HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2020 July 06.

Published in final edited form as:

Nat Biotechnol. 2020 March ; 38(3): 314–319. doi:10.1038/s41587-019-0368-8.

Accurate detection of mosaic variants in sequencing data without matched controls

Yanmei Dou[1], **Minseok Kwon**[1], **Rachel E. Rodin**[2,3,4,5], **Isidro Cortés-Ciriano**[1,†], **Ryan Doan**[2,3,4], **Lovelace J. Luquette**[1,6], **Alon Galor**[1], **Craig Bohrson**[1,6], **Christopher A. Walsh**[2,3,4], **Peter J. Park**[1,7,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[2]Division of Genetics and Genomics, Manton Center for Orphan Disease, and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA.

[3]Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA.

[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[5]Harvard/MIT MD-PhD Program, Harvard Medical School, Boston, MA, USA.

[6]Bioinformatics and Integrative Genomics PhD program, Harvard Medical School, Boston, MA, USA

[7]Ludwig Center at Harvard, Boston, MA, USA

Abstract

Detection of mosaic mutations that arise in normal development is challenging, as such mutations are typically present in only a minute fraction of cells and there is no clear matched control for removing germline variants and systematic artifacts. We present MosaicForecast, a machine-learning method that leverages read-based phasing and read-level features to accurately detect mosaic single-nucleotide variants (SNVs) and indels, achieving a multifold increase in specificity compared to existing algorithms. Using single-cell sequencing and targeted sequencing, we validated 80–90% of the mosaic SNVs and 60–80% indels detected in human brain whole-genome sequencing data. Our method should help elucidate the contribution of mosaic somatic mutations to the origin and development of disease.

*Correspondence to: peter_park@hms.harvard.edu.

†Present address: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, United Kingdom

AUTHOR CONTRIBUTIONS

Y.D. developed the algorithm and performed the analysis, under supervision by P.J.P.. M.K. generated call sets from MuTect2 and GATK haplotype callers. R.E.R. and R.D. evaluated candidate sites with targeted sequencing, supervised by C.A.W.. I.C.C., L.J.L., A.G., C.B., and M.K. helped refine the algorithm and contributed to editing of the manuscript. Y.D. and P.J.P. wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

A single individual harbors multiple populations of cells with distinct genotypes due to somatic mutations arising post-zygotically[1]. Such diversity of genotypes in an individual is referred to as somatic mosaicism. Analysis of mosaic mutations in non-disease samples enables exploration of lineage patterns during development and characterization of mutational mechanisms operative in normal cells[2–6]. Recent studies have also demonstrated that somatic mutations contribute to many diseases besides cancer[1, 7–11].

Identification of mosaic mutations in genome sequencing data remains challenging in two key aspects. First, whereas functionally-relevant cancer mutations typically confer proliferative advantage and thus have relatively high variant allele fractions (VAFs), most mosaic mutations are present in a small number of cells and have very low VAFs. In the extreme case, those occurring in post-mitotic cells are present only in a single cell and are detectable only by single cell sequencing, as we have done recently[10]. As standard cancer mutation callers typically have a lower VAF limit of 2–5%[12, 13], detection of mutations with lower VAFs requires a more sensitive bioinformatic method and/or higher sequencing depth. Second, mosaic mutations that arise early in development generally exist in multiple tissues. [2, 14] Thus, the conventional approach of using a paired control tissue for filtering germline variants and systematic errors would exclude such early-occurring mutations. Several methods have been employed to detect mosaic SNVs from non-tumor tissues, such as the use of a germline variant caller[15] with higher ploidy assumptions[8] or a combination of somatic mutation callers[3, 7, 9, 14]. Additional filtering leveraging trio data to exclude germline variants[7–9, 15] is also common. However, validation rates in these studies have been modest.

Incorporation of read-level features in a flexible framework is critical for distinguishing real mutations from artifacts[16, 17]. In place of filters with hard thresholds, recent methods such as DeepVariant[16] and Strelka2[17] use machine-learning to combine relevant read-level features to improve detection of germline and cancer somatic variants, respectively. Another component in accurate detection of mosaic SNVs in silico is read-based phasing[3, 7, 8, 18], in which a candidate mosaic mutation and a nearby germline variant are checked for haplotype consistency—i.e., a true mosaic mutation should generate one and only one additional haplotype. A major disadvantage of phasing, however, is that only a small fraction (~10–30%) of variants are phasable using short-read sequencing[18], and phasing may be ambiguous in non-diploid or low-mappability regions[6].

We developed MosaicForecast, which leverages multiple read-level features over phasable sites to build a genome-wide prediction model for finding mosaic mutations in the absence of a matched control sample. It consists of three major steps (Fig. 1a): (i) generation of a training set by read-based phasing; (ii) construction of a Random Forest (RF) model based on read-level features related to the quality and category of variants, such as variant allele fraction, read depth, mismatches per read and strand bias (Fig. 1b, Supplementary Table 1); and (iii) genome-wide prediction of mosaic SNVs. The underlying idea is similar to that of DeepVariant[16] and Strelka2[17] in that a nonlinear model that combines informative read-level features is trained using a machine-learning framework and then applied to a test set. The main difference is that, to overcome the problem that high-quality training data are not available, MosaicForecast utilizes phasable sites in building a training set. We introduce

another modeling step using multinomial logistic regression to improve the training set when some experimental validation data are available (Fig. 1c). As an illustrative example, we applied the tool to analyze whole-genome sequencing (WGS) data from brain tissues of 60 autism spectrum disorder (ASD) and 15 neurotypical individuals, sequenced at ~250X (150-bp paired-end reads).

To assemble a training set of high-confidence mosaic mutations, we first identify a lenient set of candidate mosaic variants. We used MuTect2 in its tumor-only mode for its high sensitivity, but other algorithms can be used (see Methods and Supplementary Fig. 1). To remove germline variants and recurrent artifacts, we filtered variants present in the Genome Aggregation Database (gnomAD)[19]. In addition, since the likelihood that somatic mutations occur at the same position in different individuals is vanishingly small, we also removed variants found in any other samples in the dataset (75 minus the one being analyzed). We observed that removing recurrent variants did not result in loss of sensitivity (Supplementary Fig. 2). For some experimental designs, e.g., comparing multiple tissues from the same individual, recurrent variants may be true mosaics; thus, a filtering scheme with an appropriate panel of “normals” should be chosen to remove germline variants as well as minimizing the risk of including artifacts that arise due to misalignment or index hopping[20].

We then classified the phasable variants (those for which a germline SNP is contained in the same read or its mate pair) into three categories depending on the number of observed haplotypes (hap): (i) hap=2, consistent with heterozygous germline variants; (ii) hap=3, consistent with mosaics; and (iii) hap>3, suggestive of low mappability regions, presence of copy number variations (CNVs), or sequencing-associated/other artifacts (Supplementary Fig. 3). For our brain data, ~25% of candidate mosaics were phasable with at least one germline SNP (Supplementary Table 2).

To determine whether the true genotypes can be inferred from the haplotype categories, we evaluated 483 phasable sites in selected samples for which three additional data types are available: single cell WGS, amplicon-based targeted sequencing, and trio WGS (see Methods and Supplementary Tables 2–3). The single cell WGS dataset of three individuals we have published previously[5, 10] provide an excellent resource for orthogonal validation, as the lineage information as well as allele fraction across cells allow us to distinguish mosaics from heterozygous SNPs, germline repeat/CNV variants and technical artifacts (see Methods; Supplementary Fig. 4); trio data (for two individuals) are useful for distinguishing mosaic and germline variants (Supplementary Table 3). This analysis (Fig. 1c) revealed that although the ‘hap=3’ category was enriched for true mosaic mutations (50%), the rest of the ‘hap=3’ sites turned out to be false positives, classified as repeat/CNV regions (37%), germline heterozygous (6%) and reference-homozygous (6%). Variants labeled as ‘hap=2’ were mostly germline heterozygous, and variants labeled as ‘hap>3’ were mostly false positives as expected.

To identify mosaic variants, we first built an RF model using over 30 read-level features as predictors and the haplotype number (Hap=2, Hap=3, Hap>3) as the response, at all phasable sites on diploid chromosomes (Supplementary Table 4). Then we applied this ‘Phasing prediction model’ genome-wide, excluding non-unique mapping regions[21] (see

Methods). This model resulted in modest validation rates for ‘hap=3’ sites, with 67% (55/82) in non-repeat regions and 34% (28/82) within repeat regions (Supplementary Fig. 5 and Supplementary Table 5; we define ‘repeat region’ to include interspersed repeats and low complexity sequences identified by RepeatMasker[22]). However, we noticed that many variants are clustered, in mostly repeat or non-diploid regions, when all individuals are considered together (~46% of those predicted to be ‘hap≥3’ were enriched in regions that together span only ~19MB; see Methods; Supplementary Table 6). After removing these clustered variants, the overall validation rate for the Phasing prediction model increased to 76% (55/72) in non-repeat regions and 49% (28/57) in repeat regions (Fig. 1d, Supplementary Table 5). This constitutes a 7-fold increase (non-repeat regions) and a 43-fold increase (repeat regions) in precision compared to the initial MuTect2 calls, with minimal loss of sensitivity.

With experimental validation data at phasable sites, we can further improve our prediction model. Because the haplotype number was only moderately correlated with the mosaic status (Fig. 1c, top), we reasoned that an intermediate model that defines the genotype more accurately using validation data could generate a better training set for the subsequent RF model. Visual inspection in the principal component space of the read-based features revealed that some ‘hap=3’ variants clustered with variants that were found to be repeat/CNV or reference-homozygous, suggesting that read-level data can help refine the genotype predictions (Supplementary Fig. 6a–c). With a multinomial logistic regression model incorporating the read-level features (Fig. 1a), we converted the genotyping categories from haplotype counts to ‘het’, ‘mosaic’, ‘refhom’ and ‘repeat’. The refined categories were in much better agreement with the orthogonally evaluated calls: whereas only 50% of hap=3 variants were validated mosaics, 85% of the ‘mosaic’ predictions from the regression model were validated mosaics (Fig. 1c). The resulting model was then applied to all phasable sites to generate their four-category genotype labels (Supplementary Table 4).

Using the phasable sites and their refined genotypes as a training set, we predicted mosaics genome-wide. We built an RF classification model (‘Refined genotypes prediction model’) on all phasable sites with over 30 features as covariates and the refined genotypes as the response. We then applied the RF model to the 135,250 non-phasable candidate mutations (Supplementary Table 7). Sites within non-unique mapping regions[21] (mappability score=0) as well as sites within clustered regions (Supplementary Fig. 6d) were excluded. Among the 2,220 predicted (non-phasable) mosaics, 95 randomly selected sites were evaluated using orthogonal data (same validation method as for phasable sites). As shown in Fig. 2a, 78 (82%) were confirmed as true mosaics (85% in non-repeat regions and 77% in repeat regions). Top-ranked features of the RF model are listed in Supplementary Fig. 6e.

We compared the performance of MosaicForecast with that of GATK HaplotypeCaller (GATK-HC)[23], MuTect2[12] and MosaicHunter[24] using three different approaches. First, we inspected the variants called by all methods in three 250X WGS brain samples for which single cell WGS data are available[5, 10] (both phasable and non-phasable sites; leave-one-out cross validation for three individuals). Although the lineage information in single cells provides a very useful way to benchmark algorithm performance, one limitation of this approach is that variants with low allele fraction have a proportionally low chance of being

sampled if the number of cells is small; thus, we used deep-sequencing on the IonTorrent platform to further examine those variants identified as 'refhom' by single cell data (see Methods; Supplementary Fig. 7 and Supplementary Table 8). The results show that MosaicForecast-Phase and -Refine models achieve precision that is typically several-fold higher than other tools, while maintaining high sensitivity (Fig. 2a). GATK-HC with ploidy two (GATK-HC-p2) frequently misclassified heterozygous SNPs as mosaics; MuTect2 and GATK-HC with ploidy five (GATK-HC-p5) most often misclassify repeat/CNV variants; and MosaicHunter could only detect variants within non-repeat regions, thus losing ~50% of true mosaics. At the individual level (Fig. 2b), the precision was ~92% (24/26), ~81% (25/31) and ~73% (24/33) for the MosaicForecast-Refine model, suggesting a consistently high validation rate. MosaicForecast was also able to detect more low-allele fraction variants with VAF≤0.05 (30/41) than MosaicHunter (14), GATK-HC-p5 (4) and GATK-HC-p2 (0) (Supplementary Fig. 8a).

As a second mode of validation, we evaluated candidate mosaics called by MosaicForecast-Refine in the 75 individuals using amplicon-based sequencing (~30,000X on IonTorrent). Of the 75, the IonTorrent validation rate (Supplementary Table 9) was ~94% (161/171) for the 64 samples that were sequenced using PCR-free libraries (~95%, 149/157, for diploid and ~86%, 12/14, for haploid chromosomes). For the remaining 11 sequenced using PCR-based libraries, the validation rate was ~61% (42/68; 68%, 42/62, on diploid and 0/6 for haploid chromosomes). The lower validation rate for the PCR-based samples is likely due to the PCR-induced biases, as reflected in a significantly higher proportion of G>T mutations (OR=4.1, p<1e-15, Fisher's Exact test, Supplementary Fig. 8b), which are associated with oxidative damage[25]. If we focus on non-phasable sites from diploid chromosomes (Fig. 3a), validation rates were ~95% (105/111) and ~67% (30/45) for PCR-free and PCR-based samples, respectively. In addition, the validation rates were similar in non-repeat regions (87%, 118/136) and repeat regions (85%, 17/20). Among the 177 MosaicForecast mosaics in non-repeat regions confirmed by IonTorrent, GATK-HC-p5 was only able to detect ~62% (109/177), followed by MosaicHunter (~59%, 105/177), and GATK-HC-p2 (~20%, 35/177). Among the 26 MosaicForecast mosaics in repeat regions confirmed by IonTorrent, GATK-HC-p5 and -p2 were only able to detect ~58% (15/26) and ~19% (5/26), respectively (MosaicHunter does not make calls in repeat regions). A large fraction of low VAF (≤0.05) mosaics were called by MosaicForecast but missed by both MosaicHunter and GATK HaplotypeCaller (~52%, 48/92, Supplementary Fig. 8c), indicating that MosaicForecast is particularly advantageous for detecting low VAF mutations.

Third, we tested the haplotype numbers for the extra variants identified by the other callers. Across the 75 individuals, the other methods called 1 to 80 times more mutations than MosaicForecast, but read-based phasing showed that a large proportion of phasable sites from these tools had two haplotypes or more than three haplotypes, inconsistent with mosaic variants (Fig. 1c and Fig. 2a). For example, the percentages of 'hap>3' variants were 58%, 49%, and 26% for MuTect2, GATK-HC-p5, and GATK-HC-p2, respectively; another 51% of GATK-HC-p2 were 'hap=2'. These numbers indicate that the false positive rates are indeed very high for other methods (Supplementary Fig. 9).

To determine the performance of our model across VAFs, we applied the model to a simulated dataset containing spike-in mutations (see Methods). We found that the model had similarly good performance over a relatively wide range of AFs, from 0.02 to 0.3, as reflected in the ROC curves (Supplementary Fig. 10a–b). It performed substantially worse when the AFs approach 0.5, as it becomes impossible to separate somatic variants from germline variants; in that case, a case-control scheme would be a better choice.

To examine the detection power of MosaicForecast as a function of read-depth, we simulated lower coverage data by down-sampling from the original 250X brain-WGS data, and trained one RF model at each read-depth using only the features extracted from the phasable sites in the corresponding down-sampled data. Although power decreases with coverage as expected, MosaicForecast was still able to detect ~80% (108/135) of the validated 250X variants at 50X (Fig. 3a, Supplementary Table 10). We also applied the brain WGS-trained models to the HapMap sample NA12878 to determine whether a model trained in one dataset could be applied to another dataset. For testing, we generated simulated mutations in the 300X WGS data for NA12878[26] following a realistic allele fraction distribution of early-embryonic mosaics and down-sampled to 50-250X (see Methods and Supplementary Fig. 10c–d). MosaicForecast was sensitive in detecting simulated mosaics and effective in removing non-mosaic sites across read depths: at a 5% false discovery rate, it detected ~95% of the spike-in mutations at 250X. When the training and simulation were performed at lower depths, ~90% of the spiked-in mutations were detected at 100X and ~70% at 50X (Fig. 3b). We also found that the models were robust across different read depths (see Methods, Supplementary Fig. 11).

Although MosaicForecast used variants derived from MuTect2 as an initial set, it could also start with variants identified by other tools. Validation using single cell and IonTorrent data (see Methods and Supplementary Table 11) shows that MosaicForecast (trained on MuTect2 calls) substantially raises the specificity of mosaic mutations with a minor loss in sensitivity, from 39% (38/98) to 90% (27/30) for GATK-HC-p2, from 7% (54/742) to 84% (42/50) for GATK-HC-p5, and from 47% (34/73) to 61% (31/51) for MosaicHunter (Supplementary Fig. 12). For maximal sensitivity, we could generate an input set simply by using samtools mpileup scanning, e.g., by taking all sites with ≥2% non-reference bases as potential mutations. Using single cell data to evaluate the sites called only by samtools, only a tiny fraction (1.8%, 104/5876) of a large mutation set was validated. With MosaicForecast, we achieved a striking improvement in the validation rate (45.9%, 78/170; Supplementary Fig. 12). Compared to the MuTect2 validation result (73/89, Fig. 2a), only a few more true variants were captured at the expense of many false variants. We note that the single cell-based validation strategy becomes less accurate as the VAF decreases, so the specificity and sensitivity for mpileup-based variants is likely to be an underestimate.

In addition to SNVs, MosaicForecast is capable of identifying mosaic indels (see Methods). Using the MuTect2-based approach as before, we obtained 59977 candidates (22893 deletions, 37084 insertions) from the non-repeat regions of the 75 individuals. For mosaic deletions, an RF model trained using all 831 phasable sites (Supplementary Table 12 and Supplementary Fig. 13a–c) predicted 1356 sites as “hap=3”. With additional filtering criteria (see Methods), 102 sites were classified as confident mosaic deletions. All the high-

confident mosaic deletions were from PCR-free samples. When evaluated using IonTorrent sequencing (see Methods and Supplementary Fig. 13d), ~75% (59/79) were validated as true deletions, with phasable (79%, 11/14) and non-phasable (~74%, 48/65) sites having similar validation rates (Fig. 3c and Supplementary Table 13). Two sites in the three individuals with single cell data (and at least one mutant cell) were both confirmed with lineage information (Supplementary Fig. 13e–f). Following the same approach (Supplementary Fig. 14), 134 confident mosaic insertions were found, and ~59% (32/54) from PCR-free samples were validated (Fig. 3d and Supplementary Table 14). None of the mosaic insertions from PCR-based samples were validated (Supplementary Table 14). The excessive error rate of mosaic indels in PCR-based samples are likely to be caused by replication slippage of DNA polymerase[27]. In the process of detecting mosaic SNVs and indels, we also identified several mosaic multi-nucleotide variants (e.g., two adjacent base substitutions) as well as clumped variants (e.g., a SNV and a nearby insertion). We called three such events in the three individuals with single cell sequencing data, and two of the events were found in at least one mutant cell, suggestive of true mosaics (Supplementary Fig. 15).

In summary, MosaicForecast substantially improves the detection of mosaic SNVs and indels from reference-free sequencing data, confirming that proper incorporation of various read-level features in a nonlinear classifier provides an effective way to distinguish real mosaic mutations from germline variants (especially from CNV/repeat regions) and other artifacts. The strong performance of MosaicForecast is made possible by training predictive models on phasable sites—a method for constructing a highly-confident training set of mosaic variants in-silico, without having to carry out labor-intensive experimental validations. Identification of mosaic mutations in various non-tumor tissues by the proposed method will help gain insights into the origin and propagation of somatic mutations in development and disease.

ONLINE METHODS

Data sets

WGS data from the prefrontal cortex of 75 individuals (~250X) and from the blood of two pairs of parents (~50X) were obtained by our group. Single cell WGS data from the neurons of three individuals were previously obtained[5, 10]. All data were 150bp, paired-end (PE) reads. FASTQ files and high-confidence SNV calls[28] for NA12878 (~300X, 150bp PE) generated by the Genome in a Bottle Consortium (https://ftp-trace.ncbi.nlm.nih.gov/giab/) were down-sampled to different read depths. WGS data for 30 tumor tissues (~80X, 100 bp PE) and their matched normal samples (~40X) as well as consensus variant calls were obtained from the International Cancer Genome Consortium (ICGC)[13]. Reads were aligned to GRCh37 with decoy (hs37d5) using BWA[29].

Variant calling

Five methods were used to call mosaic SNVs in the brain bulk data: MuTect2 (version 3.5 nightly-r2016-04-25-g7a7b7cd, variants tagged as “str_contraction”, triallelic_site or “t_lod_fstar” were excluded); GATK HaplotypeCaller (v3.5-0), with ploidy set to two (GATK-HC-p2) or five (GATK-HC-p5), keeping only variants tagged as ‘PASS’;

MosaicHunter (v1.1) with the minimum VAF threshold adjusted from the default 5% to 2% to enable detection of mosaics with lower VAFs; and Samtools[30] (v1.9) mpileup with sites with ≥2% non-reference bases (alt allele count ≥3, with mapping quality≥20 and base quality ≥20) called as putative mutations. Variants within segmental duplication regions according to the UCSC Genome Browser database[31] were removed. Variants at <0.02 VAF calculated by each tool were excluded. Variants with VAF ≥0.4 or present in the gnomAD database[19] were removed as likely germline variants. Variants called in multiple individuals by each method were excluded. For indels, variants present in the relevant "panel of normals" (PoN) or gnomAD as well as those present in repeat regions, including simple repeats[32], RepeatMasker regions, and segmental duplication regions were excluded. Variants with VAF <0.02 or ≥0.4 as calculated by MuTect2 were removed. Outliers with >20 deletions per sample were excluded, variants with >350X read depth, variants with ultra-low AF calculated by MosaicForecast (<0.01), variants with a high proportion (≥10%) of clipped reference reads, variants within clustered regions and variants present in the gnomAD database were removed. Further information is available in the Life Sciences Reporting Summary.

Variants generated by MuTect2 with a PoN were used as input variants to MosaicForecast, due to the high sensitivity of MuTect2. For brain data, one individual (UMB5308) likely to have contamination problems (obtain excess number of low-VAF mutations) was excluded. Sites with extra-high read depths (≥2X), sites with ≥1.5X read depths and ≥20% AF were marked as "low-confidence" and excluded. We evaluated the sensitivities of different tools in two ways: first, we generated spike-in mutations at allele fractions of 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 in the BAM files of NA12878 subsampled at 50X-250X, called variants from the BAM files with MuTect2, MosaicHunter, GATK-HC-p2 and GATK-HC-p5, and compared their sensitivities in detecting mosaics at different allele fractions and read depths. MuTect2 achieved the highest sensitivity in all circumstances (Supplementary Fig. 1a). Second, we also called candidate mosaics from real bulk sequencing data (250X) from three individuals with multiple single cells with the four different tools, and evaluated the variants using the single cell data. MuTect2 was able to detect >97% (98/101) of real mosaics called by different tools, whereas MosaicHunter, GATK-HC-p2, GATK-HC-p5 were only able to detect 34, 38 and 54 percent of real mosaics, respectively (Supplementary Fig. 1b).

Orthogonal validation of variants

SNV candidates were evaluated by single cell sequencing, trio sequencing, or ultra-deep amplicon resequencing (Supplementary Table 3). To evaluate variants using single cell data in the three individuals (1465, 4643 and 4638)[5, 10], we constructed lineage trees with single cell mutations assigned to different clades (Supplementary Table 8 and Supplementary Fig. 4). Mutations that were only present in the cells assigned to the same clade were regarded as real mosaic mutations; mutations absent from all cells were regarded as refhom, and mutations present in multiple cells and assigned to conflicting clades were regarded as germline variants (het or repeat). To further classify germline variants (repeat or het), we used an empirical threshold: if mutant cells on average have ≥20% AF, we classified them as het variants; and if mutant cells on average have <20% AF, we classified then as repeat

variants (Supplementary Fig. 4e). We further checked all “het” variants, and if a variant had ultra-high read depth in the bulk data (>300X) and the bulk allele fraction deviated significantly from 50% ($p<0.001$, two-tailed binom test), we re-classified them as repeat variants. Moreover, “linked” variants close to each other with all alt alleles on the same reads were also classified as repeat variants.

To address the concern that some variants judged as “refhom” by the limited number of single cells could be real mosaic variants, we experimentally evaluated those “refhom” sites by using IonTorrent deep sequencing. But, in order to pick an informative set for IonTorrent validation, we first categorized the sites into different groups based on the caller used and checked the read alignments using IGV. By extensively checking IGV plots (with some examples in Supplementary Fig. 7b), we found that: 1. Candidate sites called by ≥2 tools were more likely to be true mosaics than those sites called by only one tool; 2. Candidate sites called by MosaicForecast and MosaicHunter were substantially more convincing, whereas sites called by GATK-HC-P5 and MuTect2 were much less so. Almost all of candidate mosaics from the latter set came from “messy” regions with many mismatches and were unlikely to be true mosaics. Based on observation, instead of experimentally evaluating all 674 sites evaluated by single cells as refhom, we only selectively evaluated sites called by ≥2 tools. We should note that since we did not have more of the same DNA extraction for the three individuals (1465, 4643, 4638) sent out for WGS sequencing, we could only settle for the second best by using extracted DNA from nearby tissues for targeted sequencing. As a result, the validation rate could be slightly under-estimated overall. But the situation is the same for all callers.

For trio data, we considered the variants detected in the child and (i) absent from both parents as real mosaics, (ii) present in either parent with <20% VAF as CNV/repeat, and (iii) present in either parent with ~50% VAF as heterozygous. Two individuals (UMB5939 and UMB5771) had bulk WGS data from both parents (~50X), and their variant calls were evaluated with trio data.

As for IonTorrent, with additional testing of candidates in the 75 individuals, we evaluated a total of 250 candidate mosaic SNPs called by MosaicForecast (Supplementary Table 9). To compare the performance of different tools, 115 variants from different tools were evaluated by ultra-deep targeted sequencing (Supplementary Table 3). These included 54 with WGS VAF≤0.05, 24 with WGS VAF∈ (0.05,0.2] and 37 with VAF>0.2. Candidate mosaic indels were also evaluated with IonTorrent targeted sequencing. For each site, two or three different pairs of primers were designed and three sets of PCR products were generated. In addition, an experimental control was adopted as a comparison with the case samples. Given that IonTorrent sequencing has a high rate of indel errors[33], only variants present in the case and absent from control, or present in case samples with much higher allele fractions than in controls were regarded as true mosaic indels (Supplementary Fig. 13d). For mosaic insertions specifically, since the “hap=3” and “hap>3” sites were not well distinguished in the PCA space (Supplementary Fig. 14), we expected higher false positive rate, and checked the read alignments using IGV as an additional filter before IonTorrent evaluation (Supplementary Table 14). Sites from “messy” regions with excessive mismatches, sites with the mutant alleles completely linked with nearby low-AF variants in the reads, and sites

with misalignment issues were classified as “repeat” or “het” variants. These false positive sites from IGV plot were included to calculate the validation rate of mosaic insertions in Fig. 3d.

Read-based phasing

To identify germline SNPs near the SNVs detected by MuTect2, we scanned reads mapped up to 1kb away from each candidate site. After excluding reads with low mapping qualities (<20) or with low base quality at the mutant position (<20), a two-tailed binomial test was applied to remove variants whose VAFs deviated from 0.5 ($p \leq 0.05$). Variants with relatively low read depth (<20X) were also filtered.

After obtaining a set of high-confidence SNPs, we first computed the haplotype numbers along the genome using consecutive pairs of germline SNPs to determine if a region is non-diploid; if so, candidate mosaics in the region were excluded as false positives (Supplementary Fig. 3a). Next, each candidate mosaic was phased with as many nearby germline SNPs as possible and classified as follows: (i) those leading to three haplotypes were treated as potential mosaics (Fig. 1a, Supplementary Fig. 3b); (ii) those leading to two haplotypes were treated as heterozygous mutations (Supplementary Fig. 3c); and (iii) those leading to more than three haplotypes were treated as false positives, as mosaic mutations arising in a diploid organism can only define three haplotypes (Fig. 1a, Supplementary Fig. 3d).

Read-level features

The 31 features are described in Supplementary Table 1. Two of the features, ‘mapq_p’ and ‘mapq_difference’, encode mapping qualities; three account for the number of mismatches per read (‘major_mismatches_mean’, ‘minor_mismatches_mean’, ‘mismatches_p’); and six are calculated using base qualities (‘baseq_p’, ‘baseq_t’, ‘ref_baseq1b_p’, ‘ref_baseq1b_t’, ‘alt_baseq1b_p’, ‘alt_baseq1b_t’). The remaining features are read-depth, VAF, genotyping likelihoods, strand bias, biases of the read pairs towards the ref/alt alleles, bias of the sequencing cycle towards the ref/alt alleles, read mapping position bias, bias of the base query position of the ref/alt alleles, local mappability score[21], proportion of clipped reads, multiallelic examination, GC content and the three-nucleotide base context of the mutation.

Although most of these features were calculated by comparing positions or qualities of reference alleles/reads with alternative alleles/reads, we also compared the qualities of alleles at the mutant position and at 1-bp downstream of the mutant position (‘ref_baseq1b_p’, ‘ref_baseq1b_t’, ‘alt_baseq1b_p’, ‘alt_baseq1b_t’), since systematic sequencing errors have been reported to have lower base quality at the mutant position[34]. We also estimated genotype likelihoods of four different genotypes (refhom, het, althom, mosaic) based on Bernoulli sampling[24, 35] to capture sequencing errors and ref/alt allele read-depth biases, assuming that the real mutant allele fractions are 0 (refhom), 0.5 (het), 1 (althom), and uniformly distributed between 0 and 1 (Mosaic). The formulas to calculate the four genotypes are as follows:

$$L(G = \text{het} \mid Data) = \binom{\text{depth}}{\text{r}} 0.5^{\text{depth}}$$

$$L(G = \text{refhom} \mid Data) = \binom{\text{depth}}{\text{r}} \prod_{\text{i}=1}^{\text{depth}} \text{P}(\text{r}_\text{i} = \text{ref} \mid \text{q}_\text{i}, \text{o}_\text{i})$$

$$L(G = \text{althom} \mid Data) = \binom{\text{depth}}{\text{r}} \prod_{\text{i}=1}^{\text{depth}} \text{P}(\text{r}_\text{i} = \text{alt} \mid \text{q}_\text{i}, \text{o}_\text{i})$$

$$L(G = \text{mosaic} \mid Data) = \int_0^1 \theta^{\text{r}} (1-\theta)^{\text{depth}-\text{r}} \text{d}\theta = \binom{\text{depth}}{\text{r}} beta(r+1,\ depth - r + 1)$$

$$\text{r} = \sum_{\text{i}=1}^{\text{depth}} \text{P}(\text{r}_\text{i} = \text{alt} \mid \text{q}_\text{i}, \text{o}_\text{i})$$

where r_i denotes *read i*, o_i denotes observed allele on *read i* at the mutant position, q_i denotes base quality on *read i* at the mutant position, and θ denotes the real mutant allele fraction.

Compared with SNVs, indels tend to cause alignment uncertainty problems and a merely position-based method would no longer be adequate. We therefore modified several read-level features and designed several new features/filters to adapt MosaicForecast for calling mosaic indels. Specifically, “alt reads” were re-defined as reads carrying an alt allele or reads clipped at the mutant position; candidate sites within Simple Repeats and homopolymer regions were filtered; candidate sites with >=10% reference reads being soft-clipped were filtered; and when computing the difference of baseQ at the mutant position and neighboring position, the 1bp neighboring position was re-defined as read regions excluding mutant indels. All the read-level features were computed using custom Python scripts that relied on the PySam[30] library (https://github.com/pysam-developers/pysam).

Identification of regions with clustered mutations

To identify non-diploid regions that are likely to be enriched for artifacts, we applied the Phasing prediction model on all MuTect2-PoN calls from the 75 individuals and extracted variants predicted to be ‘hap=3’ or ‘hap>3’. We then selected regions with ≥3 consecutive ‘hap≥3’ sites among the 75 individuals (distance <5,000 bp between adjacent variants). The clustered ‘hap=3’ variants in these mostly repeat regions had significantly higher read-depths ($p<1e-15$, two-tailed Wilcoxon’s rank sum test) and lower mappability scores ($p<1e-15$, two-tailed Wilcoxon’s rank sum test) than non-clustered ‘hap=3’ sites, suggesting that these regions are likely to be non-diploid or otherwise error-prone regions that should be blacklisted. Validation using single cell, IonTorrent, and trio data for 540 clustered “hap=3” sites showed that ~99% of them were false positives (Supplementary Table 6).

Genotype refinement

Given that the genotype cannot be inferred accurately when the haplotype number is three or more (Fig. 1c), we first performed principal component analysis (PCA) of all variants called by MuTect2 (tumor-only mode) using all read features to determine whether real mosaics can be distinguished from false positive sites in the PCA space (Supplementary Fig. 5). When we projected the experimentally-evaluated phasable sites onto the PCA space, we found that the variants validated as mosaic, heterozygous, reference-homozygous and repeat/CNV variants form distinct clusters (Supplementary Fig. 6a,c), suggesting that the read-level features could be used to separate real mosaic mutations from germline variants and other false positive calls with higher accuracy than haplotype information alone.

Analysis of the PCA space revealed that some of the candidate mosaic variants with hap=3 clustered with hap>3 variants. Validation data showed that those hap=3 variants were repeat/CNV or reference-homozygous (Supplementary Fig. 6a). For example, genotyping likelihoods, difference of ref/alt allele query position, difference of read mapping positions and difference of ref/alt reads mapping qualities were the main features contributing to PC1; difference of mismatches per ref/alt read and difference of ref/alt allele base qualities were important features contributing to PC2 (Supplementary Fig. 6b). Repeat/CNV sites tended to have lower base qualities and more mismatches per alt read, different base query positions for ref/alt alleles, different read mapping positions and different mapping qualities for ref/alt reads, and thus were better separated from real mosaic variants along PC1 and PC2 (Supplementary Fig. 6a–b). We thus reasoned that genotype labels of phasable sites could be better predicted using these first five principal components, which collectively explained ~50% of the variance (Supplementary Fig. 5d). We used phasing as well as the first five PCs for experimentally evaluated phasable sites as covariates to model their true genotypes using multinomial linear regression (Supplementary Table 4). The resulting model was used to predict refined genotype labels for the remaining non-phasable sites. Subsequently, the labels ‘hap=2’, ‘hap=3’, and ‘hap>3’ were converted to ‘het’, ‘mosaic’, ‘repeat’ and ‘refhom’, respectively. The R package glmnet[36] was used to build the multinomial regression model, and the R package mlr[37] was used to visualize the classification as shown in Supplementary Fig. 6c.

Construction of the Random Forest (RF) model

To construct a RF classification model applicable to both phasable and non-phasable candidate mosaics, we used all phasable sites from diploid chromosomes as the training set and used the read-level features described above for phasable sites as covariates. We used the R package caret[38] to build the RF model. The model was trained to predict the haplotype numbers for the ‘Phasing prediction model’ and it was trained to predict the four refined genotypes (‘refhom’, ‘het’, ‘repeat’, and ‘mosaic’) assigned to phasable sites for the ‘Refined genotype prediction model’. To train models applicable to sequencing data with different read depths, reads for all candidate sites in the 250X training set were down-sampled to 50X, 100X, 150X and 200X respectively, and all the read-level features of phasable sites were extracted from the sampled reads to build the corresponding RF models (Supplementary Table 10).

Evaluation of brain WGS-trained model in WGS data with different read depths

We trained models with phasable sites from the brain-WGS data (down-sampled to 50-250X depth) and tested on non-phasable sites from the brain-WGS data at 50-250X as well as on the simulated data constructed using NA12878 (Supplementary Fig. 11). To evaluate the validation rate in real WGS data, reads for all candidate sites called by MuTect2 in the 250X data (from the three individuals with single cell sequencing data available) were down-sampled to 50X, 100X, 150X and 200X respectively, and all the read-level features for non-phasable sites were extracted from the sampled reads. We then applied the brain WGS-trained RF models trained with phasable sites at 50-250X read depths to non-phasable sites in the three individuals. When evaluated with single cell or IonTorrent data (Supplementary Table 3), performance was only slightly better when training and testing datasets had similar coverages. For example, ~74% of variants (40/54) were validated as true mosaics when applying a model trained at 50X WGS data to 50X test data, whereas ~66% (40/61) were validated when applying a model trained at 250X WGS data to 50X test data (Supplementary Fig. 11).

Simulation of mosaic mutations and extraction of false sites

We also evaluated the performance of MosaicForecast in simulated datasets at different read depths. The 300X WGS data for the HapMap sample NA12878[26] was down-sampled to 50X, 100X, 150X, 200X, and 250X using SAMtools[30]. Spike-in mosaic mutations with expected allele fractions of 0.02, 0.03, 0.05, 0.1 and 0.3 were generated for each case (Supplementary Fig. 10). These simulated mosaics were randomly selected and mixed in proportion (4:4:4:2:1) to mimic the real early-embryo mosaic mutations in non-tumor tissues, assuming constant mutation rate per cell division (Supplementary Fig. 10c). To simulate a set of high-quality and correctly phased mosaic variants, simulated mutations were generated in BAM files by converting alternative alleles of the high-confidence heterozygous SNPs[39] to reference alleles with Bernoulli sampling. In the 250X data, the spike-in mutations were generated at higher density at variant allele fractions (0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4) and were used to determine the performance of the models across a wider variety of VAFs.

To extract a set of false variants from real sequencing data, candidate sites were called from the down-sampled BAM files (down-sampled from the original HapMap sample NA12878, without spike-in mutations) with MuTect2 (version 3.5 nightly-r2016-04-25-g7a7b7cd). Variants at <0.02 VAF calculated by MuTect2 were excluded. Variants with VAF ≥0.4 calculated by MuTect2, or present in the gnomAD whole-genome database[19] with ≥0.1% MAF were excluded. We then applied phasing on all candidate variants, and heterozygous SNPs were chosen as those with 2 haplotypes; sites with a misalignment issue within non-diploid regions were chosen as those with >3 haplotypes. The two kinds of mutations were used as simulated false sites (Supplementary Fig. 10d). We then applied the pre-trained RF models at different read-depths to predict mosaics in simulated datasets and evaluate performance.

Statistics

13 out of 31 read level features were calculated with scipy (v1.2.1), by doing two-tailed Wilcoxon's rank sum test, two-tailed t-test or two-tailed Fisher's exact test to compare base qualities, mapping qualities, positions of ref alleles/reads and alt alleles/reads, compare base qualities at the mutant position and neighboring positions as well as evaluate strand bias, read1/read2 biases. Refer to Supplementary Table 1 for more details. Other statistical tests for each analysis are calculated with R (v3.6.1) and are described in the Method part. Further information is available in the Life Sciences Reporting Summary.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The WGS data are available at the National Institute of Mental Health Data Archive (https://nda.nih.gov/study.html?id=644).

CODE AVAILABILITY

MosaicForecast is implemented in Python and R. The source code, documentation, and examples are available at https://github.com/parklab/MosaicForecast/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants U01MH106883, R01NS032457, T32HG002295, and T32GM007753; Harvard Ludwig Center; and a Horizon 2020 grant (703543). We thank Chong Chen, Heather Gold, Chong Chu, Vinay Viswanadham and Geoff Nelson for their helpful comments.

REFERENCES

1. Biesecker LG & Spinner NB A genomic view of mosaicism and human disease. Nat Rev Genet 14, 307–320 (2013). [PubMed: 23594909]
2. Bae T et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science 359, 550–555 (2018). [PubMed: 29217587]
3. Ju YS et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543, 714–718 (2017). [PubMed: 28329761]
4. Ye AY et al. A model for postzygotic mosaicisms quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. Genome Res (2018).
5. Lodato MA et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 350, 94–98 (2015). [PubMed: 26430121]
6. Dou Y, Gold HD, Luquette LJ & Park PJ Detecting Somatic Mutations in Normal Cells. Trends Genet 34, 545–557 (2018). [PubMed: 29731376]
7. Dou Y et al. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. Hum Mutat 38, 1002–1013 (2017). [PubMed: 28503910]

8. Freed D & Pevsner J The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS Genet 12, e1006245 (2016). [PubMed: 27632392]
9. Krupp DR et al. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. Am J Hum Genet 101, 369–390 (2017). [PubMed: 28867142]
10. Lodato MA et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559 (2018). [PubMed: 29217584]
11. Yang X et al. Genomic mosaicism in paternal sperm and multiple parental tissues in a Dravet syndrome cohort. Sci Rep 7, 15677 (2017). [PubMed: 29142202]
12. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31, 213–219 (2013). [PubMed: 23396013]
13. Alioto TS et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6, 10001 (2015). [PubMed: 26647970]
14. Huang AY et al. Distinctive types of postzygotic single-nucleotide mosaicisms in healthy individuals revealed by genome-wide profiling of multiple organs. PLoS Genet 14, e1007395 (2018). [PubMed: 29763432]
15. Lim ET et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat Neurosci 20, 1217–1224 (2017). [PubMed: 28714951]
16. Poplin R et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987 (2018). [PubMed: 30247488]
17. Kim S et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 15, 591–594 (2018). [PubMed: 30013048]
18. Bohrson CL et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet 51, 749–754 (2019). [PubMed: 30886424]
19. Karczewski KJ et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv, 531210 (2019).
20. Costello M et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19, 332 (2018). [PubMed: 29739332]
21. Karimzadeh M, Ernst C, Kundaje A & Hoffman MM Umap and Bismap: quantifying genome and methylome mappability. bioRxiv, 095463 (2017).
22. Smit A, Hubley R & Green P RepeatMasker. (2013–2015).
23. Poplin R et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv (2017).
24. Huang AY et al. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. Nucleic Acids Res 45, e76 (2017). [PubMed: 28132024]
25. Chen L, Liu P, Evans TC Jr. & Ettwiller LM DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355, 752–756 (2017). [PubMed: 28209900]
26. Zook JM et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3, 160025 (2016). [PubMed: 27271295]
27. McInerney P, Adams P & Hadi MZ Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Mol Biol Int 2014, 287430 (2014). [PubMed: 25197572]

REFERENCES

28. Rimmer A et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46, 912–918 (2014). [PubMed: 25017105]
29. Li H & Durbin R Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595 (2010). [PubMed: 20080505]
30. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]

31. Haeussler M et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47, D853–D858 (2019). [PubMed: 30407534]
32. Benson G Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580 (1999). [PubMed: 9862982]
33. Bragg LM, Stone G, Butler MK, Hugenholtz P & Tyson GW Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9, e1003031 (2013). [PubMed: 23592973]
34. Meacham F et al. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 12, 451 (2011). [PubMed: 22099972]
35. Huang AY et al. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. Cell Res 24, 1311–1327 (2014). [PubMed: 25312340]
36. Friedman J, Hastie T & Tibshirani R Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1–22 (2010). [PubMed: 20808728]
37. Bernd Bischl ML, Kotthoff Lars, Schiffner Julia, Richter Jakob, Studerus Erich, Casalicchio Giuseppe, Jones Zachary M. mlr: Machine Learning in R. Journal of Machine Learning Research 17, 1–5 (2016).
38. Kuhn M Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28, 26 (2008).
39. Zook JM et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246–251 (2014). [PubMed: 24531798]

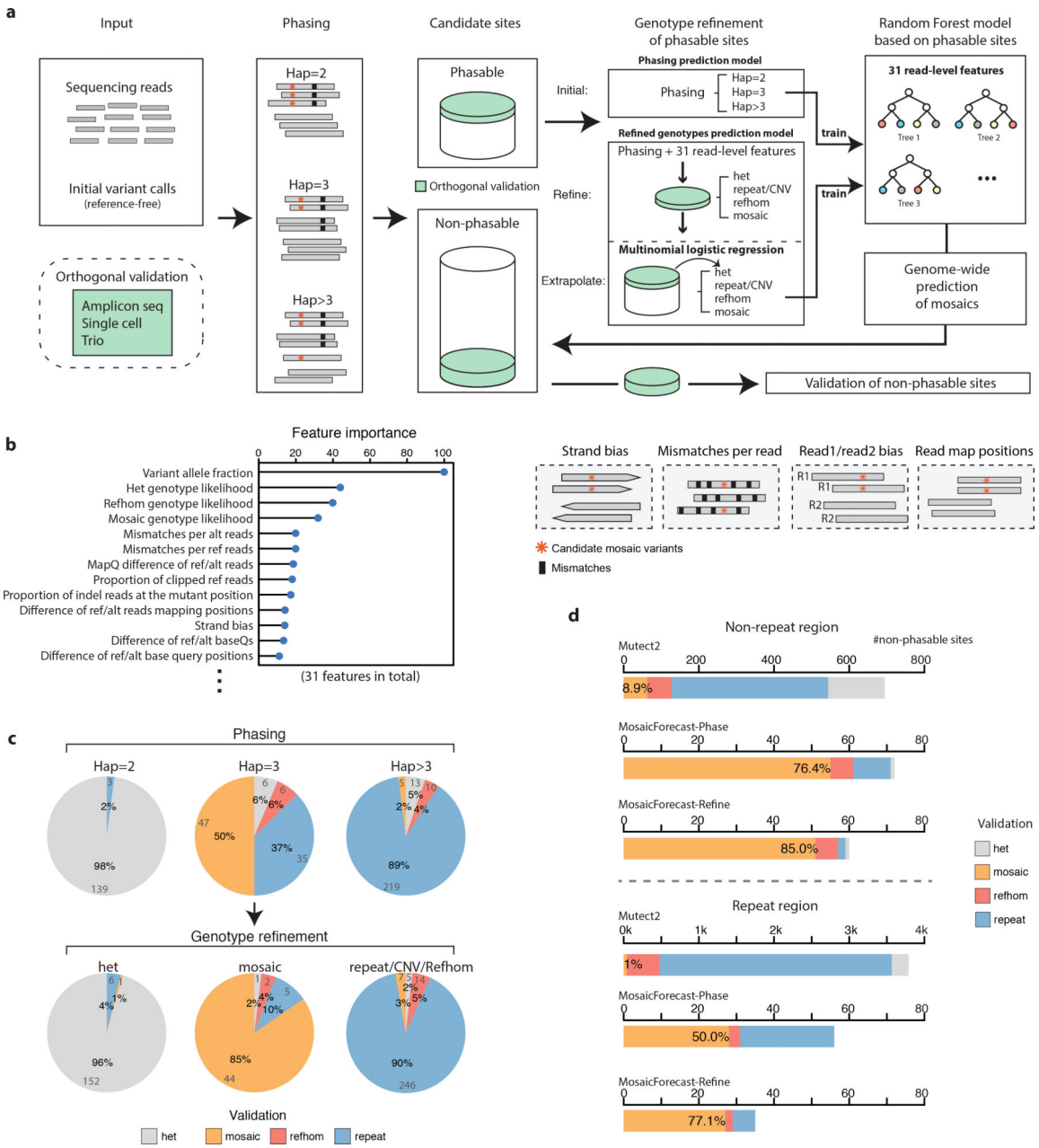

Fig. 1: Framework of MosaicForecast to detect mosaic SNVs from bulk sequencing data.

(**a**) Candidate mosaics were classified as ‘hap=2’, ‘hap=3’ or ‘hap>3’ by read-based phasing, and a Random Forest model was trained to predict the phasing by using 25 read-level features as covariates. The model was then applied to non-phasable sites to predict their genotypes. Given a list of experimentally-evaluated sites, the model could be further improved by an additional genotype-refinement step. (**b**) The relative importance of the features from the RF model for the brain WGS data, with four examples of read-level features. (**c**) 483 phasable sites were orthogonally evaluated by single cell, trio, and targeted sequencing data. After genotype refinement, the phasable sites classified as ‘hap=2’, ‘hap=3’ and ‘hap>3’ were converted to ‘het’, ‘mosaic’, ‘repeat/CNV’ and ‘refhom’ for training. (**d**)

We applied MosaicForecast to non-phasable MuTect2 candidate mosaics and evaluated them in single cell, trio, and targeted sequencing data. In non-repeat regions, the precision increased from 8.9% (MuTect2) to 76% for the Phasing prediction model and 85% for the Refined genotypes prediction model; in the RepeatMaster region, it increased from 1% (MuTect2) to 50% in the Phasing prediction model and 77% in the Refined genotypes prediction model in RepeatMasker regions.

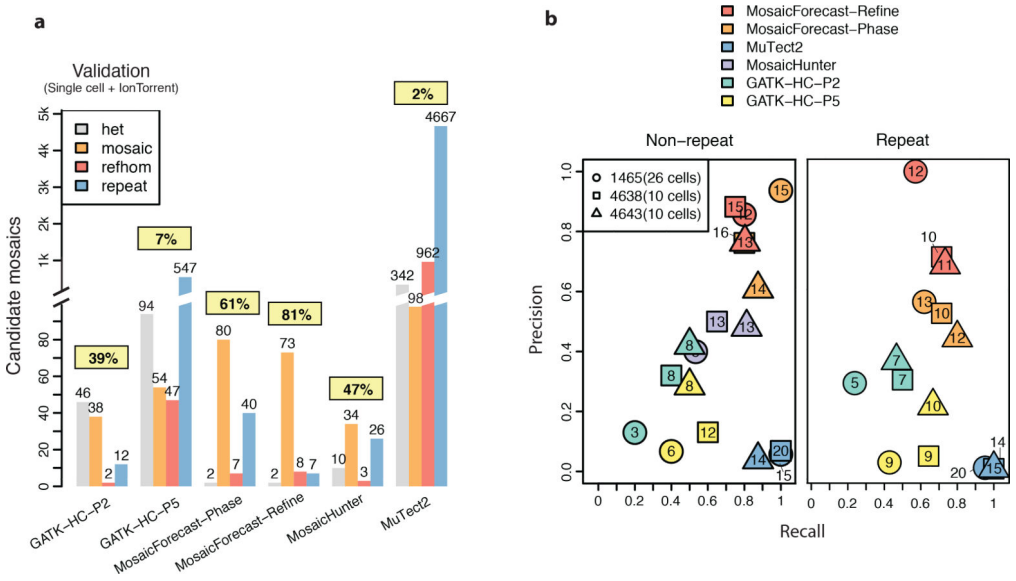

Fig. 2: Comparison among algorithms.

(**a**) Candidate mosaics (both phasable and non-phasable) in the three individuals with single cell data were evaluated (see Methods). (**b**) Precision and recall are plotted separately for the non-repeat and repeat regions (as defined by RepeatMasker) and for each individual.

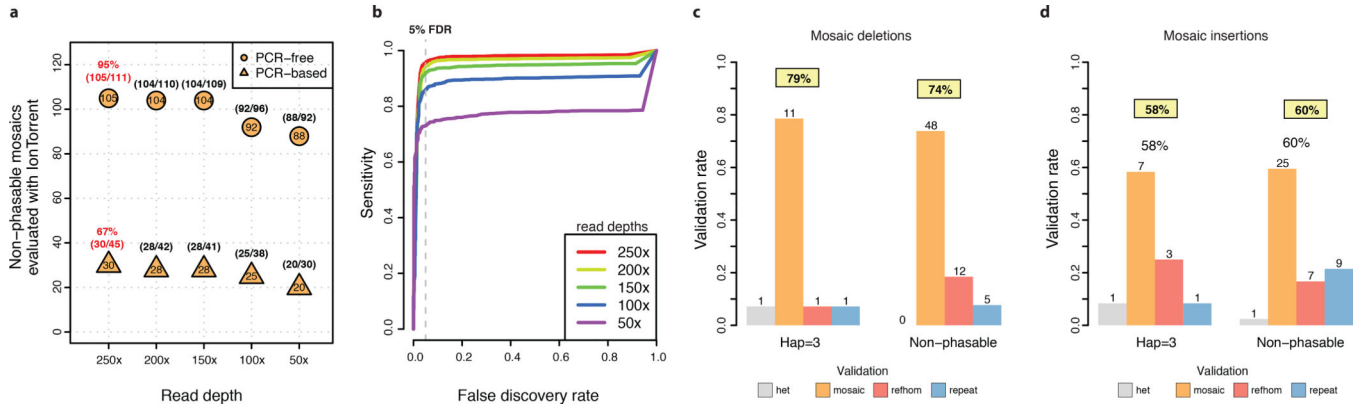

Fig. 3: Impact of read depth on sensitivity and detection of mosaic indels.

(**a**) At each coverage, a different RF model was trained on the phasable sites and predictions were made on non-phasable sites. Amplicon-sequencing data were used for validation. Although fewer true mosaics were identified at lower coverages, the sensitivity did not drop significantly (e.g., at 50X, MosaicForecast was able to detect ~80% of real variants identified at 250X). (**b**) Similar to (a) but using simulated data. The sensitivity was ~70% at 50X. (**c**) >70% of mosaic deletions called by MosaicForecast were validated by IonTorrent; the ‘hap=3’ sites and non-phasable sites had similar validation rates. (**d**) similar to (c) but for mosaic insertions.