Detecting, quantifying, and discriminating the mechanism of mosaic chromosomal aneuploidies using MAD-seq

Yu Kong,[1,4] Esther R. Berko,[1,4,5] Anthony Marcketta,[1,6] Shahina B. Maqbool,[1] Claudia A. Simões-Pires,[1,7] David F. Kronn,[2] Kenny Q. Ye,[3] Masako Suzuki,[1] Adam Auton,[1,8] and John M. Greally[1]

[1]Department of Genetics and Center for Epigenomics, Albert Einstein College of Medicine, New York, New York 10461, USA; [2]Department of Pediatrics, New York Medical College, Valhalla, New York 10595, USA; [3]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, New York 10461, USA

Current approaches to detect and characterize mosaic chromosomal aneuploidy are limited by sensitivity, efficiency, cost, or the need to culture cells. We describe the mosaic aneuploidy detection by massively parallel sequencing (MAD-seq) capture assay and the *MADSEQ* analytical approach that allow low (<10%) levels of mosaicism for chromosomal aneuploidy or regional loss of heterozygosity to be detected, assigned to a meiotic or mitotic origin, and quantified as a proportion of the cells in the sample. We show results from a multi-ethnic MAD-seq (meMAD-seq) capture design that works equally well in populations of diverse racial and ethnic origins and how the *MADSEQ* analytical approach can be applied to exome or whole-genome sequencing data, revealing previously unrecognized aneuploidy or copy number neutral loss of heterozygosity in samples studied by the 1000 Genomes Project, cell lines from public repositories, and one of the Illumina Platinum Genomes samples. We have made the meMAD-seq capture design and *MADSEQ* analytical software open for unrestricted use, with the goal that they can be applied in clinical samples to allow new insights into the unrecognized prevalence of mosaic chromosomal aneuploidy in humans and its phenotypic associations.

[Supplemental material is available for this article.]

Somatic mosaicism occurs when a single population of cells acquires a subpopulation with a different genotype. While mosaicism is an unavoidable consequence of the low level of background mutation, making every multicellular organism a mosaic to some degree, the pathogenic consequences of mosaicism are most apparent when (1) it occurs early in development and therefore affects a substantial proportion of cells forming one or more organs, and (2) the genotypic alteration is a detrimental mutation.

Chromosomal abnormalities are found at surprisingly high rates in human zygotes, with estimates that as many as three-quarters of these early embryos contain aneuploid cells (van Echten-Arends et al. 2011; Rabinowitz et al. 2012). It has been assumed that there is a selective growth and survival advantage for any subset of normal diploid cells contained within these embryos (Bazrgar et al. 2013), accounting for the generally phenotypically and chromosomally normal outcomes observed. It remains possible, however, that some of the aneuploid cells present in the zygote persist through development. This is recognized more frequently in placental than in embryonic tissues. When defined by the presence of aneuploidy in chorionic villus sampling (CVS) samples from the placenta, and the failure to detect such aneuploid cells in fetal amniocytes or cells from the newborn, it is referred to as confined placental mosaicism (CPM) (Lestou and Kalousek 1998). With current technologies, approximately 0.8%–2.0% of all CVS specimens are found to have mosaic aneuploidy, and a subset of 10%–20% of "confined" placental mosaicism is now recognized to have the same mosaic aneuploidy in the fetus (Fryburg et al. 1993; Hahnemann and Vejerslev 1997; Smith et al. 1999; Daniel et al. 2004; Lebo et al. 2015; Malvestiti et al. 2015). CPM has been found at higher rates (up to ~15% [Wilkins-Haug et al. 2006]) in cases of intra-uterine growth restriction (IUGR).

The prevalence of mosaic chromosomal aneuploidy is not known to be substantial in humans, but it is almost certainly underrecognized. The nature of the developmental event is such that it may only affect an anatomically restricted group of cells in the body, whereas routine genetic testing is almost always performed upon DNA from peripheral blood leukocytes, looking for constitutive, germline mutations. Blood is increasingly found to have age-associated mosaic chromosomal aneuploidy (Jacobs et al. 2012; Laurie et al. 2012; Schick et al. 2013; Machiela et al. 2015). Mosaicism for mutations, especially large structural variants, has been found to occur increasingly with age in blood leukocytes, reaching 2%–3% in individuals age 70 yr or older (Vattathil and Scheet 2016). When cultured, blood is especially poor as a choice of cells for detection of mosaic aneuploidy responsible for

---

---

developmental disorders, exemplified by the failure to detect tetrasomy 12p in Pallister–Killian syndrome (Ballif et al. 2006; Theisen et al. 2009; Hodge et al. 2012), with blood-derived lymphoblastoid cell lines (LCLs) likely to be even worse, given their striking oligoclonality (Ryan et al. 2006). Current large-scale studies of human phenotypes, mostly based on blood or LCL DNA, are therefore unlikely to be optimally sensitive for detection of the presence of mosaic chromosomal aneuploidy. It follows that even the studies to date involving thorough analyses of molecular genomic data are likely to have systematically missed evidence for mosaic aneuploidy occurring in human subjects. In those studies in which mosaic chromosomal aneuploidy was specifically sought, it was found to be associated with certain phenotypes, including several reports of mosaicism for chromosomal aneuploidy in peripheral blood in children with autism spectrum disorder (ASD) (Abu-Amero et al. 2010; Chen et al. 2012; Kostanecka et al. 2012). Given the limitations of using blood to detect aneuploid cells, it is likely that mosaic chromosomal aneuploidy in ASD is not limited to these specific reported individuals but is more prevalent.

Current technologies to detect and quantify mosaic chromosomal aneuploidy include karyotyping of chromosomes using large numbers of metaphase cells generated using cell culture (Kooper et al. 2009), fluorescence in situ hybridization (FISH) using probes detecting the aneuploid chromosome (Faggioli et al. 2014), microarrays to genotype the sample and measure the relative fluorescence of minor alleles (Conlin et al. 2010), single-cell sequencing (Knouse et al. 2014), and whole-genome sequencing (Dong et al. 2016). These approaches all have their relative strengths, but we currently lack an assay that combines relative ease and cost-effectiveness, sensitive detection of low proportions of aneuploid cells, no requirement for cell culture, characterization of the original mitotic or meiotic origin of the mutation, suitability for multiple racial and ethnic populations, and a supporting analytical software resource. To address this need, we developed the MAD-seq assay and its supporting, open source *MADSEQ* analytical software package (see Supplemental Material).

Results

The MAD-seq capture assay

In a single sample of cells, chromosome aneuploidy is revealed by altered dosages of minor alleles, usually referred to as alternate allele frequencies (AAF), throughout a chromosome. For example, normally an alternate (B) allele can be present at a diploid locus at 0% (AA), 50% (AB), or 100% (BB) frequencies, but a cell with a chromosome trisomy will have 0% (AAA), 33% (AAB), 67% (ABB), or 100% (BBB) frequencies for that chromosome. Mosaicism for the trisomy in a population of otherwise diploid cells is reflected by values intermediate between these extremes, with the proportion of mosaic cells reflected by the relative extent to which the minor allele frequencies resemble the diploid pattern (indicating low-level mosaicism) or the trisomic pattern (indicating high-level mosaicism). With greater numbers of loci representing each chromosome, and more of these loci heterozygous in the individual tested, there will be greater ability to detect and quantify mosaic aneuploidy in a cell sample from that person.

We therefore designed a trial customized SeqCap (Roche-NimbleGen) assay (v1MAD-seq capture) that enriches the DNA from 80,000 loci in the human genome, targeting loci with highly polymorphic single nucleotide polymorphisms (SNPs). These SNPs were selected based on being represented on the Illumina HumanOmni2.5 genotyping array and having been studied by the HapMap project (The International HapMap Consortium 2003) and by each having a minor allele population frequency of at least 0.4 across the entire HapMap cohort. The 80,000 loci were selected to be as equally spaced as possible throughout the genome, using capture oligonucleotides designed by Roche-NimbleGen, with most loci having 1–3 tiling probes (mean length 75 nt, range 59–107 nt) for a 125-base pair (bp) window centered around the interrogated SNP. Probes were successfully designed to capture SNPs at 79,605 loci.

To test this capture system, we performed an experiment simulating a variable proportion of alleles within a mixed sample. We used cell line-derived DNA samples from a male Yoruban (GM19239, Coriell Cell Repository) and a female Caucasian (GM06990), mixing the DNA to create serial dilutions of GM19239 in GM06990 in 50%, 25%, 10%, 5%, and 0.5% proportions (Supplemental Table S1). A separate sample of GM06990 on its own was prepared as a control (0%). Capture and Illumina sequencing were followed by alignment and processing using BWA (v0.7.10) (Li and Durbin 2009), and elimination of PCR duplicates using Picard (v1.119; https://broadinstitute.github.io/picard/). Variant calling was performed using the Genome Analysis Toolkit (v3.4–46) (McKenna et al. 2010), including base recalibration and variant calling using the HaplotypeCaller. We plotted the dilutions of GM19239 in GM06990 DNA against the proportion of GM19239 to GM06990 alleles in Supplemental Figure S1a, showing that the subset of GM19239 reads is clearly detectable down to 5%. We therefore proceeded to develop further an analytical approach that would allow us to detect single chromosome aneuploidy events following such capture.

The *MADSEQ* analytical approach

We provide an overview of the analytical approach in Figure 1. There are two main components to *MADSEQ*—the processing of the sequencing data and the generation and comparison of hierarchical Bayesian models. The output of the *MADSEQ* analysis consists of (1) the identification of aneuploidy for one or more chromosomes, (2) categorization of the type of aneuploidy, (3) quantification of the fraction of cells with the aneuploidy, and (4) a confidence metric in the results obtained.

We show simulated results for four types of chromosomal mutations detected by the *MADSEQ* analytical approach in Figure 2, monosomy, mitotic and meiotic trisomy, as well as copy number neutral loss of heterozygosity (cnnLOH). The chromosomal events underlying these results are represented in Supplemental Figure S2. We include cnnLOH as it could represent the consequence of trisomy rescue earlier in development (Biesecker and Spinner 2013) in currently diploid cells. However, we developed the *MADSEQ* approach so that it could also detect segmental LOH occurring in >10% of contiguous heterozygous sites tested in each chromosome. What is apparent from Figure 2 is that our ability to detect and discriminate the different types of aneuploidy events depends upon the alternate allele frequencies in combination with any deviation from the genome-wide average coverage of sequence reads for that chromosome. For example, while mosaic monosomy and mosaic mitotic trisomy have similar alternate allele frequency patterns, they differ by coverage, with trisomy generating an excess and monosomy a deficiency of sequence reads for that chromosome compared with the remainder of the genome. A meiosis II nondisjunction causing trisomy will
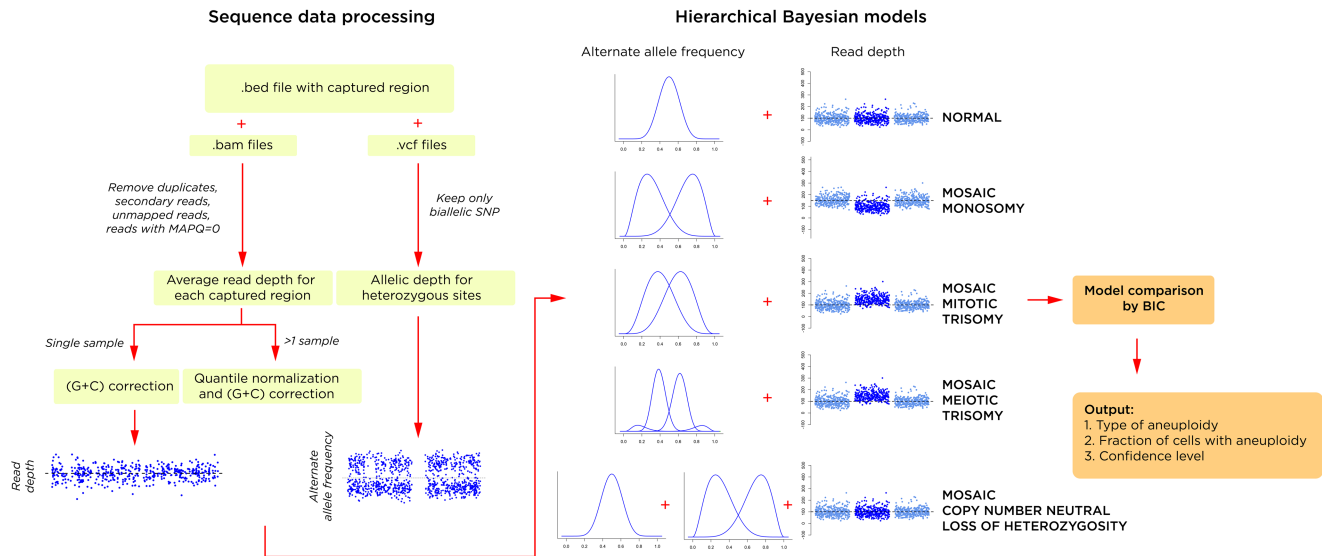
Figure 1. The *MADSEQ* analytical approach has two major components: the processing of the sequence data (*left*); and the use of alternate allele distributions and relative read depth to support a specific hierarchical Bayesian model (*right*). This model is based on fitting a mixture model of alternate allele frequencies (AAFs) at heterozygous loci (showing the additional distributions in a situation of mosaic meiotic trisomy) and read depth for the affected chromosome compared with the remainder of the genome. The winning model is selected by its significantly better Bayesian information criterion (BIC), generating an output of the presence and type of mosaic aneuploidy (or copy number neutral loss of heterozygosity), the proportion of cells with the aneuploidy, and the confidence of this prediction based on relative BIC values.

appear similar to trisomy caused by mitotic nondisjunction but can be distinguished by the presence of a chromosomal region that underwent recombination earlier in meiosis, which is flagged in the *MADSEQ* analysis when choosing the optimal model.

Evaluation of *MADSEQ* performance

We tested the performance of *MADSEQ* in two ways. The first evaluation was a re-analysis of the GM19239/GM06990 mixing experimental data, the second based on computational simulations. For the Yoruban/Caucasian mixing experiment, we used a beta distribution to fit the alternate allele frequency and measured the deviation in these samples from an expected distribution in diploid cells. We measured this deviation for each chromosome in each of the samples and plotted the relationship between this distance and the expected proportions of GM19239 DNA, as shown in Supplemental Figure S1b. We showed the expected correlation with the known mixture proportions but this time using our AAF deviation distance, indicating that the model was able to reproduce the information from known Yoruban and Caucasian genotypic differences.

We went on to explore the sensitivity, specificity, and accuracy of the *MADSEQ* approach using computational simulations. To assess sensitivity, the performance of the model was tested in terms of sequencing depth, proportion of aneuploid cells present, and the number of heterozygous sites sequenced per chromosome. For example, if 2000 heterozygous sites on a chromosome are sequenced to 100× coverage, our model should be able to detect 5% mosaicism for meiotic trisomy with >99% sensitivity at a false discovery rate (FDR) of <2% (Supplemental Fig. S3). The power to detect low proportions of mosaic aneuploidy increases with deeper sequencing and higher numbers of heterozygous sites sequenced. When 2000 heterozygous sites are sequenced to 200× coverage, we predict more than 50% power to detect all types of mosaic aneu-

ploidy occurring in 10% of cells. We also tested the conditions of whole-genome sequencing (WGS), in which coverage is much lower but the number of informative heterozygous sites per chromosome is much higher. The results from these simulations indicate that *MADSEQ* analysis of WGS data will detect aneuploidy events even to low levels of mosaicism and in the smallest human chromosomes (Supplemental Fig. S4).

Specificity issues and the generation of false positive results are important for an assay that might be used for clinical diagnostic purposes. We evaluated the FDR of our method by simulating data without aneuploidy but with the simulated introduction of noise in the sequencing data. We represented noise by replacing the alternate allele frequencies of 10% of the heterozygous sites with values distributed randomly between 0 and 1. The result (Supplemental Fig. S5) suggests that our model is very robust, with an overall FDR at <2%. The accuracy of the quantification of the fraction of abnormal cells was assessed using the root-mean-square-error (RMSE) calculated from computational simulations. In Supplemental Figure S6, we show that the quantification is accurate, with deviation from the expected proportion of less than 10% for all the conditions tested, with accuracy increasing with deeper coverage but with less effect from including more sequenced sites per chromosome.

Application of *MADSEQ* to sequencing data

We then explored whether analysis using *MADSEQ* could identify mosaic aneuploidies from publicly available sequencing data. The 1000 Genomes exome sequencing was generated to a mean of 65.7× coverage, potentially capable of discovering at least some mosaic aneuploidy events if present in these samples. Of the 2535 individuals studied in the 1000 Genomes Project, 2037 were sequenced using DNA derived from LCLs, whereas the remaining 498 were sequenced from peripheral blood leukocyte
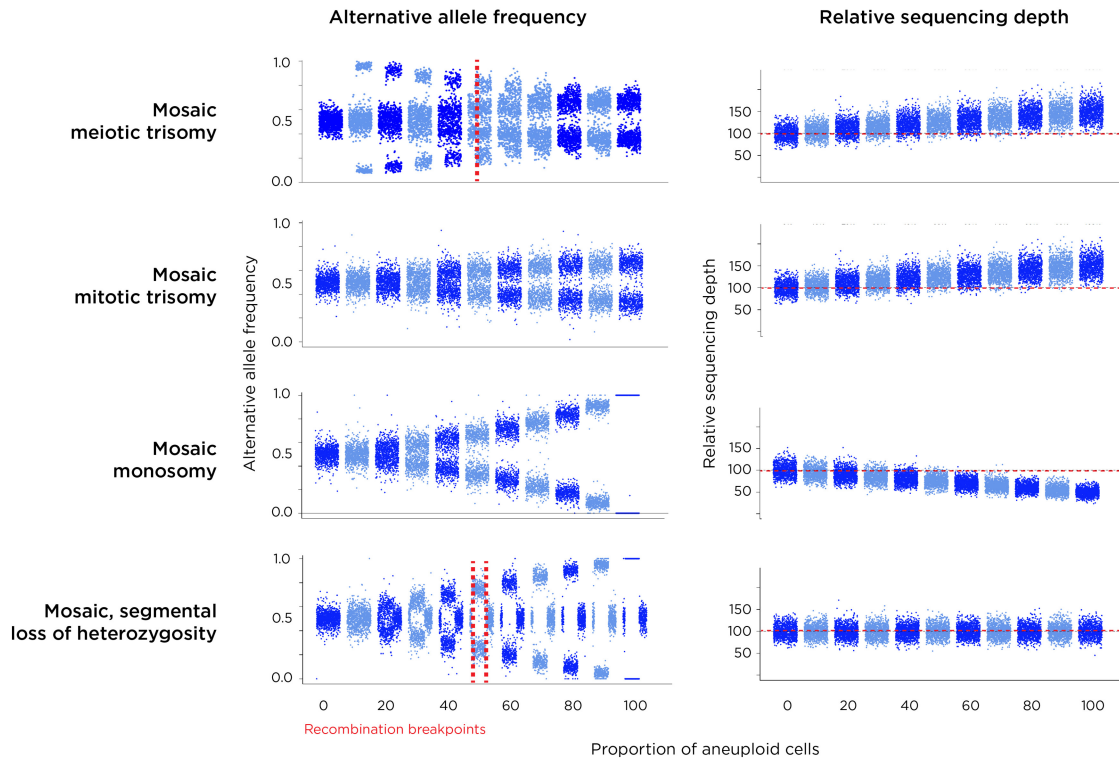
Figure 2. Results of simulations of each of the types of aneuploidy (plus copy number neutral loss of heterozygosity, cnnLOH). Each plot on the *left* shows the AAF distributions in steps of 10% proceeding from no aneuploidy (complete diploidy) to aneuploidy in all cells. On the *right* is shown the expected change of sequencing depth for the affected chromosome. Because meiotic trisomy with recombination creates a non-recombined region with LOH and a recombined area with three different haplotypes, two patterns are created within these chromosomes, the pattern in the recombined region being very distinctive (*top*). The coverage for a trisomic chromosome is increased relative to others in the genome. Trisomy occurring in mitosis (or in meiosis without recombination) generates a similar profile in terms of AAF to mosaic monosomy, but these events are distinguished by coverage being higher for trisomy and lower for monosomy. Segmental cnnLOH is shown in the *bottom* panels, the affected area of the chromosome showing an AAF pattern similar to monosomy but with no change in copy number. A region showing nonmosaic (100%) cnnLOH reflects consanguinuity, at a locus of identity by descent.

DNA. *MADSEQ* detected 83 mosaic events with high confidence (ΔBIC > 10) in 76 individuals (Supplemental Table S2). All of the detected mosaic aneuploidies were from LCL samples, and none from blood. The types of mosaic aneuploidy include 20 monosomies (0.79%), 25 mitotic trisomies (0.99%), and 37 LOH (1.46%), but no meiotic trisomies, even though the model is relatively more sensitive when detecting meiotic events. Of note, all of the cases of LOH were segmental rather than involving the whole chromosome, likely to represent the result of a repair of deletion using the remaining homologous chromosome as a template (O'Keefe et al. 2010), and not trisomy rescue.

The rate of these mosaic events among LCL samples was 3.73%. The proportions of aneuploid/segmental LOH cells in each sample varied from 4.2% to 79.9%. The most overrepresented events were mosaic mitotic trisomy of Chromosome 12 in 11 samples ($P = 1.49 \times 10^{-4}$, Binomial Test) and enrichment for mosaic segmental LOH in Chromosome 22 ($P = 7.14 \times 10^{-3}$, Binomial Test) (Supplemental Table S3). Overall, the significant lack of mosaic aneuploidy events in samples from blood ($P = 3.08 \times 10^{-34}$, Binomial Test) and the lack of meiotic events, together with the enrichment of trisomy 12, which is the most common cytogenetic abnormality in chronic B lymphocytic leukemia (Einhorn et al. 1989), combine to suggest that these mosaic aneuploidies arose during cell culture and were either neutral in effect or were promoted by positive selection for these transformed B lymphocytes. One

of the samples in which segmental LOH for distal Chromosome 11 was identified was GM12889, for which whole-genome sequencing to a mean ~50× coverage has been performed to define high-confidence, "platinum" variants (Eberle et al. 2017). We downloaded those WGS data and reran *MADSEQ*, again predicting the LOH of a 19.3-Mb region, estimated to be present in over 50% of the cells (Fig. 3). The platinum variant calling in this part of the genome in this individual should be interpreted with caution. We show representative examples of plots of the alternate allele frequencies and the comparisons of the Bayesian Information Content (ΔBIC) in Supplemental Figure S7.

We then tested a sample from a patient presenting with hemihyperplasia (OMIM 235000). The hyperplastic side of the patient demonstrated a hyperpigmented whorl pattern, following Blaschko's lines. The Blaschko's lines were only present on the hyperplastic side and did not cross the midline. Skin biopsies were performed on the normal skin of the unaffected side and from the hyperpigmented skin of the affected side. A microarray study from DNA directly extracted from these biopsies showed evidence for mosaic trisomy 12 from the affected side only. We grew fibroblasts from the remainder of the skin biopsy and extracted DNA from these cultured cells, performing exome sequencing to a mean 130× coverage. The *MADSEQ* model best fit by the results was of mosaic trisomy of mitotic origin present in 6.8% of the cells (ΔBIC = 18) (Fig. 4). Trisomy 12 has also been found in human
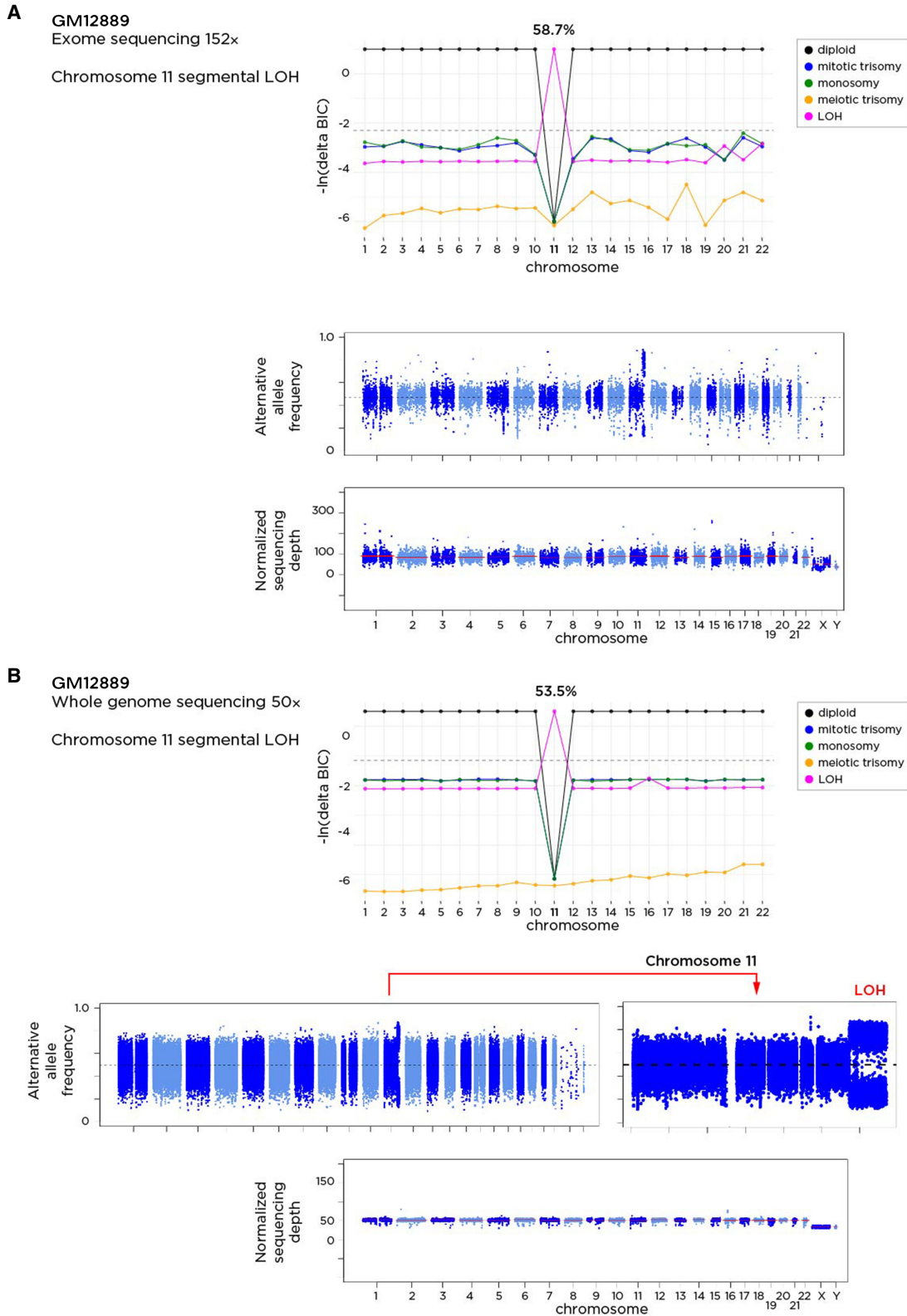
Figure 3. The GM12889 sample has been tested using exome sequencing as part of the 1000 Genomes study (*A*) and using whole-genome sequencing to mean 50× coverage (*B*). The application of the *MADSEQ* approach shows the predictions for each chromosome and the differences in Bayesian information criterion for each model. The *top* plot within each panel shows these BIC values for each model on the *y*-axis for each chromosome (*x*-axis). The horizontal dashed line represents our threshold difference for discriminating between models. In *A*, we see that a cnnLOH model for Chromosome 11 is strongly predicted, occurring in 58.7% of cells. Whole-genome sequencing data from the same DNA shows a highly concordant prediction of mosaicism for segmental cnnLOH in the same region of distal Chromosome 11 (*bottom*) in 53.5% of cells.
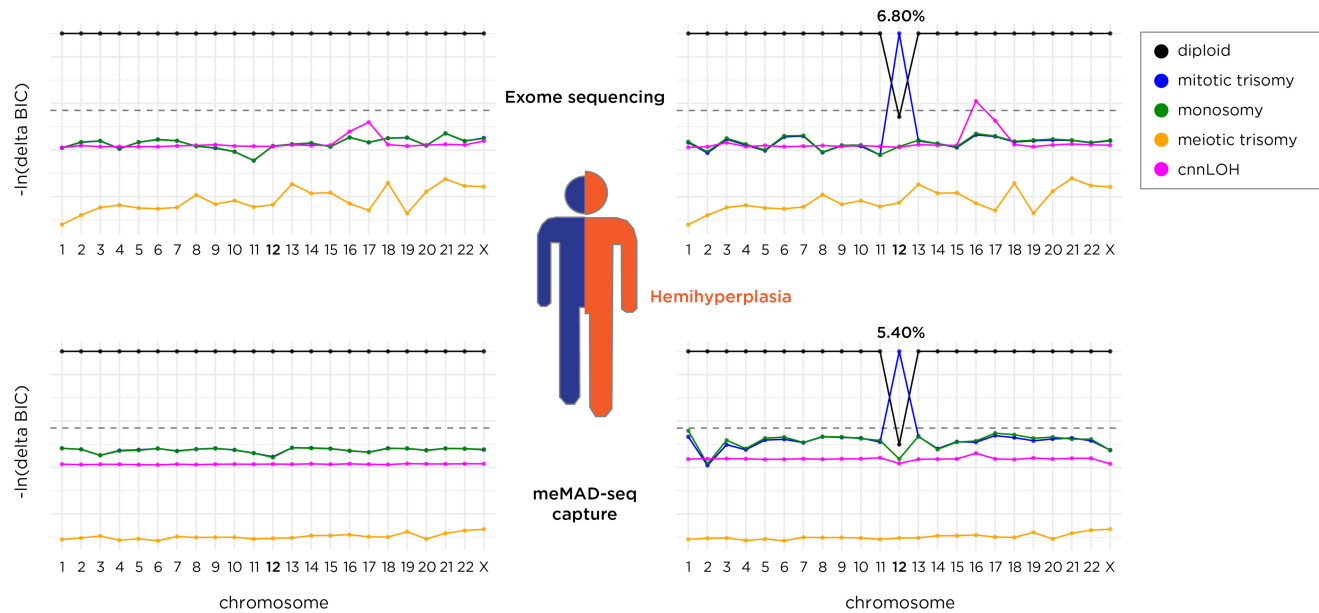
Figure 4. The results of sequencing skin biopsy samples from each side of the body of a child with hemihyperplasia. Results from the normal side are shown on the *left* and from the overgrown side on the *right*. Exome sequencing to ~100× mean coverage indicates a model of mosaic trisomy 12 of mitotic origin in 6.8% of cells, while the more targeted (multi-ethnic) meMAD-seq capture assay (described later) sequenced to a comparable depth shows evidence for the same type of mosaicism in 5.4% of cells.

embryonic stem (ES) cell lines (Draper et al. 2004) and induced pluripotent stem (iPS) cells (Taapken et al. 2011) and has been implicated in the significant increase of their cellular proliferation rate and tumorigenicity (Ben-David et al. 2014). The hemihyperplasia phenotype may therefore be the consequence of a higher cell replication rate due to the presence of the mosaic subset of cells with trisomy 12. Our categorization that this was a mosaic trisomy of mitotic origin (Fig. 4) indicates that the nondisjunction and chromosomal loss events occurred post-fertilization and are not due to meiotic trisomy during gametogenesis with later trisomy rescue, unless the rescue of a meiotic trisomy was compounded with the additional event of no recombination during meiosis.

Development of a multi-ethnic MAD-seq (meMAD-seq) capture assay design

To complement our exome sequencing results, we returned to our v1MAD-seq capture design and tested its performance with six samples, four of which were known to have autosomal trisomies, and two control cell lines apparently lacking aneuploidies. Using *MADSEQ* analysis of the results, we confirmed the four trisomies (Chromosomes 8, 13, 15, or 18), mostly concordant with prior reported proportions of trisomic cells, and defined their origins as meiotic in all cases. One of the control cell lines (GM06990) that had not previously been described to have aneuploidy was found to have a pattern consistent with 6.6% of cells having monosomy for Chromosome 6 (Table 1).

When we explored the performance of the v1MAD-seq capture and exome sequencing systems, we found that the representation of heterozygous sites on many of the smaller chromosomes was highly suboptimal. The v1MAD-seq capture design spaced loci for capture evenly throughout the genome, causing larger chromosomes to have proportionately more informative loci (Fig. 5A), while exome-seq is limited by the number of genes per chromosome, which is a function of not only the chromosome

size but also its gene content. In Figure 5B, we show this heterogeneity of representation for each chromosome for exome sequencing data. Of particular concern was the poor representation of informative loci for Chromosomes 13, 18, and 21, the most common viable full trisomies.

To create a design that is maximally efficient in sequencing informative loci, with resulting efficiencies in assay cost, we created a new multi-ethnic MAD-seq capture design. A total of 107,797 (Roche NimbleGen Catalog No. 06740260001, Design Identifier: 160407_HG19_MadSeq_EZ_HX1) common SNPs were chosen to represent each chromosome equally. We also exploited the 1000 Genomes data to identify loci that would be most likely to be polymorphic across all human populations (Supplemental Fig. S8). We show the workflow for the design of the meMAD-seq capture platform in Supplemental Figure S9. This design captures 106,402 loci in the human genome of mean length 139 bp and mean (G + C) content 44.4%.

We tested the meMAD-seq capture design on 12 samples. These included the six samples tested using the v1MAD-seq capture design, the HG01939 sample predicted to have two separate chromosomes with loss of heterozygosity from our 1000 Genomes exome sequencing re-analysis, samples from the affected and unaffected sides of the body of the individual with hemihyperplasia, and two Coriell cell repository samples, one of which was described to have mosaic trisomy 8 (GM00496), and a sample described to have mosaic trisomy 12 as well as random chromosome rearrangements (NA01454), and DNA from the H1 human embryonic stem cell line (Thomson et al. 1998). We show the results in Table 1 and Figure 5, confirming prior observations of chromosomal aneuploidies or loss of heterozygosity from Coriell's characterizations, our exome sequencing data re-analysis, or the v1MAD-seq capture results, and adding information about whether each trisomy was likely to be meiotic or mitotic in origin.

Our 1000 Genomes SNP selection strategy was designed to generate data from the same number of heterozygous sites for

Table 1. Overview of results from exome sequencing, v1MAD-seq, and (multi-ethnic) meMAD-seq capture assays, compared with information about samples known before testing

Individual	Sex	Population	Prior aneuploidy information	Exome-seq	StartPos	EndPos	v1MAD-seq (common SNPs)	meMAD-seq (1000 Genomes SNPs)	StartPos	EndPos
AG13074[a]	Female	EUR	Chr 18 Trisomy 70%				Chr 18 70% Meiotic trisomy	Chr 18 71.10% Meiotic trisomy		
F44P110[b]	Female	?	Unknown				Chr 18 49% Meiotic trisomy	Chr 18 44.64% Meiotic trisomy		
GM00496[c]	Male	?	Chr 8 Trisomy 81%					Chr 8 90.37% Meiotic trisomy		
GM06990[d]	Female	EUR	Negative				Chr 6 6.60% Monosomy	Chr 6 7.24% Monosomy		
GM19239[e]	Male	AFR	Negative				Negative	Negative		
HG01939[f]	Female	AMR	Negative	Chr 9 41.50% LOH	117,371	71,114,312		Chr 9 45.14% LOH	40,910	35,728,494
				Chr 22 17.73% LOH	19,906,511	25,145,471		Chr 22 7.83% LOH	18,288,362	27,285,813
NA00503[g]	Male	?	Chr 13 Trisomy 38%				Chr 13 37% Meiotic trisomy	Chr 13 28.02% Meiotic trisomy		
NA00682[h]	Female	?	Chr 15 Trisomy 36%				Chr 15 70% Meiotic trisomy	Chr 15 70.50% Meiotic trisomy		
NA101454[i]	Male	Asian	Chr 12 Trisomy 100%					Chr 12 53.2% Meiotic trisomy		
			Chr 13q34 Trisomy 52%					Chr 13 96.5% Mitotic trisomy		
			"random chromosome rearrangements"					Chr 1, 3–4, 6–8, 11, 14, 16–17, 20–21 ~46% Monosomy		
Hemihyperplasia (affected side)[j]	Female	?	Chr 12 Trisomy 35%	Chr12 6.80%		Mitotic trisomy		Chr 12 5.40% Mitotic trisomy		
Hemihyperplasia (unaffected side)[j]			Negative					Negative		
H1 embryonic stem cell line	Male	?	Negative					Chr 19 6.39% Mitotic trisomy		

We see that, in general when there was a known mosaic aneuploidy in the sample, we confirm it, but some samples (HG01939 and H1 embryonic stem cells) that were supposedly negative are found to have detectable mosaic aneuploidies. The proportions of cells with aneuploidy are usually, but not always, concordant with those identified using our analyses, which also identify the likely mechanism of the aneuploidy. (EUR) European; (?) ancestry unknown; (AFR) African; (AMR) Americas; (Chr) chromosome; (LOH) loss of heterozygosity. Gray shading denotes test not performed.
[a]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=AG13074&product=CC.
[b]De-identified human sample from http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004402.
[c]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM00496&product=CC.
[d]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM06990.
[e]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM19239.
[f]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=HG01939&Product=CC.
[g]Coriell: https://catalog.coriell.org/0/sections/Search/Sample_Detail.aspx?Ref=GM00503&Pgid=166.
[h]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Pgid=166&Ref=GM00682.
[i]Coriell: https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=NA01454&Product=DNA.
[j]De-identified human sample.
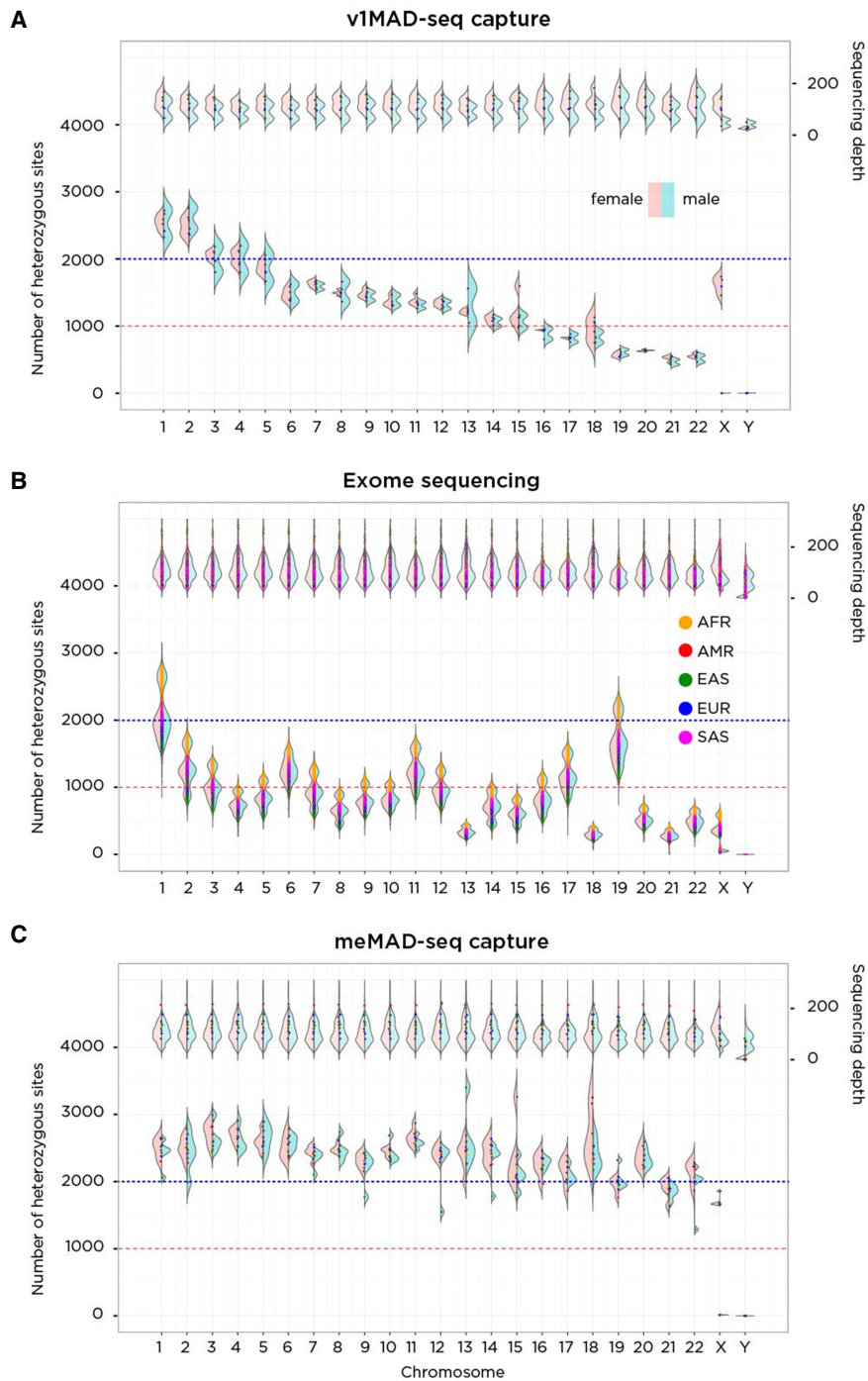
Figure 5. A performance comparison of v1MAD-seq capture (*A*), exome sequencing (*B*), and meMAD-seq capture (*C*) in terms of the number of heterozygous sites tested per chromosome. On the *A*, of each plot is a representation of sequencing depth, on the *B*, the number of heterozygous sites per chromosome. On the *left* of each violin plot is the distribution in females (pink); on the *right*, males (blue), allowing differentiation of patterns on the sex chromosomes. The v1MAD-seq capture platform spaced probes evenly throughout the genome, generating more heterozygous sites in larger chromosomes despite equivalent coverage. Exome sequencing (1000 Genomes data) generates a number of heterozygous sites that depend on chromosome size and gene density within the chromosome, so that the comparably sized Chromosomes 18 and 19 differ because of the much greater gene content of the latter. There is also an ancestry effect on heterozygosity rates, with African ancestry performing best and East Asian ancestry worst in generating heterozygous sites. The meMAD-seq capture design allows much more uniform performance of every chromosome in the genome, most exceeding 1500 heterozygous sites per chromosome.

each chromosome and across all populations. Our goal was to exceed 1000 heterozygous sites per chromosome, but we show in Figure 5C that we obtain ≥2000 heterozygous loci for every chromosome across all individuals. In Supplemental Figure S10, we show that the individuals tested using the meMAD-seq capture design were indeed from widely divergent human population groups. We confirmed mosaic aneuploidies in samples known to have these abnormalities from Coriell's characterization or from our prior studies. Two chromosomes with segmental LOH predicted from our 1000 Genomes exome sequencing data analysis were confirmed in the HG01939 cell line using the meMAD-seq capture assay. Some of the supposedly normal control samples were also revealed to have aneuploidy, including low level, previously unrecognized events in the ES cells and the GM06990 female Caucasian sample used in our initial serial dilution experiments.

For a cost comparison, we determined the reagent cost expense associated with library preparation, capture, and sequencing to comparable depth for the more mainstream exome sequencing approach (SeqCap EZ, Roche-NimbleGen) and the meMAD-seq capture alternative. We estimate that, for each assay to generate a mean ~110× coverage, the reagent cost for meMAD-seq capture would be ~40% of the cost for exome sequencing, which should be a generalizable guide for facilities with different costs. With increased production of the meMAD-seq capture kit, further cost savings may be possible.

Discussion

This study further strengthens the concern that mosaic aneuploidy is likely to be underrecognized. The 1000 Genomes Project performed extensive analysis and quality assessment of their samples and data (The 1000 Genomes Project Consortium 2015), but even these carefully studied samples have mosaic aneuploidy in several percent of the LCL samples studied, estimated by *MADSEQ* to involve as many as ~80% of cells. While we interpret the results to indicate that these cases of mosaicism arose during cell culture, this finding should increase the caution required when interpreting information from LCLs in terms of their representation of the donor's chromosomal status.

We also find that reference cell lines that have been characterized using standard techniques have evidence for chromosomal abnormalities. The human H1 ES cell line has previously been found to develop trisomies in vitro (Draper et al. 2004), requiring periodic testing to ensure that the cells being used experimentally remain diploid. Reference cell lines supplied by repositories or used in large studies also require careful characterization to ensure that they are not undergoing alterations that could lead to issues of reproducibility of results. It should be stressed that the poor representation of certain chromosomes in exome sequencing data, coupled with some of the 1000 Genomes samples having relatively lower mean coverage, combine to indicate that we are probably missing some further cases of mosaic aneuploidy. The systematic application of *MADSEQ* analysis, allied with meMAD-seq capture and sequencing, would probably reveal an even higher proportion of reference cell lines with mosaicism.

The analytical software *MADSEQ* is open source, available through Bioconductor (see Supplemental Material), and can be applied not only to meMAD-seq capture data but also, as we show, to exome sequencing and even whole-genome sequencing data. It could therefore be applied retrospectively to legacy sequence data to look for aneuploidy, generating preliminary results that could then prompt the application of meMAD-seq capture for more systematic studies. It appears that more development will be needed to extend the use of *MADSEQ* to cancer samples. When multiple chromosomes have abnormal copy numbers, determining what coverage value represents diploidy becomes difficult, weakening a foundational component of the analysis. Our NA01454 sample (Supplemental Fig. S7i) is not from a cancer but has multiple chromosomes with abnormal patterns of AAFs and copy numbers and helps to illustrate how the approach starts to have difficulties when many chromosomes are affected. This will be a focus of further algorithm development, but in the short term, *MADSEQ* is valuable for detecting the presence of aneuploidy in even these complex samples. With further development, *MADSEQ* could also be used to detect mosaicism for copy number variants (CNVs). However, the resolution of detection will differ based on spacing of the captured heterozygous loci which, in the meMAD-seq capture design, is higher for larger chromosomes, with more physical clustering of captured loci in smaller chromosomes. The most appropriate future application of *MADSEQ* for mosaic CNV identification may be from WGS data.

Ideally, prospective studies will exploit existing WGS data or the targeted sequencing option of meMAD-seq capture, allowing optimal cost and performance. The meMAD-seq capture design, which will be made publicly available through Roche-NimbleGen, shows excellent performance not only in terms of maximizing the number of informative sites per chromosome but also testing each chromosome comparably. We were careful to ensure that the design could be applied equally effectively to people of widely differing ancestries, allowing it to be used worldwide and in our local, highly diverse clinical population.

We anticipate several areas of human disease research that would immediately benefit from *MADSEQ*. In prenatal genetics care, screening is performed looking for chromosomal aneuploidies, increasingly using noninvasive prenatal testing (NIPT) of cell-free DNA in the maternal blood. Positive results from this screening approach can be followed up with invasive tests of fetal cells (chorionic villus sampling, amniocentesis), which can then, in a proportion of cases, lead to discordance between aneuploid NIPT and normal fetal chromosomal results. This discordance is presumed to be due to confined placental mosaicism for the aneu-

ploidy, but this diagnosis can only be made with the certainty afforded by the sensitivity of the test used on the fetal cells. There is potential for WGS or meMAD-seq capture and sequencing allied with *MADSEQ* analysis to enhance the sensitivity of these diagnostic studies. A second potential area worth exploring for covert aneuploidy is in individuals with autism spectrum disorder. We have noted in a prior study (Berko et al. 2014) the association between advanced maternal age and the risk of having a child with ASD (Sandin et al. 2012). The increased nondisjunction rate in oocytes from older mothers suggests that chromosomal aneuploidy should be tested as a possible mediator of this association, but at present, there is little evidence for aneuploidy in individuals with ASD. A more sensitive assay like meMAD-seq capture applied to samples other than blood from individuals with ASD born to older mothers may be helpful in exploring a potential cause of this heterogeneous condition.

Sequencing data from WGS or the meMAD-seq capture assay combined with the *MADSEQ* analytical approach can be used on uncultured cells, detects low levels of aneuploidy, identifies the likely mechanism of the initial causative event, is relatively cost-efficient, and can be used in any ancestral background. It combines many of the advantages of existing assays to detect aneuploidy and should be suitable for high-throughput studies. The eventual goal should be to associate different types of mosaic chromosomal events with human phenotypes.

Methods

Molecular assays

Cell line mixing experiments

DNA extracted from two individuals' LCLs (GM06990, CEU, female and GM19239, YRI, male) was mixed at different proportions, choosing 100%:0%, 99.5%:0.5%, 95%:5%, 90%:10%, 75%:25%, and 50%:50% to mimic different levels of mosaicism. The DNA was sequenced following a capture protocol using the SeqCap EZ Choice system from Roche-NimbleGen. The list of targeted regions and probes used for the v1MAD-seq capture design can be found in our dbGaP submission. Knowing the genotype of both samples, we were able to extract sites that mimic the different types of mosaic aneuploidy. Specifically, to simulate mitotic aneuploidy, we first extracted loci with different genotypes in the two cell lines, 0/1 in CEU and 1/1 in YRI to mimic overrepresentation of the alternate allele, and 0/1 in CEU and 0/0 in YRI to mimic underrepresentation of the alternate allele. The two mixtures separated further when a higher proportion of YRI DNA was mixed with CEU DNA (Supplemental Table S1). Because these mixtures of DNA alter the distribution of alternate allele frequency without changing the actual copy number of chromosomes, we applied the model without the coverage module.

Exome sequencing of a patient with hemihyperplasia

DNA was extracted from fibroblasts cultured from skin biopsies from the affected and unaffected sides of the body of a patient with hemihyperplasia (OMIM 235000). We performed exome sequencing of the sample from the affected side of this patient using the SeqCap EZ Exome Enrichment kit v3.0 (Roche-NimbleGen) and 100-base pair paired-end sequencing on the Illumina HiSeq 2500 system. The average coverage was 142.1×.

Targeted sequencing

DNA was purchased from Coriell for the samples HG01939, NA00682, and NA01454, and DNA was extracted from fibroblasts

(AG13074, GM00496, and GM00503), LCLs (GM06990, GM19239) buccal epithelial cells of one patient (F44P110), and a human embryonic cell line (H1). DNA extracted from the normal and abnormal cultured fibroblasts of the hemihypertrophy patient was also included. We used the Roche NimbleGen SeqCap EZ Choice system to capture the multi-ethnic design of the 105,703 common SNPs described below. All of the samples were sequenced with 100-bp paired-end sequencing using the Illumina HiSeq 2500 (Illumina), generating an average coverage of 134.6×.

Sequencing depth correction

G + C content can vary in the genome and influence the number of reads generated at each captured region, potentially introducing bias into aneuploidy detection. We therefore used LOESS correction to correct in our package for such bias. Given the targeted region and BAM file, the average coverage for each targeted region (raw_cov_i) is calculated by a coverage function from the *MADSEQ* R package called *GenomicAlignments*. If more than one sample is sequenced during the same capture protocol, quantile normalization is first applied to the coverage across all samples. The G + C content (gc_i) for each targeted region is then calculated as the G + C percentage of the reference genome (excluding Ns). Coverage for each targeted region was grouped by 0.1% increments of G + C content, and the average coverage for each level of G + C content was calculated. The scatterplot representing the G + C content plotted against the average coverage for each G + C level can be produced as part of the *MADSEQ* pipeline (Supplemental Fig. S11). The regression curve between coverage and G + C content was fitted by LOESS. The GC content for the ith region is gc_i, the fitted coverage for this region denoted as cov_{gc_i}. The expected coverage (cov_{exp}) is set to the median of read depth across all regions. The corrected coverage for the ith region ($corrected\_cov_i$) is then calculated as

$$corrected\_cov_i = raw\_cov_i - (cov_{gc_i} - cov_{exp})$$

Bayesian models

Our statistical model for detecting mosaic aneuploidy consists of two parts. For each chromosome, we first consider the distribution of the alternate allele frequencies at the heterozygous sites; secondly, we consider the distribution of sequencing depth at these loci. If mosaic aneuploidy is present in the sample, we expect the distribution of both the alternate allele frequency and sequencing depth to deviate from that expected in a simple diploid sample. Here, we describe each part of the model separately:

1. Detection of aneuploidy from alternate allelic fractions.

    The alternate allele fraction is the proportion of reads carrying the alternate allele at a given heterozygous site, calculated as the alternate allelic depth divided by the total read depth. If there is no aneuploidy in any of the sampled cells, then the AAFs at heterozygous sites are expected to be centered around 0.5 (ignoring confounding biases such as reference bias) (Castel et al. 2015). However, if a fraction of cells within the sample are aneuploid, then the distribution of AAFs will deviate from the expected midpoint and instead be better described by a mixture of distributions, where the number and parameters of the mixture components depend on the origin of the aneuploidy and the degree of mosaic aneuploidy.

    1. Model$_0$: diploid chromosome. For a normal, diploid chromosome state, the AAF at heterozygous sites is expected to be a single distribution centered around the midpoint (average AAF across all heterozygous sites). In this situation, we model the alternate allelic depth (AD_i) for biallelic heterozygous site

i as a simple beta-binomial distribution, given the read depth for the ith site N_i

$$AD_i \sim betabinomial(N_i, \alpha_n = \mu\kappa, \beta_n = (1-\mu)\kappa)$$

Here, α and β are determined by the prior μ and κ. μ denotes the midpoint, namely the average AAF across all the heterozygous sites. κ represents the variance of the AAF. We model κ as a gamma distribution

$$\kappa \sim gamma\left(shape = \frac{m^2}{\sigma^2}, rate = \frac{m}{\sigma^2}\right)$$

In our model, the larger the κ is, the smaller the variance for the beta distribution. For the purpose of Bayesian inference, we assigned the prior for this gamma distribution as $m = 10$, $\sigma = 10$ to represent a flat prior distribution.

To account for noise normally present in high-throughput sequencing data, we added an additional outlier component weighted as 1% ($\omega_0 = 0.01$; $\omega_n + \omega_o = 1$) of all heterozygous reads. The outlier component is modeled by a uniform beta-binomial distribution as

$$AD_o \sim betabinomial(N_o, \alpha_o = 1, \beta_o = 1)$$

If there are K heterozygous sites from one chromosome, and mu and Kappa are constant for all $n = 1,...K$, the likelihood of the data is then given by

$$P(AD|N, \alpha, \beta) = \prod_{i=1}^{K} (\omega_n dbetabinom(AD_i, N_i, \alpha_n, \beta_n)$$
$$+ \omega_o dbetabinom(AD_i, N_i, \alpha_o, \beta_o))$$

where $P(AD|N, \alpha, \beta)$ are vectors of the parameters.

2. Model$_1$: mosaic monosomy. Mosaic monosomy is the consequence of the loss of one chromosome in a subset of cells. In mosaic monosomy, the AAF separates from the midpoint into two mixtures (Fig. 2). One mixture is shifted toward lower values due to the overrepresentation of the reference allele, and the other shifted toward higher values due to the over-representation of the alternate allele. We assume that the two mixtures have equal weight ($\omega_1 = \omega_2 = 0.495$) and variance ($\kappa$), with the same outlier component ($\omega_o = 0.01$) described in model$_0$; the allelic depth of heterozygous sites can be modeled as

$$AD_i \sim (\omega_1 betabinomial(N_i, \alpha_1 = \mu_1\kappa, \beta_1 = (1-\mu_1)\kappa)$$
$$+ \omega_2 betabinomial(N_i, \alpha_2 = \mu_2\kappa, \beta_2 = (1-\mu_2)\kappa)$$
$$+ \omega_o betabinomial(N_i, \alpha_o = 1, \beta_o = 1))$$

The more monosomic cells in the sample, the further these two mixtures will separate (Fig. 2). The priors μ_1 and μ_2, which are the average AAF of the two separated mixtures, are determined by the fraction of aneuploidy cells, f. Given the expected midpoint (calculated as average AAF for all heterozygous sites from the whole genome) m, the expected mean AAF of the two mixtures (μ_1, μ_2) can be calculated as

$$\mu_1 = m + \frac{mf(1-m)}{1-f+mf}; \mu_2 = m - \frac{mf(1-m)}{1-mf}$$

The hyperprior on f is modeled by a uniform beta distribution, which means the fraction of abnormal cells ranges from 0% to 100% with equal prior probability before

inference

$$f \sim beta(\alpha_f = 1, \beta_f = 1)$$

Thus, under this model, the likelihood of the data is given by

$$P(AD|N, \alpha, \beta) = \prod_{i=1}^{K}(\omega_1 dbetabinom(AD_i, N_i, \alpha_1, \beta_1)$$
$$+ \omega_2 dbetabinom(AD_i, N_i, \alpha_2, \beta_2)$$
$$+ \omega_0 dbetabinom(AD_i, N_i, \alpha_o, \beta_o))$$

3. Model$_2$: mosaic mitotic trisomy. Mosaic mitotic trisomy arises from nondisjunction during mitotic cell division, resulting in an extra copy of one of the normal chromosomes. As a result, the AAF at heterozygous sites will be separated into two mixtures, in a qualitatively similar pattern to that of mosaic monosomy case (Fig. 2). However, the expected average AAFs of the mixtures for a given fraction of aneuploidy differ, with the means expected to be given by

$$\mu_1 = m + \frac{mf(1-m)}{1+mf}; \mu_2 = m - \frac{mf(1-m)}{1+f-mf}$$

The hyperprior f and the weights for separated mixtures ω_1 and ω_2 and the outlier component are the same as described in mosaic monosomy.

4. Model$_3$: mosaic meiotic trisomy. Trisomy can also be acquired during meiotic cell division. Mosaic meiotic trisomy can be distinguished from mitotic trisomy by the presence of two additional mixtures for part of the chromosome near the boundaries (Fig. 2), which are the consequence of recombination during meiosis. Based on the assumption that the four separated mixtures have the same variance (κ), the allelic depth of heterozygous sites can be modeled as

$$AD_i \sim (\omega_1 betabinomial(N_i, \alpha_1 = \mu_1\kappa, \beta_1 = (1-\mu_1)\kappa)$$
$$+ \omega_2 betabinomial(N_i, \alpha_2 = \mu_2\kappa, \beta_2 = (1-\mu_2)\kappa)$$
$$+ \omega_3 betabinomial(N_i, \alpha_3 = \mu_3\kappa, \beta_3 = (1-\mu_3)\kappa)$$
$$+ \omega_4 betabinomial(N_i, \alpha_4 = \mu_4\kappa, \beta_4 = (1-\mu_4)\kappa)$$
$$+ \omega_0 betabinomial(N_i, \alpha_0 = 1, \beta_0 = 1)$$

Among the four separated mixtures, the two mixtures in the center are expected to have the same means as described in model$_2$ (μ_1, μ_2). The means of the two additional mixtures near the boundaries are given by

$$\mu_3 = \frac{mf}{2+mf-2m}; \mu_4 = 1 - \frac{f(1-m)}{f+2m-fm}$$

We assume that the two central mixtures have the same weight and that the two edge mixtures also have the same proportion. As the edge mixtures can have smaller or equal weight compared to the center mixtures, we therefore model the prior of the weight of the edge mixtures by a truncated uniform beta distribution

$$\omega_3 \sim beta(\alpha = 1, \beta = 1)I(0.1, 0.25);$$
$$\omega_1 = \omega_2; \omega_3 = \omega_4; \omega_0 = 0.01;$$
$$\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_0 = 1$$

The hyperprior f is modeled the same way as described above.

5. Model$_4$: mosaic loss of heterozygosity (LOH). Mosaic copy number neutral loss of heterozygosity can be due to multiple reasons, for example, due to trisomy rescue when the whole chromosome is involved, or recombination-mediated repair when the LOH is segmental. As a result, the AAFs of the heterozygous sites or all or some of the chromosome will also be split into two mixtures (Fig. 2). To characterize such regional effects of LOH, we introduced a reversible jump model, which contains two change points (cgp_s, cgp_e) to account for the start and end of the LOH status, to describe the combination of normal and abnormal regions on the same chromosome.

In the normal regions, the model is the same as for a normal chromosome

$$AD_i \sim \omega_1 betabinomial(N_i, \alpha_1 = \mu_1\kappa, \beta_1 = (1-\mu_1)\kappa)$$
$$+ \omega_0 betabinomial(N_i, \alpha_0 = 1, \beta_0 = 1); (i < cgp_s \ or \ i > cgp_e)$$

In the LOH region, the distribution of AAFs is separated into two mixtures:

$$AD_i \sim \omega_2 betabinomial(N_i, \alpha_2 = \mu_2\kappa, \beta_1 = (1-\mu_2)\kappa)$$
$$+ \omega_3 betabinomial(N_i, \alpha_3 = \mu_3\kappa, \beta_3 = (1-\mu_3)\kappa)$$
$$+ \omega_0 betabinomial(N_i, \alpha_0 = 1, \beta_0 = 1); (cgp_s \leq i \leq cgp_e)$$

For the normal region ($i < cgp_s \ or \ i > cgp_e$), the distribution is modeled in the same way as for a normal chromosome, with μ_1 calculated here as the average AAF for heterozygous sites. For the LOH region, the weights for the two mixtures are assumed to be the same. The means of the separated mixtures are calculated as

$$\mu_2 = \frac{m(1+f)}{(1-f+2mf)}; \mu_3 = \frac{m(1-f)}{(f-2mf+1)}$$

The hyperprior f is modeled the same way as described above.

Given that there is a total of K heterozygous sites on each chromosome, the priors of the changing points, which are the starting locus and ending locus of the abnormal region, are modeled by two uniform distributions ranging from 1 to K. In order to be robust against noise in the data, we require that the LOH region spans at least 10% of the total number of loci (K) on one chromosome

$$cgp_s \sim uniform(1, 0.90K);$$
$$cgp_e \sim uniform(cgp_s + 0.10K, K)$$

2. Inference of the type of aneuploidy from sequencing depth.

While the distribution of AAF is informative of mosaic aneuploidy, it is difficult to distinguish between the cases without additional information, as, for example, mitotic trisomy and mosaic monosomy have similar distribution of AAFs. In order to improve our differentiation of different types of aneuploidy, we augmented our model with information about sequencing depth.

In our model, the expected coverage for normal chromosome is denoted by m_g. If there is only one sample, m_g is calculated as the median of GC corrected coverage for the whole genome. If there are multiple samples, m_g is calculated as the median across the normalized coverage for that chromosome across all samples.

For the sequencing depth, the total number of targeted regions from one chromosome is $nRegion$; the coverage of the ith region is cov_i. In order to characterize the overdispersion of the depth observed in massively parallel sequencing data, we model the

coverage as a negative binomial distribution

$$cov_i \sim NegativeBinomial(prob = p, size = r)$$

where the prior of r is modeled by a weakly informative gamma distribution

$$r \sim gamma(shape = 0.01, rate = 0.01)$$

and p is taken as

$$p = \frac{r}{m_{cov} + r}$$

Here, m_{cov}, which is the mean of the coverage for this chromosome, is determined by the expected normal coverage m_g and the fraction of aneuploid cells f

$$m_{cov} = m_g \frac{(2+f)}{2} (type = trisomy);$$
$$m_{cov} = m_g \frac{(2-f)}{2} (type = monosomy);$$
$$m_{cov} = m_g (type = normal \ or \ type = LOH)$$

In this way, we could further optimize the estimation for the fraction f through the coverage information, while at the same time better inferring the type of aneuploidy. The likelihood of the coverage data over all sites is

$$P(cov|p,r) = \prod_{i=1}^{nRegion} NegativeBinomial(cov_i, prob = p, size = r)$$

To combine information from the AAF and coverage models, we take the combined likelihood as

$$L = L_{AAF} \times L_{Cov} = P(Z|N, \alpha, \beta) \times P(cov|p,r)$$

MCMC sampling

Having the likelihood function and prior set for each model, the posterior distribution is sampled through Markov Chain Monte Carlo (MCMC). The sampling process is done using JAGS in the R package *rjags* (https://cran.r-project.org/web/packages/rjags/index.html). The script for the model is included in the *MADSEQ* package at Bioconductor (see Supplemental Material). For all the sample and computational simulation, we set the burn-in steps to 10,000, and we sampled two chains, each with a total of 10,000 steps and each step sampled at every two steps. The convergence of the two chains is checked using the Gelman and Rubin diagnostic with the *coda* package in R (https://cran.r-project.org/web/packages/coda/index.html).

Model comparison

After we get the posterior distribution from MCMC sampling, the goodness of fit of models is compared using the Bayesian information criterion. The exact maximum likelihood of each model cannot be calculated directly from the MCMC procedure because of the hierarchical nature of the model, so we take a point estimate of the likelihood for each model using the median of each parameter from the posterior distribution (Ntzoufras 2009).

Ultimately, the model with the lowest BIC is preferred as the best model, and the type and fraction of aneuploidy are estimated from the posterior distribution of the best model. If the ΔBIC between the selected model and other models is less than 10, then we consider the chosen model to be low confidence.

Computational simulations

We aimed to evaluate the performance of our model as a function of sequencing depth, the type and fraction of aneuploidy cells,

and the number of heterozygous sites sequenced for one chromosome. We randomly generate data as follows:

1. **Simulation of coverage**. Given the expected normal coverage m_{cov} and the fraction of the aneuploid cells f, the average coverage for the chromosome m_{cov} can be calculated as described above. The sequencing depth cov_i for the ith site was randomly drawn from the negative binomial distribution

$$Cov_i \sim nbinom\left(m_{cov}, prob = \frac{m_{cov}}{var_{cov}}, r = \frac{m_{cov}^2}{var_{cov} - m_{cov}}\right)$$

We set the variance of the coverage (var_{cov}) to 30× m_{cov} based on what we observed from the actual sequencing data. The total number of targeted regions ($nRegions$) was fixed as 1.5 times the total number of heterozygous sites (K).

2. **Simulation of AAFs**. Given the type of aneuploidy, the fraction of abnormal cells (f), the total number of heterozygous sites (K) on one chromosome, and the midpoint AAF (m) across all heterozygous sites. The mean AAF for each mixture (μ_j) is easily calculated using the formula described in the model section. The weight for each mixture (ω_j) is given in the same way as in the model; we randomly assigned $\omega_j K$ sites into the jth mixture. Knowing the average coverage for the simulated chromosome (m_{cov}), the read depth for the ith heterozygous site (N_i) is also random drawn from the negative binomial distribution

$$N_i \sim nbinom\left(m_{cov}, prob = \frac{m_{cov}}{var_{cov}}, r = \frac{m_{cov}^2}{var_{cov} - m_{cov}}\right)$$

The alternate allelic depth for the ith site (AD_i) is randomly generated from the binomial distribution as

$$AD_{ij} \sim binom(N_{ij}, prob = \mu_j)$$

3. **Simulation of noise**. To account for the influence of noise in real sequencing data, we randomly selected 1% of the sites to have an alternate allele frequency drawn from a uniform beta distribution. We also randomly picked 1% of the regions to have random coverage uniformly spanning from one read to the maximum amount of coverage. When testing the false positive rate, we increased the noise level to 10% instead.

Since we only use sites that are genotyped as heterozygous to estimate the distribution of AAFs, we have to consider the capacity of the genotyping algorithms to call heterozygous sites. In general, genotyping algorithms will call a site as heterozygous if there are multiple reads supporting each allele. We therefore filtered out sites with fewer than three reads supporting both the alternate and reference alleles and sites whose AAFs are less than 0.02 or greater than 0.98 from the simulated data. We simulated 500 sets of data for each aneuploidy scenario.

Exome sequencing data from the 1000 Genomes Project

The BAM files of exome sequencing data of 2535 individuals from the 1000 Genomes Project were downloaded from the FTP site of the 1000 Genomes Project: ftp://ftp.1000genomes.ebi.ac.uk/. The BED file containing the targets of the exome sequencing was downloaded from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/exome_pull_down_targets/.

Design of multi–ethnic targeted loci

We show the steps involved in Supplemental Figure S9. Genotyping data of the 1000 Genomes Project were downloaded from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

First, we kept only biallelic loci. The heterozygosity rate for each locus was then calculated as

$$\text{heterozygosity rate} = \frac{\text{individuals genotyped as heterozygous}}{\text{total number of individuals}}$$

Loci with a heterozygosity rate greater than 0.4 were retained. Loci located in repetitive regions annotated by the RepeatMasker (rmsk) track from the UCSC Genome Browser were excluded. Loci located within ±20 bp of known indels and within ±500 bp of gaps were excluded. We used the hg19 version of the human genome as this is the version used by Roche-NimbleGen for their probe capture designs. In order to decrease (G + C) content bias, we removed loci located within extreme (G + C) content regions ([G + C] < 0.3 or [G + C] > 0.65, 200 bp context).

According to the computational simulation, the model can achieve very high sensitivity when there are ≥2000 heterozygous sites sequenced on each chromosome. As the mean heterozygosity for the loci we retained was 0.45, we aimed to keep 5000 loci per chromosome for the targeted sequencing.

In order to make loci evenly distributed along the chromosome instead of forming clusters, we binned each chromosome into 500 equal-sized windows using BEDTools (Quinlan and Hall 2010). We then randomly selected ~10 loci from each window. In total, we created a list containing 105,703 common SNPs for further capture. We used a Q-Q plot to show that there was no clustering of loci compared to randomly selected loci.

Alignment, genotyping, and data processing

Raw FASTQ files from the sequencing were aligned to the GRCh37 human reference genome using BWA-MEM (v0.7.10) in default paired-end mode (Li 2013). Picard (v1.119) was used to mark duplicates, and *GATK* (v3.4-46) (McKenna et al. 2010) was used for indel re-alignment and base recalibration following best practices. HaplotypeCaller was used to call variants and to genotype all of the targeted sites.

Estimation of ancestries of samples

We performed principal component analysis (PCA) using EIGENSTRAT (Price et al. 2006) across 12 samples sequenced by meMAD-seq assay and 2054 samples (26 populations) in the 1000 Genomes Project. SNPs used for PCA analysis are 1000G SNPs located in the captured regions of the meMAD-seq assay. VCFtools (Danecek et al. 2011) is used to process the SNPs.

MADSEQ model application

Having aligned the BAM file, genotyped the VCF file, and prepared the BED file containing the targeted regions, we used the *MADSEQ* package to correct for GC bias and filter noise, running the *MADSEQ* model as described in the documentation on Bioconductor.

Statistical analysis

We performed a binomial test and χ^2 test to test enrichment and association between detected aneuploidy and other factors. All the statistical testing was performed in R (v3.2) (R Core Team 2012).

Software availability

The *MADSEQ* package is available from Bioconductor: http://bioconductor.org/packages/MADSEQ/. *MADSEQ* source code is also available in the Supplemental Material. The meMAD-seq capture design is available from Roche-NimbleGen: Catalog No. 067402 60001, Design Identifier: 160407_HG19_MadSeq_EZ_HX1.

Data access

Acknowledgments

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74.

Abu-Amero KK, Hellani AM, Salih MA, Seidahmed MZ, Elmalik TS, Zidan G, Bosley TM. 2010. A *de novo* marker chromosome derived from 9p in a patient with 9p partial duplication syndrome and autism features: genotype-phenotype correlation. *BMC Med Genet* **11:** 135.

Ballif BC, Rorem EA, Sundin K, Lincicum M, Gaskin S, Coppinger J, Kashork CD, Shaffer LG, Bejjani BA. 2006. Detection of low-level mosaicism by array CGH in routine diagnostic specimens. *Am J Med Genet A* **140:** 2757–2767.

Bazrgar M, Gourabi H, Valojerdi MR, Yazdi PE, Baharvand H. 2013. Self-correction of chromosomal abnormalities in human preimplantation embryos and embryonic stem cells. *Stem Cells Dev* **22:** 2449–2456.

Ben-David U, Arad G, Weissbein U, Mandefro B, Maimon A, Golan-Lev T, Narwani K, Clark AT, Andrews PW, Benvenisty N, et al. 2014. Aneuploidy induces profound changes in gene expression, proliferation and tumorigenicity of human pluripotent stem cells. *Nat Commun* **5:** 4825.

Berko ER, Suzuki M, Beren F, Lemetre C, Alaimo CM, Calder RB, Ballaban-Gil K, Gounder B, Kampf K, Kirschen J, et al. 2014. Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. *PLoS Genet* **10:** e1004402.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14:** 307–320.

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16:** 195.

Chen CP, Lin SP, Su JW, Lee MS, Wang W. 2012. Phenotypic features associated with mosaic tetrasomy 9p in a 20-year-old female patient include autism spectrum disorder. *Genet Couns* **23:** 335–338.

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.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158.

Daniel A, Wu Z, Darmanian A, Malafiej P, Tembe V, Peters G, Kennedy C, Adès L. 2004. Issues arising from the prenatal diagnosis of some rare trisomy mosaics—the importance of cryptic fetal mosaicism. *Prenat Diagn* **24:** 524–536.

Dong Z, Zhang J, Hu P, Chen H, Xu J, Tian Q, Meng L, Ye Y, Wang J, Zhang M, et al. 2016. Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med* **18:** 940–948.

Draper JS, Smith K, Gokhale P, Moore HD, Maltby E, Johnson J, Meisner L, Zwaka TP, Thomson JA, Andrews PW. 2004. Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat Biotechnol* **22:** 53–54.

Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang H-Y, Humphray SJ, Halpern AL, et al. 2017. A reference data set of 5.4 million phased human variants validated by genetic

inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27:** 157–164.

Einhorn S, Burvall K, Juliusson G, Gahrton G, Meeker T. 1989. Molecular analyses of chromosome 12 in chronic lymphocytic leukemia. *Leukemia* **3:** 871–874.

Faggioli F, Vijg J, Montagna C. 2014. Four-color FISH for the detection of low-level aneuploidy in interphase cells. *Methods Mol Biol* **1136:** 291–305.

Fryburg JS, Dimaio MS, Yang-Feng TL, Mahoney MJ. 1993. Follow-up of pregnancies complicated by placental mosaicism diagnosed by chorionic villus sampling. *Prenat Diagn* **13:** 481–494.

Hahnemann JM, Vejerslev LO. 1997. European collaborative research on mosaicism in CVS (EUCROMIC)—fetal and extrafetal cell lineages in 192 gestations with CVS mosaicism involving single autosomal trisomy. *Am J Med Genet* **70:** 179–187.

Hodge JC, Hulshizer RL, Seger P, St Antoine A, Bair J, Kirmani S. 2012. Array CGH on unstimulated blood does not detect all cases of Pallister–Killian syndrome: a skin biopsy should remain the diagnostic gold standard. *Am J Med Genet A* **158A:** 669–673.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

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:** 651–658.

Knouse KA, Wu J, Whittaker CA, Amon A. 2014. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci* **111:** 13409–13414.

Kooper AJA, Faas BHW, Feuth T, Creemers JWT, Zondervan HH, Boekkooi PF, Quartero RWP, Rijnders RJP, van der Burgt I, van Kessel AG, et al. 2009. Detection of chromosome aneuploidies in chorionic villus samples by multiplex ligation-dependent probe amplification. *J Mol Diagn* **11:** 17–24.

Kostanecka A, Close LB, Izumi K, Krantz ID, Pipan M. 2012. Developmental and behavioral characteristics of individuals with Pallister–Killian syndrome. *Am J Med Genet A* **158A:** 3018–3025.

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:** 642–650.

Lebo RV, Novak RW, Wolfe K, Michelson M, Robinson H, Mancuso MS. 2015. Discordant circulating fetal DNA and subsequent cytogenetics reveal false negative, placental mosaic, and fetal mosaic cfDNA genotypes. *J Transl Med* **13:** 260.

Lestou VS, Kalousek DK. 1998. Confined placental mosaicism and intrauterine fetal growth. *Arch Dis Child Fetal Neonatal Ed* **79:** F223–F226.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

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.

Malvestiti F, Agrati C, Grimi B, Pompilii E, Izzi C, Martinoni L, Gaetani E, Liuti MR, Trotta A, Maggi F, et al. 2015. Interpreting mosaicism in chorionic villi: results of a monocentric series of 1001 mosaics in chorionic villi with follow-up amniocentesis. *Prenat Diagn* **35:** 1117–1127.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.

Ntzoufras I. 2009. *Bayesian modeling using WinBUGS*. Wiley, Hoboken, NJ.

O'Keefe C, McDevitt MA, Maciejewski JP. 2010. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood* **115:** 2731–2739.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

R Core Team. 2012. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rabinowitz M, Ryan A, Gemelos G, Hill M, Baner J, Cinnioglu C, Banjevic M, Potter D, Petrov DA, Demko Z. 2012. Origins and rates of aneuploidy in human blastomeres. *Fertil Steril* **97:** 395–401.

Ryan JL, Kaufmann WK, Raab-Traub N, Oglesbee SE, Carey LA, Gulley ML. 2006. Clonal evolution of lymphoblastoid cell lines. *Lab Invest* **86:** 1193–1200.

Sandin S, Hultman CM, Kolevzon A, Gross R, MacCabe JH, Reichenberg A. 2012. Advancing maternal age is associated with increasing risk for autism: a review and meta-analysis. *J Am Acad Child Adolesc Psychiatry* **51:** 477–486.e1.

Schick UM, McDavid A, Crane PK, Weston N, Ehrlich K, Newton KM, Wallace R, Bookman E, Harrison T, Aragaki A, et al. 2013. Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS One* **8:** e59823.

Smith K, Lowther G, Maher E, Hourihan T, Wilkinson T, Wolstenholme J. 1999. The predictive value of findings of the common aneuploidies, trisomies 13, 18 and 21, and numerical sex chromosome abnormalities at CVS: experience from the ACC U.K. Collaborative Study. Association of Clinical Cytogeneticists Prenatal Diagnosis Working Party. *Prenat Diagn* **19:** 817–826.

Taapken SM, Nisler BS, Newton MA, Sampsell-Barron TL, Leonhard KA, McIntire EM, Montgomery KD. 2011. Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. *Nat Biotechnol* **29:** 313–314.

Theisen A, Rosenfeld JA, Farrell SA, Harris CJ, Wetzel HH, Torchia BA, Bejjani BA, Ballif BC, Shaffer LG. 2009. aCGH detects partial tetrasomy of 12p in blood from Pallister–Killian syndrome cases without invasive skin biopsy. *Am J Med Genet A* **149A:** 914–918.

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* **282:** 1145–1147.

van Echten-Arends J, Mastenbroek S, Sikkema-Raddatz B, Korevaar JC, Heineman MJ, van der Veen F, Repping S. 2011. Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Hum Reprod Update* **17:** 620–627.

Vattathil S, Scheet P. 2016. Extensive hidden genomic mosaicism revealed in normal tissue. *Am J Hum Genet* **98:** 571–578.

Wilkins-Haug L, Quade B, Morton CC. 2006. Confined placental mosaicism as a risk factor among newborns with fetal growth restriction. *Prenat Diagn* **26:** 428–432.