DeepMosaic: Control-independent mosaic single nucleotide variant detection using deep convolutional neural networks

Xiaoxu Yang[1,2,*,#], Xin Xu[1,2,*], Martin W. Breuss[1,2], Danny Antaki[1,2], Laurel L. Ball[1,2], Changuk Chung[1,2], Chen Li[1,2], Renee D. George[1,2], Yifan Wang[3], Taejeoing Bae[3], Alexej Abyzov[3], Jonathan Sebat[4,5,6,7], NIMH Brain Somatic Mosaicism Network[†], Joseph G. Gleeson[1,2,#]

1. Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA.
2. Rady Children's Institute for Genomic Medicine, San Diego, CA, USA.
3. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.
4. Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, CA, USA.
5. Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA.
6. Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA.
7. Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA.

* These authors contributed equally to this work

[#] Correspondence: jogleeson@health.ucsd.edu and xiy010@health.ucsd.edu

[†] Full membership of the Brain Somatic Mosaicism Consortium Network is listed in the Supplementary Text.

Introductory paragraph

Mosaic variants (MVs) reflect mutagenic processes during embryonic development[1] and environmental exposure[2], accumulate with aging, and underlie diseases such as cancer and autism[3]. The detection of MVs has been computationally challenging due to sparse representation in non-clonally expanded tissues. While heuristic filters and tools trained on clonally expanded MVs with high allelic fractions are proposed, they showed relatively lower sensitivity and more false discoveries[4-9]. Here we present DeepMosaic, combining an image-based visualization module for single nucleotide MVs, and a convolutional neural networks-based classification module for control-independent MV detection. DeepMosaic achieved a higher accuracy compared with existing methods on biological and simulated sequencing data, with a 96.34% (158/164) experimental validation rate. Of 932 mosaic variants detected by DeepMosaic in 16 whole genome sequenced samples, 21.89-58.58% (204/932-546/932) MVs were overlooked by other methods. Thus, DeepMosaic represents a highly accurate MV classifier that can be implemented as an alternative or complement to existing methods.

Main text

Postzygotic mosaicism describes a phenomenon whereby cells arising from one zygote harbor distinguishing genomic variants[1, 10]. MVs can act as recorders of embryonic development, cellular lineage and environmental exposure. They accumulate with aging, play important roles in human cancer progression[3, 10], and are implicated in over 200 noncancerous disorders[11, 12]. Collectively, estimates are that MVs contribute to 5-10% of the 'missing genetic heritability' in more than 100 human disorders[11, 13].

Compared with the higher allelic fractions (AF) of 5-10% found in clonal tumors or pre-cancerous mosaic conditions, AFs found in non-clonal disorders, or neutral variants used for lineage studies, are typically present at much lower AFs. Existing methods, however, based on classic statistical models like Mutect2[9] and Strelka2[7] and heuristic filters are often optimized for the high fraction variants in cancer with relatively high variant AFs. Similarly, because of their conceptual origin in cancer, most existing programs including the more recent NeuSomatic[14], also require matched control samples. This can be problematic when mutations are present across different tissues ('tissue shared' mosaicism), or when only one sample is available.

Newer methods that aim to overcome these limitations, such as MosaicHunter[5] or MosaicForecast[4], are based conceptually similarly on the use of features extracted from raw data, rather than the sequence and alignment themselves, or replace the filters with traditional machine-learning methods. While these are a useful proxy, they only represent a limited window into the sheer wealth of information. Because of these limitations, researchers often resort to visual inspection of raw sequence alignment in a genome browser, a so-called 'pileup', to distinguish artifacts from true positive variants[15]. This is a laborious and low-throughput process that allows spot checking, but cannot be implemented on a large scale for variant lists numbering in the thousands for programs like MuTect2.

Image-based representation of pileups and the application of deep convolutional neural networks represents a potential solution for these limitations. Previous attempts like DeepVariant[14] were successful in detecting heterozygous or alternative homozygous single nucleotide variants (SNVs) from direct representation of aligned reads by using deep neural networks. The DeepVariant genotype model, unfortunately, did not consider a mosaic genotype, and lacked orthogonal validation experiments. Here we introduce DeepMosaic comprising two modules: a visualization module for image-based representation of single nucleotide variants, which forms the basic input for a convolutional neural network (CNN)-based classification module for mosaic variant detection. Five different biological and computationally simulated dataset as well as amplicon validation were used to train and benchmark DeepMosaic.

To automatically generate a useful visual representation similar to a browser snapshot, we developed the visualization module of DeepMosaic (DeepMosaic-VM, Fig.1a-d). The input for this visualization is short-read WGS data, processed with a GATK current best-practice pipeline (insertion/deletion, or INDEL, realignment and base quality recalibration). DeepMosaic-VM processes this input into an ‘RGB’ image, representing a pileup at each genomic position. In contrast to a regular browser snapshot, we encode sequences as different intensities within one channel, and use other channels for base qualities and strand orientations. We further split the pileup of reference reads and alternative reads based on the reference genome information (Fig. 1a-d), to improve visualization and allow assessment of mosaicism at a glance.

The classification module of DeepMosaic (DeepMosaic-CM) is a CNN-based classifier for MVs. We trained 10 different CNN models with more than 180,000 image-based representations from both true-positive and true-negative biological variants in several recently published high-quality experimentally validated public datasets[16, 17], and computationally simulated reads with added MVs (employing Illumina HiSeq error models) across a range of AFs and depths (Fig. 1e, Methods and Supplementary Fig. 1a-b) to select a model with optimal performance. To ensure its resemblance of real data, we controlled the distribution of AFs in the training set (Supplementary Fig. 1c). In addition, a range of expected technical artifacts, including multiple alternative alleles, homopolymers, and alignment artifacts, were manually curated and labeled negative in the training set to represent expected pitfalls that often result in false positive mosaic calls for other programs (Supplementary Fig. 1d).

To further expand training across a range of different read depths, the biological training data were also up- and down-sampled to obtain data at read depths ranging from 30x to 500x (Supplementary Fig. 1e), which includes the most commonly used depths for WGS in current clinical and scientific settings. In addition to the output from DeepMosaic-VM, we further incorporated population genomic and sequence features (e.g. population allele frequency, genomic complexity, ratio of read depth), which are not easily represented in an image, as input for the classifier (Fig. 1f). Depth ratios were calculated from the expected depth and used to exclude false positive detections from potential copy number variations. gnomAD population allelic frequencies were used to exclude common variants. Segmental duplication and repeat masker regions were used to exclude 24% of low complexity regions genome-wide.

Ten different CNN architectures were trained on 180,000 training variants described above. The CNN models included Inception-v3[18], which was used in DeepVariant; Deep Residual Network[19] (Resnet) which was used in the control-dependent caller NeuSomatic; Densenet[20] and 7 different builds of EfficientNet[21], for its high performance on rapid image classification (Methods, Supplementary Fig. 2a). Each model was trained on the data described above with 5 to 15 epochs to optimize the hyper-parameters until training accuracies plateaued (>0.90).

To compare the different models after training, we employed an independent gold-standard validation dataset of ~400 MVs from one brain sample[22] (Methods). On these, EfficientNet-b4 showed the highest accuracy, Matthews's correlation coefficient, and true positive rate when trained for 6 epochs (Supplementary Fig. 2b). We thus select this model as the default model of DeepMosaic-CM (Supplementary Fig. 3 and Fig. 1f).

To uncover the information prioritized by the selected default model, we used a gradient visualization technique with guided backpropagation[23] to highlight the pixels with guiding classification decisions (Supplementary Fig. 4). The results suggested that the algorithm not only recognized the edges for reference and alternative alleles, but also integrated additional available information, such as insertion/deletions in the sequences, overall base qualities, alignment artifacts, and other features which may not be extracted by digested feature-based methods.

We evaluated the performance of DeepMosaic using 20,265 variants from the above training data that was hidden from model training and selection. The receiver operating characteristic (ROC) curve and precision-recall curves on the hidden validation dataset showed >0.99 area under curve for a range of coverage (30x ~500x, Fig. 2a and 2b) across a range of AFs (Supplementary Fig. 3a and 3b), demonstrating high sensitivity and specificity.

Next, we benchmarked DeepMosaic's performance relative to other detection software and used data generated from a distinct sequencing error model to test for its utility on general sequencing data. We compared the performance of DeepMosaic with the widely used Mutect2 (paired mode), Strelka2 (somatic mode) with heuristic filters, MosaicHunter (single mode), and MosaicForecast (Methods). We generated an additional computationally simulated dataset of 439,200 variants based on the error model of a different Illumina sequencer (NovaSeq, Methods), with AF ranges from 1% to 25%, and depth ranges from 50x to 500x. Mutect2 paired methods and Strelka2 somatic mode used simulated mutated samples as “tumor” and simulated reference samples as “normal” for their paired modes. DeepMosaic showed equal or better performance than all other methods tested, especially for low allelic fraction variants (Fig. 2c), noticeably, even for low read depth data; and it performed better than methods that have additional information from paired samples. Overall DeepMosaic showed almost a 1.5-3 fold increase of the detection sensitivity for AFs under 3% compared with other methods (Fig. 2c). This is likely because our models integrate additional genomic sequence and quality information from the original BAM file (Supplementary Fig. 4), and are capable of distinguishing mosaic variants from different Illumina error models.

To exclude limitations resulting from benchmarking with simulated data and demonstrate that models trained on PCR-amplified libraries are also useful for PCR-free sequencing libraries, we extended benchmarking to biological data. We performed the same comparison on our recently published 200x WGS dataset[12] with 16 samples (blood and sperm) from 8 healthy individuals[24]. Paired methods compared two samples from the same individual, and control-independent samples used a published dataset of a panel of normals[25]. Variants detected by Mutect2 (paired mode), Strelka2 (somatic mode) and MosaicHunter (single mode) were subjected to a series of published heuristic filters[24, 25]. As we had access to the biological samples, we also performed orthogonal validation, using deep amplicon sequencing of 241 randomly selected MVs with a representative AF distribution compared to the complete candidate variant list (Methods, Fig. 3a and 3b, Supplementary Table 1).

As expected from the test of the computationally generated data, DeepMosaic showed the highest overall validation rate (96.34%, 158/164) among all 5 methods (Fig. 3c), demonstrating the power of DeepMosaic that models trained on PCR-amplified biological data and simulated data can accurately classify these PCR-free biological data. Of the 932 MVs detectable by DeepMosaic, 21.89% (204/932, 33/34 experimentally validated) were overlooked by MosaicForecast, 58.58% (546/932, 96/98 validated) overlooked by MosaicHunter, 50.32% (469/932, 90/94 validated) overlooked by Strelka2 (somatic mode) with heuristic filters, 43.13% (402/932, 81/85 validated) overlooked by Mutect2 (paired mode) with heuristic filters[24]. DeepMosaic also accurately detected variants with relatively low AFs (Fig. 3d). Finally, DeepMosaic outperformed other methods across most of the AF bins (Fig. 3e).

In current practice, researchers often combine multiple programs in one variant detection pipeline to detect different categories of MVs[24-26]. We thus further compared DeepMosaic with different pipelines used in recent publications, using data from 200x WGS of the 16 samples[24]: 1] With the MosaicForecast pipeline[4], which uses Mutect2 single mode (each sample compared with the publicly available panel of normal) as input; 2] With what we call the M2S2MH pipeline, which we recently published[24], combining Mutect2 paired mode (i.e. compared between different samples from a same individual), Strelka2 somatic mode and MosaicHunter single mode followed by a series of heuristic filters (Supplementary Fig. 5a). Of the 932 MVs identified by DeepMosaic, 78.11% (728/932, 125/130 validated) overlapped with MosaicForecast and 60.09% (560/932, 87/91 validated) overlapped with M2S2MH. In contrast, 21.89% (204/932, 33/34 validated) were undetected by MosaicForecast, and 39.91% (372/932, 71/73 validated) were overlooked by M2S2MH. These variants uniquely detected by DeepMosaic all showed validation rate > 97% (Supplementary Fig. 5b and 5c), demonstrating that DeepMosaic can accurately detect a considerable number of variants undetectable by widely used methods.

To test the performance of these samples on data widely curated clinically, we compared detection sensitivity for genome samples with standard WGS read depth, by down-sampling blood-derived data from a 70-year old healthy individual, in whose blood we observed the highest number of mosaic variants (due to clonal hematopoiesis[24]). As all programs had high validation on this sample at 200x, the recovery rate was used to distinguish the ability of different programs to detect clonal hematopoiesis variants. DeepMosaic showed similar recovery in the down-sampled data (Supplementary Fig. 6d) as M2S2MH and slightly outperformed MosaicForecast at 100x and 150x. We found that the performance of DeepMosaic was not substantially influenced by the read depth according to the down-sampling benchmark on biological data.

To understand whether different pipelines had unique strengths or weaknesses, we separated all the detected variants into 7 groups (G1-G7) based upon sharing between different pipelines, Supplementary Fig. 6a). DeepMosaic specific variants showed similar base substitution features compared with other methods (Supplementary Fig. 6b). Similar to the computationally derived data, we found that DeepMosaic recovered additional low AF MVs with high accuracy (validation rate 95%, Supplementary Fig. 6c). Finally, we summarized the genomic features of variants detected by DeepMosaic and other pipelines. All caller groups report similar ratios of intergenic and intronic variants (Supplementary Fig. 7a). Analysis of other genomic features showed DeepMosaic-specific variants (G1) expressed consistency with other groups (Supplementary Fig. 7b), reflecting that the low-fraction variants detectable only by DeepMosaic do not represent technical artifacts.

While we propose DeepMosaic as a powerful tool for mosaic variant detection, it currently is underpowered for mosaic INDELs and mosaic repetitive variant detection which might be error-prone in genome. In practice, MosaicForecast can detect variants in genomic repeat regions with high accuracy, while M2S2MH has good performance for tissue-specific variants. Thus different methods complement one another.

DeepMosaic is the first image-based tool for the accurate detection of mosaic SNVs from short-read sequencing data and does not require a matched control sample. Compared with NeuSomatic that compresses all the bases in a genomic position into 10 features[6], DeepMosaic-VM provides complete representation of information present in the BAM file. Compared with other re-coding methods like DeepVariant[14], DeepMosaic-CM has the ability to define MVs as an independent genotype and DeepMosaic-VM can be applied as an independent variant visualization tool for the user's convenience. To further integrate population information not present in the raw BAM, 4 different features are also integrated in DeepMosaic to facilitate classification.

Despite the unique features from image representation and a neural network based variant classifier, DeepMosaic can reproducibly identify the majority (~70%) of variants detectable by conventional methods; in addition, however, this unique architecture results in higher sensitivity, and the detection of variants with relatively lower AF both in simulated and experimentally derived data validated by orthogonal experiments.

Both down-sampled biological data in blood of an individual with advanced age and computationally generated data showed that DeepMosaic has the potential to identify variants at relatively high sensitivity and high accuracy for WGS at depths as low as 30x. Clonal hematopoiesis in blood without a known driver mutation is reported[27], but can be difficult to detect because of technical limitations induced by noise and lower supporting read counts[28]. For the past 10-15 years, hundreds of thousands of whole-genome sequencing datasets from clinical, commercial, or research labs have been generated at relatively low depth, but most have not been subjected to unbiased mosaicism detection due to lack of sufficiently sensitive methods. DeepMosaic could enable a genome-level unbiased detection of mutations that requires only conventional sequencing data.

By using a training set comprising representative technical artifacts such as homopolymers and truncated reads, DeepMosaic acquired the power to distinguish biologically true positive variants from false positives, which might be filtered out directly by rule-based methods like MosaicHunter[5] or MosaicForecast[4]. We also demonstrated that DeepMosaic works for various Illumina short read sequencing platforms and different library preparation strategies (PCR-amplified and PCR-free).

Although the EfficientNet-b4 at epoch6 performed best, we provide all pre-trained models (Densenet, EfficientNet, Inception-v3, and Resnet) as DeepMosaic-CM modules on GitHub. Users are provided with the options to prepare their own data with labeled genotypes for model training for DeepMosaic, to generate data-specific, personalized models, and to increase the detection specificity for DeepMosaic on specialized data sets. For instance, homopolymers and tandem repeats are increasingly recognized as important in disease and development, but are currently not detected with DeepMosaic, because of the limited training data; however, users with specialized data sets could remedy this problem by re-training.

Likewise, gnomAD population AF features used in this study also rely on rely on a matched ancestry background to avoid population stratification. Annotations such as gene names, variant functional annotations, gnomAD allelic frequency, homopolymer and dinucleotide repeat annotation, as well as segmental duplication and UCSC repeat masker regions are provided in the final output to facilitate customization, as described at the GitHub homepage of DeepMosaic (https://github.com/Virginiaxu/DeepMosaic). Finally, apart from Mutect2 single mode, DeepMosaic can also process variant lists generated by multiple methods such as the GATK HaplotypeCaller with ploidy 50 or 100[22]. Thus, DeepMosaic can be used directly as is, or can be customized to the needs of the end users, providing an adaptable and efficient mosaic variant detection workflow.

Methods

Curation of training and benchmark data

SimData1:

For the initial training procedure, 10,000 variants were randomly generated on chromosome 22 to get the list of alternative bases. Pysim[29] was then used to simulate paired-end sequencing reads with random errors generated from the Illumina HiSeq sequencer error model. Alternative reads were generated by replacing the genomic bases with the alternative bases in the list, with the same error model. Alternative and reference reads were randomly mixed to generate an alternative AF of 0, 1, 2, 3, 4, 5, 10, 15, 20, 25, and 50%. The data was randomly sampled for a targeted depth of 30, 50, 100, 150, 200, 250, 300, 400, and 500x. FASTQ files were aligned to the GRCh37d5 human reference genome with BWA (v0.7.17) *mem* command. Aligned data were processed by GATK (v3.8.1) and Picard (v2.18.27) for marking duplicate, sorting, INDEL realignment, base quality recalibration, and germline variant calling. The up- and down-sampling expanded this dataset into a pool of 990,000 different variants. Depth ratios were calculated as defined. To avoid the situation that a randomly generated mutations falls on a common SNP position in the genome, which would bias the training and benchmarking, gnomAD allele frequencies were randomly assigned from 0 to 0.001 for simulated mosaic positive and from 0 to 1 for simulated negative variants, which were established as homozygous or heterozygous.

BioData1:

Variant information and raw sequencing reads from 80-120x PCR-amplified PE-150 WGS data of 29 samples from 6 normal individuals were extracted from published data[16, 17] on SRA (SRP028833, SRP100797, and SRP136305). 921 variants identified from WGS of samples from different organs of the donors and validated by orthogonal experiments were selected and labeled as mosaic positive. 492 genomic positions from the control samples validated with 0% AF were selected and labeled as negative. 162 variants with known sequencing artifacts were first filtered by MosaicHunter, manually selected and labeled as negative. The 1575 genomic positions were also down-sampled and up-sampled for a targeted depth of 30, 50, 100, 150, 200, 250, 300, 400, and 500x, to expand this dataset into a pool of 14,175 different conditions. Depth ratios were calculated accordingly, gnomAD allele frequencies, segmental duplication, and repeat masker information was annotated.

Random subsampling from SimData1 and the entire BioData1 were assembled together to generate a training and validation dataset with approximately 200,000 variants from the 1,000,000 training variants. 180,000 variants were selected for model training. This dataset was used for the model training and evaluation of the sensitivity and specificity of the selected model, and their features including AF distribution and biological appearances were very similar to published biological data (Supplementary Fig. 1).

BioData2:

To estimate the performance of the pre-trained models and select the model with the best performance for DeepMosaic-CM, we introduced an independent gold-standard dataset. Variants were computationally detected from replicated sequencing experiments generated from 6 distinct sequencing centers and validated in 5 different centers, known as the common reference tissue project from the Brain Somatic Mosaicism Network[22]. 400 variants underwent multiple levels of computational validations including haplotype phasing, CNV exclusion, population shared exclusion, as well as experimental validation such as whole-genome single cell sequencing, Chromium Linked-read sequencing (10X Genomics), PCR amplicon sequencing, and droplet digital PCR. After validation, 43 true positive MVs and 357 false positive variants were determined as gold-standard evaluation set for low-fraction single nucleotide MVs from the 250x WGS data[22]. We extracted deep whole-genome sequences for those variants, labeled them accordingly and used them as gold standard validation set for model selection (Supplementary Fig. 2).

SimData2:
To compare the performance of DeepMosaic and other software to detect mosaicism on simulated data, we randomly generated another simulation dataset, with the following modifications: 1] only 7610 variants on non-repetitive region of chromosome 22 were considered true positive genomic positions; 2] random errors were generated from the Illumina NovaSeq sequencer error model. 3] Data was randomly down-sampled and up-sampled for a targeted depth of 50, 100, 200, 300, 400, and 500x. A total of 439,200 different variants were generated. FASTQ files were aligned and processed with BWA (v0.7.17), SAMtools (v1.9), and Picard (v2.18.27). The data was subjected to DeepMosaic as well as Mutect2 (GATK v4.0.4, both paired mode and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with different models trained for different read depth (250x model for depth≥300x).

BioData3:
This additional dataset is used to compare the performance of DeepMosaic and other mosaic variant callers on biological samples. 16 WGS samples from blood and sperm of 8 individuals were sequenced at 200x[24] (PRJNA588332). WGS was performed using an Illumina TrueSeq PCR-free kit with 350bp insertion size and sequenced on an Illumina HiSeq sequencer. Reads were aligned to the GRCh37 genome with BWA (v0.7.15) *mem* and duplicates were removed with sambamba (v0.6.6) and base quality recalibrated by GATK (v3.5.0). Processed BAM files were subjected to DeepMosaic as well as Mutect2 (GATK v4.0.4, both paired mode and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with 200x models trained for the specific depth. Data from one of the individuals (F02) was down sampled to 150x, 100x, 50x, and 30x with the SAMtools (v1.9) *view* command for the further benchmark of DeepMosaic.

Neural network building and model training

For the 10 neural network architectures, Inception-v3, Resnet and Densenet were imported from PyTorch's (v1.4.0) built-in library, while the 7 different builds of EfficientNet were imported from the efficientnet_pytorch (v0.6.1) Python (v3.7.1) package. The final fully connected layer of each model was replaced to be fed into 3 output units representing intermediate results instead of the default 1,000 output units for the 1,000 ImageNet classes to significantly reduce the total images required to extract basic features such as edges, stripes from raw images. A transfer-learning method was adopted for model training. Each model's initial pre-trained weights provided by Pytorch and efficientnet_pytorch packages were trained on the ImageNet dataset. Before model training, we randomly divided the entire training dataset (including down-sampling and up-sampling of SimData1 and BioData1) to 80% "training" and 20% "evaluation" sets and fixed the split during model training while shuffling the order within the training set and evaluation set for each training epoch to form mini-batches for gradient descent. Each network architecture was trained using a batch size of 20 with a stochastic gradient descent (SGD) optimizer with learning rate of 0.01, and momentum of 0.9. The training was terminated until the training losses plateaued and evaluation accuracy reached 90% for each model architecture. The training was conducted on NVIDIA Kepler K80 GPU Nodes on San Diego Supercomputer Centre's Comet computational clusters.

Network selection

For selecting the "best-performing" neural network architecture among the trained Inception-v3, Resnet, Densenet and 7 different builds of EfficientNet, the gold standard evaluation dataset (BioData2) was used to test each model's performance on biological (non-simulated) MVs determined by the dataset. Accuracy, MCC, True positive rates were calculated for each model and in the end EfficientNet-b4 at epoch 6 with the highest Accuracy, MCC and True positive rate among all model architectures was selected as our DeepMosaic model. The performance of DeepMosaic model (EfficientNet-b4 architecture) was further evaluated.

Usage of DeepMosaic

Detailed instructions for users as well as the demo input and output is provided on GitHub (https://github.com/Virginiaxu/DeepMosaic).

Orthogonal validation with deep amplicon sequencing

Deep amplicon sequencing analysis was applied to 241 variants from the 1355 candidates detected by all 5 mosaic variant callers from the 200× WGS of 16 samples[24] to experimentally confirm the validation rate of DeepMosaic as well as other methods. PCR products for sequencing were designed with a target length of 160-190 bp with primers being at least 60 bp from the base of interest. Primers were designed using the command-line tool of Primer3[30] with a Python (v3.7.3) wrapper. PCR was performed according to standard procedures using GoTaq Colorless Master Mix (Promega, M7832) on sperm, blood, and an unrelated control. Amplicons were enzymatically cleaned with ExoI (NEB, M0293S) and SAP (NEB, M0371S) treatment. Following normalization with the Qubit HS Kit (ThermoFisher Scientific, Q33231), amplification products were processed according to the manufacturer's protocol with AMPure XP Beads (Beckman Coulter, A63882) at a ratio of 1.2x. Library preparation was performed according to the manufacturer's protocol using a Kapa Hyper Prep Kit (Kapa Biosystems, KK8501) and barcoded independently with unique dual indexes (IDT for Illumina, 20022370). The libraries were sequenced on a NovaSeq platform with 100 bp paired-end reads. Reads from deep amplicon sequencing were mapped to the GRCH37d5 reference genome by BWA mem and processed according to GATK (v3.8.2) best practices without removing PCR duplicates. Putative mosaic sites were retrieved using SAMtools (v1.9) mpileup and pileup filtering scripts described in previous TAS pipelines[24]. Variants were considered positively validated for mosaicism if 1] their lower 95% exact binomial CI boundary was above the upper 95% CI boundary of the control; 2] their AF was >0.5%. The number of reference and alternative alleles calculated from the Amplicon validation was provided in Supplementary Table 1.

Analysis of different categories of variants overlap with different genomic features

In order to assess the distribution of MVs and their overlap with genomic features, an equal number of variants (mSNVs/INDELs as in group G1-G7 in Supplementary Fig. 6) was randomly generated with the BEDtools (v2.27.1) shuffle command within the region from Strelka2 without the subtracted regions (e.g. repeat regions). This process was repeated 10,000 times to generate a distribution and their 95% CI. Observed and randomly subsampled variants were annotated with whole-genome histone modifications data for H3k27ac, H3k27me3, H3k4me1, and H3k4me3 from ENCODE v3 downloaded from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/)—specifically for the overlap with peaks called from the H1 human embryonic cell line (H1), as well as peaks merged from 10 different cell lines (Mrg; Gm12878, H1, Hmec, Hsmm, Huvec, K562, Nha, Nhek, and Nhlf). Gene region, intronic, and exonic regions from NCBI RefSeqGene (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz); 10 Topoisomerase 2A/2B (Top2a/b) sensitive regions from ChIP-seq data (Samples: GSM2635602, GSM2635603, GSM2635606, and GSM2635607); CpG islands: data from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/); genomic regions with annotated early and late replication timing[31]; high nucleosome occupancy tendency (>0.7 as defined in the source, all values were extracted and merged) from GM12878; enhancer genomic regions from the VISTA Enhancer Browser (https://enhancer.lbl.gov/); and DNase I hypersensitive regions and transcription factor binding sites from Encode v3 tracks from the UCSC genome browser (wgEncodeRegDnaseClusteredV3 and wgEncodeRegTfbsClusteredV3, respectively).

Data availability

WGS data used to generate the training set are available at the Sequence Read Archive (SRA, Accession No. SRP028833 and SRP100797). The gold standard WGS data is available at the National Institute of Mental Health Data Archive (NIMH NDA Study ID 792: https://dx.doi.org/10.15154/1504248) and the Brain Somatic Mosaicism Consortium Data Portal (https://bsmn.synapse.org/Explore/Studies/DetailsPage?id=syn21781120). The independent sperm and blood deep WGS data are available at SRA (Accession No. PRJNA588332).

Code availability

DeepMosaic is implemented in Python; the code, documentation and demos are available at https://github.com/Virginiaxu/DeepMosaic.

References

1. Dou, Y., Gold, H.D., Luquette, L.J. & Park, P.J. Detecting Somatic Mutations in Normal Cells. *Trends in Genetics* (2018).

2. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272 (2020).

3. Lee, J.H. et al. Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243-247 (2018).

4. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched controls. *Nature biotechnology* (2020).

5. 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 research* **45**, e76 (2017).

6. Sahraeian, S.M.E. et al. Deep convolutional neural networks for accurate somatic mutation detection. *Nature communications* **10**, 1041 (2019).

7. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018).

8. Wood, D.E. et al. A machine learning approach for somatic mutation discovery. *Sci Transl Med* **10** (2018).

9. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219 (2013).

10. Biesecker, L.G. & Spinner, N.B. A genomic view of mosaicism and human disease. *Nature reviews. Genetics* **14**, 307-320 (2013).

11. Yang, X. et al. MosaicBase: A Knowledgebase of Postzygotic Mosaic Variants in Noncancer Disease-related and Healthy Human Individuals. *Genomics Proteomics Bioinformatics* (2020).

12. Poduri, A., Evrony, G.D., Cai, X. & Walsh, C.A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758 (2013).

13. Freed, D., Stevens, E.L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes* **5**, 1064-1094 (2014).

14. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* **36**, 983-987 (2018).

15. McNulty, S.N. et al. Diagnostic Utility of Next-Generation Sequencing for Disorders of Somatic Mosaicism: A Five-Year Cumulative Cohort. *Am J Hum Genet* **105**, 734-746 (2019).

16. Huang, A.Y. et al. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24**, 1311-1327 (2014).

17. 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).

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in Proceedings of the IEEE conference on computer vision and pattern recognition 2818-2826 (2016).

19. He, K., Zhang, X., Ren, S. & Sun, J. in Proceedings of the IEEE conference on computer vision and pattern recognition 770-778 (2016).

20. Iandola, F. et al. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* (2014).

21. Tan, M. & Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

22. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human brain tissue. *bioRxiv* (2020).

23. Springenberg, J.T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

24. Breuss, M.W. et al. Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat Med* **26**, 143-150 (2020).

25. Yang, X. et al. Temporal stability of human sperm mosaic mutations results in life-long threat of transmission to offspring. *bioRxiv* (2020).

26. Pelorosso, C. et al. Somatic double-hit in MTOR and RPS6 in hemimegalencephaly with intractable epilepsy. *Hum Mol Genet* **28**, 3755-3765 (2019).

27. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742-752 (2017).

28. Lawson, A.R.J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75-82 (2020).

29. Xia, Y., Liu, Y., Deng, M. & Xi, R. Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC bioinformatics* **18**, 53 (2017).

30. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289-1291 (2007).

31. Hansen, R.S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139-144 (2010).

Acknowledgement

The authors thank Dr. Yanmei Dou for helping to set up the MosaicForecast pipeline. We thank Profs. Peter J. Park, Garrison W. Cottrell, John V. Moran, Melissa Gymrek, Drs. Patrick J. Reed, August Y. Huang, and Si-Jin Cheng for their valuable comments and suggestions. This work was supported by the National Institute of Mental Health (U01MH108898), Rady Children's Institute for Genomic Medicine and Howard Hughes Medical Institute. The authors thank San Diego Supercomputer Center (SDSC) (TG-IBN190021). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq

6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929).

Contributions:

X.Y., X. X., and J.G.G. conceived this project with input from M.B. and D.A.; X.Y. designed the study workflow and managed the project. X.X. implemented the image representation and neural network classifier under supervision and instruction by X.Y.; X.Y., C.L. and X.X. generated the training data with the help from D. A. and R.D.G.; X.X. performed the training and model selection under supervision by X.Y.; independent dataset were processed by M.B., D.A., and R.D.G. under supervision by J.S. and J.G.G.; X.Y. and M.B. performed the validation experiments with help from L.L.B. and C.C.; X.Y. and X.X. wrote the original manuscript with input from all listed authors; X.Y. and J.G.G. edited the manuscript. DeepMosaic is benchmarked on part of the Brain Somatic Mosaicism Network (BSMN) common brain experiment and common analysis pipeline for SNVs contributed by Y.W., T.B. under supervision by A.A.; J.G.G. supervised this project. All authors discussed the results and contributed to the final manuscript.

Competing interests:

The authors declare no competing interests.

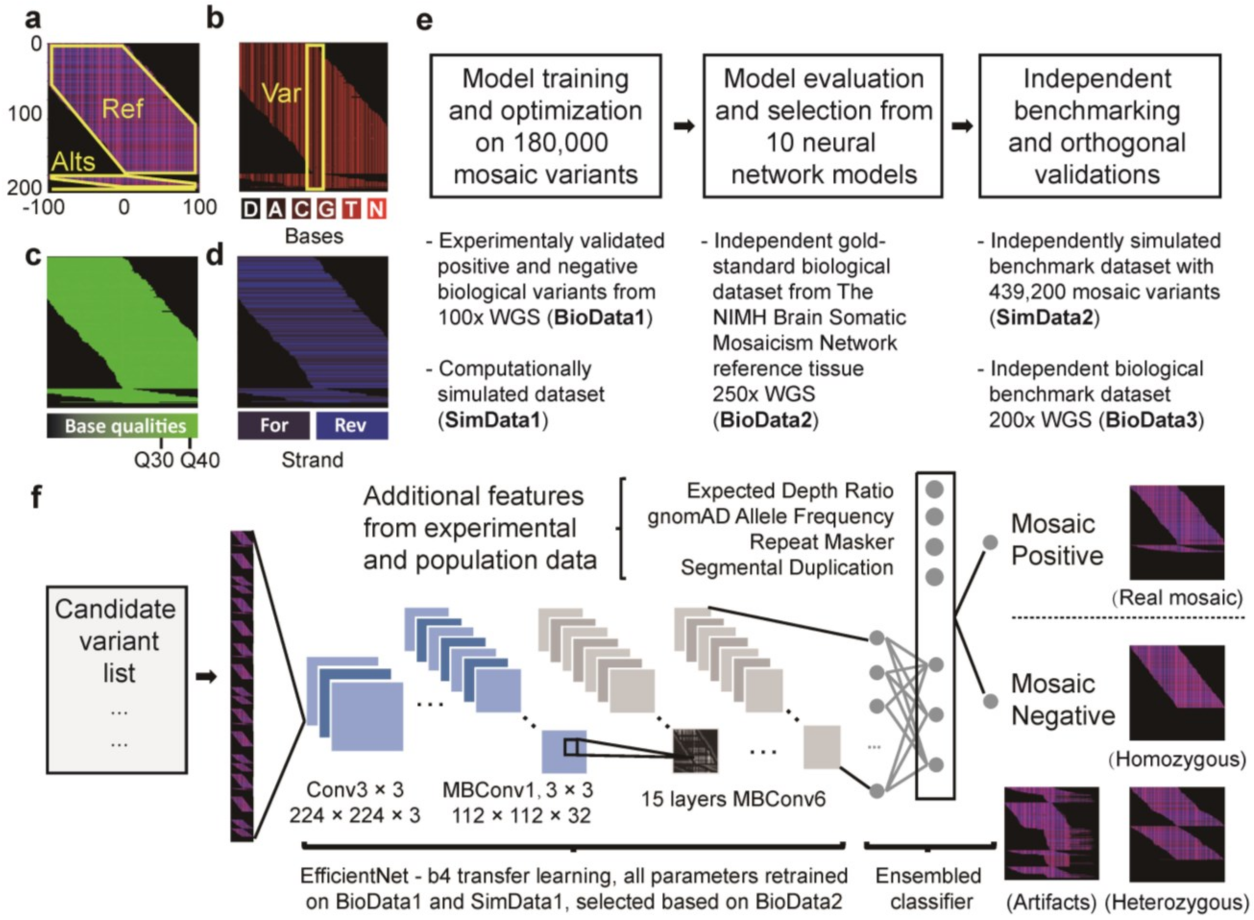
a
0
100
200
-100
0
100
Ref
Alts
b
Var
D A C G T N
Bases
c
Base qualities
Q30 Q40
d
For
Rev
Strand
e
Model training and optimization on 180,000 mosaic variants
Model evaluation and selection from 10 neural network models
Independent benchmarking and orthogonal validations
- Experimentaly validated positive and negative biological variants from 100x WGS (BioData1)
- Computationally simulated dataset (SimData1)
- Independent gold-standard biological dataset from The NIMH Brain Somatic Mosaicism Network reference tissue 250x WGS (BioData2)
- Independently simulated benchmark dataset with 439,200 mosaic variants (SimData2)
- Independent biological benchmark dataset 200x WGS (BioData3)
f
Candidate variant list
...
...
Additional features from experimental and population data
Expected Depth Ratio
gnomAD Allele Frequency
Repeat Masker
Segmental Duplication
Conv3 × 3
224 × 224 × 3
MBConv1, 3 × 3
112 × 112 × 32
15 layers MBConv6
Mosaic Positive
(Real mosaic)
Mosaic Negative
(Homozygous)
(Artifacts)
(Heterozygous)
EfficientNet - b4 transfer learning, all parameters retrained on BioData1 and SimData1, selected based on BioData2
Ensembled classifier

Fig. 1| Image representation, model training strategies, and framework of DeepMosaic.

a, DeepMosaic-VM: Composite RGB image representation of sequenced reads separated into "Ref" - reads supporting the reference allele; or "Alts" - reads supporting alternative alleles; each outlined in yellow. **b**, Red channel of compound image contains base information from BAM file. "D" - deletion; "A" – Adenine; "C" – cytosine; "G" – guanine; "T" – thymine; "N" – low-quality base. Yellow box: Var: candidate position, centered in the image. **c**, Green channel: base quality information. Note that channel intensity was modulated in this example for better visualization. **d**, Blue channel: strand information (i.e. forward or reverse). **e**, Model training, model selection, and overall benchmark strategy for DeepMosaic-CM (Methods and Supplementary Fig. 1). Ten different convolutional neural network models were trained on 180,000 experimentally validated positive and negative biological variants from 29 WGS data from 6 individuals sequenced at 100x[16, 17] (BioData1), as well as simulated data with different AFs (SimData1) resampled to different depth. Models were evaluated based upon an independent gold-standard biological dataset from the 250x WGS data of the Brain Somatic Mosaicism Network Reference Tissue Project[22] (BioData2). DeepMosaic was further benchmarked on 16 independent biological datasets from 200x WGS data[24] (BioData3) as well as 439,200 independently simulated variants (SimData2). Deep amplicon sequencing was carried out as an independent evaluation on variants detected by different software

(Supplement Table 1). **f**, Application of DeepMosaic-CM in practice. Input images are generated from the candidate variants. 16 convolutional layers extracted information from input images. Population genomic features were assembled for final output. Images of positive and negative variants are shown as examples. Conv: convolutional layers; MBConv: mobile convolutional layers.

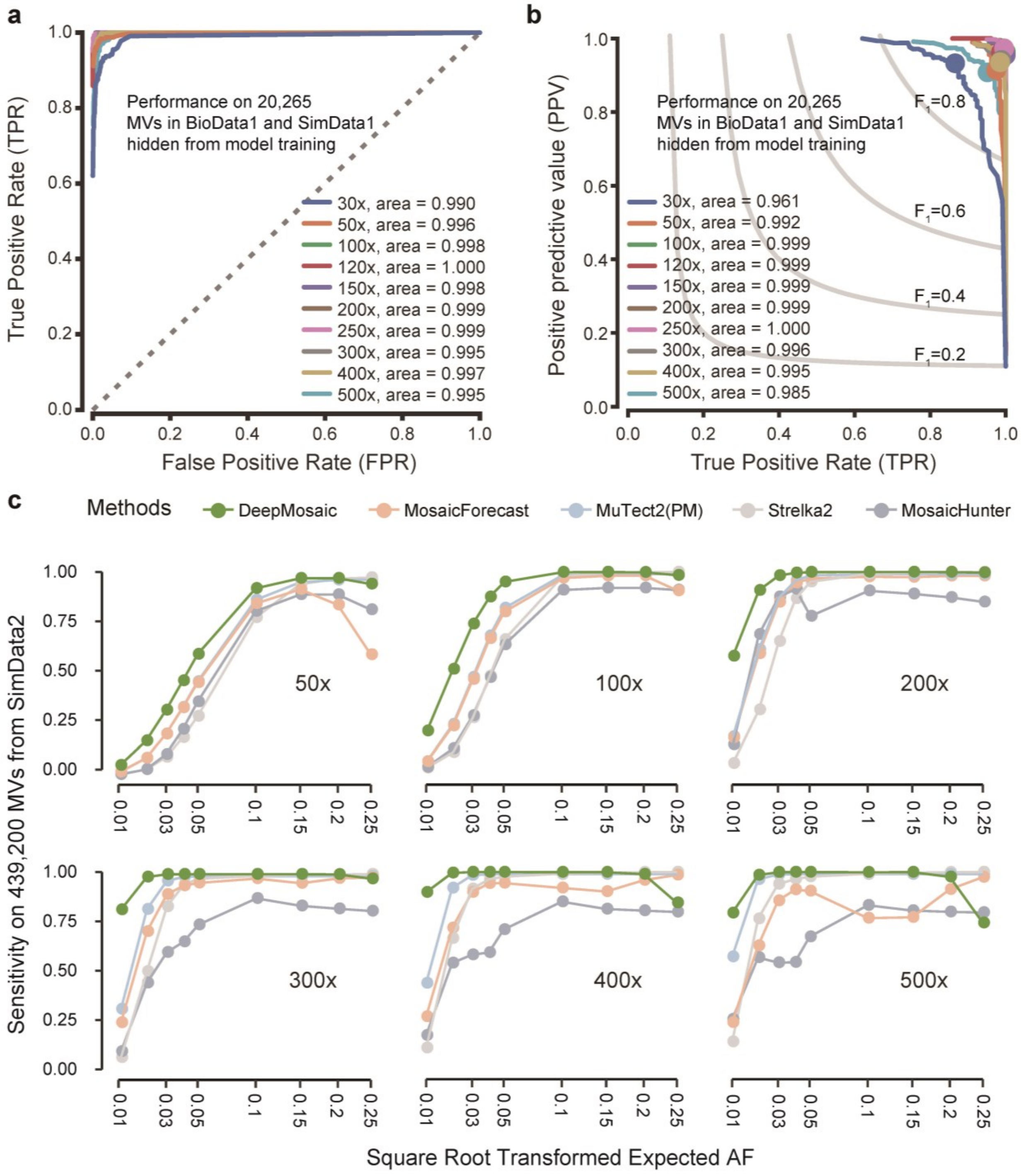

Fig. 2| DeepMosaic showed high performance on simulated benchmark variants.

a, Receiver operating characteristic (ROC) curve for DeepMosaic. True positive rates (TPR) and false positive rates (FPR) were evaluated from 20,265 variants (BioData1 and SimData1) hidden from model training and model selection. Colors show groups of intended read depth. **b**, Precision-recall curves for DeepMosaic, evaluated from the 20,265 hidden variants, dots showed the performance of the default parameters for DeepMosaic-CM. ROC and precision-recall curves for DeepMosaic on different AFs are provided in Supplementary Fig. 3. **c**, Sensitivity of DeepMosaic and other mosaic callers on 439,200 independently simulated benchmark variants (SimData2) at simulated read depths and AFs. DeepMosaic performed equally well or better than other tested methods, especially at lower read-depths and lower expected AFs.

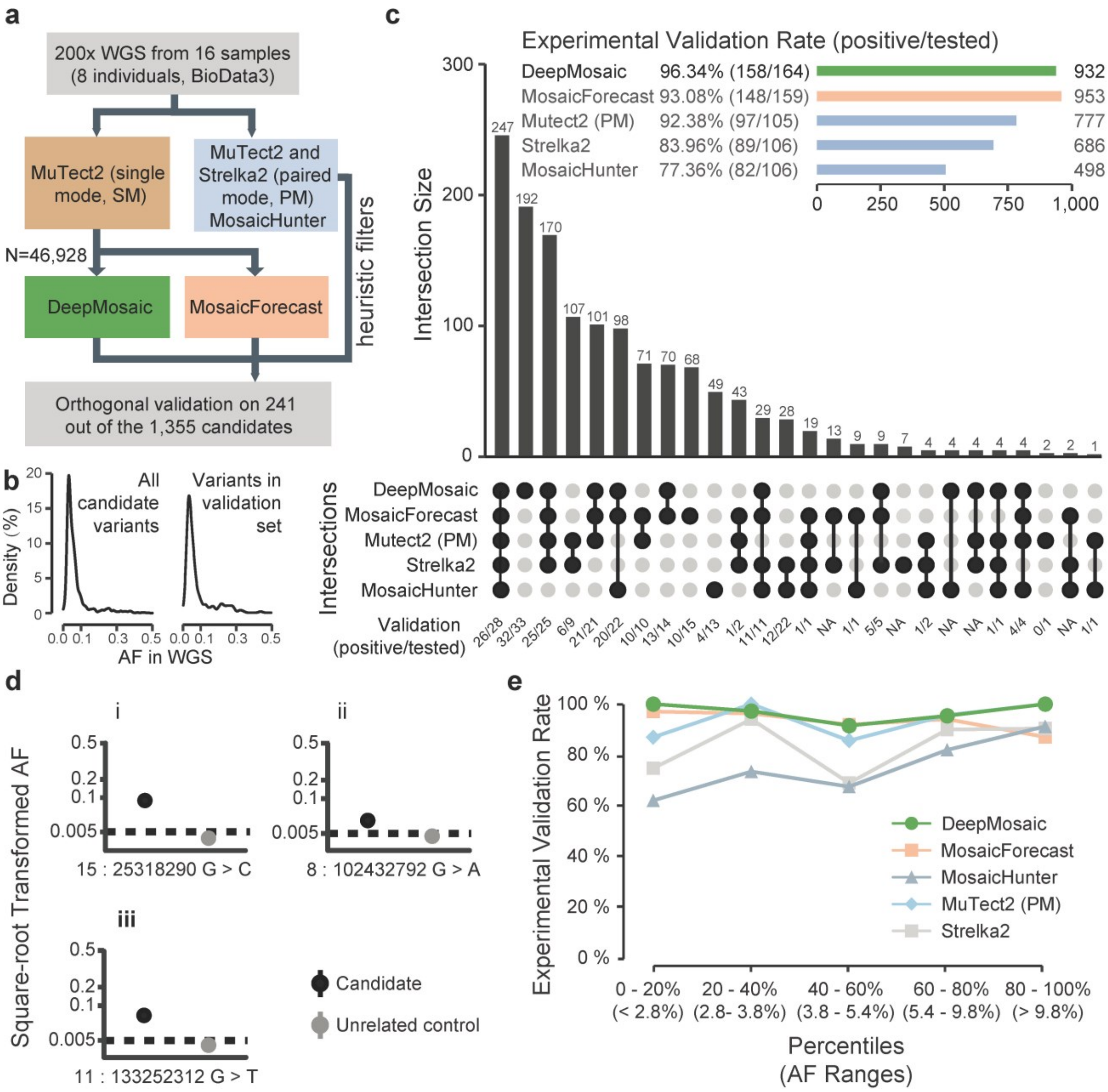

Fig. 3| DeepMosaic performance validated on biological data.

a, DeepMosaic and other mosaic variant detection methods were applied to 200x whole-genome sequencing data from 16 samples, which were not used in the training or validation stage for any of the listed methods (BioData3). Raw variant lists were either obtained by comparing samples using a panel-of-normal[25] strategy with MuTect2 single mode, between different samples from a same individual using MuTect2 paired mode or Strelka2 somatic mode, or detected directly without control with MosaicHunter single mode with heuristic filters[24]. A total of 46,928 candidate variants from MuTect2 single mode were analyzed by DeepMosaic and MosaicForecast. Orthogonal validation with deep amplicon sequencing was carried out on a total of 241 variants out of the 1355 candidates called by at least one method. **b**, Distribution of AFs of the whole candidate mosaic variant list and the 241 randomly selected variants. **c**, Comparison of validation results between different mosaic variant calling methods, 'UpSet' plot shows the intersection of different mosaic detection methods and the validation result of each category. Variants identified by DeepMosaic showed the highest validation rate on biological data. **d**, Examples of validated variants called by DeepMosaic and MosaicForecast (i), only by DeepMosaic (ii), or by DeepMosaic and other traditional methods (iii). **e**, Comparison of validation rate in different AF range percentage bins of variants. DeepMosaic showed the highest validation rate at a range of AFs, approximately 48 experimentally validated variants are shown in each AF bin.