nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01559-w

Control-independent mosaic single nucleotide variant detection with DeepMosaic

Received: 13 November 2020

Accepted: 10 October 2022

Published online: 02 January 2023

Check for updates

Xiaoxu Yang[1,2,27] ✉, Xin Xu[1,2,27], Martin W. Breuss[1,2,3], Danny Antaki[1,2], Laurel L. Ball[1,2], Changuk Chung[1,2], Jiawei Shen[1,2], Chen Li[1,2], Renee D. George[1,2], Yifan Wang[4], Taejeong Bae[4], Yuhe Cheng[5,6,7], Alexej Abyzov[4], Liping Wei[8], Ludmil B. Alexandrov[5,6,7], Jonathan L. Sebat[9,10,11,12], NIMH Brain Somatic Mosaicism Network* & Joseph G. Gleeson[1,2] ✉

Mosaic variants (MVs) reflect mutagenic processes during embryonic development and environmental exposure, accumulate with aging and underlie diseases such as cancer and autism. The detection of noncancer MVs has been computationally challenging due to the sparse representation of nonclonally expanded MVs. Here we present DeepMosaic, combining an image-based visualization module for single nucleotide MVs and a convolutional neural network-based classification module for control-independent MV detection. DeepMosaic was trained on 180,000 simulated or experimentally assessed MVs, and was benchmarked on 619,740 simulated MVs and 530 independent biologically tested MVs from 16 genomes and 181 exomes. DeepMosaic achieved higher accuracy compared with existing methods on biological data, with a sensitivity of 0.78, specificity of 0.83 and positive predictive value of 0.96 on noncancer whole-genome sequencing data, as well as doubling the validation rate over previous best-practice methods on noncancer whole-exome sequencing data (0.43 versus 0.18). DeepMosaic represents an accurate MV classifier for noncancer samples that can be implemented as an alternative or complement to existing methods.

Postzygotic mosaicism describes a phenomenon whereby cells arising from one zygote harbor distinguishing genomic variants[1,2]. MVs can act as recorders of embryonic development, cellular lineage and environmental exposure. They accumulate with aging[2,3] and are implicated in over 200 noncancerous disorders[4,5]. Collectively, MVs are estimated to contribute to 5–10% of the 'missing genetic heritability' in more than 100 human disorders[4,6–8].

Compared with the higher allelic fractions (AF) of 5–10% found in cancer or precancerous mosaic conditions, AFs found in nonclonal disorders, or neutral variants used for lineage studies, are frequently an order of magnitude lower. Existing methods such as MuTect2 (ref. [9]) and Strelka2 (ref. [10]), however, are based on classic statistical models, using heuristic filters that are often optimized for higher AF MVs seen in cancer. Similarly, because of their conceptual origin in cancer, most existing programs, including the recent NeuSomatic[11], also require matched 'noncancer' control samples. This can be problematic when mutations are present across different tissues ('tissue shared' mosaicism), or when only one sample is available.

Recent methods that aim to overcome these limitations in noncancer MV detection, MosaicHunter[12] and MosaicForecast[13], are based on conceptually similar uses of features extracted from raw data, rather than the sequence and alignment themselves, or replace heuristic

A full list of affiliations appears at the end of the paper. ✉e-mail: xiy010@health.ucsd.edu; jogleeson@health.ucsd.edu

filters with traditional machine-learning methods. While these are useful proxies, they represent a limited window into the sheer wealth of information contained in raw sequencing data. Furthermore, performance on whole-exome sequencing (WES) data is limited due to the intrinsically biased experimental nature of exome capture[13,14]. To address 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 MVs[15]. This is a laborious and low-throughput process that allows spot checking, but cannot be implemented on a large scale for putative MV lists generated from programs such as MuTect2 or GATK HaplotypeCaller using 'ploidy' setting of 50 (ref. [16]) often numbering in the thousands.

Image-based representation of raw sequencing reads and the application of deep convolutional neural networks (CNNs) represent a potential solution for these limitations for non-MV detection. An example approach such as DeepVariant[11] was designed and trained for detecting only heterozygous or homozygous single nucleotide variants (SNVs) from direct representation of aligned reads. DeepVariant, however, is not designed for noncancer MV detection. Here, we introduce DeepMosaic (https://github.com/Virginiaxu/DeepMosaic), comprising two modules—a visualization module for image-based representation of SNVs separating reference and alternative supporting reads—as input for a CNN-based classification module for MV detection. Nine different biological and computationally simulated datasets were used to train and benchmark DeepMosaic. Finally, large-scale deep amplicon sequencing (Methods) provided an orthogonal experimentally validated set of DeepMosaic-detected variants and allowed for direct comparison with other state-of-the-art methods.

Results

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 sequencing data, processed with a GATK current best-practice pipeline (insertion/deletion, or INDEL, realignment, followed by base quality score recalibration). DeepMosaic-VM exports this data into an RGB (red, green, blue) image, representing a pileup at each genomic position (Methods). In contrast to a regular browser snapshot, DeepMosaic encodes sequences as different intensities within one channel and uses other channels for base quality and strand orientation. DeepMosaic further separates the pileups into 'ref' reads and 'alt' reads on the basis of the reference genome information (Fig. 1a–d). This improved pileup visualization allows the assessment of mosaicism at a glance for humans, and converts the biological variant detection problem into an edge- and shape-detection problem, which is more suitable for image-based classification by CNN models.

The classification module of DeepMosaic (DeepMosaic-CM) is based on CNN transfer learning for MVs. We trained ten different CNN models with more than 180,000 image-based representations from both true-positive and true-negative biological variants derived from several previously published high-quality public datasets with orthogonal experimental validations by amplicon sequencing or droplet digital PCR techniques[17–19]. To subsidize the requirement for CNN training, we included roughly 50% computationally simulated reads with spiked-in MVs (using Illumina HiSeq error models) across a range of AFs and depths (Fig. 1e, Methods and Extended Data Fig. 1a,b). This training dataset (BioData1 and SimData1) was aimed to train a model with optimal performance on noncancer MV detection. To ensure performance in 'real-world' settings, we matched the distribution of AFs in the training set with experimentally determined AFs (Extended Data Fig. 1c). In addition, a range of expected technical artifacts, including false-positive variants with multiple alternative alleles, false-positive variants located near homopolymers >5 bp or on the edge of dinucleotide repeats and variants that are detected as alignment artifacts after the validation experiments[17,18], were manually curated and labeled as negative in the training set to represent expected pitfalls that often result in false-positive noncancer MVs for other programs (Extended Data 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 30× to 500× (Extended Data Fig. 1e), covering the range of the most commonly used whole-genome sequencing (WGS) and WES read depths in current clinical and scientific settings. In addition to the output from DeepMosaic-VM, we further incorporated population genomic and sequence features (for example, population allele frequency from population study, genomic complexity from field-acceptable database, the ratio of read depth by experimental design), 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 (CNVs) or aneuploidies. gnomAD population allelic frequencies were used to exclude common variants[20]. Segmental duplication and repeat masker regions were used to exclude 24% of the genome consisting of low complexity regions.

Ten different CNN architectures were trained on the 180,000 variants described above. The CNN models included Inception-v.3 (ref. [21]), which was retrained and used by DeepVariant; Deep Residual Network[22] (Resnet), which was retrained and used in the control-dependent caller NeuSomatic; Densenet[23] and seven different builds of EfficientNet[24], to optimize performance on rapid image classification (details are documented in the Methods, network structure and dimension of convolutional layers are provided in Extended Data Fig. 2a). Each model was trained with 5–15 epochs to optimize the hyper-parameters until training accuracies plateaued (>0.90).

To compare the performance of posttraining models and to contrast models trained with distinct datasets, we used an independent gold-standard validation dataset of roughly 400 MVs from one gold-standard brain sample generated by the Brain Somatic Mosaicism Network[16] (BSMN) (BioData2, Methods) and another amplicon-validated dataset from 18 samples from a publicly available dataset we recently published[19] (BioData3, Methods). On these, EfficientNet-b4 showed the highest accuracy, Matthews's correlation coefficient and true-positive rate when trained for six epochs (Extended Data Fig. 2b). We thus selected this as the default model of DeepMosaic-CM (Fig. 1f). Additional EfficientNet-b4 models trained on the 1:1 mixture of biological and simulated data showed a similar performance compared with the biological-data-only training set but much higher specificity compared with models trained on simulated-data-only training set (Extended Data Fig. 2c).

To uncover the information prioritized by the selected default model, we used a gradient visualization technique with guided backpropagation[25] to highlight the pixels guiding classification decisions (Extended Data Fig. 3). The results indicated that the algorithm not only recognized the edges for reference and alternative alleles, but also integrated additional available information, such as insertion/deletions, 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 training data that were hidden from model training and selection. The receiver operating characteristic curve and precision-recall curves on the hidden validation dataset showed >0.99 area under the curve for a range of coverages (30× to roughly 500×, Extended Data Fig. 4a,b) across a range of AFs (Extended Data Fig. 4c,d), demonstrating good sensitivity and specificity.

Next, we benchmarked DeepMosaic's performance relative to other detection software, using data generated from two distinct sequencing error models to test for its use on general sequencing data. We compared the performance of DeepMosaic with the widely used MuTect2 (paired mode) and Strelka2 (somatic mode) followed by heuristic filters, MosaicHunter (single mode) and MosaicForecast

https://doi.org/10.1038/s41587-022-01559-w

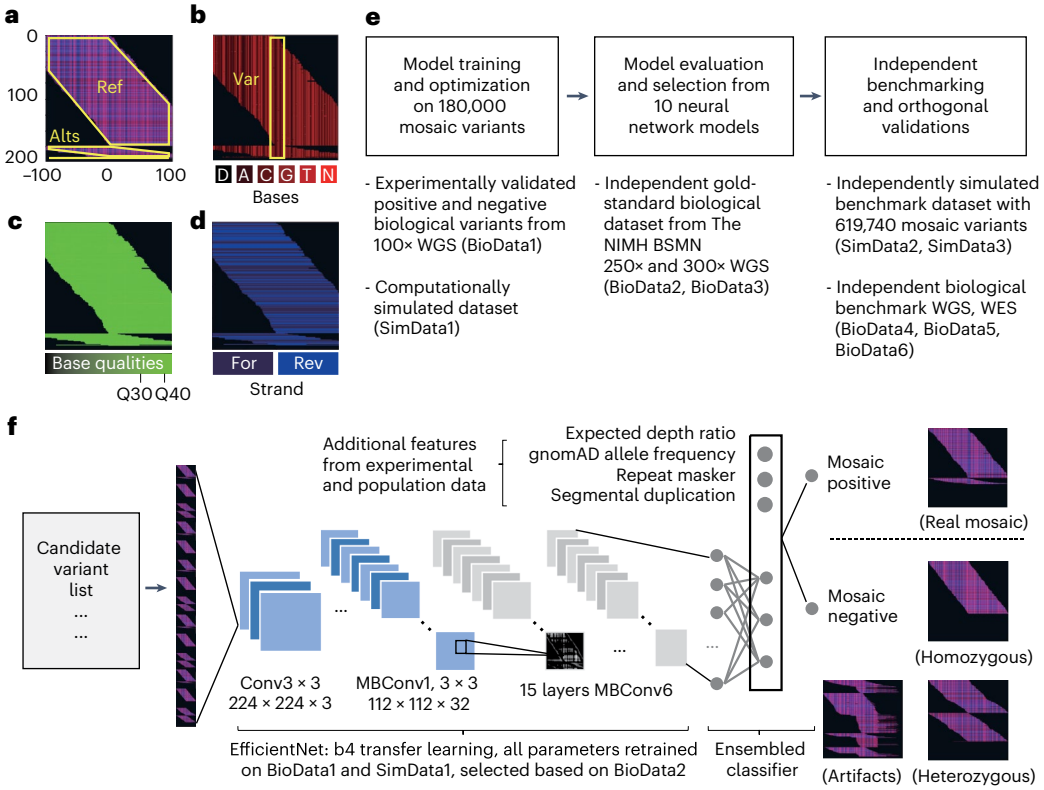


Fig. 1 | Image representation, model training strategies and framework of DeepMosaic. a, DeepMosaic-VM. A composite RGB image representation of sequenced reads separated into Ref, reads supporting the reference allele, or Alts, reads supporting alternative alleles, each outlined in yellow. **b**, Red channel of the compound image contains base information from the BAM file. D, deletion; A, adenine; C, cytosine; G, guanine; T, thymine; N, low-quality base. Yellow box shows Var, the candidate variant position centered in the image. **c**, Green channel showing base quality information. Note that channel intensity was modulated in this example for better visualization. **d**, Blue channel showing strand information (that is, forward or reverse). **e**, Model training, model selection and overall benchmark strategy for DeepMosaic-CM (Methods and Extended Data Fig. 1). Ten different CNN models were trained on 180,000 experimentally validated positive and negative biological variants from 29 whole-genome sequencing (WGS) data from six individuals sequenced at 100× (refs. [17,18]) (BioData1), as well as simulated data with different AFs (SimData1) resampled to a different depth. Models were evaluated on the basis of an independent gold-standard biological dataset from the 250× WGS data of the Reference Tissue Project of the BSMN[16] (BioData2) as well as an independent 300× WGS dataset from the BSMN capstone project[19] (Biodata3). DeepMosaic was further benchmarked on 16 independent biological datasets from 200× WGS data[28] (BioData4), on 181 independently generated 300× noncancer WES data (BioData5), 2,430 TCGA-MC3 WES samples (BioData6), as well as 619,740 independently simulated variants (SimData2 and SimData3). Deep amplicon sequencing was carried out as an independent evaluation on variants detected by different software (Supplementary 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.

(Methods). We generated two additional computationally simulated datasets of 439,200 and 180,540 positions on the basis of the error model of a different Illumina sequencer with similar methods to the training set (NovaSeq, SimData2, Methods) or a similar ratio of true-positive and true-negative labels to real biological data[19] by replacing reads from the 'Genome in a Bottle' sample HG002 (NA24345, SimData3, Methods)[26,27], with AF ranges from 1 to 25%, and depth ranges from 50× to 500×. MuTect2 paired mode and Strelka2 somatic mode used simulated mutated samples as 'tumor' and simulated reference or original HG002 samples as 'normal' for their paired modes. DeepMosaic showed equal or better performance than all other methods tested, especially for low AF variants (Fig. 2 and Extended Data Fig. 5), noticeably, even for low read depth data (50×), and it performed better than methods that use the additional information from paired samples. Overall DeepMosaic showed a 1.5–3-fold increased detection sensitivity for AFs under 3% compared with other methods, with comparable specificity (Fig. 2). This is probably because our models integrate additional genomic sequence and quality information from the original BAM file and are capable of distinguishing MVs from false-positive variants resulting from different sequencing errors.

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 the previously published 200× WGS dataset with 16 samples (blood and sperm) from eight healthy individuals[7,28] (BioData4). Paired methods compared two samples from the same individual, and control-independent samples used a published dataset of a panel of normal people[7]. Variants detected by MuTect2 (paired mode), Strelka2 (somatic mode) and MosaicHunter (single mode) were subjected to a series of published heuristic filters[7,28]. As we had access to the biological samples, we also performed orthogonal validation, using deep amplicon sequencing of 239 MVs with a representative AF distribution compared to the complete candidate variant list (Methods, Fig. 3a,b and Supplementary Table 1).

As expected from the test of the computationally generated data, DeepMosaic showed a high sensitivity (0.78), specificity (0.83), accuracy (0.79) and overall validation rate (96.3%, 158 of 164) among all five methods (Fig. 3c), demonstrating that DeepMosaic, trained on PCR-amplified biological data and simulated data, can accurately

https://doi.org/10.1038/s41587-022-01559-w

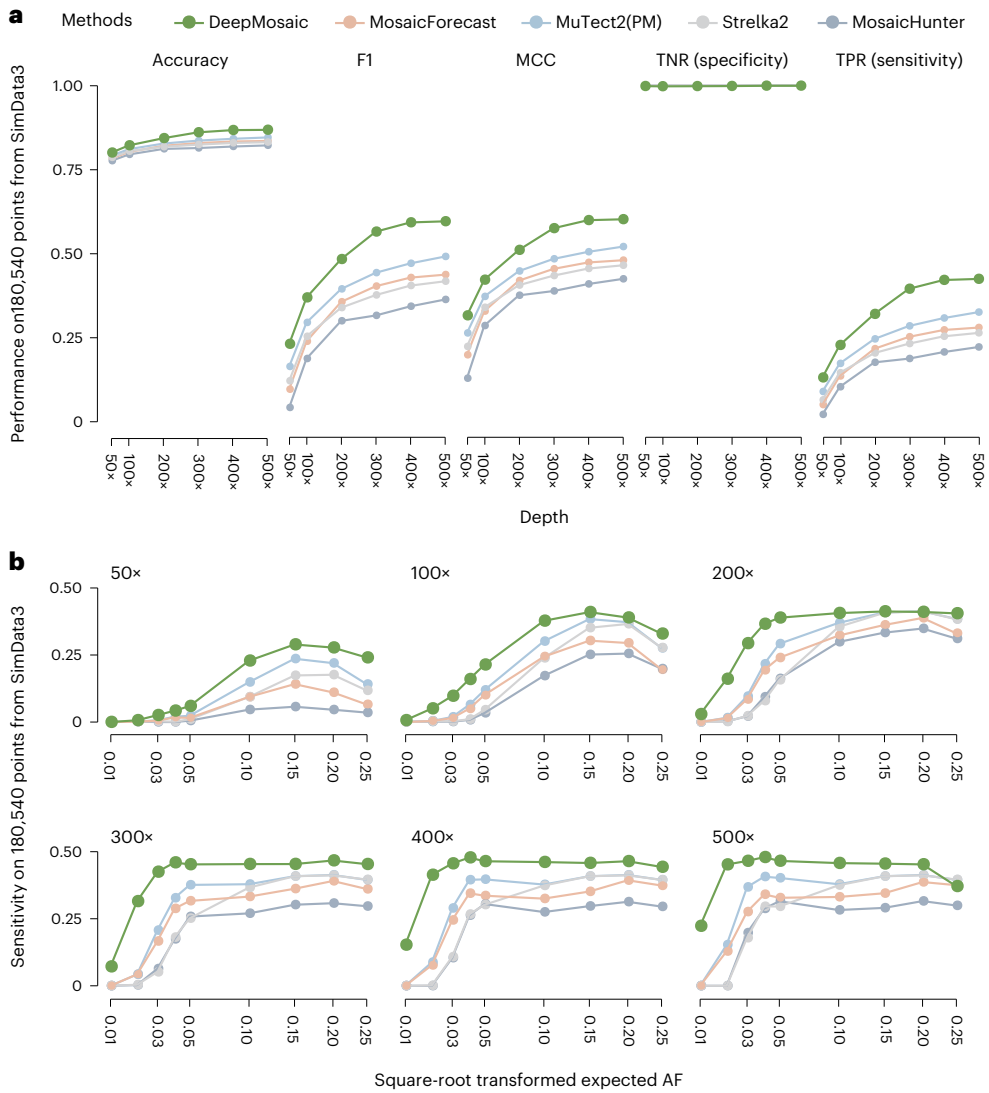


Fig. 2 | DeepMosaic performance on simulated benchmark variants. a, Benchmark test on 180,540 genomic positions (SimData3) generated by replacing reads from biological data with simulated MVs. DeepMosaic showed higher accuracy, F1 score, MCC (Matthews correlation coefficient), TPR (true positive rate, sensitivity) and comparable TNR (true negative rate, specificity) compared with widely accepted methods for MV detection, specificity of all callers are close to 1. **b**, Sensitivity of DeepMosaic and other mosaic callers on SimData3 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. Variant-stabilized square-root transformation was used for visualization purposes.

classify PCR-free biological data. Of the 819 WGS MVs detectable by DeepMosaic, 21.0% (172 of 819, 33 of 34 experimentally validated as positive) were overlooked by MosaicForecast, 30.1% (247 of 819, 96 of 98 validated) by MosaicHunter, 26.7% (219 of 819, 90 of 94 validated) by Strelka2 (somatic mode) with heuristic filters and 42.9% (351 of 819, 81 of 85 validated) by MuTect2 (paired mode) with heuristic filters[28]. DeepMosaic also accurately detected variants with relatively low AF and outperformed other methods across most of the AF bins (Fig. 3d). We additionally tested the performance of NeuSomatic on the same dataset, and it showed higher specificity (0.92) but much lower sensitivity (0.33) on the orthogonally validated dataset (Extended Data Fig. 6). Also, 49.9% DeepMosaic variants (409 of 819, 99 of 105 validated) were missed by NeuSomatic, indicating that neural networks trained on cancer data might underperform on a noncancer biological dataset.

In current practice, researchers often combine multiple programs in one variant detection pipeline to detect different categories of MVs[7,28,29]. We thus further compared DeepMosaic with different WGS pipelines used in recent publications, using data from 200× WGS of the 16 samples[28]: (1) with the MosaicForecast pipeline[13], which uses MuTect2 single mode (each sample compared with the publicly available panel of normal) as input; (2) with what we recently published as the M2S2MH pipeline[28], combining MuTect2 paired mode (that is, compared between different samples from the same individual), Strelka2 somatic mode and MosaicHunter single mode followed by a series of heuristic filters (Extended Data Fig. 7a). Of the 819 MVs identified by DeepMosaic, 79.0% (647 of 819, 125 of 130 validated) overlapped with MosaicForecast and 68.4% (560 of 819, 87 of 91 validated) overlapped with M2S2MH. By contrast, 21.0% (172 of 819, 33 of 34 validated) were undetected by MosaicForecast, and 33.0% (271 of 819, 71 of 73 validated) were overlooked by M2S2MH. These variants, uniquely detected by DeepMosaic, all showed validation rates >95% (Extended Data Fig. 7b–d), demonstrating accurate detection of a considerable number of variants undetectable by widely used methods.

To test the performance of DeepMosaic on data widely curated clinically, we compared detection sensitivity for genome samples with standard WGS read depth, by down-sampling blood-derived data

https://doi.org/10.1038/s41587-022-01559-w

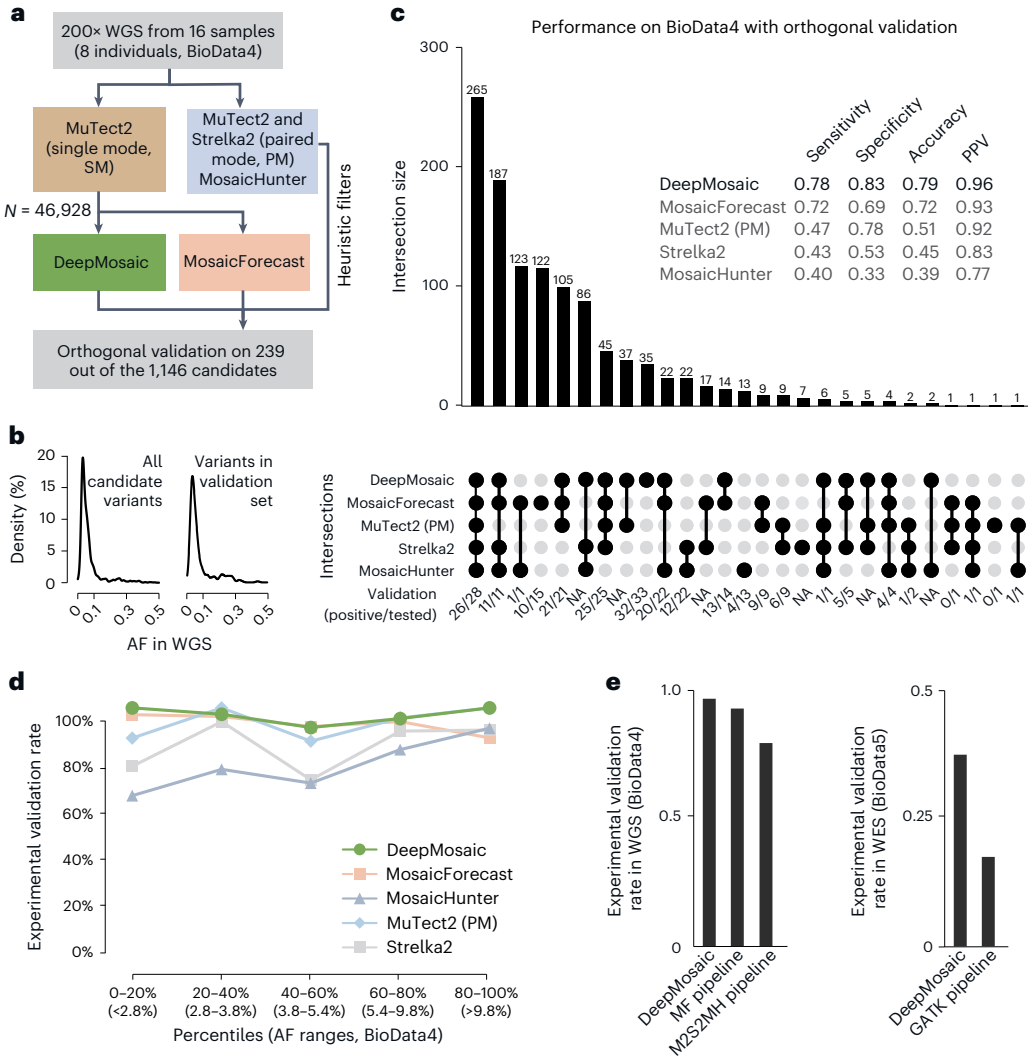


Fig. 3 | DeepMosaic performance validated on biological data. a, DeepMosaic and other MV detection methods were applied to 200× WGS data from 16 samples, which were not used in the training or validation stage for any of the listed methods (BioData4). Raw variant lists were either obtained by comparing samples using a panel-of-normal[7] strategy with MuTect2 single mode, between different samples from the same individual using MuTect2 paired mode or Strelka2 somatic mode or detected directly without control with MosaicHunter single mode with heuristic filters[28]. 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 239 variants out of the 1,146 candidates called by at least one method. **b**, Distribution of AFs of the whole candidate MV list and the 239 experimentally quantified variants. **c**, Comparison of validation results between different MV calling methods. The UpSet plot shows the intersection of different mosaic detection methods and the validation result of each category. Variants identified by DeepMosaic showed high sensitivity and specificity to biological data. **d**, 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. **e**, Comparision of experimental validation rate of DeepMosaic on WGS (BioData4) and WES (BioData5) outperforms other computational pipelines. PPV: positive predictive value; PM: paired mode; MF: MosaicForecast.

from a 70-year-old healthy individual, in whose blood we observed the highest number of MVs (due to clonal hematopoiesis[28]). As all programs had high validation on this sample at 200×, the recovery rate was used to distinguish the ability of different programs to detect clonal hematopoiesis variants. DeepMosaic showed a similar recovery in the down-sampled data (Extended Data Fig. 8) as M2S2MH and slightly outperformed MosaicForecast at 100× and 150×. 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 WGS pipelines had unique strengths or weaknesses, we separated all the detected variants into seven groups (G1-G7) on the basis of sharing between different pipelines (Extended Data Fig. 8a). DeepMosaic-specific variants showed similar base-substitution features compared with other methods (Extended Data Fig. 8b). Similar to the computationally derived data, we found that DeepMosaic recovered additional low AF MVs with high accuracy (validation rate 95%, Extended Data Fig. 8c). Finally, we summarized the genomic features of variants detected by DeepMosaic and other pipelines. All caller groups reported similar ratios of intergenic and intronic variants (Extended Data Fig. 9a). Analysis of other genomic features showed DeepMosaic-specific variants (G1) expressed consistency with other groups (Extended Data Fig. 9b), reflecting that the low-fraction variants detectable by only DeepMosaic did not represent technical artifacts. Further detailed single base substitution (SBS) signature analyses with a larger number of variants should shed light on genetic mechanisms detected by DeepMosaic and other callers at different AF.

Compared to WGS, the selection of exonic regions in WES is prone to capture bias and other artifacts, which could affect MV detection, often

exhibiting relatively low detection rates and high false-positive rates[14]. Furthermore, methods such as MosaicForecast do not explicitly support WES[13]. As the image representation of WES—unlike the classical feature extraction methods—is expected to be similar to WGS, we postulated that DeepMosaic, even as currently trained on genome data, could perform satisfactorily on WES data as well. Thus, we benchmarked DeepMosaic on our recent noncancer WES dataset. Of 181 samples from 101 individuals who underwent 300× WES (BioData5, Methods and Extended Data Fig. 10a,b), candidate MVs were detected by DeepMosaic and the BSMN best-practice pipeline[16]. Experimental validation on 291 of the 585 candidate variants showed a higher validation rate for DeepMosaic (43.1%) compared with the best-practice pipeline (17.6%, Fig. 3e and Supplementary Table 2). DeepMosaic also consistently demonstrated high specificity (0.86) and accuracy (0.78), with compromised sensitivity (0.43) probably due to differences in the nature of exonic and genomic sequencing. Thus, DeepMosaic has the ability to complement and potentially improve on existing pipelines on large-scale existing noncancer WES data.

Compared to noncancer MVs, the clonal expansion of cancer-related MVs will result in a much higher portion of cancer MVs in the genome, and different mutation features for tools to recognize. NeuSomatic was trained on cancer samples and demonstrated lower sensitivity on noncancer MVs (Extended Data Fig. 6). We expect similar performance for DeepMosaic for cancer samples, and thus systematically estimated the performance of DeepMosaic, with its current models, on cancer WES data. We reprocessed 2,430 WES samples from the TCGA-MC3 data collection (BioData6 and Methods), used DeepMosaic, and compared the result with five other callers (MuSE[30], MuTect[9], SomaticSniper[31], VarScan2 (ref. [32]) and Radia[33]) provided in the original publications[34]. Benchmarked by the gold-standard dataset[34], DeepMosaic demonstrated high specificity (0.97) and accuracy (0.77), similar to noncancer WES and WGS. Due to the different nature of somatic mutations described above, however, the sensitivity was lower (0.08, Supplementary Table 3 and Supplementary Table 4). We found that some variants defined as technical artifacts in the DeepMosaic training set, for example, 24% variants with AFs higher than 0.5 (674,175 of 2,814,168 total variants), variants with multiple alternative alleles and/or with copy number alternations or aneuploidies, were defined as true positive in tumor samples. As such, we do not recommend the use of DeepMosaic for cancer in its current form, unless further training sets are used to optimize the detection of these mutation types.

While the cancer WES datasets were not suitable for DeepMosaic, they were suitable for the estimation of computational resources. We further estimated the computational resources for DeepMosaic on the basis of 1,215 WES and 48 WGS datasets, DeepMosaic consumed an average of 1,403.8 s (range 9.1–50,168.9) on one WES sample and 22,718.2 s (range 6,565.8–60,800.0) for a 300× genome, an average of 1.3 Gb (range 0.9–1.8 Gb) maximum memory for an exome and an average of 1.2 Gb (range 1.1–1.3 Gb) for a genome, which could be accelerated by more CPU nodes or using graphical processing unit (GPU) nodes (Supplementary Table 5 and Extended Data Fig. 10c,d).

Discussion

DeepMosaic detects noncancer mosaic SNVs from short-read sequencing data and does not require a matched control sample. Compared with NeuSomatic, which compresses all the bases in a genomic position into ten features[35], DeepMosaic-VM provides a complete representation of information present in the BAM file and showed higher sensitivity on noncancer MVs. Compared with other recoding methods such as DeepVariant[11], DeepMosaic-CM can distinguish between MVs and other genotypes. DeepMosaic-VM can be applied as an independent variant visualization tool for users' convenience. To further improve accuracy, DeepMosaic integrates four genomic features and population information absent in the raw BAM files.

Both biological and simulated data showed that DeepMosaic has the potential to identify MVs at relatively high sensitivity and accuracy for WGS at depths as low as 50×. For the past 10–15 years, hundreds of thousands of WGS datasets from clinical, commercial or research laboratories have been generated at relatively low depth, but most have not been subjected to unbiased mosaicism detection due to the lack of sufficiently sensitive methods. DeepMosaic could enable a genome-level unbiased MV detection that requires only conventional sequencing data. For instance, clonal hematopoiesis without a known driver mutation is reported[36] but can be difficult to detect because of technical limitations induced by noise and lower supporting read counts[37].

By using a training set comprising representative technical artifacts such as homopolymers and truncated reads, DeepMosaic acquired the power to distinguish biologically true positive from false-positive MVs. These might have otherwise been filtered out by rule-based methods such as MosaicHunter[12] or MosaicForecast[13]. We demonstrated that training the models on a mixture of roughly 1/1 simulated and biological data did not adversely affect performance on an independent biological evaluation set. We also demonstrated that DeepMosaic worked well for various Illumina short-read sequencing platforms applying different library preparation strategies (PCR-amplified and PCR-free).

Although the EfficientNet-b4 model performed best, we provide all pretrained CNN models (Densenet, EfficientNet, Inception-v.3 and Resnet) on GitHub. DeepMosaic users can prepare their own data with labeled genotypes for training, generate data-specific, personalized models, test other potential factors influencing detection sensitivity such as the ratio of positive to negative labels and increase the detection specificity on specialized datasets. For instance, homopolymers and tandem repeats are increasingly recognized in disease and development, but, because of the limited training data, are currently not detected with DeepMosaic; however, users could retrain with such specialized datasets. Likewise, although detecting MVs from WES can be challenging, DeepMosaic outperformed the existing best-practice pipeline. We propose that further training on large-scale experimentally validated WES data could further improve performance.

While we propose DeepMosaic as a tool for MV detection in WGS and WES, it is not designed to detect mosaic INDELs and mosaic repetitive variants, regions known to be fraught with errors, nor is DeepMosaic, in its current form, suitable for cancer samples. In practice, MosaicForecast can detect mosaic INDEL variants with reasonable accuracy, while M2S2MH has good performance for tissue-specific variants due to the inclusion of additional information from the 'normal' comparison sample. And methods such as MuTect2 paired mode showed a higher sensitivity for cancer samples. Thus, different methods complement one another and should be selected for the purpose.

Despite the features from image representation and a neural network-based variant classifier, DeepMosaic can reproducibly identify most (roughly 70%) WGS MVs detectable by conventional methods. This unique architecture results in higher sensitivity, and the detection of variants with relatively lower AF, both in simulated and experimentally derived, and orthogonally validated data. DeepMosaic shows a drop of sensitivity at higher AF, probably due to the inclusion of depth ratio, which helps to avoid false-positive calls from CNV. DeepMosaic showed consistently high accuracy in noncancer WGS and WES data (Supplementary Table 4) and thus is suitable for high-specificity variant detection. Nevertheless, the higher accuracy at lower AFs should make it a good complement to other methods.

Population allele frequencies used in this study also 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 University of Californian, Santa Cruz (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 WGS

and WES variant lists generated by multiple methods such as the GATK HaplotypeCaller with 'ploidy' 50 or 100[16]. Thus, DeepMosaic can be used directly as is or can be customized to the needs of the end-users, providing an adaptable MV detection workflow.

Online content

Any methods, additional references, Nature Portfolio 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/s41587-022-01559-w.

References

1. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting somatic mutations in normal cells. *Trends Genet.* **34**, 545–557 (2018).
2. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
3. Lee, J. H. et al. Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243–247 (2018).
4. Yang, X. et al. MosaicBase: a knowledgebase of postzygotic mosaic variants in noncancer disease-related and healthy human individuals. *Genom. Proteom. Bioinform.* **18**, 140–149 (2020).
5. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758 (2013).
6. Freed, D., Stevens, E. L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes* **5**, 1064–1094 (2014).
7. Yang, X. et al. Developmental and temporal characteristics of clonal sperm mosaicism. *Cell* **184**, 4772–4783 e4715 (2021).
8. Breuss, M. W., Yang, X. & Gleeson, J. G. Sperm mosaicism: implications for genomic diversity and disease. *Trends Genet.* **37**, 890–902 (2021).
9. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
10. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
11. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
12. 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).
13. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched controls. *Nat. Biotechnol.* **38**, 314–319 (2020).
14. 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).
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. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biol.* **22**, 92 (2021).
17. Huang, A. Y. et al. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.* **24**, 1311–1327 (2014).
18. 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).
19. Breuss, M. W. et al. Somatic mosaicism reveals clonal distributions of neocortical development. *Nature* **604**, 689–696 (2022).
20. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds. Bajcsy, R., Li, F.F., & Tuytelaars, T.) 2818–2826 (IEEE, 2016).
22. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds. Bajcsy, R., Li, F.F., & Tuytelaars, T.) 770–778 (IEEE, 2016).
23. Iandola, F. et al. Densenet: implementing efficient convnet descriptor pyramids. Preprint at *arXiv* arXiv:1404.1869 (2014) https://arxiv.org/abs/1404.1869
24. Tan, M. & Le, Q. V. Efficientnet: rethinking model scaling for convolutional neural networks. *PMLR* **97**, 6105–6114 (2019).
25. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: the all convolutional net. Preprint at *arXiv* arXiv:1412.6806 (2014) https://arxiv.org/abs/1412.6806
26. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
27. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
28. Breuss, M. W. et al. Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat. Med.* **26**, 143–150 (2020).
29. Pelorosso, C. et al. Somatic double-hit in MTOR and RPS6 in hemimegalencephaly with intractable epilepsy. *Hum. Mol. Genet.* **28**, 3755–3765 (2019).
30. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
31. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
32. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
33. Radenbaugh, A. J. et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS ONE* **9**, e111516 (2014).
34. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281 e277 (2018).
35. Sahraeian, S. M. E. et al. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).
36. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
37. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).

[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 Pediatrics, Section of Genetics and Metabolism, University of Colorado School of Medicine, Aurora, CO, USA. [4]Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA. [5]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, USA. [6]Department of Bioengineering, UC San Diego, La Jolla, CA, USA. [7]Moores Cancer Center, UC San Diego, La Jolla, CA, USA. [8]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China. [9]Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, CA, USA. [10]Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. [11]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. [12]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [27]These authors contributed equally: Xiaoxu Yang, Xin Xu. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: xiy010@health.ucsd.edu; jogleeson@health.ucsd.edu

NIMH Brain Somatic Mosaicism Network

Xiaoxu Yang[1,2,27], Martin W. Breuss[1,2,3], Danny Antaki[1,2], Laurel L. Ball[1,2], Changuk Chung[1,2], Dan Averbuj[1,2], Subhojit Roy[1,2], Renee D. George[1,2], Eric Courchesne[1,2], Yifan Wang[4], Taejeong Bae[4], Alexej Abyzov[4], August Y. Huang[13], Alissa D'Gama[13], Caroline Dias[13], Christopher A. Walsh[13], Javier Ganz[13], Michael Lodato[13], Michael Miller[13], Pengpeng Li[13], Rachel Rodin[13], Robert Hill[13], Sara Bizzotto[13], Sattar Khoshkhoo[13], Zinan Zhou[13], Alice Lee[14], Alison Barton[14], Alon Galor[14], Chong Chu[14], Craig Bohrson[14], Doga Gulhan[14], Eduardo Maury[14], Elaine Lim[14], Euncheon Lim[14], Giorgio Melloni[14], Isidro Cortes[14], Jake Lee[14], Joe Luquette[14], Lixing Yang[14], Maxwell Sherman[14], Michael Coulter[14], Minseok Kwon[14], Peter J. Park[14], Rebeca Borges-Monroy[14], Semin Lee[14], Sonia Kim[14], Soo Lee[14], Vinary Viswanadham[14], Yanmei Dou[14], Andrew J. Chess[15], Attila Jones[15], Chaggai Rosenbluh[15], Schahram Akbarian[15], Ben Langmead[16], Jeremy Thorpe[16], Sean Cho[16], Andrew Jaffe[17], Apua Paquola[17], Daniel Weinberger[17], Jennifer Erwin[17], Jooheon Shin[17], Michael McConnell[17], Richard Straub[17], Rujuta Narurkar[17], Yeongjun Jang[4], Cindy Molitor[18], Mette Peters[18], Fred H. Gage[19], Meiyan Wang[19], Patrick Reed[19], Sara Linker[19], Alexander Urban[20], Bo Zhou[20], Xiaowei Zhu[20], Aitor S. Amero[21], David Juan[21], Inna Povolotskaya[21], Irene Lobon[21], Manuel S. Moruno[21], Raquel G. Perez[21], Tomas Marques-Bonet[21], Eduardo Soriano[22], Gary Mathern[23], Diane Flasch[24], Trenton Frisbie[24], Huira Kopera[24], Jeffrey Kidd[24], John Moldovan[24], John V. Moran[24], Kenneth Kwan[24], Ryan Mills[24], Sarah Emery[24], Weichen Zhou[24], Xuefang Zhao[24], Aakrosh Ratan[25], Alexandre Jourdon[26], Flora M. Vaccarino[26], Liana Fasching[26], Nenad Sestan[26], Sirisha Pochareddy[26], Soraya Scuderi[26] & Joseph G. Gleeson[1,2]

[13]Boston Children's Hospital, Boston, MA, USA. [14]Harvard University, Cambridge, MA, USA. [15]Icahn School of Medicine at Mt. Sinai, New York, NY, USA. [16]Kennedy Krieger Institute, Baltimore, MD, USA. [17]Lieber Institute for Brain Development, Baltimore, MD, USA. [18]Sage Bionetworks, Camarillo, CA, USA. [19]Salk Institute for Biological Studies, La Jolla, CA, USA. [20]Stanford University, Stanford, CA, USA. [21]Universitat Pompeu Fabra, Barcelona, Spain. [22]University of Barcelona, Barcelona, Spain. [23]University of California, Los Angeles, Los Angeles, CA, USA. [24]University of Michigan, Ann Arbor, MI, USA. [25]University of Virginia, Charlottesville, VA, USA. [26]Yale University, New Haven, CT, USA.

Methods

Visualization of MVs

Visualization of MVs was based on Python (v.3.7.81) packages Pysam (v.0.11.2.2, https://github.com/pysam-developers/pysam) and NumPy (v.1.16.2, https://numpy.org/). The input for this visualization is short-read sequencing data in the format of a BAM file, processed with a GATK (v.3.8.1) best-practice pipeline (with insertion/deletion, or INDEL, realignment, followed by base quality score recalibration). Inspired by DeepVariant[11], DeepMosaic-VM exports this data into an RGB image, representing a pileup at each genomic position, as well as generating a NumPy object for the classification module. In contrast to presenting scattered reads by DeepVariant, DeepMosaic encodes separated sequences piling up into 'ref' reads and 'alt' reads on the basis of the reference genome information. This improves pileup visualization and allows the assessment of mosaicism at a glance for humans, and converted the biological variant detection problem into an edge- and shape-detection problem, which is more suitable for image-based classification by CNN models (https://github.com/Virginiaxu/DeepMosaic/blob/master/deepmosaic/featureExtraction.py).

Curation of training and benchmark data

SimData1. For the initial training procedure, 10,000 variants were randomly generated on chromosome 22 to obtain the list of alternative bases. Pysim[38] 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 were randomly sampled for a targeted depth of 30, 50, 100, 120, 150, 200, 250, 300, 400 and 500×. FASTQ files were aligned to the GRCh37d5 human reference genome with BWA (v.0.7.17) mem command. Aligned data were processed by GATK (v.3.8.1) and Picard (v.2.18.27) for marking duplicates, 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 randomly generated mutations fall on a common single nucleotide polymorphism 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.

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 7,610 variants on the nonrepetitive region of chromosome 22 were considered true-positive genomic positions; (2) random errors were generated from the Illumina NovaSeq sequencer error model and (3) data were randomly down-sampled and up-sampled for a targeted depth of 50, 100, 200, 300, 400 and 500×. A total of 439,200 different variants were generated. FASTQ files were aligned and processed with BWA (v.0.7.17), SAMtools (v.1.9) and Picard (v.2.18.27). The data were subjected to DeepMosaic as well as MuTect2 (GATK v.4.0.4, both paired mode and single mode), Strelka2 (v.2.9.2), MosaicHunter (v.1.0.0) and MosaicForecast (v.8-13-2019) with different models trained for different read depth (250× model for depth ≥300×).

SimData3. We further generated another simulation dataset in a way that was fundamentally different from the training data with a positive to negative ratio similar to real data[19] to compare the performance of DeepMosaic and other software for the detection of MVs. We selected 30,090 genomic positions with reference homozygous genotype from a different genomic region (the entire Chromosome 1) of the whole-genome deep sequences from the 'Genome in a Bottle' sample HG002 (NA24345)[27]. The genomic positions from the 30,090 positions were genotyped as homozygous and fulfilled additional criteria: (1) zero alternative bases in the raw sequencing data; (2) no detectable insertions/deletions in the position of interest; (3) have a genomic distance of at least 1,000 bases between each other. On this clear background, 15,471 of them were labeled as 'true negative' with reference homozygous genotype, 6,868 were labeled as 'true-positive' MVs with expected alternative AF 0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20 and 0.25 (on average 763 variants for each genotype) and 7,751 were labeled as 'true-negative' heterozygous variants with alternative AF 0.50; the latest version of a different software BAMSurgeon (updated 24 December 2020) was used to generate this simulation dataset and retain the sequencing errors from the original biological samples. The original BAM file was first up-sampled, and alternative reads were replaced to generate the expected AF, mapped back to the genome and merged back to the BAM file, according to the software manual[26]. BAM files with and without simulated data were down-sampled to 500×, 400×, 300×, 200×, 100× and 50×. The data were subjected to DeepMosaic as well as MuTect2 (GATK v.4.0.4, both paired mode, and single mode), Strelka2 (v.2.9.2), MosaicHunter (v.1.0.0) and MosaicForecast (v.8-13-2019) with different models trained for different read depth (250× model for depth ≥300×), the performance of the 180,540 points were evaluated.

BioData1. Variant information and raw sequencing read from 80–120× PCR-amplified PE-150 WGS data of 29 samples from six normal individuals were extracted from published data[17,18] on the Sequence Read Archive (SRA) (SRP028833, SRP100797 and SRP136305). Then 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. Next, 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, then manually selected and labeled as negative. The 1,575 genomic positions were also down-sampled and up-sampled for a targeted depth of 30, 50, 100, 150, 200, 250, 300, 400 and 500×, to expand this dataset into a pool of 14,175 different conditions. Depth ratios were calculated accordingly and gnomAD allele frequencies, segmental duplication and repeat masker information were annotated.

Categories of technical artifacts include (1) variants with multiple alternative alleles, as from our experiences, the chance of a noncancer MV occurring twice at the same genomic position at the early embryonic development stage is rare; (2) alignment artifacts, evidenced by short truncated or hard-clipped reads mapping to a certain genomic region, resulting in small truncated mapped reads piled up; (3) ultra-low mapping quality and base reads and (4) ultra-high AF variants because they are not expected in postzygotic noncancer situations.

The entire BioData1 and random subsampling from SimData1 were combined to generate a training and validation dataset with approximately 200,000 variants from the 1,000,000 training variants. Next, 180,000 variants were selected for model training, 45% from SimData1 and 55% from resampling of BioData1. 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 (Extended Data Fig. 1).

BioData2. To estimate the performance of the pretrained 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 six distinct sequencing centers and validated in five different centers, known as the Reference Tissue Project from the BSMN[16]. Here, 400 variants underwent multiple levels of computational validation 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 sets for low-fraction single nucleotide MVs from the 250× WGS data[16]. We extracted deep WGSs for those variants, labeled them accordingly and used them as gold-standard validation set for model selection (Extended Data Fig. 2).

BioData3. To evaluate the performance of DeepMosaic-CM trained on a different portion of biological variants, we included another large-scale validation experiment we recently generated. Variant information and raw sequencing read of 300× PCR-free PE-150-only WGS of 18 samples from nine different brain regions, cerebellum, heart, liver and both kidneys of one individual was extracted from the capstone project of the BSMN[19]. Then 1,400 genomic positions with variants identified from the WGS sample and reference homozygous/heterozygous controls validated by orthogonal experiments were selected and labeled as positive and negative according to the experimental validation result. The 1,400 genomic positions were also down- and up-sampled for a targeted depth of 30, 50, 100, 150, 200, 250, 300, 400 and 500×. Depth ratios were calculated accordingly, and gnomAD allele frequencies, segmental duplication and repeat masker information were annotated.

BioData4. This additional WGS dataset was used to compare the performance of DeepMosaic and other MV callers on biological samples. Here, 16 WGS samples from the blood and sperm of eight individuals were sequenced at 200× (ref. [28]) (PRJNA588332). WGS was performed using an Illumina TrueSeq PCR-free kit with 350 bp insertion size and sequenced on an Illumina HiSeq sequencer. Reads were aligned to the GRCh37d5 genome with BWA (v.0.7.15) mem and duplicates were removed with sambamba (v.0.6.6) and base quality recalibrated by GATK (v.3.5.0). Processed BAM files were subjected to DeepMosaic as well as MuTect2 (GATK v.4.0.4, both paired mode and single mode), Strelka2 (v.2.9.2), MosaicHunter (v.1.0.0) and MosaicForecast (v.8-13-2019) with 200× models trained for the specific depth. Data from one of the individuals (F02) were down-sampled to 150×, 100×, 50× and 30× with the SAMtools (v.1.9) view command for the further benchmark of DeepMosaic.

BioData5. We included an additional WES dataset that was used to compare the performance of DeepMosaic and other MV calling pipelines on WES data. Here, 181 WES samples from the brain and blood/saliva of 101 individuals were sequenced at roughly 300× (NDA). Genomic DNA was extracted from pulverized brain and white blood cells/buccal epithelial samples using Qiagen Miniprep and Maxiprep kits according to the protocols provided by the manufacturer. Genomic DNA samples were prepared for WES using the Agilent SureSelect XT Human All Exon v.5 kits and sequenced on an Illumina HiSeq 2500 sequencer at a targeted depth of roughly 300×. Reads were aligned to the GRCh37d5 genome with BWA (v.0.7.17) mem and duplicates were removed and base quality recalibrated by GATK (v.4.0.4) according to the established best-practice pipeline[16]. Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v.4.0.4) single mode as well as GATK (v.4.0.4) HaplotypeCaller ('polidy' 50) and previously established filters[16].

BioData6. We assessed the performance of DeepMosaic on a large-scale tumor dataset. We downloaded and analyzed 2,430 WES samples from 1,215 individuals from six different cancer types from the TCGA-MC3 collection[34]. From this, 468 were patients with skin cutaneous melanoma, 406 with bladder urothelial carcinoma, 157 with glioblastoma multiforme, 112 with breast invasive carcinoma, 50 with lung squamous cell carcinoma and 23 with colon adenocarcinoma. Performance was compared with call sets provided in their respective original publications. Data were downloaded from the Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/, sample IDs provided with variants in Supplementary Table 3). Fastq files were generated using Picard SAMTOFASTQ and aligned to GRCh37d5 genome with BWA (v.0.7.17) mem. Duplicates were removed, reads near INDEL regions were realigned and base quality scores were recalibrated with GATK v.3.8.1 and Picard v.2.20.7. Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v.4.0.4) single mode, then the final call set was compared with the TCGA-MC3 call set detected by MuSE[30], MuTect[9], SomaticSniper[31], VarScan2 (ref. [32]) and Radia[33] using the publicly released gold standard (https://gdc.cancer.gov/about-data/publications/mc3-2017) from the same dataset[34]. Part of the computing resources and CPU consumption also were estimated from this dataset with Linux command time.

Neural network building and model training

For the ten neural network architectures, Inception-v.3, Resnet and Densenet were imported from PyTorch's (v.1.4.0) built-in library, while the seven different builds of EfficientNet were imported from the efficientnet_pytorch (v.0.6.1) Python (v.3.7.1) package. The final fully connected layer of each model was replaced to be fed into three output units representing intermediate results instead of the default 1,000 output units for the 1,000 ImageNet classes to substantially reduce the total images required to extract basic features such as edges, and stripes from raw images. A transfer-learning method was adopted for model training. Each model's initial pretrained 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- and up-sampling of SimData1 and BioData1) into 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 optimizer with a 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 Center's Comet computational clusters. Codes, scripts and functions used for training, together with the guidance and annotations were provided on the DeepMosaic GitHub page (https://github.com/Virginiaxu/DeepMosaic/tree/master/deepmosaic/trainModel.py).

Network selection

To select the 'best-performing' neural network architecture among the trained Inception-v.3, Resnet, Densenet and seven different builds of EfficientNet, the gold-standard evaluation dataset (BioData2) were used to test each model's performance on biological (nonsimulated) MVs determined by the dataset. ACC (accuracy), MCC (Matthews correlation coefficient (MCC)) and 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 the DeepMosaic model (EfficientNet-b4 architecture) was further evaluated.

Independent model training and evaluation for DeepMosaic-CM

To evaluate the performance of DeepMosaic-CM when trained on a different portion of biological variants, 15 epochs were trained for the EfficientNet-b4 architecture on five different training sets consisting of 122,424 genomic positions. EfficientNet was imported from the efficientnet_pytorch (v.0.6.1) Python (v.3.7.1) package. The five different training sets were generated based on SimData1, BioData1, SimData2 and BioData3. (1) BioData only: 40,808 variants from the entire BioData1 and BioData3 were pooled. The overall positive/negative ratio was 26.8/73.2%. (2) SimData only for SimData1: 40,808 variants were selected from SimData1 with the matched number of positive and negative labels as BioData only. (3) SimData only for SimData2: 40,808

variants that were agreed by both MuTect2 and Strelka2 as 'positive' or agreed by both methods as 'negative' were selected from SimData2 with the matched number of positive and negative labels compared to BioData only. (4) BioData+SimData for SimData1: 40,808 variants half from BioData and half from SimData only for SimData1 were selected with the matched number of positive and negative labels compared to BioData only. (5) BioData+SimData for SimData2: 40,808 variants half from BioData and half from SimData only for SimData2 were selected with the matched number of positive and negative labels compared to BioData only. Each network architecture was trained using a batch size of four with a stochastic gradient descent optimizer with a learning rate of 0.01, and momentum of 0.9. Fifteen different epochs were trained on each of the five training sets described above, and the model after each epoch was saved for performance evaluation. The training was conducted on NVIDIA GTX 980 GPU Nodes in San Diego Supercomputer Center's Triton Shared Computing Cluster. The training performance of the models was further evaluated on BioData2, which has not been used for any of the training procedures. The above models were also trained using codes described in https://github.com/Virginiaxu/DeepMosaic/tree/master/deepmosaic/trainModel.py.

Usage of DeepMosaic

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

Orthogonal validation with deep amplicon sequencing method

Deep amplicon sequencing analysis[7] was applied to 239 variants from the 1,146 candidates detected by all five MV callers from the 200× WGS of 16 samples from BioData4 (ref. [28]) as well as 291 out of 585 candidates detected by both WES pipelines from the 181 samples from BioData5 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 Primer3 (ref. [39]) with a Python (v.3.7.3) wrapper[7]. 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 1.2×. 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 (v.3.8.2) best practices without removing PCR duplicates. Putative mosaic sites were retrieved using SAMtools (v.1.9) mpileup and pileup filtering scripts described in previous target amplicon-sequencing pipelines[28]. Variants were considered positively validated for mosaicism if (1) their lower 95% exact binomial confidence interval boundary was above the upper 95% confidence interval boundary of the control and (2) their AF was >0.5%. The number of references and alternative alleles calculated from the amplicon validation was provided in Supplementary Table 1. Codes for primer design and data analyses were available on GitHub (https://github.com/shishenyxx/PASM).

Analysis of different categories of variants overlap with different genomic features

To assess the distribution of MVs and their overlap with genomic features across the genome, an equal number of variants (mSNVs/INDELs as in group G1-G7 in Extended Data Fig. 9) was randomly generated with the BEDtools (v.2.27.1) shuffle command within the region from Strelka2 without the subtracted regions (for example repeat regions). This process was repeated 10,000 times to generate distribution and their 95% confidence interval. Observed and randomly subsampled variants were annotated with whole-genome histone modifications data for H3k27ac, H3k27me3, H3k4me1 and H3k4me3 from ENCODE v.3 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 ten different cell lines (Mrg; merged from Gm12878, H1, Hmec, Hsmm, Huvec, K562, Nha, Nhek and Nhlf). Gene region, intronic and exonic regions from National Center for Biotechnology Information RefSeqGene (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz); ten Topoisomerase 2A/2B (Top2a/b) sensitive regions from chromatin immunoprecipitation with sequencing 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[40]; 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 v.3 tracks from the UCSC genome browser (wgEncodeRegDnaseClusteredV3 and wgEncodeRegTfbsClusteredV3, respectively).

Parameters and codes used for different MV callers and pipelines

Raw WGS sequencing data were aligned and processed with a GATK snakemake pipeline with INDEL realignment and base quality score recalibration added, previously deposited on GitHub (codes and parameters available at https://github.com/shishenyxx/Adult_brain_somatic_mosaicism/tree/master/pipelines/WGS_processing_pipeline): the genome version is described above. WGS and simulated variant calling using MuTect2 (paired model, GATK v.4.0.4) and Strelka2 (somatic model, v.2.9.2) were carried out on previously deposited pipeline (codes available on GitHub: https://github.com/shishenyxx/Adult_brain_somatic_mosaicism/tree/master/pipelines/WGS_SNV_indel_calling_pipeline/Mutect2_PM_Strelka2). MuTect2 single mode (GATK v.4.0.4) was carried out with a 300× panel of normal analysis, the control panel was collected from healthy individuals and not used in generating any of the training, testing or validation data described in this paper codes previously deposited (https://github.com/shishenyxx/Adult_brain_somatic_mosaicism/tree/master/pipelines/WGS_SNV_indel_calling_pipeline/Mutect2_single_mode). MosaicForecast (v.8-13-2019) analysis was carried out with help from the original authors, model 250xRFmodel_addRMSK_Refine.rds was used for >200× WGS, the other models (brain_MT2-PON.50x.rds, brain_MT2-PON.100x.rds, brain_MT2-PON.150x.rds and brain_MT2-PON.200x.rds) were used according to different depth for the simulated and biological data, respectively. Codes for the MosaicForecast pipeline is publically available (https://github.com/shishenyxx/Adult_brain_somatic_mosaicism/tree/master/pipelines/WGS_SNV_indel_calling_pipeline/MosaicForecast_pipeline). MosaicHunter (v.1.0.0) single mode was ran under following the user guide (https://github.com/zzhang526/MosaicHunter/tree/master/docs/ MosaicHunterUserGuide.pdf), parameters (30X_genome_b37_ctrl_cohort_2020_09_22.properties, 50X_genome_b37_ctrl_cohort_2020_04_07.properties, 100X_genome_b37_ctrl_cohort_2020_04_07.properties, 150X_genome_b37_ctrl_cohort_2020_10_15.properties, 200X_genome_b37_ctrl_cohort_2020_04_07.properties, 300X_genome_b37_ctrl_cohort_2018_11_29.properties, 400X_genome_b37_ctrl_cohort_2020_04_07.properties, 500X_genome_b37_ctrl_cohort_2020_04_07.properties) and codes for WGS variant calling are provided on GitHub (https://github.com/shishenyxx/Adult_brain_somatic_mosaicism/tree/master/pipelines/WGS_SNV_indel_calling_pipeline/

MosaicHunter_single_mode_pipeline). The packages, parameters and performance analysis of all the above pipelines were described in our recent publications[7,19]. DeepMosaic (v.1.0.0) analyses were carried out with default parameters on the GitHub page (https://github.com/Virginiaxu/DeepMosaic). NeuSomatic was established on the basis of the website, a singularity container was used to carry out the NeuSomatic analysis. Codes and parameters were available on GitHub (https://github.com/shishenyxx/DeepMosaic/tree/master/For_publication).

The WES data from our malformation of cortical development (MCD) cohort (BioData5)[41] and the TCGA-MC3 collection (BioData6) were collected and BAM files are processed using a pipeline based on the data processing part of the BSMN common pipeline (https://github.com/shishenyxx/MCD_mosaic/tree/main/Pipelines/Alignment), followed by GATK HaplotypeCaller with polidy 2 to call the germline variants for the indel annotations. The BSMN common pipeline for WES was carried out following the official release (https://github.com/bsmn/bsmn-pipeline) from mapping to GATK (v.3.8.2) HaplotyperCaller polidy 50 variant calling and the BSMN common filtering. Details for the parameters and their benchmarks are described in the original publication[16].

Other software and versions

BAM and variant processing software also include Picard v.2.18.27, BCFtools v.1.10.32, sambamba v.0.6.6, iFish, Define. Plotting and visualization software include R v.3.5.1, ggplot2 v.3.3.1, Rcpp v.1.03, PyTorch v.1.6.0, Pysam v.0.11.2.2, Python v.3.7.1 and v.3.7.81, SciPy v.1.3.1, pandas v.0.24.2, matplotlib v.3.1.1, NumPy v.1.16.2 and seaborn v.0.9.0.

Reporting summary

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

Data availability

WGS data used to generate the training set are available at the SRA (accession nos. SRP028833 and SRP100797, BioData1). The gold-standard WGS data and validated capstone project data are available at the National Institute of Mental Health Data Archive (NIMH Data Archive ID 792 and 919: https://nda.nih.gov/study.html?id=792, BioData2, and https://nda.nih.gov/study.html?id=919, BioData3) and the Brain Somatic Mosaicism Consortium Data Portal, independent benchmark brain genotyping is also part of the SRA accession no. PRJNA736951 (BioData3). Simulated data generated from NA24385 (HG002) are available at https://humanpangenome.org/hg002/. The independent sperm and blood deep WGS data are available at SRA (accession nos. PRJNA588332 and PRJNA660493, BioData4). Independent WES data from brain, blood and saliva samples were available in NIMH Data Archive under study number 1484 (https://nda.nih.gov/study.html?id=1484, BioData5). TCGA-MC3 data are available on the GDC portal (https://portal.gdc.cancer.gov/, sample IDs provided with variants in Supplementary Table 3). Annotations downloaded from UCSC genome browser (https://genome.ucsc.edu/) and ANNOVAR (https://annovar.openbioinformatics.org/en/latest/).

Code availability

DeepMosaic is currently implemented in Python; the source code, documentation and demos are available at https://github.com/Virginiaxu/DeepMosaic. Codes for running different MV callers are documented in the Methods section.

References

38. Xia, Y., Liu, Y., Deng, M. & Xi, R. Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinf.* **18**, 53 (2017).
39. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
40. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107**, 139–144 (2010).
41. Chung, C. et al. Comprehensive multiomic profiling of somatic mutations in malformations of cortical development. *Nat. Genet.* (in the press).

Acknowledgements

We thank Y. Dou for helping to set up the MosaicForecast pipeline. We thank M. K. Gilson for the help with computational resources. We thank P. J. Park, G. W. Cottrell, J. V. Moran, M. Gymrek, P. J. Reed, A. Y. Huang, S.-J. Cheng and Y. Chen for their valuable comments, help and suggestions. This work was supported by the National Institute of Mental Health (NIMH) (grant nos. U01MH108898 and R01MH124890 to J.G.G.), Rady Children's Institute for Genomic Medicine and the Howard Hughes Medical Institute. We thank San Diego Supercomputer Center (grant no. TG-IBN190021 to X.Y. and J.G.G.) for computational help. This publication includes data generated at the UC San Diego IGM Genomics Center using an Illumina NovaSeq 6000 platform that was purchased with funding from a National Institutes of Health SIG grant (no. S10OD026929 X.Y. and J.G.G.).

Author contributions

X.Y., X.X. and J.G.G. conceived this project with input from M.W.B. and D.A. X.Y. designed the study 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., X.X., J.S. and Y.C. generated and collected all the training and benchmark data with the help from D.A., R.D.G., L.W. and L.B.A. X.X. performed the training and model selection under supervision by X.Y. The independent dataset was processed by M.W.B., D.A. and R.D.G. under supervision by J.L.S. and J.G.G. X.Y. and M.W.B. performed the validation experiments with help from L.L.B. and C.C. X.Y. and X.X. wrote the original and revised manuscript with input from all listed authors. X.Y. and J.G.G. revised and edited the manuscript. DeepMosaic is benchmarked on part of the BSMN Reference Tissue Project and common analysis pipeline for SNVs contributed by Y.W., T.B. under supervision by A.A. and the BSMN capstone project contributed by M.W.B., X.Y., D.A. and X.X. under supervision by J.G.G. All authors discussed the results and contributed to the final manuscript.

Competing interests

L.B.A. is a compensated consultant and has equity interest in io9, LLC. His spouse is an employee of Biotheranostics, Inc. L.B.A. is an inventor of a US Patent 10,776,718 and he also declares US provisional applications with serial numbers: 63/289,601; 63/269,033; 63/366,392 and 63/367,846. All other authors declare no competing interests.

Additional information

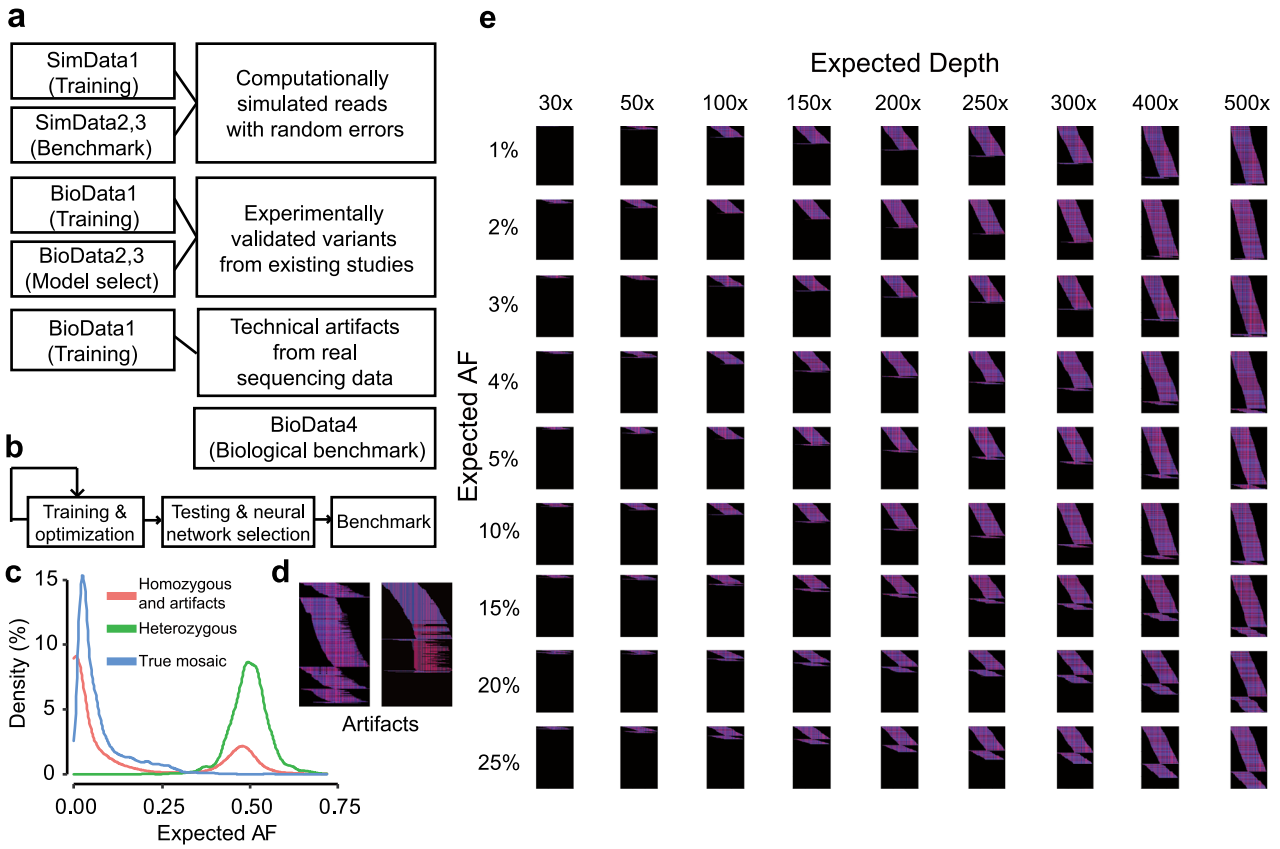
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01559-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01559-w.

Correspondence and requests for materials should be addressed to Xiaoxu Yang or Joseph G. Gleeson.

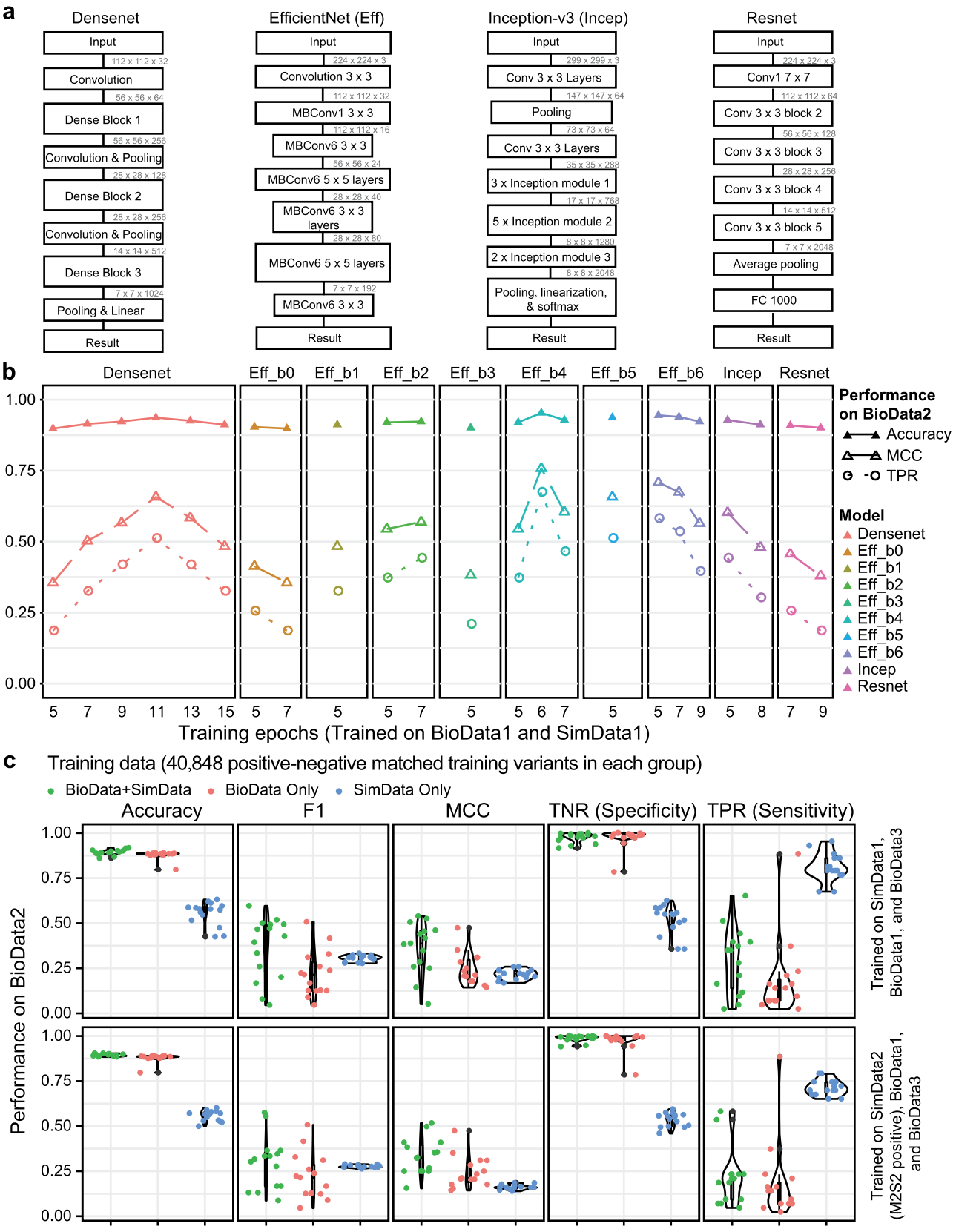
Peer review information *Nature Biotechnology* thanks Anders Skanderup, Moritz Gerstung and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41587-022-01559-w



Extended Data Fig. 1 | Training strategies and examples of training data for DeepMosaic. (a) More than 200,000 training and validation variants were generated for DeepMosaic, including computational simulations (SimData1), and biologically validated variants from existing studies with manually curated technical artifacts (BioData1). We further included 1 gold-standard dataset for testing and model selection (BioData2); all selected positive or negative variants underwent amplicon sequencing in at least one tissue sample according to the publication. We further included independent simulated data (SimData2 and SimData3) and validated independent biological data (BioData3-WGS, BioData4-WGS, and BioData5-WES) to benchmark DeepMosaic. **(b)** The overall strategies of model training and benchmarking for each tested model. **(c)** The distribution of probability density of expected AFs for different variants from the training set. Red: Reference homozygous variants and technical artifacts are labeled 'Negative' in the training set. Green: Heterozygous variants are also labeled 'Negative' in the training set. Blue: True mosaic variants are labeled 'Positive' in the training set. **(d)** Two examples of false positive variants with different sequencing artifacts, left: multiple alternative alleles from sequencing bias or alignment artifacts; right: reads truncated because of sequencing or alignment artifacts. **(e)** All training images were down-sampled and up-sampled into 30×, 50×, 100×, 150×, 200×, 250×, 300×, 400× and 500×, mutant allelic fractions (AFs) from the simulated data that were set as 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25% and shown.

https://doi.org/10.1038/s41587-022-01559-w

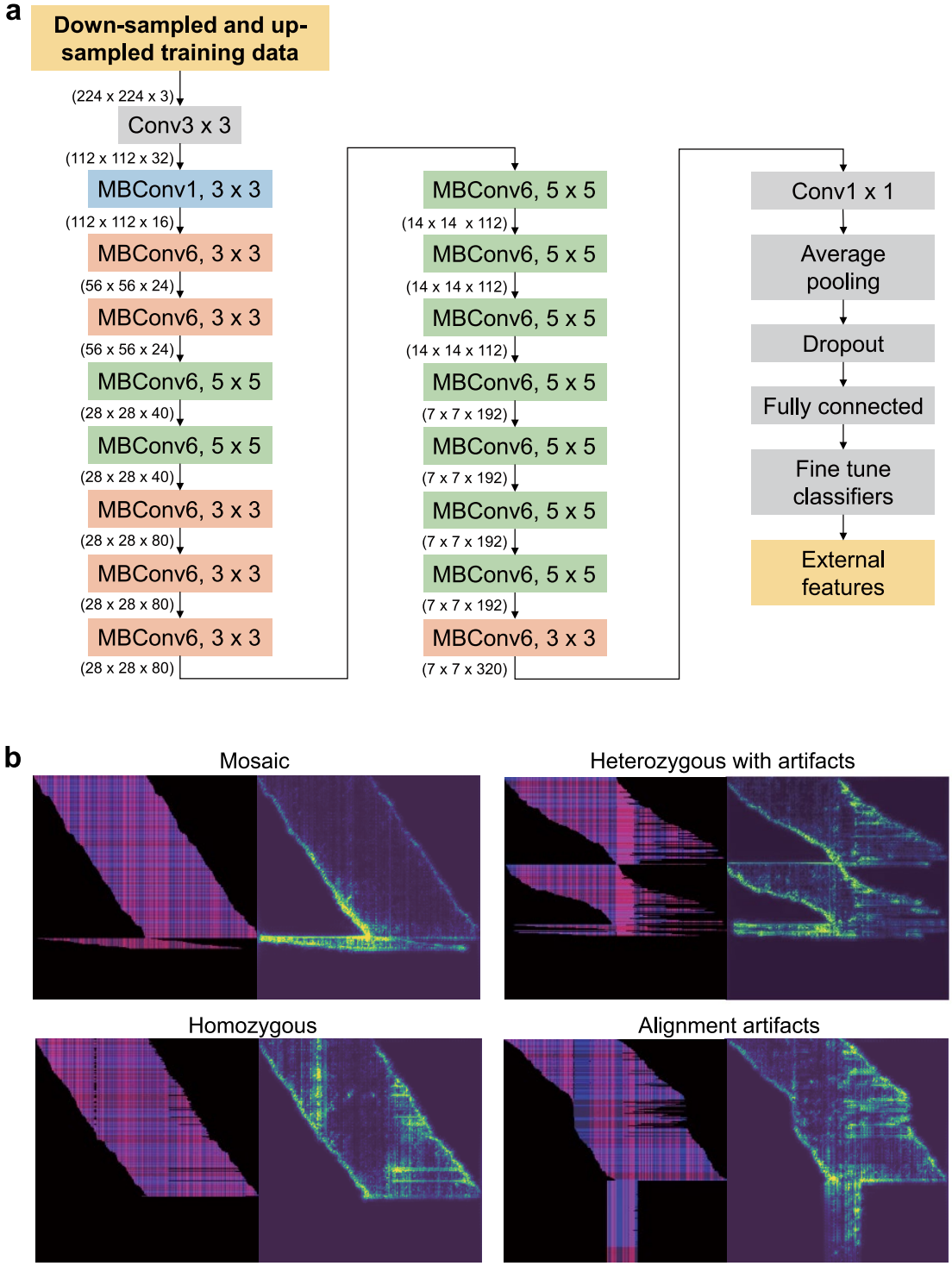


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41587-022-01559-w

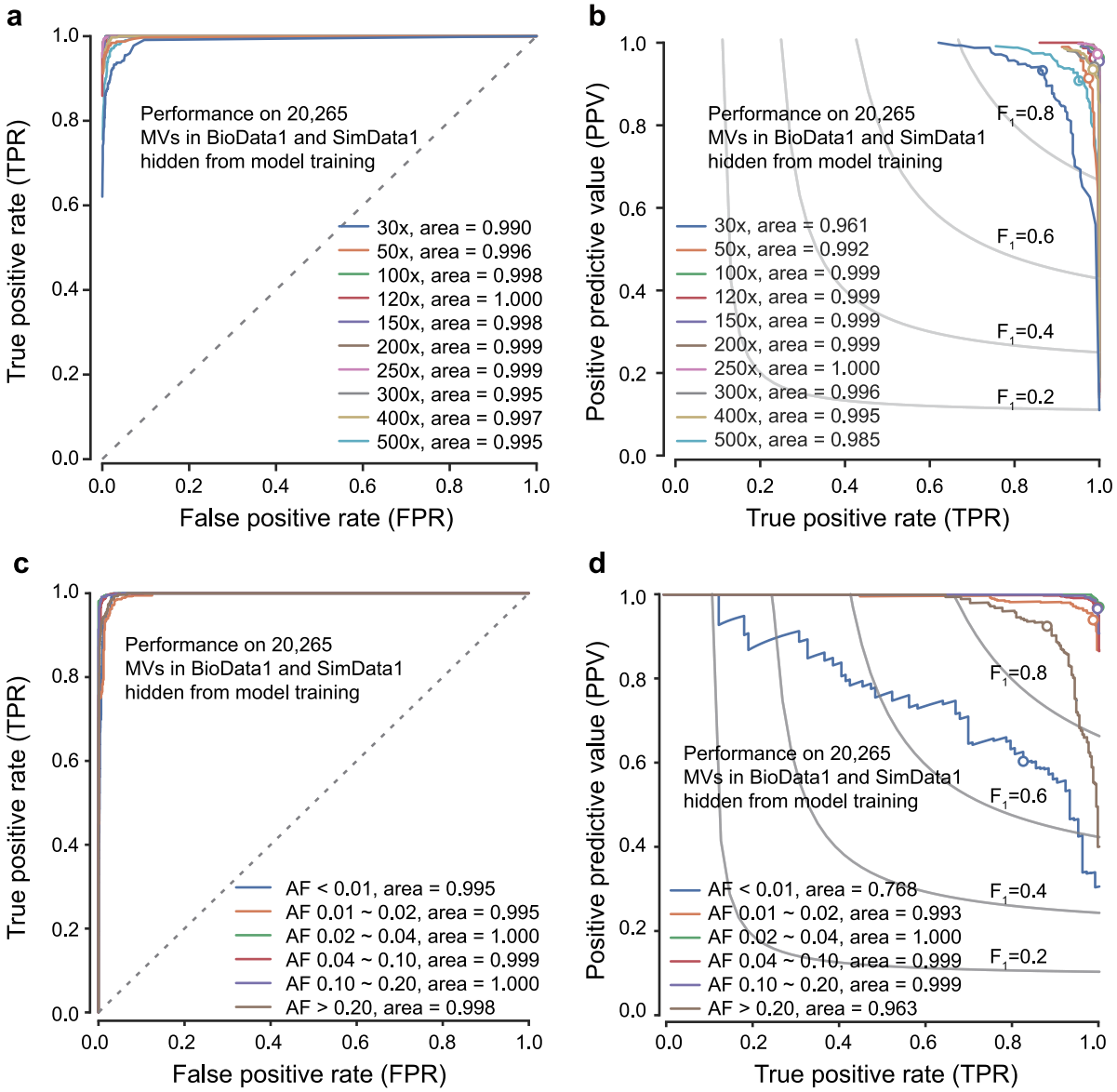
Extended Data Fig. 2 | Network model selection based on an independent gold-standard testing set. (a) Comparison of network structures implementing a variety of classification algorithms. For different build versions of EfficientNet, only a general structure is shown. Inception v3 was used in DeepVariant, and Resnet was used in NeuSomatic. **(b)** All models were trained on 180,000 training variants from BioData1 and SimData1 until the models reach training accuracy > 0.9. Accuracy, Matthews's correlation coefficient (MCC), and Sensitivity of different network structures trained with the same data with different epochs. EfficientNet-b4 trained at 6 epochs demonstrated the highest Accuracy, MCC, and TPR (true positive rate, sensitivity) on the gold standard validation set16 (BioData2); thus it was used as the default core model for DeepMosaic. We additionally provide an option for experienced users to train their own models with self-labeled training data. **(c)** EfficientNet-b4 models were trained on 5 additional datasets, each for 15 epochs. The training datasets were generated with different compositions of biologically validated data and simulated data. Models trained only on simulated data showed overall higher sensitivity but much lower specificity on the gold standard evaluation set (BioData2) due to the high fraction of false-positive calls. Models trained only on biological data showed similar overall performance compared with models trained on a mixture of biological and simulated data. All three training sets are generated with the same number of positive and negative data points as the biological data and with the same number of total variants. M2S2 Positive: training variants were labeled positive by both MuTect2 and Strelka2. n = 15, boundaries are the range for each violin plot, for data in the inner boxplot, the center is the median, upper bound is the upper hinge/75% quantile, the lower bond is the lower hinge/25% percentile, lower whisker represents lower hinge − 1.5*IQR, upper whisker represents upper hinge + 1.5*IQR.

https://doi.org/10.1038/s41587-022-01559-w



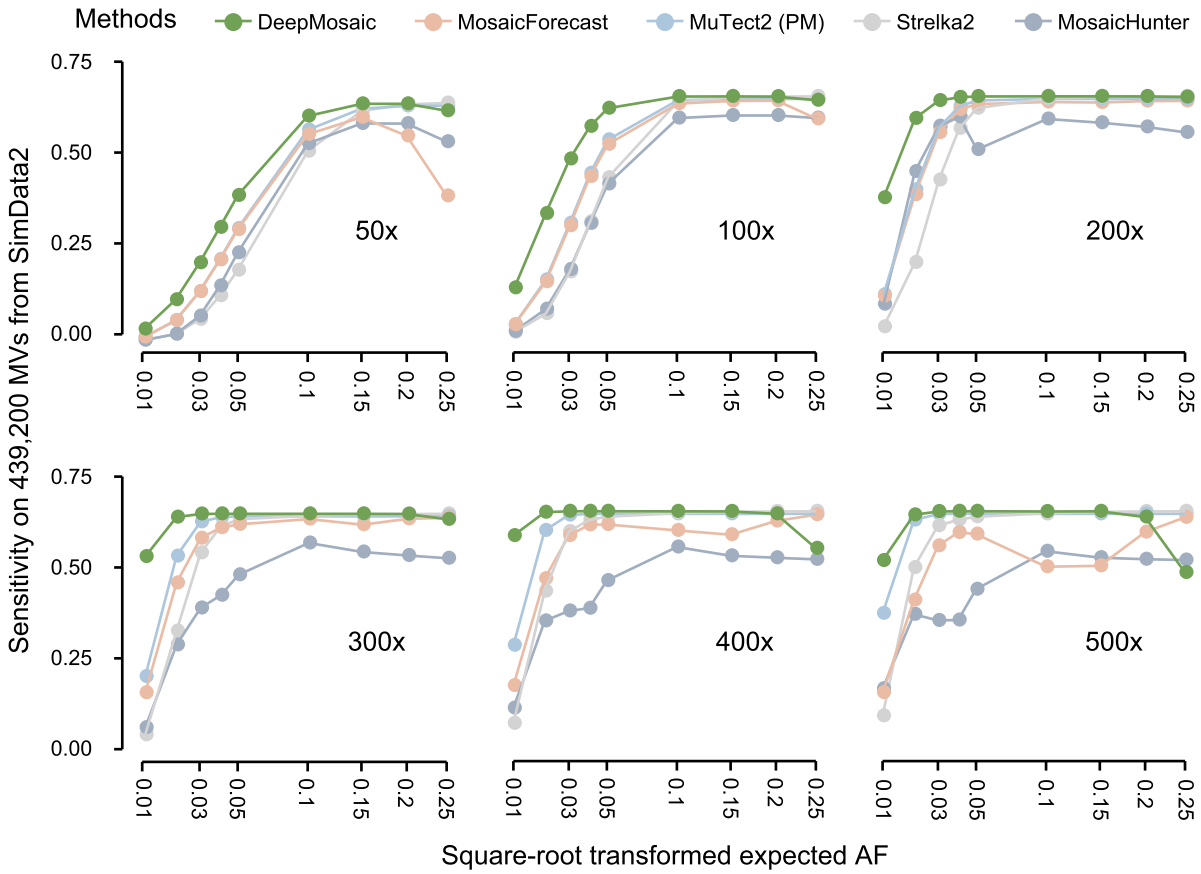
Extended Data Fig. 3 | The convolutional neural network of the DeepMosaic default model and gradient visualization with guided backpropagation for the DeepMosaic default model (EfficientNet-b4). (a) Down-sampled and up-sampled image files coded from the original BAM files were used as input. 16 mobile convolutional layers were adapted from EfficientNet-b4, with optimized parameter size and structures. Numbers represent the dimensions of trained hyperparameters. **(b)** A mosaic, a homozygous, and a heterozygous variant with artifacts, as well as a technical artifact, are shown here for the gradient visualization with guided backpropagation method[25] implemented for the DeepMosaic core model, EfficientNet-b4 trained at epoch 6, left: image coding, right: gradient heatmap. The edges of bases, the sequence information, as well as other high-dimensional information, are highlighted by the model.

https://doi.org/10.1038/s41587-022-01559-w



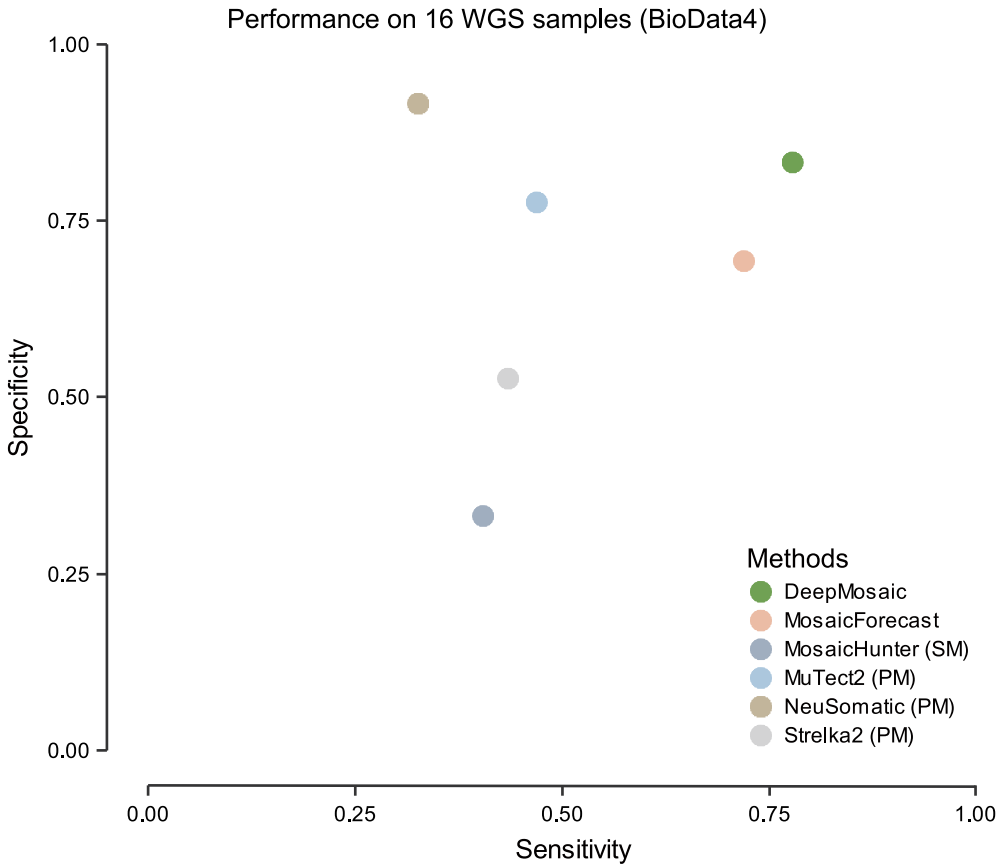
Extended Data Fig. 4 | Performance of DeepMosaic default model (EfficientNet-b4) on data hidden from training. **(a)** Receiver operating characteristic (ROC) curve for DeepMosaic. True positive rates (TPR) and false-positive rates (FPR) were evaluated from 20,265 variants (BioData1 and SimData1) hidden 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. **(c)** ROC curve for DeepMosaic. TPR and FPR were evaluated from 20,265 variants (BioData1 and SimData1) hidden from model training and model selection. Colors show groups of bins of different expected AFs. **(d)** Precision-recall curves for DeepMosaic, evaluated from the 20,265 hidden variants, dots showed the performance of the default parameters for DeepMosaic-CM for different AF bins. Iso-F1 curves were shown for each precision-recall pair with identical F1 scores labeled in **(b)** and **(d)**.

https://doi.org/10.1038/s41587-022-01559-w



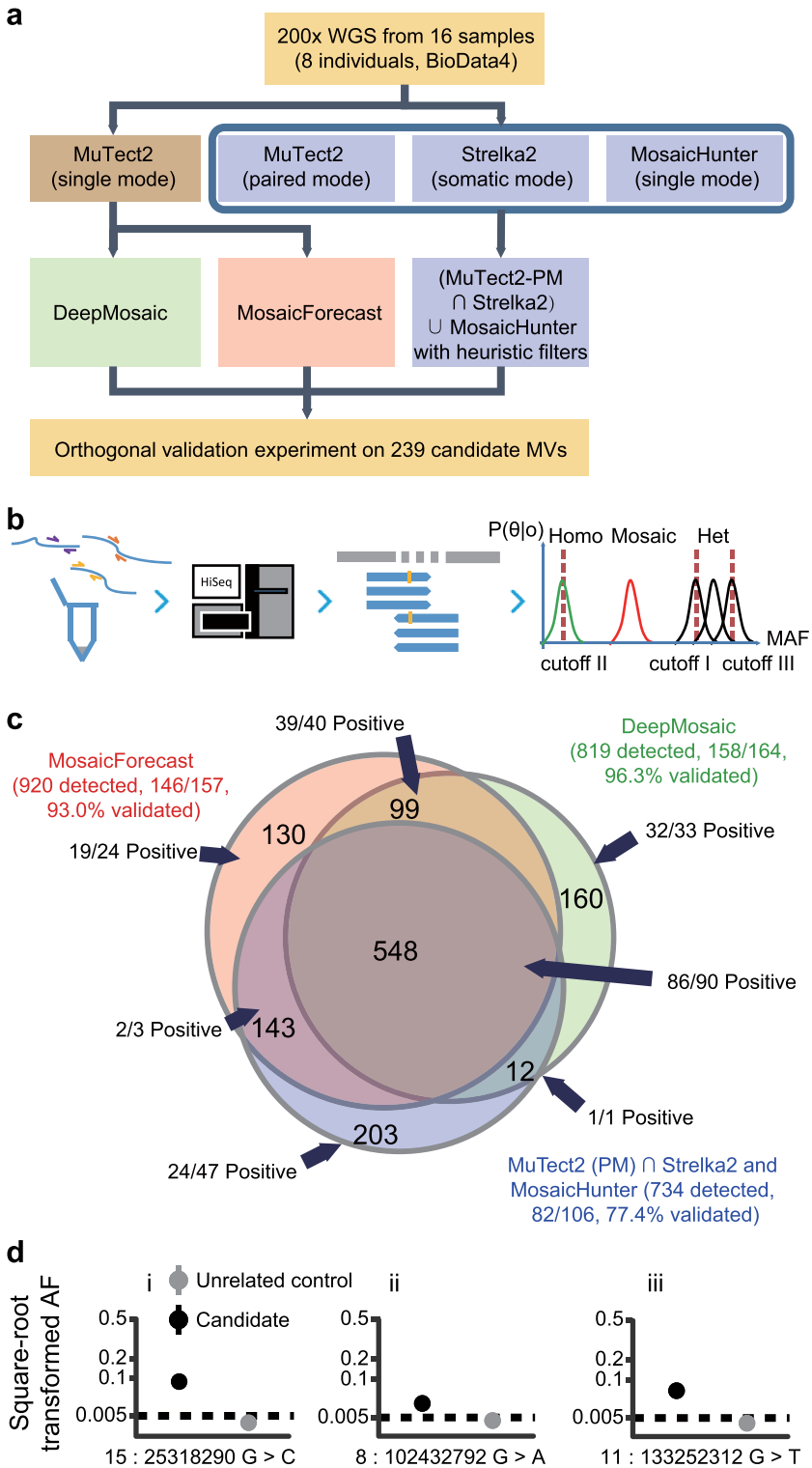
Extended Data Fig. 5 | Performance of DeepMosaic and other mosaic variant callers on SimData2. Sensitivity of DeepMosaic and other mosaic callers on 439,200 independently simulated benchmark variants (SimData2) at simulated read depths and AFs. DeepMosaic performed equally well or better than other tested methods, especially at lower expected AFs. The true positive sites to calculate sensitivity do not include variants that fall into genomic repetitive regions.

https://doi.org/10.1038/s41587-022-01559-w



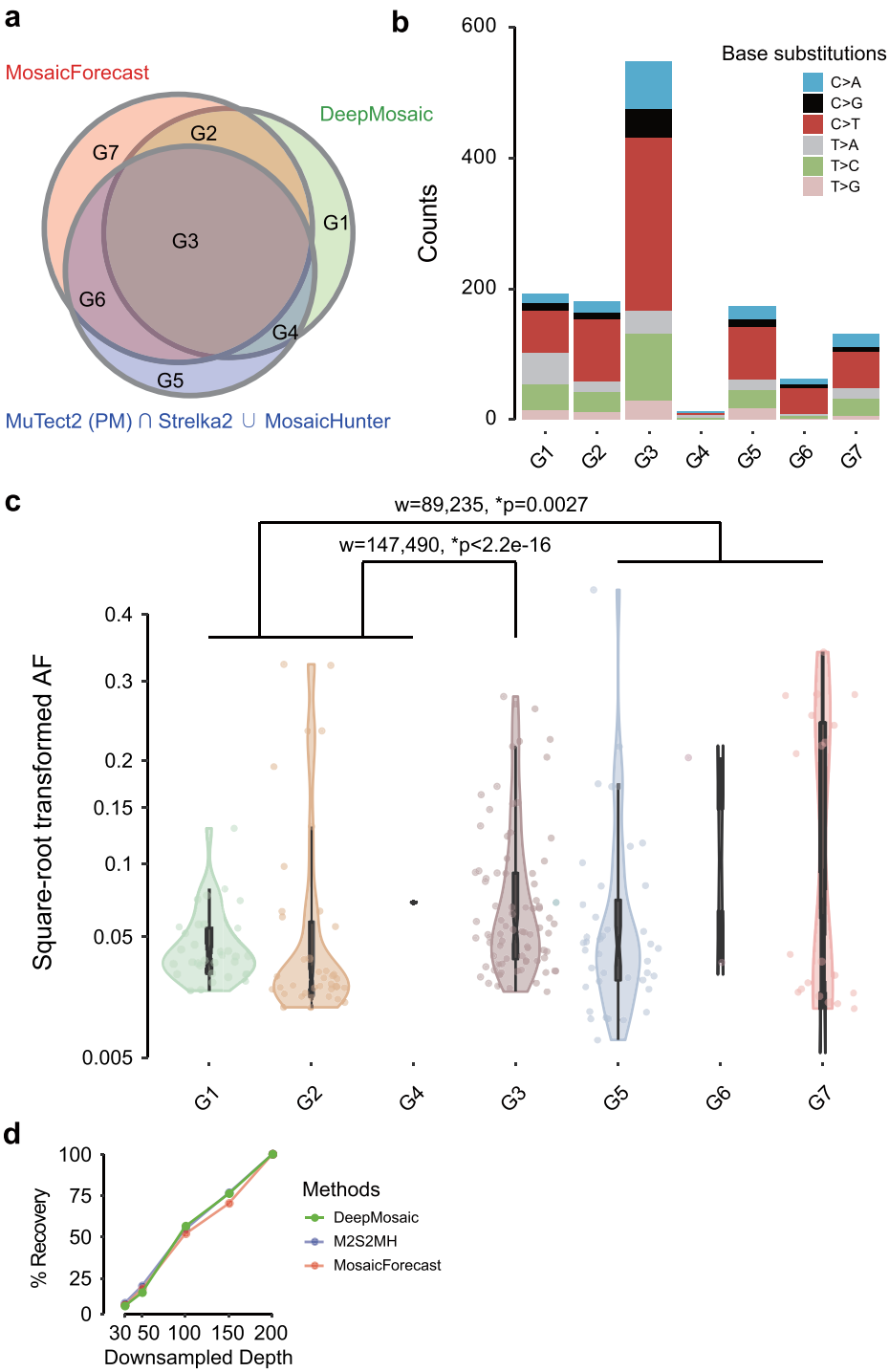
Extended Data Fig. 6 | Sensitivity and specificity of DeepMosaic and other mosaic variant callers on BioData4. Sensitivity and specificity were calculated from the orthogonal validation experiment of 239 variants from BioData4. Mosaic variant detection was carried out with DeepMosaic, MosaicForecast, MosaicHunter, MuTect2, NeuSomatic, and Strelka2 on 16 WGS samples sequenced at 200×. Raw variant calls are provided in Supplementary Table 1, and a summary of performance is provided in Supplementary Table 3. SM: single mode, variant calling without control; PM: paired mode, variant calling by comparing the sequences between two samples. PM: paired mode; SM: single mode.

https://doi.org/10.1038/s41587-022-01559-w



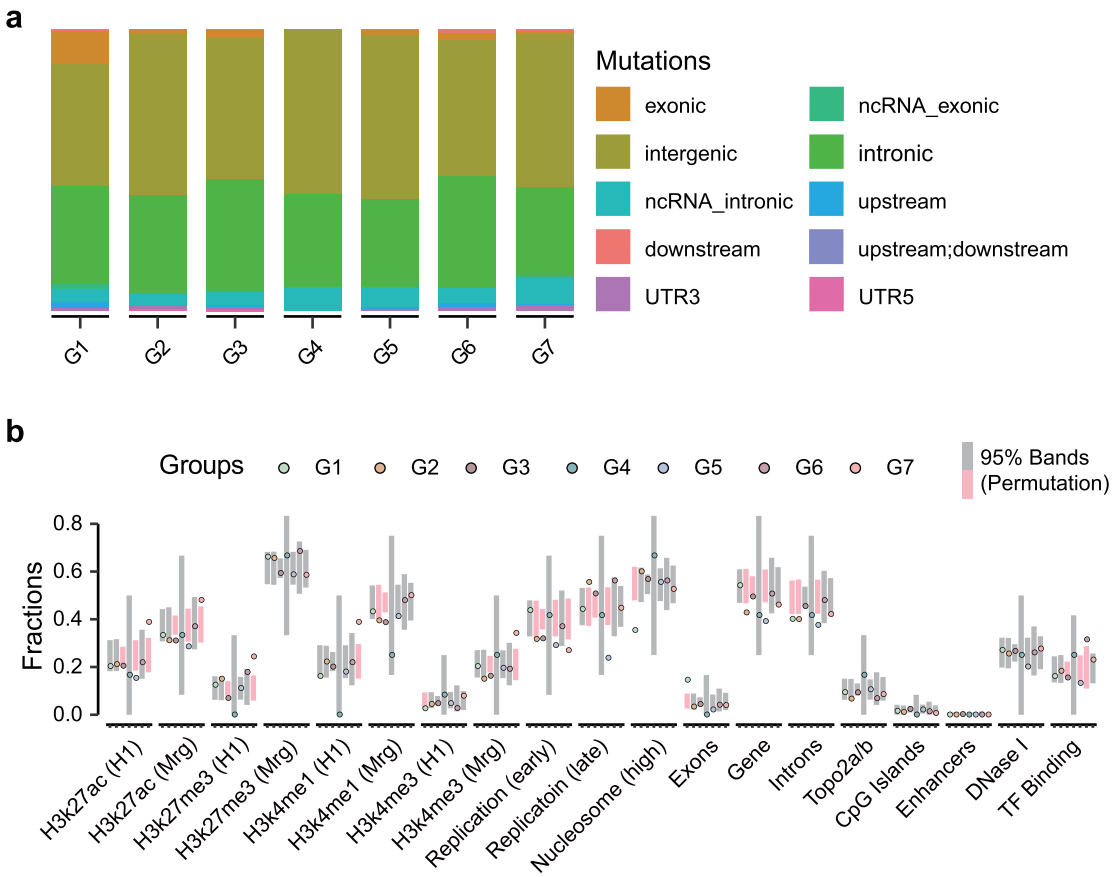
Extended Data Fig. 7 | Comparison of DeepMosaic and traditional mosaic variant calling strategies on a WGS biological dataset (BioData4). (a) Compared with the mosaic variant calling strategy (M2S2MH) used in a previous publication[28], DeepMosaic, and MosaicForecast[13] strategies are also listed. **(b)** Schematics for amplicon validation. Primers were designed for different candidates and amplicons were collected for Illumina sequencing. Information from aligned reads was calculated and genotypes were determined. **(c)** Venn diagram of the experimentally validated results and the portions of variants from different study strategies. DeepMosaic demonstrated a 96.3% (158/164) validation rate. Of all the 819 variants identified by DeepMosaic, 33.0% (271/819) were missed by the MuTect2 Strelka2 MosaicHunter pipeline with a validation rate of 97.26 (71/73) and 21.0% (172/819) were missed by the MosaicForecast pipeline with validation rate 97.06 (33/34). **(d)** Examples of validated variants are called by DeepMosaic and MosaicForecast (i), only by DeepMosaic (ii), or by DeepMosaic and other traditional methods (iii).

https://doi.org/10.1038/s41587-022-01559-w



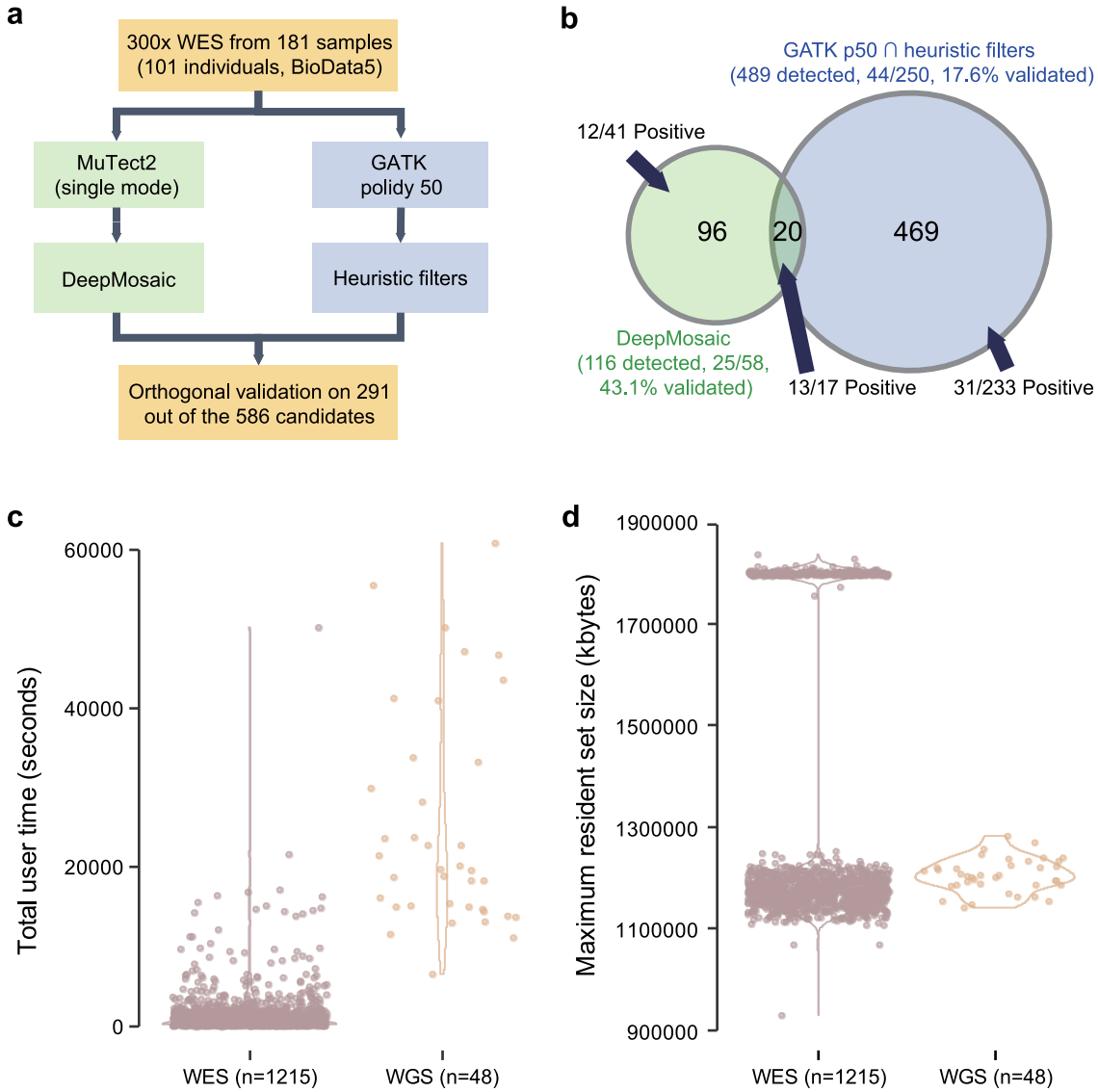
Extended Data Fig. 8 | Comparison of features of variants called by DeepMosaic and other pipelines. (a) Different overlapping groups of variants detected by the 3 pipelines were separated into 7 groups. **(b)** DeepMosaic-specific (G1) variants present similar base-substitution features compared with variants detected by the MuTect2-Strelka2-MosaicHunter combined pipeline as well as the MosaicForecast pipeline (G2-G7). **(c)** Allelic fractions of the variants detected in the original WGS sample showed that DeepMosaic-specific variants (G1, G2, and G4) showed a significantly lower average AF than variants detectable by all 3 pipelines (G3, $p < 2.2e\text{-}16$ by a two-tailed Wilcoxon rank sum test with continuity correction) and lower than variants detectable only in other pipelines (G5, G6, and G7, $p = 0.0027$ by a two-tailed Wilcoxon rank sum test with continuity correction; n = 160 for G1; n = 99 for G2; n = 548 for G3; n = 12 for G4; n = 203 for G5; n = 143 for G6; n = 130 for G7; for data in the inner boxplot, centre is the median, upper bound is the upper hinge/75% quantile, lower bond is the lower hinge/25% percentile, lower whisker represent lower hinge − 1.5*IQR, upper whisker represent upper hinge + 1.5*IQR, boundry of the violin plot is the range). **(d)** Recovery rate of DeepMosaic, M2S2MH, and MosaicForecast at different depths from downsampling of BioData3. DeepMosaic showed a similar variant recovery rate compared with M2S2MH and MosaicForecast, even when considering the lower AF variants detected by DeepMosaic.

https://doi.org/10.1038/s41587-022-01559-w



Extended Data Fig. 9 | Enrichment of genomic features for variants called by DeepMosaic and conventional methods. (a) Variants called from different pipelines shared similar variant types and contributions. The groups are defined the same as Extended Data Fig. 8a. The relative contribution of different types of MVs is stable between different variant groups. **(b)** Enrichment analysis of variants in different genomic features. Unlike the variants shared with other callers, DeepMosaic-specific (G1) variants present depletion in high nucleosome occupancy regions. 10,000 permutation was carried out on randomly selected gnomAD variants, significant comparisons are shown in pink. Overall DeepMosaic-specific variants (G1) do not show significantly different genomic features compared with permutation intervals.

https://doi.org/10.1038/s41587-022-01559-w



Extended Data Fig. 10 | Comparison of DeepMosaic and traditional mosaic variant calling strategies on a WES biological dataset (BioData5), and the computational resources required for WES (BioData6) and WGS (BioData4). **(a)** Compared with the mosaic variant calling strategy (GATK Haplotypecaller 'polidy' 50 with Heuristic filters) established in the previous publication and DeepMosaic strategies. **(b)** Venn diagram of the experimentally validated results and the portions of variants from different study strategies. DeepMosaic demonstrated a 43.1% (25/58) validation rate, significantly overperforming the 17.6% (44/250) validation rate established before[16]. **(c)** DeepMosaic consumes on average 1403.8 (range 9.1 – 50168.9) seconds to run an exome and 22718.2 (range 6565.8–60800.0) seconds for a 300× genome, respectively, on a 12-core CPU node. **(d)** DeepMosaic consumes an average of 1.3 Gb (range 0.9 Gb–1.8 Gb) maximum memory for an exome and an average of 1.2 Gb (range 1.1 Gb–1.3 Gb) for a genome. Some exomes required more resources than others and formed a bimodal distribution, but the cause for this was not explored. Results were calculated from real data run at the San Diego Supercomputer Center. For data in(c) and (d), upper and lower boundary of the violin plot is the range.

nature research

Corresponding author(s): Joseph G. Gleeson, Xiaoxu Yang

Last updated by author(s): Sep 30, 2022

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 our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒ A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Visualization of mosaic variants was based on Python (v3.7.81) packages Pysam (v0.11.2.2, https://github.com/pysam-developers/pysam) and NumPy (v1.16.2, https://numpy.org/), The input for this visualization is short-read sequencing data in the format of a BAM file, processed with a GATK (v3.8.1) best-practice pipeline (with insertion/deletion, or INDEL, realignment, followed by base quality score recalibration). Codes available as https://github.com/Virginiaxu/DeepMosaic/blob/master/deepmosaic/featureExtraction.py.
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.
SimData1: Pysim was then used to simulate paired-end sequencing reads.
SimData2: Pysim was then used to simulate paired-end sequencing reads.
SimData3: BAMSurgeon (updated 24 Dec 2020) was used to generate this simulation dataset.
BioData4: Reads were aligned to the GRCh37d5 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).
BioData5: Reads were aligned to the GRCh37d5 genome with BWA(v0.7.17) mem and duplicates were removed and base quality recalibrated by GATK (v4.0.4) according to the established best-practice pipeline.
BioData6: Fastq files were generated using Picard SAMTOFASTQ and aligned to GRCh37d5 genome with BWA (v0.7.17) mem. Duplicates were removed, reads near INDEL regions were realigned, and base quality scores were recalibrated with GATK v3.8.1 and picard v2.20.7.

Data analysis

DeepMosaic is implemented in Python (3.7.81); the code, documentation and demos are available at https://github.com/Virginiaxu/DeepMosaic. Data analysis is described in detail in the methods, and codes used to produce the data are also provided in https://github.com/shishenyxx/Adult_brain_somatic_mosaicism and https://github.com/shishenyxx/Sperm_control_cohort_mosaicism. Specifically we used the following software for data analysis: BWA v0.7.17, GATK 3.5.0, and v3.8.1 and v3.8.2, Strelka2 v2.9.2, Mutect2 from GATK v4.0.4, MosaicForecast v8-13-2019, MosaicHunter v1.0.0, Pysim, SAMtools v1.9, BEDTools v2.27.1, efficientnet_pytorch v0.6.1.
SimData1: 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 duplicates, sorting, INDEL realignment, base quality recalibration, and germline

variant calling.
SimData2: 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 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 different models trained for different read depth (250x model for depth≥300x).
SimData3: Bam files with and without simulated data were downsampled to 500x, 400x, 300x, 200x, 100x, and 50x. The data 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 different models trained for different read depth (250x model for depth≥300x).
BioData4: 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.
BioData5: Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v4.0.4) single mode as well as GATK (v4.0.4) Haplotypecaller ("polidy" 50) and previously established filters.
BioData6: Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v4.0.4) single mode, then the final call set was compared with the TCGA-MC3 call set detected by MuSE(PMID: 27557938), MuTect(PMID: 23396013), SomaticSniper(PMID: 22155872), VarScan2 (PMID: 22300766), and Radia(PMID: 25405470) using the publicly released gold standard (https://gdc.cancer.gov/about-data/publications/mc3-2017) from the same dataset.
Variant processing software includes Picard v2.18.27, BCFtools v1.10.32, sambamba v0.6.6, iFish, Define. Plotting and visulization software include R v3.5.1, ggplot2 v3.3.1, Rcpp v1.03, PyTorch v 1.6.0, pysam v0.11.2.2, Python v3.7.1 and v3.7.81, SciPy v1.3.1, pandas v0.24.2, matplotlib v3.1.1, numpy v1.16.2, and seaborn v0.9.0.

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 and reviewers. 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

WGS data used to generate the training set are available at the Sequence Read Archive (SRA, Accession No. SRP028833 and SRP100797, BioData1). The gold standard WGS data and validated capstone project data are available at the National Institute of Mental Health Data Archive (NIMH Data Archive ID 792 and 919: https://nda.nih.gov/study.html?id=792, BioData2, and https://nda.nih.gov/study.html?id=919 BioData3) and the Brain Somatic Mosaicism Consortium Data Portal, independent benchmark brain genotyping is also part of the SRA accession PRJNA736951 (BioData3). Simulated data generated from NA24385 (HG002) are available at https://humanpangenome.org/hg002/. The independent sperm and blood deep WGS data are available at SRA (Accession No. PRJNA588332 and PRJNA660493, BioData4). Independent WES data from brain, blood, and saliva samples were available in NIMH Data Archive under study number 1484 (https://nda.nih.gov/study.html?id=1484, BioData5). TCGA-MC3 data are available on the GDC portal (https://portal.gdc.cancer.gov/, sample IDs provided with variants in Supplementary Table 3).

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Eight different group of biological data and experimentally generated data were used in this study to justify the performance of the same software for biological and simulated data. Size of training data in each epoch was determined by the processing capacity of computational nodes we used. Total number of variants used for training were based on all available training data and the all possible combinations of depth and expected allelic fractions. As described in the main text, 180,000 variants were used in training and 20,265 for validation (BioData1 and SimData1), 400 variants were used for model evaluation (BioData2), 619,740 simulated variants were used for benchmark (SimData2 and SimData3). 40,848 variants from BioData1, BioData3, SimData1, and SimData2 were used for the additional model evaluation. Sixteen samples sequenced at 300x WGS were used for biological benchmark (BioData4) for DeepMosaic. One-hundred and eighty one samples sequenced at 300x WES were used as independent biological benchmark for non-cancer WES (BioData5). Data from the TCGA-MC3 collection were collected for 2430 samples from 1215 individual were used as independent biological benchmark for cancer WES (BioData6).
Data exclusions	We included all available data for this study.
Replication	1 simulation dataset and 2 biological datasets are included in the model training, 2 simulation and 4 biological datasets are included for the validation of performance, thus we did not include further complete biological replication experiments.
Randomization	No experimental groups were allocated by the scientists. Training and validation data were chosen as described in the Methods.
Blinding	Randomization and cross-validation in model training was designed in the codes, public available data are used for model training and evaluation, no further experimental blinding was assigned by the researchers.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The participants of BioData4 were previously recruited and the study included males at different ages (Breuss et al. 2020). The participants of BioData5 were patients with focal cortical dysplasia and underwent surgical treatment affected brain samples after surgery as well as control saliva or blood samples were collected. The participants of BioData6 were de-identified according to GDC regulations.
Recruitment	For BioData4, the procedure followed the Human Research Protections Programs at University of California, San Diego approval #161151. All recruited males are healthy but with at least one child diagnosed with ASD; see also Breuss et al. 2020. For BioData5, the procedure followed the Human Research Protections Programs at University of California, San Diego approval #140028. For BioData6, the data were de-identified according to GDC regulations.
Ethics oversight	IRB at UC, San Diego

Note that full information on the approval of the study protocol must also be provided in the manuscript.