saliva	1.18	0.6[@]	absent
259003	22	loss	42,912,136	50,717,129	0.131	-0.129	blood	0.42	0.54	0.34
274013	10	loss	121,717,932	134,916,366	0.159	-0.324	saliva	0.48	0.44	absent
274600	18	loss	48,458,662	76,870,586	0.190	-0.434	saliva	0.55	0.49	absent
260462	18	loss	662,103	2,740,714	0.171	-0.339	saliva	0.51	0.46	absent
260462*	18	gain	12,702,610	15,323,214	0.118	0.263	saliva	0.41	0.5	absent
260462	18	loss	48,466,843	74,962,645	0.153	-0.345	saliva	0.47	0.45	absent
257978	5	LOH	146,077,526	179,731,635	0.167	-0.002	blood	0.33	0.24	0.26
274396	11	LOH	66,834,252	134,126,612	0.255	-0.0047	saliva	0.51	0.28	0.17

[@]adjusted tetrasomy clonality.

***located in peri-centromeric region and detected during *post hoc* analysis.**

Validation of the 11 mosaic abnormalities using SNP microarrays on DNA derived from both blood and saliva successfully detected all abnormalities in at least one tissue (Table 1). 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 (Supplemental Table S2). Examination of the raw aCGH data in one case (Supplemental Fig S17) showed that only small fragments of one of the events were detected but these called segments were individually much smaller than the actual event.

Detection of the mosaic events was largely dependent on the assayed tissue, suggesting the importance of tissue-specificity (present in only a subset of tissues) in mosaicism detection. Out of the 11 mosaic events, 3 were detected in blood and in saliva samples while the remaining eight were only observed in saliva (Table 1, Supplemental Fig S8-S18). There were 2 abnormalities detected from 1,652 blood samples and 9 detected from 3,246 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 (Supplemental Fig S16 and S19). This gradient of increasing clonality along the chromosome is compatible with LOH-mediated mosaic reversion, characterised by distinct cell populations carrying partially overlapping independent LOH events, as reported recently (Choate et al. 2015). Nevertheless, despite generation and analysis of high-depth (~400×) WES data for this sample, and the identification of several strong candidate genes, including *CEP57* (the cause of mosaic aneuploidy syndrome (Snape et al. 2011)) in the reversion-localised region, no plausibly pathogenic rare (below 1% minor allele frequency) coding sequence variants were identified (Supplemental Table S4). It may be that the gene of interest is several megabases distal to the breakpoint region.

We assessed the pathogenicity of the events detected in these nine children based on their phenotypes and known genomic disorders whose phenotypes matched those found in these children. Note that of the nine children presented here, four (Decipher IDs: 261373, 259003, 260462, and 257978) had been discovered and examined for pathogenicity during an earlier study (see Table 2, Supplemental Note S1, and King 2015 et al). The mosaic events identified in seven of nine children were considered definitely pathogenic on the basis of being multi-megabase CNVs that overlap known genomic-disorder regions (Supplemental Note S1). 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 (Supplemental Table S4).

Table 2: Phenotypes for children with identified structural mosaicism.

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

Empirical evaluation of detection of mosaicism from WGS data

One sample, with three mosaic abnormalities detected on a single chromosome, which had also been detected during an earlier analysis (King *et. al.* 2015), provided a valuable opportunity to use whole-genome sequencing data to clarify rearrangement architecture and to demonstrate MrMosaic performance on whole-genome sequence data. After the whole-genome sequencing data were generated and analyzed, MrMosaic easily detected these multi-megabase mosaic events, found with 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. Analysis of the WGS read pair data using BreakDancer (Chen et al. 2009) 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 (Guilherme et al. 2011) and inverted duplications (Knijnenburg et al. 2007). Additionally, all three mosaic components arose from a single parental origin (paternal, in this case) (King et al. 2015) which would be expected in a ring chromosome. We suspected that the underlying abnormality in this child is a ring chromosome, although we were unable to access the cellular material required to generate the cytogenetic data to prove this hypothesis (Supplemental Fig S21).

DISCUSSION

Structural mosaic abnormalities are multi-megabase, post-zygotic mutations that have previously been associated with developmental disorders (Conlin et al. 2010; King et al. 2015). This work introduces a novel method to detect these mutations from next generation sequencing data.

In an extensive simulation study we show adequate discriminative ability 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. We also compare our method to the popular array-based mosaic detection method, MAD, and show 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 duplications involving chromosome-arm events, such as 12p and 18p mosaic tetrasomy as MrMosaic easily detected events of this size. Given the dimensionality of the simulation parameter space (i.e. clonality, event size, coverage) and the computational cost of running these simulations we restricted our analysis to parameter values in line with previous observations of structural mosaicism and reasonable experimental parameters at current sequencing costs. Additionally, we also chose more extreme parameter values for size and clonality in order to illustrate the dynamic range of the method in high-depth whole genome sequence data (e.g. performance at <100kb resolution and low clonality events), even though few previous pathogenic variants with these charactristics have previously been described. These simulated performances only serve as illustrations for the selected parameter set and are not readily generalizable to other combinations of parameters, given the non-linear interaction between parameters. Overall, simulation results show that MrMosaic is able to detect variants similar to previously described pathogenic variants with good performance.

We used MrMosaic to uncover pathogenic structural mosaicism in a large exome study of children with undiagnosed developmental disorders. Applying our method to the exome data of 4,911 enrolled children, we identified nine individuals with structural mosaicism; the majority of these mutations were considered pathogenic. Assessment of pathogenicity was largely based on identifying substantial overlap between the known syndromic manifestations of large, well-known syndromic disorders, and the predominant phenotypes seen in each child. In one child with LOH-mosaicism, no pathogenic mutations were identified in the mosaic LOH region suggesting that a pathogenic allele may lie outside of this mosaic region.. In this WES-based analysis we recovered 8 of the 10 abnormalities previously detected in a subset of 1,226 samples previously analysed with SNP genotyping chip data suggesting that exome-analysis alone is sensitive to detecting large-scale mosaicism. 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.

The overall frequency of mosaicism detected in this study, 0.18%, is lower and significantly different ($p < 10^{-4}$, binomial test) from the 0.59% structural mosaicism frequency estimated

from previous studies. One likely explanation for the discrepancy in these frequencies is ascertainment bias, as some classes of structural mosaicism (e.g. mosaic trisomies) are likely to have been diagnosed by prior diagnostic testing (e.g. karyotype or microarray) and not enrolled into the DDD study. Another component of this discordance may be due to decreased sensitivity, as mosaicism smaller than 2 Mb is challenging to detect by exome and these small events account for ~25% (9/36) of mosaic copy number events described previously (Pham et al. 2014). Given the low number of mosaic events in our cohort, due to the low mosaicism rate and tissue specificity, and the lack of publicly available large-scale developmental disorder datasets, this study only provides a limited estimate of the real-life performance of MrMosaic on non-simulated datasets. As developmental disorder studies increase in sample size and scope we envision that screening for mosaicism will provide additional explanatory power, increasing the number of diagnosed cases.

In one sample we observed a gradient of mosaicism, a phenomenon likely associated with mosaic reversion of a *de novo* mutation dominantly inducing genome instability. Analysis of the mosaic LOH region with high-depth exome data did not identify a strong candidate coding variant and a further WGS-based search for candidate pathogenic *de novo* mutations 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.

As expected, 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. 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 while 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 prove 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 (Woods et al. 1994), 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 WES or WGS, which are likely to become more common in the future.

The majority of the mosaic events we observed in saliva-derived DNA were not observed in blood. 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 we detected in saliva. DNA derived from saliva has a mixed origin, mainly lymphocytes (derived from mesoderm) and epithelium (derived from epiderm) (Endler et al. 1999); 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 (Conlin et al. 2012), and tissue-limited mosaicism has been observed in mosaic tetrasomies of chromosomes 5p, 8p, 9p and 18p (Choo et al. 2002) ; 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 structural mosaicism, resulting in missed diagnoses. Studying the saliva tissue in these children permitted the identification of their mosaic abnormalities and ended for them and their families, their quest for diagnosis.

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.

METHODS

MrMosaic

Implementing mosaic detection requires generating an input file and executing the algorithm; the latter consists of several steps: statistical testing, segmentation, filtering, and results visualisation. 'BAF' is used below as an alias 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 were di-allelic single-nucleotide polymorphic (1%-99% MAFs among European individuals in the UK10K project (Walter et al. 2015)) 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 is extracted using fast_pileup function in the perl module Bio::DB::Sam (Stajich et al. 2002), 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 total of reference bases and non-reference bases. Heterozygous sites were defined as loci with a BAF between 0.06 and 0.94, inclusive. The B_{dev} is calculated at heterozygous sites as the absolute difference between the BAF and 0.5. Only loci with sufficient read coverage (at least 7 reads) are used for analysis.

C_{dev} values were generated using the following method: read depths from each target region was collected, 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 (Fitzgerald et al. 2014)). Lastly, using the Aberration Detection Algorithm v2 (ADM2) method by Agilent® a final error-weighted value, is produced, which we use as the C_{dev} value.

The statistical testing step of the MrMosaic algorithm begins by data smoothing, using a rolling median (width of 5) across heterozygote and homozygous sites, so as to utilize the depth information in homozygous sites to reduce variance. From this point forward, only heterozygote sites are 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 we compute 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 is greater (has a higher median rank) than the distribution of the background. We use 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 we apply a distance-weighted resampling scheme, to down-weight distant points from the chosen site. The tri-cube distance, inspired by Loess smoothing, was chosen as a decay function for the resampling weights and considers data

points up to 0.5 Mb upstream and downstream of the given position. An equal number of data points is then sampled around the chosen site and from the background (n=100) and the Mann-Whitney U test is performed. Finally, we combine 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 operates on the combined p value generated above. Segmentation is performed using the GADA algorithm (González et al. 2011a), 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 generates contiguous segments of putative chromosomal abnormalities. Segments in close proximity (within 1Mb) that show the same signal direction (loss, gain, LOH) are merged to reduce over-segmentation.

The filtering step is required to assess which of the segments generated above are likely reflective of true mosaicism. While testing MrMosaic in exome simulation analyses we observed that true-positive detections (those overlapping simulated events) tended to be larger (greater number of probes) and have stronger evidence of deviation (GADA amplification value) than putative segments that did not overlap simulated regions (i.e. false-positive, spurious calls) (Supplemental Fig S22-S24). We captured these two features in a scoring metric calculated from the cumulative empirical 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. We used the Mscore to filter those events least likely to represent false positives. We selected events with an Mscore of 8 or greater for analysis because we observed that this appeared to provide a good balance between sensitivity and specificity (Supplemental Fig S24).

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 (Supplemental Table Table S6). The detection plots are png files 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.

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.

Simulating Mosaicism

We 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. We compared performance to a modified version of MAD we 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 WES and WGS simulations and are described first below. The remaining steps 4-10 were executed consistently for WES and WGS simulations and are described next.

For WES simulations, loci selection (1) was based on di-allelic single nucleotide polymorphic positions (between 1% and 99% UK10K (Walter et al. 2015) European minor allele frequency) in the V3 version of the target-region design. 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 >= 10, baseQ >=10, properly mapped read-pairs), where parental exomes had a mean average sequencing output of 67× (calculated where × 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 ×) ∈ {50, 75, 100}, event clonality m ∈ {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 WGS simulations, the loci selection (1) was based on di-allelci single nucleotide polymorphic (1% - 99% European MAFs from the 1000 genomes project (Abecasis et al. 2012) May-2013 release) autosomal positions. To calculate depth at these loci (2), we 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 1000 genomes project (Abecasis et al. 2012). 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 25×, event clonality m ∈ {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 ×) ∈ {30, 40, 50, 60} for the three mosaic types {loss, gain, LOH}. One hundred trials (4) were conducted per WGS simulation.

The remaining simulations steps 4-10 described below were performed consistently for WES and WGS 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 (Supplemental Table S6). To calculate the final simulated read depth (5) for each site i (SDP_i), we sampled from a Poisson distribution with λ_i equal to the scaled read depth. Only positions with a final read depth (6) of at least 7 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 assignment of genotypes (8) (AA, AB, or BB) at each position *i* was randomly 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), we adjusted the expected heterozygote proportion of 0.5 with respect to the chosen event type and clonality, and sampling 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 WES and WGS 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.

Description of Samples & Sequencing

The samples used in this analysis derived from the Deciphering Developmental Disorders study, a proband two-parent trio-based investigation of children with undiagnosed developmental disorders from the UK and Ireland (King et al. 2015; Firth and Wright 2011; Wright et al. 2014; Fitzgerald et al. 2014). DNA was extracted from blood and saliva and was processed at the Wellcome Trust Sanger Institute by array CGH and exome sequencing. 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.

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 using the Illumina® HiSeq 2500 platform with a target of at least 50× mean coverage using paired-end sequence reads of 75-bp read-length. Measured exome coverage ranged from 14× to 155× with a mean of 69× (Supplemental Fig S24). Alignment to the reference genome GRCh37-hs37d was performed by BWA version 0.5.9 (Li and Durbin 2009) and saved in BAM-format files (Li et al. 2009).

Additionally, two exome samples were processed *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 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-hs37d was performed by BWA version 0.5.9 (Li and Durbin 2009) and saved in BAM-format files (Li et al. 2009). Re-alignment to GRCh38 was not done as this method avoids mitochondrial regions, and harnesses exonic regions, whose mapping is unlikely to be affected by alternate scaffolds. Average coverage was calculated 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 = 24.8×. Rearrangement analysis was carried out using BreakDancer v1.0 (Chen et al. 2009).

Additional filtering implemented in addition to Mscore quality score

Some events with very high Mscores appeared to represent real, but constitutive, abnormalities. There were two failure modes we identified: constitutive duplications and homozygosity by descent (HBD). Constitutive duplications genuinely produce strong signals in MrMosaic, but also constitutive deletion and ROH events may produce putative detections if individual probes had mapping artefacts that resulted in spurious signals. We used BCFtools ROH to identify and filter HBD regions and flagged as suspicious events with greater than 25% reciprocal overlap with CNVs detected through constitutive copy-number detection. In addition, we 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 we filtered such systematic errors by filtering putative mosaic events seen in more than 2.5% of samples. Remaining putative detections were each manually reviewed. Tissue specificity was assessed through manual review of each detection in both saliva and blood (Supplemental Note S1).

SNP genotyping chip validation

Illumina® HumanOmniExpress-24 Beadchips (713,014 markers) were used to ensure that both saliva and blood tissue were analysed using SNP microarray. To complete dual-tissue SNP microarray for the validation experiment, SNP microarray chips were run on blood samples for IDs 261373, 273553, 259003, 260462, and 257978. Illumina GenomeStudio software was used to generate log R ratio and BAF metrics and Illumina® Gencall software was used to calculate genotypes. Structural mosaic detection was performed using MAD (González et

al. 2011a). 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.

DATA ACCESS

The complete raw exome sequencing data has been submitted to the European Genome-phenome Archive (EGA; https://www. ebi.ac.uk/ega/) under accession number EGAS00001000775, and is available following Data Access Committee (DAC) approval.

MrMosaic is primarily written in the R language, available as an open source tool at Github (https://github.com/asifrim/mrmosaic). MrMosaic source code is available in the Supplemental Material.

ACKNOWLEDGMENTS

This study relied on the generous participation of DDD patients and their parents. We are grateful to the DDD informatics and HGI pipeline staff for generating the data, DDD laboratory staff for sample handling, and Sanger genotyping core for running the validation arrays. Yanick Crow, Helen Firth, David Fitzpatrick, and Wendy Jones provided invaluable clinical expertise. We thank Jeff Barrett for his sharp, constructive feedback. Rolph Pfundt and James Lupski aided the interpretation of the revertant mosaic mutation. The DDD is supported by the Health Innovation Challenge Fund (HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (WT098051).

AUTHOR CONTRIBUTIONS

DK, AS and MH developed the algorithm and wrote the manuscript; TF generated the ADM2 exome scores; YC, EH, TH, SM, SM, MS, ST, PV recruited and phenotyped the DDD patients with detected mosaicism.

DISCLOSURE DECLARATION

MEH is a co-founder, shareholder and consultant to Congenica Ltd, a company providing diagnostic decision support software.

REFERENCES

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65. http://dx.doi.org/10.1038/nature11632 (Accessed December 13, 2013).

Amarasinghe KC, Li J, Hunter SM, Ryland GL, Cowin PA, Campbell IG, Halgamuge SK. 2014. Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15**: 732. http://www.biomedcentral.com/1471-2164/15/732?utm_content=buffer65663&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4162913&tool=pmcentrez&rendertype=abstract.

Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y. 2014. CANOES: Detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res* **42**.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14**: 307–20. http://www.ncbi.nlm.nih.gov/pubmed/23594909.

Bruno DL, White SM, Ganesamoorthy D, Burgess T, Butler K, Corrie S, Francis D, Hills L, Prabhakara K, Ngo C, et al. 2011. Pathogenic aberrations revealed exclusively by single nucleotide polymorphism (SNP) genotyping data in 5000 samples tested by molecular karyotyping. *J Med Genet* **48**: 831–839.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677–81. http://dx.doi.org/10.1038/nmeth.1363.

Choate KA, Lu Y, Zhou J, Elias PM, Zaidi S, Paller AS, Farhi A, Nelson-Williams C, Crumrine D, Milstone LM, et al. 2015. Frequent somatic reversion of KRT1 mutations in ichthyosis with confetti. *J Clin Invest* **125**: 1703–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4396494&tool=pmcentrez&rendertype=abstract (Accessed April 29, 2016).

Choo S, Teo SH, Tan M, Yong MH, Ho LY. 2002. Tissue-limited mosaicism in Pallister-Killian syndrome -- a case in point. *J Perinatol* **22**: 420–3. http://www.ncbi.nlm.nih.gov/pubmed/12082482.

Conlin LK, Kaur M, Izumi K, Campbell L, Wilkens A, Clark D, Deardorff MA, Zackai EH, Pallister P, Hakonarson H, et al. 2012. Utility of SNP arrays in detecting, quantifying, and determining meiotic origin of tetrasomy 12p in blood from individuals with Pallister-Killian syndrome. *Am J Med Genet Part A* **158 A**: 3046–3053.

Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB. 2010. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum Mol Genet* **19**: 1263–1275.

Cooper GM, Coe BP, Girirajan S, Rosenfeld J a, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846. http://dx.doi.org/10.1038/ng.909.

Eggermann T, Soellner L, Buiting K, Kotzot D. 2015. Mosaicism and uniparental disomy in prenatal diagnosis. *Trends Mol Med* **21**: 77–87.

Endler G, Greinix H, Winkler K, Mitterbauer G, Mannhalter C. 1999. Genetic fingerprinting in mouthwashes of patients after allogeneic bone marrow transplantation. *Bone Marrow Transplant* **24**: 95–98.

Firth H V, Wright CF. 2011. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* **53**: 702–3. http://www.ncbi.nlm.nih.gov/pubmed/21679367 (Accessed April 29, 2016).

Fitzgerald TW, Gerety SS, Jones WD, van Kogelenberg M, King DA, McRae J, Morley KI, Parthiban V, Al-Turki S, Ambridge K, et al. 2014. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**: 223–228. http://www.ncbi.nlm.nih.gov/pubmed/25533962 (Accessed December 24, 2014).

Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, Diaz de Ståhl T, Zaghlool A, Giedraitis V, et al. 2014. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet* **46**: 624–8. http://www.ncbi.nlm.nih.gov/pubmed/24777449 (Accessed March 22, 2016).

Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D, et al. 2012. Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* **90**: 217–228.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, et al. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91**: 597–607.

Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, Chambert K, Mick E, Neale BM, Fromer M, et al. 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**: 2477–87. http://www.nejm.org/doi/abs/10.1056/NEJMoa1409405 (Accessed February 8, 2017).

González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, Rothman N, Chanock SJ, Armengol L, Pérez-Jurado LA. 2011a. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics* **12**: 166. http://www.biomedcentral.com/1471-2105/12/166 (Accessed January 24, 2014).

González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, Rothman N, Chanock SJ, Armengol L, Pérez-Jurado L a. 2011b. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics* **12**: 166. http://www.biomedcentral.com/1471-2105/12/166.

Guilherme RS, Meloni VFA, Kim CA, Pellegrino R, Takeno SS, Spinner NB, Conlin LK, Christofolini DM, Kulikowski LD, Melaragno MI. 2011. Mechanisms of ring chromosome formation, ring instability and clinical consequences. *BMC Med Genet* **12**: 171. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3309960&tool=pmcentrez&rendertype=abstract.

Hook EB. 1977. Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use. *Am J Hum Genet* **29**: 94–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1685228&tool=pmcentrez&rendertype=abstract.

Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner M-J, et al. 2012. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* **44**: 651–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3372921&tool=pmcentrez&rendertype=abstract.

Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman P V., Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A, et al. 2014. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N Engl J Med* **371**: 2488–2498. http://www.nejm.org/doi/abs/10.1056/NEJMoa1408617 (Accessed November 21, 2016).

King DA, Fitzgerald TW, Miller R, Canham N, Clayton-Smith J, Johnson D, Mansour S, Stewart F, Vasudevan P, Hurles ME. 2014. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res*

24: 673–687.

King DA, Jones WD, Crow YJ, Dominiczak AF, Foster NA, Gaunt TR, Harris J, Hellens SW, Homfray T, Innes J, et al. 2015. Mosaic structural variation in children with developmental disorders. *Hum Mol Genet* **24**: 2733–2745.

Knijnenburg J, van Haeringen A, Hansson KBM, Lankester A, Smit MJM, Belfroid RDM, Bakker E, Rosenberg C, Tanke HJ, Szuhai K. 2007. Ring chromosome formation as a novel escape mechanism in patients with inverted duplication and terminal deletion. *Eur J Hum Genet* **15**: 548–55. http://www.ncbi.nlm.nih.gov/pubmed/17342151.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. XThe next-generation sequencing revolution and its impact on genomics. *Cell* **155**.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**: 1525–1532.

Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, et al. 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* **44**: 642–50. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3366033&tool=pmcentrez&rendertype=abstract.

Lee C, Iafrate AJ, Brothman AR. 2007. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* **39**: S48-54. http://www.ncbi.nlm.nih.gov/pubmed/17597782 (Accessed April 29, 2016).

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract (Accessed December 11, 2013).

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract (Accessed April 20, 2011).

Lonigro RJ, Grasso CS, Robinson DR, Jing X, Wu Y-M, Cao X, Quist MJ, Tomlins SA, Pienta KJ, Chinnaiyan AM. 2011. Detection of Somatic Copy Number Alterations in Cancer Using Targeted Exome Capture Sequencing. *Neoplasia* **13**: 1019–IN21. http://www.neoplasia.com/article/S1476558611800886/fulltext.

Machiela MJ, Zhou W, Sampson JN, Dean MC, Jacobs KB, Black A, Brinton LA, Chang I-S, Chen C, Chen C, et al. 2015. Characterization of Large Structural Genetic Mosaicism in Human Autosomes. *Am J Hum Genet* **96**: 487–497. http://linkinghub.elsevier.com/retrieve/pii/S0002929715000191 (Accessed November 21, 2016).

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, et al. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* **14**: R120. http://genomebiology.com/2013/14/10/R120%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053953%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053953&tool=pmcentrez&rendertype=abstract.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**: 247. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4122774&tool=pmcentrez&render

type=abstract (Accessed April 23, 2016).

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, et al. 2010. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am J Hum Genet* **86**: 749–764.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*. http://www.ncbi.nlm.nih.gov/pubmed/26826718 (Accessed February 1, 2016).

Pham J, Shaw C, Pursley A, Hixson P, Sampath S, Roney E, Gambin T, Kang S-HL, Bi W, Lalani S, et al. 2014. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10 362 consecutive cases. *Eur J Hum Genet* **22**: 969–78. http://www.ncbi.nlm.nih.gov/pubmed/24398791.

Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. 2008. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**: 309–18. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2704547&tool=pmcentrez&rendertype=abstract (Accessed February 17, 2014).

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, et al. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**: 2747–2754.

Rampášek L, Arbabi A, Brudno M. 2014. Probabilistic method for detecting copy number variation in a fetal genome using maternal plasma sequencing. *Bioinformatics* **30**: i212–i218. http://bioinformatics.oxfordjournals.org/content/30/12/i212.short?rss=1.

Robinson WP. 2000. Mechanisms leading to uniparental disomy and their clinical consequences. *Bioessays* **22**: 452–9. http://www.ncbi.nlm.nih.gov/pubmed/10797485.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**: 2648–2654.

Snape K, Hanks S, Ruark E, Barros-Núñez P, Elliott A, Murray A, Lane AH, Shannon N, Callier P, Chitayat D, et al. 2011. Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. *Nat Genet* **43**: 527–529. http://dx.doi.org/10.1038/ng.822.

Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, Juliusson G, Rosenquist R, Höglund M, Borg A, Ringnér M. 2008. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol* **9**: R136. http://genomebiology.com/2008/9/9/R136.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**: 1611–8. http://genome.cshlp.org/content/12/10/1611 (Accessed March 18, 2016).

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JRB, Xu C, Futema M, Lawson D, et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82–90. http://www.nature.com/doifinder/10.1038/nature14962 (Accessed June 16, 2016).

Woods CG, Bankier A, Curry J, Sheffield LJ, Slaney SF, Smith K, Voullaire L, Wellesley D. 1994. Asymmetry and skin pigmentary anomalies in chromosome mosaicism. *J Med Genet* **31**: 694–701. http://www.ncbi.nlm.nih.gov/pubmed/7815438 (Accessed June 16, 2016).

Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T, et al. 2014. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**: 1305–14.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4392068&tool=pmcentrez&rendertype=abstract (Accessed December 17, 2014).

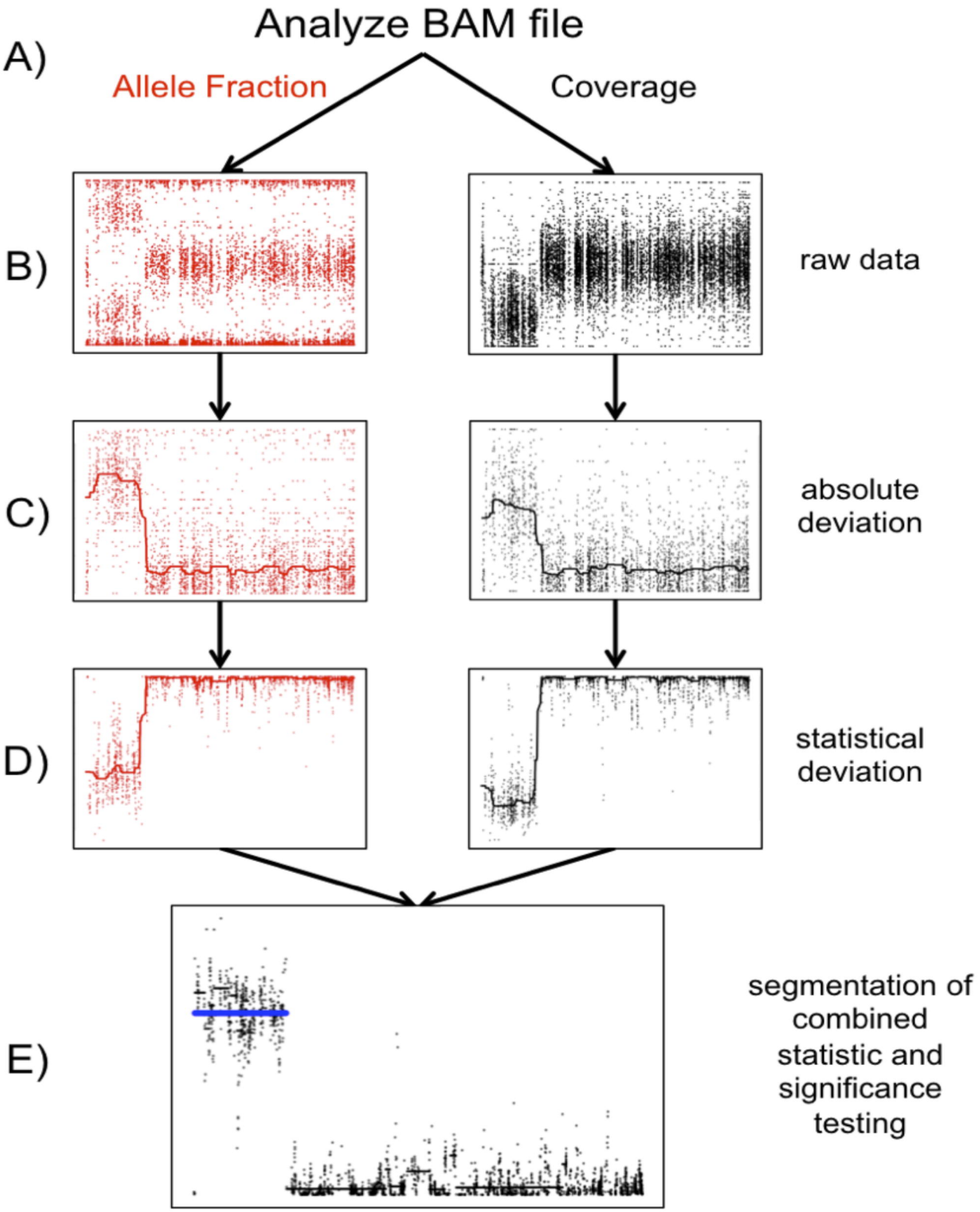
Yamazawa K, Ogata T, Ferguson-Smith AC. 2010. Uniparental disomy and human disease: an overview. *Am J Med Genet C Semin Med Genet* **154C**: 329–34. http://www.ncbi.nlm.nih.gov/pubmed/20803655 (Accessed April 29, 2016).

FIGURE LEGENDS

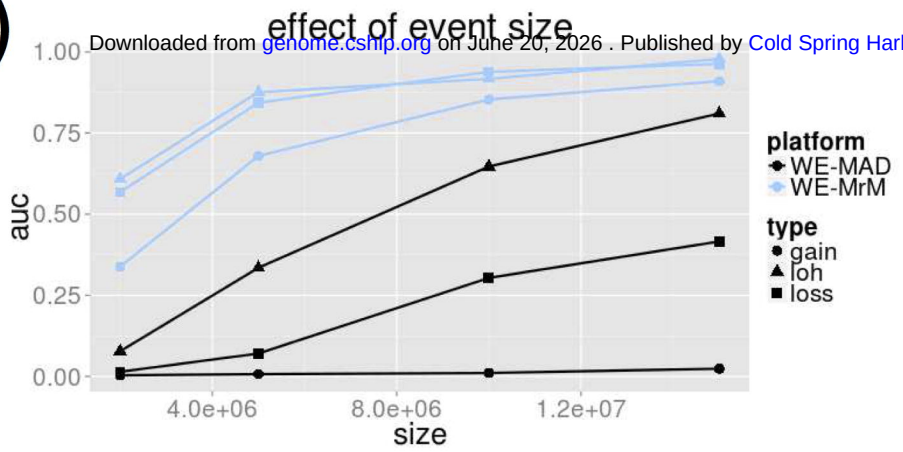
Figure 1: Detecting structural mosaicism using MrMosaic: A) Exome data are stored in a BAM file from which allele fraction (left column) and coverage (right column) are measured at polymorphic positions within or near target regions. A simulated mosaic deletion is depicted. B) The raw data, consisting of BAFs (Y axis: B allele frequency) and normalized coverage (Y axis: log ratio of normalized coverage) are plotted across chromosome space (X axis) for a simulated mosaic deletion. C) Absolute deviation of BAF (Y axis: B_{dev}) and normalized coverage (Y axis: C_{dev}) at heterozygous sites are analyzed. A smoothed median has been included. 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 randomly selected sites from other chromosomes, generating a test statistic (Y axis). A smoothed median has been included. E) The test statistics are depicted in log scale. 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.

Figure 2: Simulation performance summarised by AUC: We measured the average precision (area under the precision recall curve) for MrMosaic implemented on whole-exome (WE) simulations (panels A,C,E), and MrMosaic & MAD implemented on whole-genome (WG) simulations (panels B,D,F). The depth, size, and coverage measured for WES and WGS simulations were selected to accentuate informative differences in performance. AUC across size: Simulated events of 50% clonality were studied for WES (A) and WGS (B) simulations. Whereas for WES simulations, simulated exome depth was 75× depth, for WGS simulations it was 30× depth. MrMosaic on whole-genome data (WG-MrM) outperforms MrMosaic on exome data (WE-MrM), which outperforms MAD on exome data (WE-MAD). AUC across clonality: Whereas for WES (C) simulations the simulated size and coverage was 5 Mb & 75×, for WGS (D) simulations it was 100 kb & 30×. AUC across average coverage: Simulated events of 50% were studied for both WES (E) and WGS (F) simulations. Whereas for WES simulations, simulated event size was 5 Mb, for WGS simulations it was 100 kb.

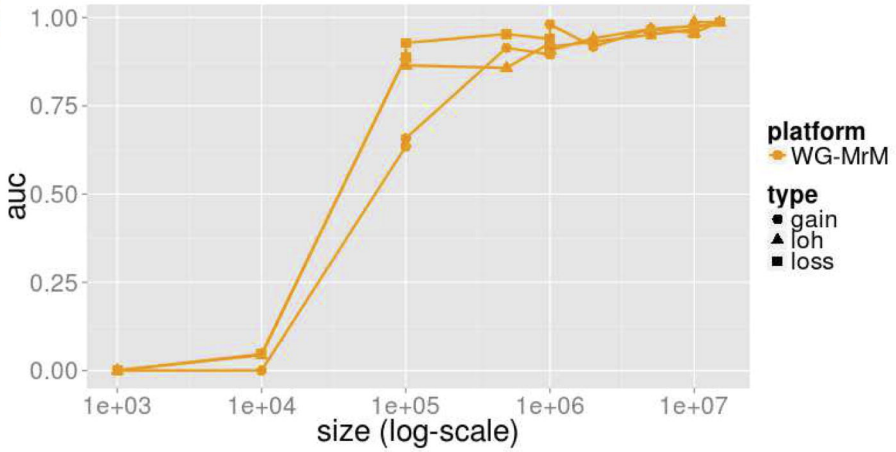
Figure 3: Structural mosaicism detected from exome data: Structural Mosaicism Detected by MrMosaic in the DDD (Deciphering Developmental Disorders) study. Black and red dots represent copy-number and allele fraction, respectively. C_{dev} and B_{dev} are plotted in black and red trend lines. The blue line represents statistically significant segmented detections passing a threshold. Different classes of events are found: A-C) Mosaic gains, D-F) mosaic losses, G) mixed copy-number, and H-I) loss-of-heterozygosity events



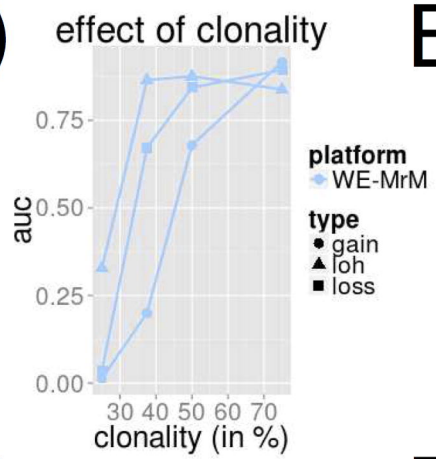
A)



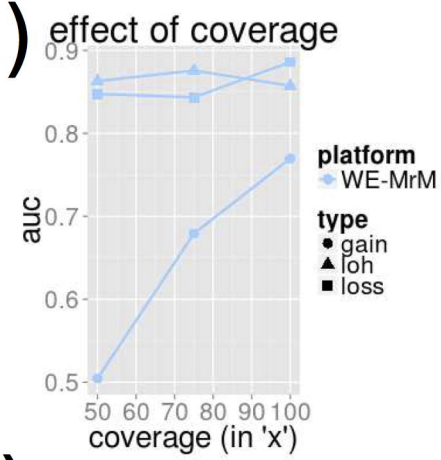
B)



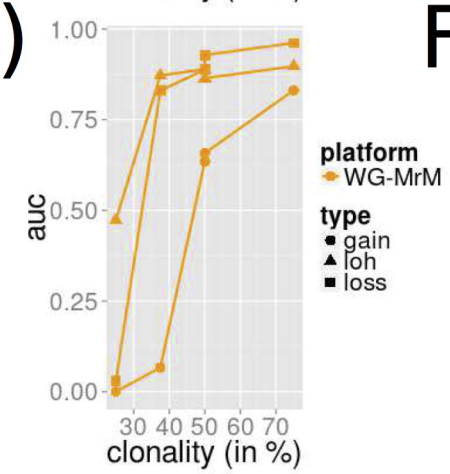
C)



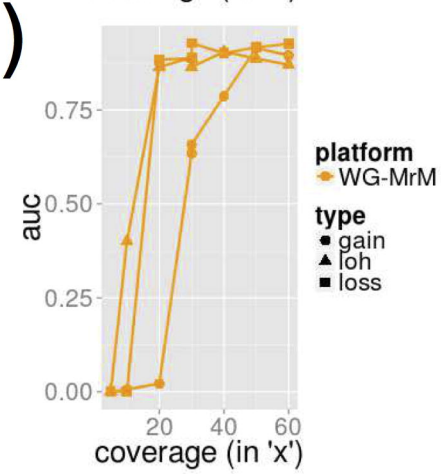
E)



D)



F)



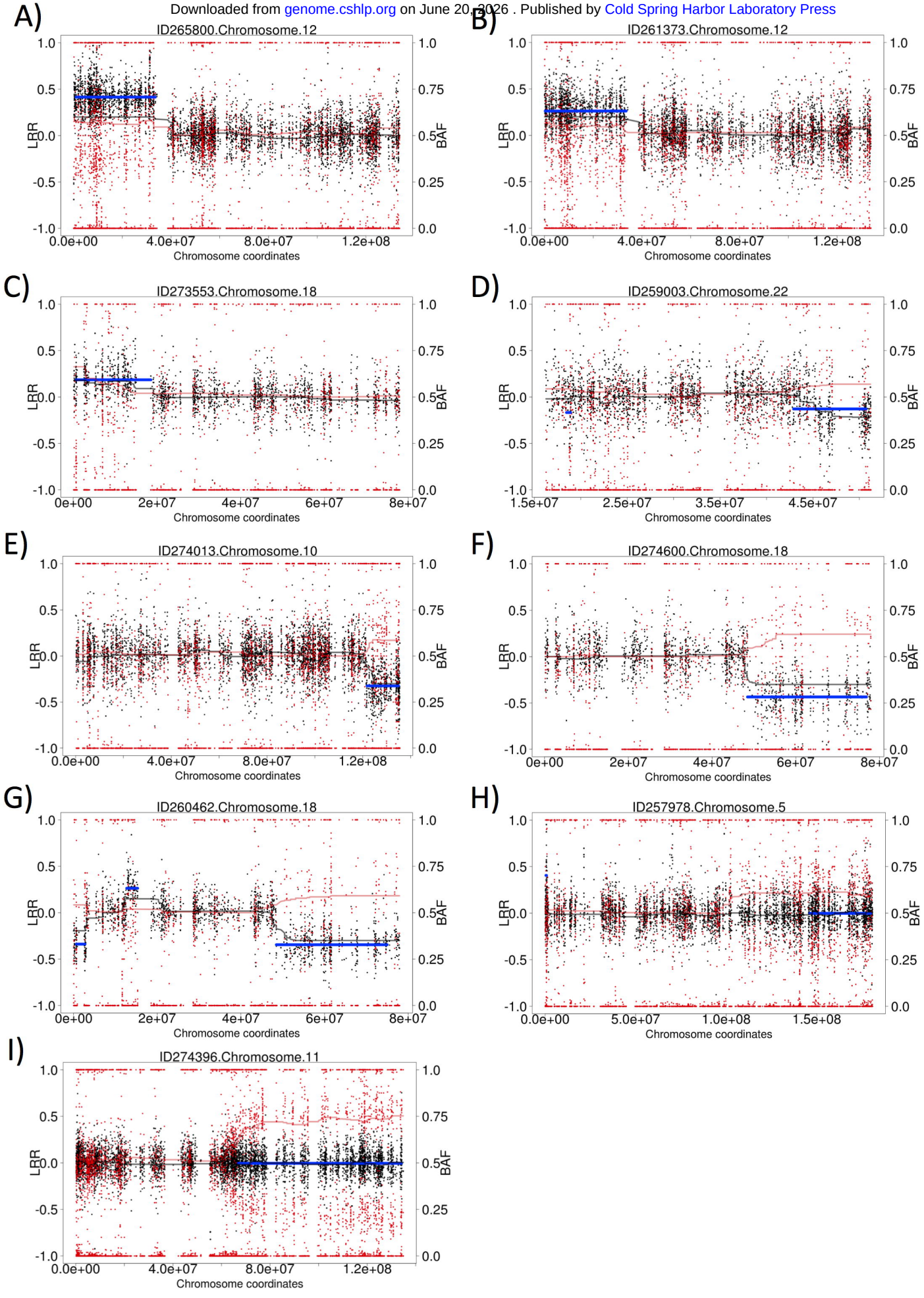


FIGURE LEGENDS

Figure 1: 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. A simulated mosaic deletion is depicted. B) The raw data, consisting of BAFs (Y axis: B allele frequency) and normalized coverage (Y axis: log ratio of normalized coverage) are plotted across chromosome space (X axis) for a simulated mosaic deletion. C) Absolute deviation of BAF (Y axis: B_{dev}) and normalized coverage (Y axis: C_{dev}) at heterozygous sites are analyzed. A smoothed median has been included. 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 randomly selected sites from other chromosomes, generating a test statistic (Y axis). A smoothed median has been included. E) The test statistics are depicted in log scale. 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.

Figure 2: Simulation performance summarised by AUC: We measured the average precision (area under the precision recall curve) for MrMosaic implemented on whole-exome (WE) simulations (panels A,C,E), and MrMosaic & MAD implemented on whole-genome (WG) simulations (panels B,D,F). The depth, size, and coverage measured for WES and WGS simulations were selected to accentuate informative differences in performance. AUC across size: Simulated events of 50% clonality were studied for WES (A) and WGS (B) simulations. Whereas for WES simulations, simulated exome depth was 75x depth, for WGS simulations it was 30x depth. MrMosaic on whole-genome data (WG-MrM) outperforms MrMosaic on exome data (WE-MrM), which outperforms MAD on exome data (WE-MAD). AUC across clonality: Whereas for WES (C) simulations the simulated size and coverage was 5 Mb & 75x, for WGS (D) simulations it was 100 kb & 30x. AUC across average coverage: Simulated events of 50% were studied for both WES (E) and WGS (F) simulations. Whereas for WES simulations, simulated event size was 5 Mb, for WGS simulations it was 100 kb.

Figure 3: Structural mosaicism detected from exome data: Structural Mosaicism Detected by MrMosaic in the DDD (Deciphering Developmental Disorders) study. 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. Different classes of events are found: A-C) Mosaic gains, D-F) mosaic losses, G) mixed copy-number, and H-I) loss-of-heterozygosity events