Accurate detection of mosaic variants in sequencing data without matched controls

Yanmei Dou1, Minseok Kwon1, Rachel E. Rodin2,3,4,5, Isidro Cortés-Ciriano1,8, Ryan Doan2,3,4, Lovelace J. Luquette6,6, Alon Galor1, Craig Bohrson1,6, Christopher A. Walsh2,3,4 and Peter J. Park1,6*

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 and indels, achieving a multifold increase in specificity compared with existing algorithms. Using single-cell sequencing and targeted sequencing, we validated 80–90% of the mosaic single-nucleotide variants and 60–80% of 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.

A single individual harbors multiple populations of cells with distinct genotypes due to somatic mutations arising postzygotically. Such diversity of genotypes in an individual is referred to as somatic mosaicism. Analysis of mosaic mutations in nondisease samples enables exploration of lineage patterns during development and characterization of mutational mechanisms operative in normal cells. Recent studies have also demonstrated that somatic mutations contribute to many diseases besides cancer.

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 postmitotic cells are present only in a single cell and are detectable only by single-cell sequencing, as we have done recently. As standard cancer mutation callers typically have a lower VAF limit of 2–5% (refs. 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 and thus may be ambiguous in nondiploid or low-mappability regions. Incorporation of read-level features in a flexible framework is critical for distinguishing real mutations from artifacts. In place of filters with hard thresholds, recent methods such as DeepVariant and Strelka2 (ref. 15) 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, in which a candidate mosaic mutation and a nearby germline variant are checked for haplotype consistency—that is, 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 is phasable using short-read sequencing, and phasing may be ambiguous in nondiploid or low-mappability regions.

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: (1) generation of a training set by read-based phasing; (2) construction of a random forest (RF) model based on read-level features related to the quality and category of variants, such as VAF, read depth, mismatches per read and strand bias; and (3) genome-wide prediction of mosaic SNVs. The underlying idea is similar to that of DeepVariant and Strelka2 (ref. 15) 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 uses 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 and 15 neurotypical individuals, sequenced at ~250x (150-base pair (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)16. 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.

*Corresponding author. E-mail: peter_park@hms.harvard.edu

1Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. 2Division of Genetics and Genomics, Manton Center for Orphan Disease, and Howard Hughes Medical Institute, Boston Children’s Hospital, Boston, MA, USA. 3Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA. 4Broad Institute of MIT and Harvard, Cambridge, MA, USA. 5Harvard/MIT MD–PhD Program, Harvard Medical School, Boston, MA, USA. 6Bioinformatics and Integrative Genomics PhD program, Harvard Medical School, Boston, MA, USA. 7Ludwig Center at Harvard, Boston, MA, USA. 8Present address: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. - e-mail: peter_park@hms.harvard.edu
Letters

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 an RF model was trained to predict the phasing by using over 30 read-level features as covariates. The model was then applied to nonphasable 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, In total, 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 nonphasable MuTect2 candidate mosaics and evaluated them in single-cell, trio and targeted sequencing data. In nonrepeat regions, the precision increased from 8.9% (MuTect2) to 76% for the phasing prediction model and 85% for the refined genotypes prediction model; in RepeatMasker regions, it increased from 1% (MuTect2) to 50% in the phasing prediction model and 77% in the refined genotypes prediction model. The unit k stands for 1000.

(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, for example, 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. 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): (1) hap = 2, consistent with heterozygous germline variants; (2) hap = 3, consistent with mosaics; and (3) 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 and 3). The single-cell WGS dataset of three individuals we have published previously provides 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 and 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 ‘phasching prediction model’ genome-wide, excluding nonunique mapping regions (see Methods). This model resulted in modest validation rates for hap = 3 sites, with 67% (55 of 82) in nonrepeat regions and 34% (28 of 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). However, we noticed that many variants are clustered, in mostly repeat or nondiploid regions, when all individuals are considered together (~46% of those predicted to be hap ≥ 3 were enriched in regions that together span only ~19 Mb; see Methods and Supplementary Table 6). After removing these clustered variants, the overall validation rate for the phasing prediction model increased to 76% (55 of 72) in nonrepeat regions and 49% (28 of 57) in repeat regions (Fig. 1d and Supplementary Table 5). This constitutes a sevenfold increase (nonrepeat regions) and a 43-fold increase (repeat regions) in precision compared with 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

---

**Fig. 2 | Comparison among algorithms.** a, Candidate mosaics (both phasable and nonphasable) in the three individuals with single-cell data were evaluated (see Methods). b, Precision and recall are plotted separately for the nonrepeat and repeat regions (as defined by RepeatMasker) and for each individual. The number inside each symbol corresponds to the number of validated mosaics.

---

**Fig. 5 and Supplementary Table 5; we define ‘repeat region’ to include interspersed repeats and low-complexity sequences identified by RepeatMasker.** However, we noticed that many variants are clustered, in mostly repeat or nondiploid regions, when all individuals are considered together (~46% of those predicted to be hap ≥ 3 were enriched in regions that together span only ~19 Mb; see Methods and Supplementary Table 6). After removing these clustered variants, the overall validation rate for the phasing prediction model increased to 76% (55 of 72) in nonrepeat regions and 49% (28 of 57) in repeat regions (Fig. 1d and Supplementary Table 5). This constitutes a sevenfold increase (nonrepeat regions) and a 43-fold increase (repeat regions) in precision compared with 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...
nonphasable candidate mutations (Supplementary Table 7). Sites within nonunique mapping regions\(^1\) (mappability score = 0) as well as sites within clustered regions (Supplementary Fig. 6d) were excluded. Among the 2,220 predicted (nonphasable) 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 nonrepeat 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)\(^2\), MuTect2\(^1\) and MosaicHunter\(^3\) 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 were used for validation. Although fewer true mosaics were identified at lower coverages, the sensitivity did not drop substantially (for example, at 50x, MosaicForecast was able to detect ~80% of real variants identified at 250x). 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 nonphasable sites had similar validation rates. d, Similar to c but for mosaic insertions. FDR, false discovery rate.

Using PCR-based libraries, the validation rate was ~61% (42 of 68; 68%, 42 of 62, on diploid and 0 of 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 (odds ratio = 4.1, P < 1 × 10\(^{-15}\), Fisher’s exact test; Supplementary Fig. 8b), which are associated with oxidative damage\(^5\). If we focus on nonphasable sites from diploid chromosomes (Fig. 3a), validation rates were ~95% (105 of 111) and ~67% (30 of 45) for PCR-free and PCR-based samples, respectively. In addition, the validation rates were similar in non-repeat regions (87%, 118 of 136) and repeat regions (85%, 17 of 20). Among the 177 MosaicForecast mosaics in nonrepeat regions confirmed by IonTorrent, GATK-HC-p5 was only able to detect ~62% (109 of 177), followed by MosaicHunter (~59%, 105 of 177) and GATK-HC-p2 (~20%, 35 of 177). Among the 26 MosaicForecast mosaics in repeat regions confirmed by IonTorrent, GATK-HC-p5 and GATK-HC-p2 were only able to detect ~58% (15 of 26) and ~19% (5 of 26), respectively (MosaicHunter does not make calls in repeat regions). A large fraction of low-V AF (≤0.05) mosaics were called by MosaicForecast but missed by both MosaicHunter and GATK–HC (~52%, 48 of 92; Supplementary Fig. 8c), indicating that MosaicForecast is particularly advantageous for detecting low-V AF mutations.

Third, we tested the haplotype numbers for the extra variants identified by the other callers. Across the 75 individuals, the other methods called 1–80 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 (Figs. 1c and 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 VAFs, from 0.02 to 0.3, as reflected in the receiver operating characteristic (ROC) curves (Supplementary Fig. 10a,b). It performed substantially worse when the VAFs approached 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 250× 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 of 135) of the validated 250× variants at 50× (Fig. 3a and 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 300× WGS data for NA12878 (ref. 26) following a realistic allele fraction distribution of early embryonic mosaics and down-sampled to 50–250× (see Methods and Supplementary Fig. 10c,d). MosaicForecast was sensitive in detecting simulated mosaics and effective in removing nonmosaic sites across read depths: at a 5% false discovery rate, it detected ~95% of the spike-in mutations at 250×. When the training and simulation were performed at lower depths, ~90% of the spliced-in mutations were detected at 100× and ~70% at 50× (Fig. 3b). We also found that the models were robust across different read depths (see Methods and 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 of 98) to 90% (27 of 30) for GATK-HC-p2, from 7% (54 of 742) to 84% (42 of 50) for GATK-HC-p5 and from 47% (34 of 73) to 61% (31 of 51) for MosaicHunter (Supplementary Fig. 12). For maximal sensitivity, we could generate an input set simply by using SAMTools mpileup scanning, for example, by taking all sites with ≥2% nonreference bases as potential mutations. Using single-cell data to evaluate the sites called only by SAMTools, only a tiny fraction (1.8%, 104 of 5,876) of a large mutation set was validated. With MosaicForecast, we achieved a striking improvement in the validation rate (45.9%, 78 of 170; Supplementary Fig. 12). Compared with the MuTect2 validation result (73 of 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 59,977 candidates (22,893 deletions, 37,084 insertions) from the nonrepeat 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 1,356 sites as hap = 3. With additional filtering criteria (see Methods), 102 sites were classified as confident mosaic deletions. All of the high-confidence mosaic deletions were from PCR-free samples. When evaluated using IonTorrent sequencing (see Methods and Supplementary Fig. 13d), ~75% (59 of 79) were validated as true deletions, with phasable (79%, 11 of 14) and nonphasable (~74%, 48 of 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 of 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 rates of mosaic indels in PCR-based samples are likely to be caused by replication slippage of DNA polymerase. In the process of detecting mosaic SNVs and indels, we also identified several mosaic multi-nucleotide variants (for example, two adjacent base substitutions) as well as clumped variants (for example, an 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 nontumor tissues by the proposed method will help gain insights into the origin and propagation of somatic mutations in development and disease.

Online content
Any additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-019-0368-8.

Received: 4 December 2018; Accepted: 23 November 2019; Published online: 6 January 2020

References
1. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. Nat. Rev. Genet. 14, 307–320 (2013).
2. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregestation and neurogenesis. Science 359, 550–555 (2018).
3. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543, 714–718 (2017).
4. Ye, A. Y. et al. A model for postzygotic mosaicisms quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. Genome Res. 28, 943–951 (2018).
5. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 350, 94–98 (2015).
6. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting somatic mutations in normal cells. Trends Genet. 34, 543–557 (2018).
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).
8. Freed, D. & Pevsner, J. The contribution of mosaic variants to autism spectrum disorder. PLoS Genet. 12, e1006245 (2016).
9. Krupp, D. R. et al. Exonic mosaic mutations contribute risk for autism spectrum disorder. Am. J. Hum. Genet. 101, 369–390 (2017).
10. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559 (2018).
11. Yang, X. et al. Genomic mosaicism in paternal sperm and multiple parentage in a Dravet syndrome cohort. Sci. Rep. 7, 15677 (2017).
12. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
13. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. 6, 10001 (2015).
14. Huang, A. Y. 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).
15. Lim, E. T. et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat. Neurosci. 20, 1217–1224 (2017).
16. Hovi, S. N. & et al. A universal SNP and small-indel caller using deep neural networks. Nat. Biotechnol. 36, 983–987 (2018).
17. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594 (2018).
18. Rohrssen, C. L. et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat. Genet. 51, 749–754 (2019).
19. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at bioRxiv https://doi.org/10.1101/531210 (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).

21. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).

22. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2013–2015).

23. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2018).

24. Huang, A. Y. 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).

25. Chen, L., Liu, P., Evans, T. C. Jr. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752–756 (2017).

26. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).

27. McInerney, P., Adams, P. & Hadi, M. Z. Error rate comparison during polymerase chain reaction by DNA polymerase. *Mol. Biol. Int.* **2014**, 287430 (2014).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
Methods

Datasets. 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 obtained1. All data were 150 bp paired-end reads. FASTQ files for high-confidence SNV calls (by NA12878 (300x), 150 bp, paired-end) generated by the Genome in a Bottle Consortium (https://ftp-trace.ncbi.nlm.nih.gov/igv/b37) were down-sampled to different read depths. WGS data for 30 tumor tissues (~80x, 100bp, paired-end) and their matched normal samples (~40x) as well as consensus variant calls were obtained from the International Cancer Genome Consortium2. Reads were aligned to GRCh37 with bowtie (v2.3.5) with Burrows-Wheeler Aligner (BWA)3.

Variant calling. Five methods were used to call mosaic SNVs in the brain bulk data: MuTect2 (v3.5 nightly r2014-06-25 g47a7b7cd, variants tagged as ‘str_contraction’, ‘triallelic_site’ or ‘_lod_star’ were excluded), GATK-HC (v3-5-0), with the minimum VAF threshold adjusted from the default 5 to 2% to enable detection of mosaics with lower VAFs; and SAMTools (v1.9) mpileup with sites with ≥20% nonreference 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 database4 were removed. Variants at <0.02 VAF calculated by each tool were excluded. Variants with VAF ≥0.4 or present in gnomAD5 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’6 were regarded as refhom. If two variants were close to each other with all alt alleles on the same reads were also classified as repeat variants. 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, and variants with >350x read depth, variants with ultra-low VAF calculated by MosaicForecast (v<0.01), variants with a high proportion (≥10%) of clipped reference reads, variants within clustered regions and variants present in gnomAD were removed. Further information is available in the Nature Research 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 (obtained excess number of low-VAF mutations) was excluded. Sites with extra-high read depths (≥2 fold) and sites with ≥1.5 fold read depths and ≥20% VAF 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 50–250x; 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 of 101) of real mosaics called by different tools, whereas MosaicHunter, GATK-HC-p2 and GATK-HC-p5 were only able to detect 34, 38 and 54% of real mosaics, respectively (Supplementary Fig. 1b). Further information is available in the Nature Research Reporting Summary.

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)16,17, we constructed lineage trees with single-cell mutations assigned by different tools. We then evaluated how many of the mosaic insertions in Fig. 3d.

Read-based phasing. To identify germline SNPs near the SNVs detected by MuTect2, we scanned reads mapped up to 1 kbp 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 ≤ 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 whether a region was 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: (1) those leading to three haplotypes were treated as potential mosaics (Fig. 1a and Supplementary Fig. 3b); (2) those leading to two haplotypes were treated as heterozygous mutations (Supplementary Fig. 3c); and (3) 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 and Supplementary Fig. 3d).

Read-level features. The 31 features are described in Supplementary Table 1. Two of the features, ‘mapp_p’ and ‘mapp_difference’, encode mapping qualities; three account for the number of mismatches per read (‘major_mismatches_mean’, ‘minor_mismatches_mean’, ‘mismatches_p’ and ‘mismatches_t’), 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 read mapping position of the SNVs detected by MosaicForecast, local mapability and centroid 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 position11. We also estimated genotype likelihoods of four different genotypes (ref/hom, alt/hom, mosaic) based on Bemoulli sampling18 to capture sequencing errors and ref/alt allele read depth biases, assuming that the real mutant allele fractions are 0 (ref/hom), 0.5 (het), 1 (alt/hom) and uniformly distributed between 0 and 1 (mosaic). The formulas to calculate the four genotypes are as follows:

\[ L(G = \text{het/alt/mosaic}) = \frac{\text{depth} \times [1 - F(G)] + \text{depth} \times F(G)}{0.02 \times r + 0.98 \times R(G)} \]

where \( G \) represents the genotype (heterozygote, homozygote or mosaic), \( F(G) \) is the frequency of the genotype within the sample, \( \text{depth} \) is the coverage of the read, \( r \) is the read length and \( R(G) \) is the read length distribution within the sample.
Mosaic Forecast for calling mosaic indels. Specifically, ‘alt reads’ were re-defined as modified several read-level features and designed several new features/filters to adapt to the mutant allele fraction, \( G_i \), denotes genotype, \( \beta \) denotes beta function and \( L \) denotes likelihood.

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 to mosaic Forecast for calling mosaic indels. Specifically, ‘alt reads’ were re-defined as modified several read-level features and designed several new features/filters to adapt to the mutant allele fraction, \( G_i \), denotes genotype, \( \beta \) denotes beta function and \( L \) denotes likelihood.

To construct an RF classification model 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 of 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 nonphasable sites from the brain WGS data (reference sequences) as well as on the 250x 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 of the read-level features of nonphasable sites were extracted from the sampled reads. We then applied the brain WGS-trained RF models trained with phasable sites at 50–250x depth to nonphasable 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 of 54) were validated as true mosaics when applying a model trained at 50x WGS data to 50x test data, whereas ~66% (40 of 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 30x WGS sample data for the HapMap sample NA12878 (ref. 35) were down-sampled to 50x, 100x, 150x, 200x and 250x using SAMTools41. 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 nontumor 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 SNPs42 to reference alleles with Bernoulli sampling. In the 250x data, the spike-in mutations were generated at higher density at VAFs (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-r2016-04-25-7a75f7c0d). 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 database43 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 nondiploid 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 pretrained RF models at different read depths to predict mosaics in simulated datasets and evaluate performance.

Statistics. Thirty-one 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, and positions of ref alleles/reads and alt alleles/reads; to compare base qualities at the mutation position and neighboring positions; and to evaluate strand bias and read1/read2 biases. Refer to Supplementary Table 1 for more details. Other statistical tests for each analysis were calculated with R (v3.6.1) and are described in the Methods. Further information is available in the Nature Research Reporting Summary.

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 can be found at https://github.com/parklab/MosaicForecast/.

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).
29. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
30. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
32. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
33. Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* **9**, e1003031 (2013).
34. Meacham, F. et al. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
35. Huang, A. Y. et al. Postzygotic single-nucleotide mosaicsisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.* **24**, 1311–1327 (2014).
36. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
37. Bischl, B. et al. mlr: Machine Learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
38. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 26 (2008).
39. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

**Acknowledgements**
This work was supported by National Institutes of Health grants (nos. U01MH106883, R01NS032457, T32HG002295 and T32GM007753); by the Harvard Ludwig Center; and by a Horizon 2020 grant (no. 703543). We thank C. Chen, H. Gold, C. Chu, V. Viswanadham and G. Nelson for their helpful comments.

**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 to 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 interests**
The authors declare no competing interests.

**Additional information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0368-8.

**Correspondence and requests for materials** should be addressed to P.J.P.

**Reprints and permissions information** is available at www.nature.com/reprints.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

☐ n/a  Confirmed
  ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
  ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
  ☐ The statistical test(s) used AND whether they are one- or two-sided
  ☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.
  ☐ A description of all covariates tested
  ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
  ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
  ☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  ☐ Give P values as exact values whenever suitable.
  ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
  ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
  ☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
  ☐ Clearly defined error bars
  ☐ State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No data collection software was used in this study.

Data analysis

We aligned FASTQ files using BWA-MEM version 0.7.8. We did data preprocessing of the bam files using GATK best practices, including Picard mark duplicates (v1.130), GATK indel realignment, GATK base quality recalibration (v3.4-46). We used MuTect2 Panel-Of-Normals strategy (GATK version 3.5 nightly-r2016-04-23-g7a7b7cd) to call putative variants as input of MosaicForecast. We also used MosaicHunter (v1.1), GATK HaplotypeCaller (v3.5-0) and Samtools (v1.3) to call variants as comparisons of MosaicForecast.

The code of MosaicForecast is available publicly in a Github repository, https://github.com/parklab/MosaicForecast. R packages caret (v6.0-83), randomForest (v4.6-14), mlf (v2.15.0), ggplot2 (3.1.1) and python packages numpy (v1.16.1), pandas (v0.25.1), pysam (v0.15.2), pyfaidx (v0.5.3), pysamstats (v1.1.2), scipy (v1.2.1) as well as bedtools (v2.25.0) and ANNOVAR (v2017-07-17) were used in the data analysis process. conda (v4.6.14) and docker (v19.03.4) were used to set up the dependencies.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/reportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No sample size calculation was needed for this methods paper.

Data exclusions

Variants within segmental duplication regions according to the UCSC Genome Browser database were removed. Variants at <0.02 VAF calculated by each tool (MuTect2, GATK-HC-p2, GATK-HC-p5, MosaicHunter) were excluded. Variants with VAF ≥0.4 or present in the gnomAD database were removed as likely germline variants. One individual (UM85308) was excluded from the analysis due to likely contamination (“100 time higher number of predicted mosaic sSNVs compared with others, and 93% of the sites have 2-5% AF). As a result, although all successfully targeted mosaic indels from UM85308 were validated, we did not use the variants from UM85308 to calculate validation rate. For candidate mosaic indels specifically, variants present in the relevant “panel of normals” (PoN) or gnomAD as well as those present in repeat regions, including simple repeats, 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 or insertions per sample were excluded, variants with >50X 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. All 11 samples sequenced using PCR-based libraries contain extra-high number of indels and were excluded. We used IonTorrent to evaluate mosaic insertions from PCR-based samples, but none of the mosaic insertions from PCR-based samples were validated. The excessive error rate of mosaic indels in PCR-based samples are likely to be caused by replication slippage of DNA polymerase. While doing read-level phasing to identify germline SNPs near the SNVs detected by MuTect2, we scanned reads mapped up to 1kb away from each candidate site and excluded reads with low mapping qualities (<20) or with low base quality at the mutant position (<20), a binomial test was applied to remove variants whose VAFs deviated from 0.5. Variants with relatively low read depth (<2X) were also filtered.

Replication

Our code is available on GitHub. To allow for reproducibility of our work and to avoid over-fitting of our models, we used (i) cross-validation using training and testing sets, (ii) used multiple validation methods: amplicon-based resequencing, single cell sequencing, and family trio data, and (iii) performed simulation studies using a HapMap sample.

Randomization

N/A

Blinding

N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | Unique biological materials |
| ☒   | Antibodies |
| ☒   | Eukaryotic cell lines |
| ☒   | Palaeontology |
| ☒   | Animals and other organisms |
| ☒   | Human research participants |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | ChIP-seq |
| ☒   | Flow cytometry |
| ☒   | MRI-based neuroimaging |