

1 DeepMosaic: Control-independent mosaic single nucleotide variant

2 detection using deep convolutional neural networks

3 Xiaoxu Yang^{1,2,*,#}, Xin Xu^{1,2,*}, Martin W. Breuss^{1,2,3}, Danny Antaki^{1,2}, Laurel L. Ball^{1,2}, Changuk
4 Chung^{1,2}, Chen Li^{1,2}, Renee D. George^{1,2}, Yifan Wang⁴, Taejeoing Bae⁴, Alexej Abyzov⁴, Liping
5 Wei⁵, Jonathan Sebat^{6,7,8,9}, NIMH Brain Somatic Mosaicism Network[†], Joseph G. Gleeson^{1,2,#}

6

7 1. Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA.

8 2. Rady Children's Institute for Genomic Medicine, San Diego, CA, USA.

9 3. Department of Pediatrics, Section of Genetics and Metabolism, University of Colorado School
10 of Medicine, Aurora, CO, USA.

11 4. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

12 5. Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School
13 of Life Sciences, Peking University, Beijing, China.

14 6. Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La
15 Jolla, CA, USA.

16 7. Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA.

17 8. Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla,
18 CA, USA.

19 9. Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA.

20

21 * These authors contributed equally to this work

22 # Correspondence: jogleeson@health.ucsd.edu and xiy010@health.ucsd.edu

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

25

26 **Introductory paragraph**

27 **Mosaic variants (MVs) reflect mutagenic processes during embryonic development¹ and**
28 **environmental exposure², accumulate with aging, and underlie diseases such as cancer**
29 **and autism³. The detection of MVs has been computationally challenging due to sparse**
30 **representation in non-clonally expanded tissues. While heuristic filters and tools trained**
31 **on clonally expanded MVs with high allelic fractions are proposed, they show relatively**
32 **lower sensitivity and more false discoveries⁴⁻⁹. Here we present DeepMosaic, combining**
33 **an image-based visualization module for single nucleotide MVs, and a convolutional**
34 **neural networks-based classification module for control-independent MV detection.**
35 **DeepMosaic achieved higher accuracy compared with existing methods on biological and**
36 **simulated sequencing data, with a 96.34% (158/164) experimental validation rate. Of 932**
37 **mosaic variants detected by DeepMosaic in 16 whole genome sequenced samples, 21.89-**
38 **58.58% (204/932-546/932) MVs were overlooked by other methods. Thus, DeepMosaic**
39 **represents a highly accurate MV classifier that can be implemented as an alternative or**
40 **complement to existing methods.**

41
42 Postzygotic mosaicism describes a phenomenon whereby cells arising from one zygote harbor
43 distinguishing genomic variants^{1, 10}. MVs can act as recorders of embryonic development, cellular
44 lineage, and environmental exposure. They accumulate with aging, play important roles in human
45 cancer progression^{3, 10}, and are implicated in over 200 non-cancerous disorders^{11, 12}. Collectively,
46 estimates are that MVs contribute to 5-10% of the 'missing genetic heritability' in more than 100
47 human disorders^{11, 13}.

48
49 Compared with the higher allelic fractions (AF) of 5-10% found in clonal tumors or pre-cancerous
50 mosaic conditions, AFs found in non-clonal disorders, or neutral variants used for lineage studies,
51 are typically present at much lower AFs. Existing methods, however, based on classic statistical
52 models like MuTect² and Strelka² and heuristic filters are often optimized for the high fraction
53 variants in cancer with relatively high variant AFs. Similarly, because of their conceptual origin in
54 cancer, most existing programs including the more recent NeuSomatic¹⁴, also require matched
55 control samples. This can be problematic when mutations are present across different tissues
56 ('tissue shared' mosaicism), or when only one sample is available.

57
58 Newer methods that aim to overcome these limitations, such as MosaicHunter⁵ or
59 MosaicForecast⁴, are based conceptually similarly on the use of features extracted from raw data,
60 rather than the sequence and alignment themselves, or replace the filters with traditional
61 machine-learning methods. While these are a useful proxy, they only represent a limited window
62 into the sheer wealth of information. Because of these limitations, researchers often resort to
63 visual inspection of raw sequence alignment in a genome browser, a so-called 'pileup', to
64 distinguish artifacts from true positive variants¹⁵. This is a laborious and low-throughput process

65 that allows spot checking, but cannot be implemented on a large scale for variant lists numbering
66 in the thousands for programs like MuTect2.

67
68 Image-based representation of pileups and the application of deep convolutional neural networks
69 represents a potential solution for these limitations. Previous attempts like DeepVariant¹⁴ were
70 successful in detecting heterozygous or alternative homozygous single nucleotide variants
71 (SNVs) from direct representation of aligned reads by using deep neural networks. The
72 DeepVariant genotype model, unfortunately, did not consider a mosaic genotype, and lacked
73 orthogonal validation experiments. Here we introduce DeepMosaic comprising two modules: a
74 visualization module for image-based representation of SNVs, which forms the basic input for a
75 convolutional neural network (CNN)-based classification module for mosaic variant detection.
76 Seven different biological and computationally simulated dataset as well as amplicon validation
77 were used to train and benchmark DeepMosaic.

78
79 To automatically generate a useful visual representation similar to a browser snapshot, we
80 developed the visualization module of DeepMosaic (DeepMosaic-VM, Fig.1a-d). The input for
81 this visualization is short-read WGS data, processed with a GATK current best-practice pipeline
82 (insertion/deletion, or INDEL, realignment and base quality recalibration). DeepMosaic-VM
83 processes this input into an 'RGB' image, representing a pileup at each genomic position. In
84 contrast to a regular browser snapshot, we encode sequences as different intensities within one
85 channel, and use other channels for base quality and strand orientation. We further split the
86 pileup of reference reads and alternative reads based on the reference genome information (Fig.
87 1a-d), to improve visualization and allow assessment of mosaicism at a glance.

88
89 The classification module of DeepMosaic (DeepMosaic-CM) is a CNN-based classifier for MVs.
90 We trained 10 different CNN models with more than 180,000 image-based representations from
91 both true-positive and true-negative biological variants in several recently published high-quality
92 experimentally validated public datasets¹⁶⁻¹⁸, and computationally simulated reads with added
93 MVs (employing Illumina HiSeq error models) across a range of AFs and depths (Fig. 1e,
94 Methods and Supplementary Fig. 1a-b) to select a model with optimal performance. To ensure
95 its resemblance of real data, we controlled the distribution of AFs in the training set
96 (Supplementary Fig. 1c). In addition, a range of expected technical artifacts, including multiple
97 alternative alleles, homopolymers, and alignment artifacts, were manually curated and labeled
98 negative in the training set to represent expected pitfalls that often result in false positive mosaic
99 calls for other programs (Supplementary Fig. 1d).

100
101 To further expand training across a range of different read depths, the biological training data
102 were also up- and down-sampled to obtain data at read depths ranging from 30x to 500x
103 (Supplementary Fig. 1e), which includes the most commonly used depths for WGS in current
104 clinical and scientific settings. In addition to the output from DeepMosaic-VM, we further
105 incorporated population genomic and sequence features (e.g. population allele frequency,

106 genomic complexity, ratio of read depth), which are not easily represented in an image, as input
107 for the classifier (Fig. 1f). Depth ratios were calculated from the expected depth and used to
108 exclude false positive detections from potential copy number variations (CNVs). gnomAD
109 population allelic frequencies were used to exclude common variants. Segmental duplication and
110 repeat masker regions were used to exclude 24% of low complexity regions genome-wide.
111

112 Ten different CNN architectures were trained on 180,000 training variants described above. The
113 CNN models included Inception-v3¹⁹, which was used in DeepVariant; Deep Residual Network²⁰
114 (Resnet) which was used in the control-dependent caller NeuSomatic; Densenet²¹ and 7 different
115 builds of EfficientNet²², for its high performance on rapid image classification (Methods,
116 Supplementary Fig. 2a). Each model was trained on the data described above with 5 to 15
117 epochs to optimize the hyper-parameters until training accuracies plateaued (>0.90).
118

119 To compare the different models after training and to contrast models trained with distinct
120 datasets, we employed an independent gold-standard validation dataset of ~400 MVs from one
121 brain sample²³ (BioData2, Methods) and another amplicon-validated dataset from 18 samples
122 from one individual¹⁸ (BioData3, Methods). On these, EfficientNet-b4 showed the highest
123 accuracy, Matthews's correlation coefficient, and true positive rate when trained for 6 epochs
124 (Supplementary Fig. 2b). We thus selected this model as the default model of DeepMosaic-CM
125 (Supplementary Fig. 3a and Fig. 1f). Additional EfficientNet-b4 models trained on the 1:1 mixture
126 of biological data and simulated data showed similar performance compared with biological data
127 only training set but much higher specificity compared with models trained only on simulated data
128 (Supplementary Fig. 2c).
129

130 To uncover the information prioritized by the selected default model, we used a gradient
131 visualization technique with guided backpropagation²⁴ to highlight the pixels with guiding
132 classification decisions (Supplementary Fig. 3b). The results suggested that the algorithm not
133 only recognized the edges for reference and alternative alleles, but also integrated additional
134 available information, such as insertion/deletions in the sequences, overall base qualities,
135 alignment artifacts, and other features which may not be extracted by digested feature-based
136 methods.
137

138 We evaluated the performance of DeepMosaic using 20,265 variants from the above training
139 data that were hidden from model training and selection. The receiver operating characteristic
140 (ROC) curve and precision-recall curves on the hidden validation dataset showed >0.99 area
141 under the curve for a range of coverages (30x ~500x, Supplementary Fig. 4a and 4b) across a
142 range of AFs (Supplementary Fig. 4c and 4d), demonstrating high sensitivity and specificity.
143

144 Next, we benchmarked DeepMosaic's performance relative to other detection software, using
145 data generated from two distinct sequencing error models to test for its utility on general
146 sequencing data. We compared the performance of DeepMosaic with the widely used MuTect2

147 (paired mode), Strelka2 (somatic mode) with heuristic filters, MosaicHunter (single mode), and
148 MosaicForecast (Methods). We generated two additional computationally simulated datasets of
149 439,200 and 180,540 positions based on the error model of a different Illumina sequencer with
150 similar methods as the training set (NovaSeq, SimData2, Methods) or a similar ratio of true
151 positive and true negative labels as real biological data¹⁸ by replacing reads from the ‘Genome
152 in a Bottle’ sample HG002 (NA24345, SimData3, Methods)^{25, 26}, with AF ranges from 1% to 25%,
153 and depth ranges from 50x to 500x. MuTect2 paired methods and Strelka2 somatic mode used
154 simulated mutated samples as “tumor” and simulated reference or original HG002 samples as
155 “normal” for their paired modes. DeepMosaic showed equal or better performance than all other
156 methods tested, especially for low allelic fraction variants (Fig. 2 and Supplementary Fig. 5),
157 noticeably, even for low read depth data; and it performed better than methods that have
158 additional information from paired samples. Overall DeepMosaic showed a 1.5-3 fold increase of
159 the detection sensitivity for AFs under 3% compared with other methods (Fig. 2b), with
160 comparable specificity (Fig 2a). This is likely because our models integrate additional genomic
161 sequence and quality information from the original BAM file (Supplementary Fig. 3b), and are
162 capable of distinguishing mosaic variants from different Illumina error models.

163
164 To exclude limitations resulting from benchmarking with simulated data and demonstrate that
165 models trained on PCR-amplified libraries are also useful for PCR-free sequencing libraries, we
166 extended benchmarking to biological data. We performed the same comparison on our recently
167 published 200x WGS dataset¹² with 16 samples (blood and sperm) from 8 healthy individuals²⁷.
168 Paired methods compared two samples from the same individual, and control-independent
169 samples used a published dataset of a panel of normals²⁸. Variants detected by MuTect2 (paired
170 mode), Strelka2 (somatic mode) and MosaicHunter (single mode) were subjected to a series of
171 published heuristic filters^{27, 28}. As we had access to the biological samples, we also performed
172 orthogonal validation, using deep amplicon sequencing of 241 randomly selected MVs with a
173 representative AF distribution compared to the complete candidate variant list (Methods, Fig. 3a
174 and 3b, Supplementary Table 1).

175
176 As expected from the test of the computationally generated data, DeepMosaic showed the
177 highest overall validation rate (96.34%, 158/164) among all 5 methods (Fig. 3c), demonstrating
178 the power of DeepMosaic that models trained on PCR-amplified biological data and simulated
179 data can accurately classify these PCR-free biological data. Of the 932 MVs detectable by
180 DeepMosaic, 21.89% (204/932, 33/34 experimentally validated) were overlooked by
181 MosaicForecast, 58.58% (546/932, 96/98 validated) overlooked by MosaicHunter, 50.32%
182 (469/932, 90/94 validated) overlooked by Strelka2 (somatic mode) with heuristic filters, 43.13%
183 (402/932, 81/85 validated) overlooked by MuTect2 (paired mode) with heuristic filters²⁷.
184 DeepMosaic also accurately detected variants with relatively low AFs (Fig. 3d). Finally,
185 DeepMosaic outperformed other methods across most of the AF bins (Fig. 3e).

186

187 In current practice, researchers often combine multiple programs in one variant detection pipeline
188 to detect different categories of MVs²⁷⁻²⁹. We thus further compared DeepMosaic with different
189 pipelines used in recent publications, using data from 200x WGS of the 16 samples²⁷: 1] With
190 the MosaicForecast pipeline⁴, which uses MuTect2 single mode (each sample compared with
191 the publicly available panel of normal) as input; 2] With what we call the M2S2MH pipeline, which
192 we recently published²⁷, combining MuTect2 paired mode (i.e. compared between different
193 samples from a same individual), Strelka2 somatic mode and MosaicHunter single mode
194 followed by a series of heuristic filters (Supplementary Fig. 6a). Of the 932 MVs identified by
195 DeepMosaic, 78.11% (728/932, 125/130 validated) overlapped with MosaicForecast and 60.09%
196 (560/932, 87/91 validated) overlapped with M2S2MH. In contrast, 21.89% (204/932, 33/34
197 validated) were undetected by MosaicForecast, and 39.91% (372/932, 71/73 validated) were
198 overlooked by M2S2MH. These variants uniquely detected by DeepMosaic all showed validation
199 rate > 97% (Supplementary Fig. 6b and 6c), demonstrating that DeepMosaic can accurately
200 detect a considerable number of variants undetectable by widely used methods.
201

202 To test the performance of these samples on data widely curated clinically, we compared
203 detection sensitivity for genome samples with standard WGS read depth, by down-sampling
204 blood-derived data from a 70-year old healthy individual, in whose blood we observed the highest
205 number of mosaic variants (due to clonal hematopoiesis²⁷). As all programs had high validation
206 on this sample at 200x, the recovery rate was used to distinguish the ability of different programs
207 to detect clonal hematopoiesis variants. DeepMosaic showed similar recovery in the down-
208 sampled data (Supplementary Fig. 7) as M2S2MH and slightly outperformed MosaicForecast at
209 100x and 150x. We found that the performance of DeepMosaic was not substantially influenced
210 by the read depth according to the down-sampling benchmark on biological data.
211

212 To understand whether different pipelines had unique strengths or weaknesses, we separated
213 all the detected variants into 7 groups (G1-G7) based upon sharing between different pipelines,
214 Supplementary Fig. 7a). DeepMosaic specific variants showed similar base substitution features
215 compared with other methods (Supplementary Fig. 7b). Similar to the computationally derived
216 data, we found that DeepMosaic recovered additional low AF MVs with high accuracy (validation
217 rate 95%, Supplementary Fig. 7c). Finally, we summarized the genomic features of variants
218 detected by DeepMosaic and other pipelines. All caller groups report similar ratios of intergenic
219 and intronic variants (Supplementary Fig. 8a). Analysis of other genomic features showed
220 DeepMosaic-specific variants (G1) expressed consistency with other groups (Supplementary Fig.
221 8b), reflecting that the low-fraction variants detectable only by DeepMosaic do not represent
222 technical artifacts.
223

224 While we propose DeepMosaic as a powerful tool for mosaic variant detection, it currently is
225 underpowered for mosaic INDELs and mosaic repetitive variant detection which might be error-
226 prone in the genome. In practice, MosaicForecast can detect mosaic INDEL variants with high
227 accuracy, while M2S2MH has good performance for tissue-specific variants due to the inclusion

228 of additional information from the “normal” comparison sample. Thus different methods
229 complement one another.

230
231 DeepMosaic is the first image-based tool for the accurate detection of mosaic SNVs from short-
232 read sequencing data and does not require a matched control sample. Compared with
233 NeuSomatic that compresses all the bases in a genomic position into 10 features⁶, DeepMosaic-
234 VM provides complete representation of information present in the BAM file. Compared with other
235 re-coding methods like DeepVariant¹⁴, DeepMosaic-CM has the ability to define MVs as an
236 independent genotype and DeepMosaic-VM can be applied as an independent variant
237 visualization tool for the user’s convenience. To further integrate population information not
238 present in the raw BAM, 4 different features are also integrated in DeepMosaic to facilitate
239 classification.

240
241 Despite the unique features from image representation and a neural network based variant
242 classifier, DeepMosaic can reproducibly identify the majority (~70%) of variants detectable by
243 conventional methods; in addition, however, this unique architecture results in higher sensitivity,
244 and the detection of variants with relatively lower AF both in simulated and experimentally derived
245 data validated by orthogonal experiments. DeepMosaic shows a drop of sensitivity at higher AF,
246 likely due to our inclusion of depth ratio which help to avoid false-positive calls from CNV. The
247 higher sensitivity at lower AFs will make it a good complement for other methods.

248
249 Both down-sampled biological data in blood of an individual with advanced age and
250 computationally generated data showed that DeepMosaic has the potential to identify variants at
251 relatively high sensitivity and high accuracy for WGS at depths as low as 50x. Clonal
252 hematopoiesis in blood without a known driver mutation is reported³⁰, but can be difficult to detect
253 because of technical limitations induced by noise and lower supporting read counts³¹. For the
254 past 10-15 years, hundreds of thousands of whole-genome sequencing datasets from clinical,
255 commercial, or research labs have been generated at relatively low depth, but most have not
256 been subjected to unbiased mosaicism detection due to lack of sufficiently sensitive methods.
257 DeepMosaic could enable a genome-level unbiased detection of mutations that requires only
258 conventional sequencing data.

259
260 By using a training set comprising representative technical artifacts such as homopolymers and
261 truncated reads, DeepMosaic acquired the power to distinguish biologically true positive variants
262 from false positives, which might be filtered out directly by rule-based methods like
263 MosaicHunter⁵ or MosaicForecast⁴. We demonstrated that training the models on a mixture of
264 ~1:1 simulated and biological data does not adversely affect performance on an independent
265 biological evaluation set. We also demonstrated that DeepMosaic works well for various Illumina
266 short read sequencing platforms applying different library preparation strategies (PCR-amplified
267 and PCR-free).

268

269 Although the EfficientNet-b4 performed best, we provide all pre-trained models (Densenet,
270 EfficientNet, Inception-v3, and Resnet) as DeepMosaic-CM modules on GitHub. Users are
271 provided with the options to prepare their own data with labeled genotypes for model training for
272 DeepMosaic, to generate data-specific, personalized models, to test other potential factors
273 influencing detection sensitivity such as the ratio of positive:negative labels, and to increase the
274 detection specificity for DeepMosaic on specialized data sets. For instance, homopolymers and
275 tandem repeats are increasingly recognized as important in disease and development, but are
276 currently not detected with DeepMosaic, because of the limited training data; however, users
277 with specialized data sets could remedy this problem by re-training.

278

279 Likewise, gnomAD population AF features used in this study also rely on a matched ancestry
280 background to avoid population stratification. Annotations such as gene names, variant
281 functional annotations, gnomAD allelic frequency, homopolymer and dinucleotide repeat
282 annotation, as well as segmental duplication and UCSC repeat masker regions are provided in
283 the final output to facilitate customization, as described at the GitHub homepage of DeepMosaic
284 (<https://github.com/VirginiaXu/DeepMosaic>). Finally, apart from MuTect2 single mode,
285 DeepMosaic can also process variant lists generated by multiple methods such as the GATK
286 HaplotypeCaller with ploidy 50 or 100²³. Thus, DeepMosaic can be used directly as is, or can be
287 customized to the needs of the end users, providing an adaptable and efficient mosaic variant
288 detection workflow.

289 **References**

- 290 1. Dou, Y., Gold, H.D., Luquette, L.J. & Park, P.J. Detecting somatic mutations in normal
291 cells. *Trends in Genetics* **34**, 545-557 (2018).
- 292 2. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium.
293 *Nature* **578**, 266-272 (2020).
- 294 3. Lee, J.H. et al. Human glioblastoma arises from subventricular zone cells with low-level
295 driver mutations. *Nature* **560**, 243-247 (2018).
- 296 4. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched
297 controls. *Nature biotechnology* (2020).
- 298 5. Huang, A.Y. et al. MosaicHunter: accurate detection of postzygotic single-nucleotide
299 mosaicism through next-generation sequencing of unpaired, trio, and paired samples.
300 *Nucleic acids research* **45**, e76 (2017).
- 301 6. Sahraeian, S.M.E. et al. Deep convolutional neural networks for accurate somatic
302 mutation detection. *Nature communications* **10**, 1041 (2019).
- 303 7. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat
304 Methods* **15**, 591-594 (2018).
- 305 8. Wood, D.E. et al. A machine learning approach for somatic mutation discovery. *Sci Transl
306 Med* **10** (2018).

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

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

311 11. Yang, X. et al. MosaicBase: a knowledgebase of postzygotic mosaic variants in
312 noncancer disease-related and healthy human individuals. *Genomics Proteomics
313 Bioinformatics* **18**, 140-149 (2020).

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

316 13. Freed, D., Stevens, E.L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes*
317 **5**, 1064-1094 (2014).

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

320 15. McNulty, S.N. et al. Diagnostic Utility of Next-Generation Sequencing for Disorders of
321 Somatic Mosaicism: A Five-Year Cumulative Cohort. *Am J Hum Genet* **105**, 734-746
322 (2019).

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

325 17. Huang, A.Y. et al. Distinctive types of postzygotic single-nucleotide mosaisms in healthy
326 individuals revealed by genome-wide profiling of multiple organs. *PLoS Genet* **14**,
327 e1007395 (2018).

328 18. Breuss, M.W. et al. Somatic mosaicism in the mature brain reveals clonal cellular
329 distributions during cortical development. *bioRxiv*, doi:10.1101/2020.08.10.244814
330 (2020).

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

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

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

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

339 23. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human
340 brain tissue. *Genome biology* **22**, 92 (2021).

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

343 25. Ewing, A.D. et al. Combining tumor genome simulation with crowdsourcing to benchmark
344 somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623-630 (2015).

345 26. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human
346 genomes. *Nature biotechnology* **37**, 555-560 (2019).

347 27. Breuss, M.W. et al. Autism risk in offspring can be assessed through quantification of

348 male sperm mosaicism. *Nat Med* **26**, 143-150 (2020).

349 28. Yang, X. et al. Temporal stability of human sperm mosaic mutations results in life-long
350 threat of transmission to offspring. *bioRxiv*, doi:10.1101/2020.10.14.339796 (2020).

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

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

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

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

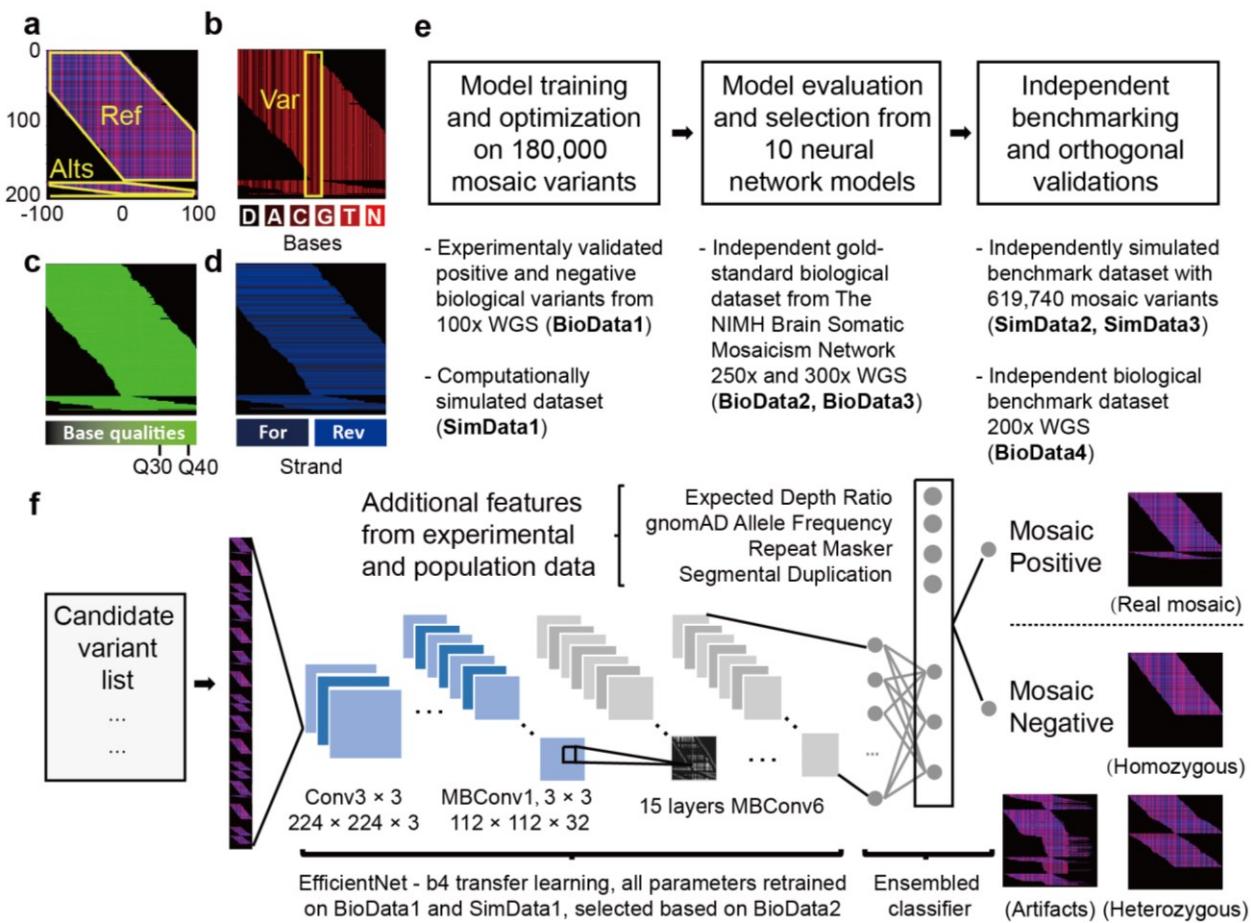
359 33. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program
360 Primer3. *Bioinformatics* **23**, 1289-1291 (2007).

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

363

364

365



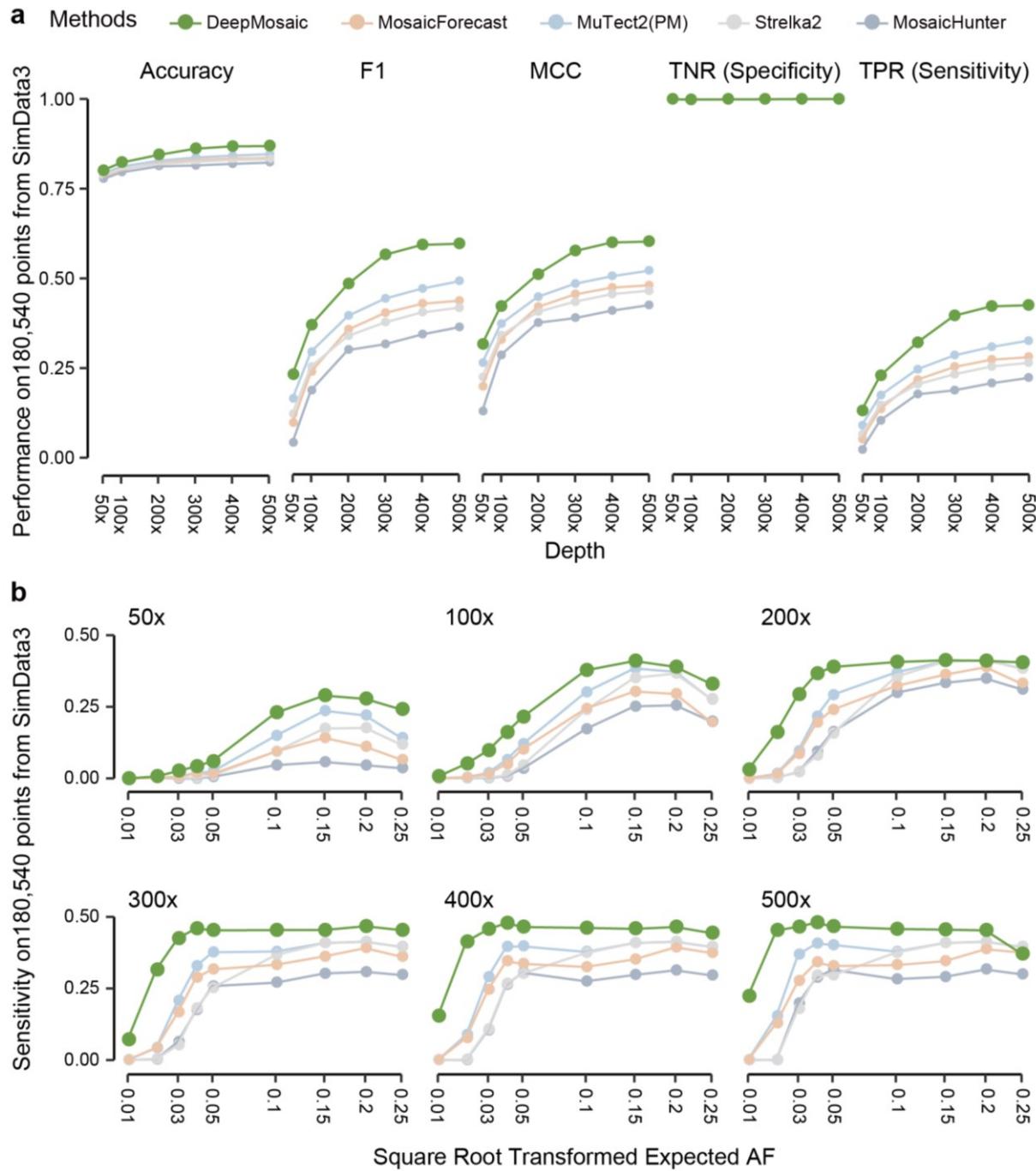
366

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

368 **a**, DeepMosaic-VM: Composite RGB image representation of sequenced reads separated into
369 “Ref” - reads supporting the reference allele; or “Alts” - reads supporting alternative alleles; each
370 outlined in yellow. **b**, Red channel of the compound image contains base information from BAM
371 file. “D” - deletion; “A” – Adenine; “C” – cytosine; “G” – guanine; “T” – thymine; “N” – low-quality
372 base. Yellow box: Var: candidate position, centered in the image. **c**, Green channel: base quality
373 information. Note that channel intensity was modulated in this example for better visualization.
374 **d**, Blue channel: strand information (i.e. forward or reverse). **e**, Model training, model selection,
375 and overall benchmark strategy for DeepMosaic-CM (Methods and Supplementary Fig. 1). Ten
376 different convolutional neural network models were trained on 180,000 experimentally validated
377 positive and negative biological variants from 29 WGS data from 6 individuals sequenced at
378 100x^{16, 17} (BioData1), as well as simulated data with different AFs (SimData1) resampled to a
379 different depth. Models were evaluated based upon an independent gold-standard biological
380 dataset from the 250x WGS data of the Reference Tissue Project of the Brain Somatic Mosaicism
381 Network²³ (BioData2) as well as an independent 300x WGS dataset from the Brain Somatic

382 Mosaicism Network Capstone project¹⁸ (Biodata3). DeepMosaic was further benchmarked on 16
383 independent biological datasets from 200x WGS data²⁷ (BioData4) as well as 619,740
384 independently simulated variants (SimData2 and SimData3). Deep amplicon sequencing was
385 carried out as an independent evaluation on variants detected by different software (Supplement
386 Table 1). **f**, Application of DeepMosaic-CM in practice. Input images are generated from the
387 candidate variants. 16 convolutional layers extracted information from input images. Population
388 genomic features were assembled for final output. Images of positive and negative variants are
389 shown as examples. Conv: convolutional layers; MBConv: mobile convolutional layers.

390



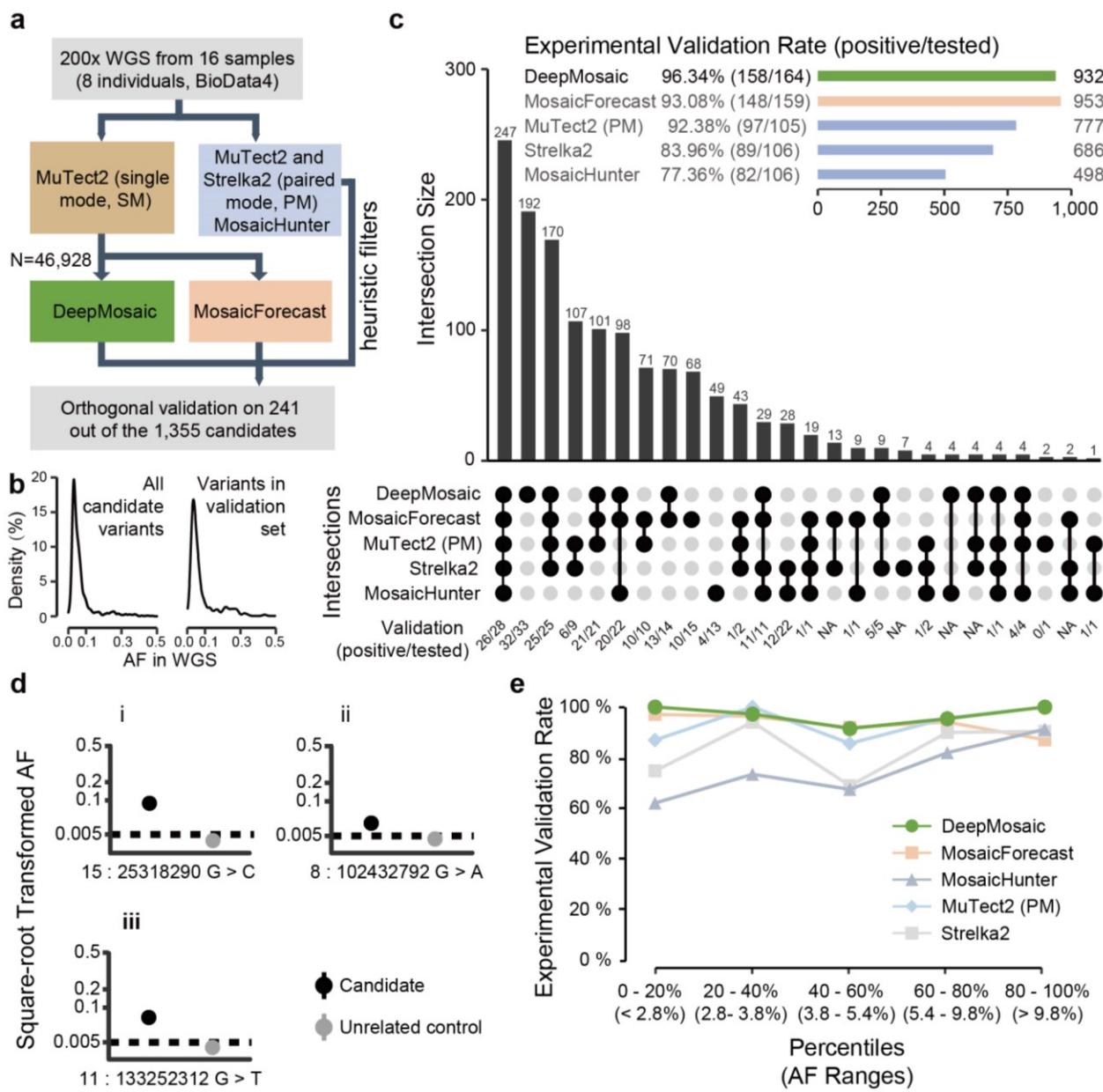
391

392 **Fig. 2| DeepMosaic performance on simulated benchmark variants.**

393 **a**, Benchmark test on 180,540 genomic positions (SimData3) generated by replacing reads from
 394 biological data with simulated variants. DeepMosaic showed higher accuracy, F1 score, MCC
 395 (Matthews correlation coefficient), sensitivity, and comparable specificity compared with widely
 396 accepted methods for mosaic variant detection. **b**, Sensitivity of DeepMosaic and other mosaic
 397 callers on SimData3 at simulated read depths and AFs. DeepMosaic performed equally well or

398 better than other tested methods, especially at lower read-depths and lower expected AFs.
399

400



401

402 **Fig. 3| DeepMosaic performance validated on biological data.**

403 **a**, DeepMosaic and other mosaic variant detection methods were applied to 200x whole-genome
 404 sequencing data from 16 samples, which were not used in the training or validation stage for any
 405 of the listed methods (BioData4). Raw variant lists were either obtained by comparing samples
 406 using a panel-of-normal²⁸ strategy with MuTect2 single mode, between different samples from a
 407 same individual using MuTect2 paired mode or Strelka2 somatic mode, or detected directly
 408 without control with MosaicHunter single mode with heuristic filters²⁷. A total of 46,928 candidate
 409 variants from MuTect2 single mode were analyzed by DeepMosaic and MosaicForecast.

410 Orthogonal validation with deep amplicon sequencing was carried out on a total of 241 variants
411 out of the 1355 candidates called by at least one method. **b**, Distribution of AFs of the whole
412 candidate mosaic variant list and the 241 randomly selected variants. **c**, Comparison of validation
413 results between different mosaic variant calling methods, ‘UpSet’ plot shows the intersection of
414 different mosaic detection methods and the validation result of each category. Variants identified
415 by DeepMosaic showed the highest validation rate on biological data. **d**, Examples of validated
416 variants called by DeepMosaic and MosaicForecast (i), only by DeepMosaic (ii), or by
417 DeepMosaic and other traditional methods (iii). **e**, Comparison of validation rate in different AF
418 range percentage bins of variants. DeepMosaic showed the highest validation rate at a range of
419 AFs, approximately 48 experimentally validated variants are shown in each AF bin.
420

421 Methods

422 **Curation of training and benchmark data**

423 **SimData1:**

424 For the initial training procedure, 10,000 variants were randomly generated on chromosome 22
425 to get the list of alternative bases. Pysim³² was then used to simulate paired-end sequencing
426 reads with random errors generated from the Illumina HiSeq sequencer error model. Alternative
427 reads were generated by replacing the genomic bases with the alternative bases in the list, with
428 the same error model. Alternative and reference reads were randomly mixed to generate an
429 alternative AF of 0, 1, 2, 3, 4, 5, 10, 15, 20, 25, and 50%. The data were randomly sampled for
430 a targeted depth of 30, 50, 100, 120, 150, 200, 250, 300, 400, and 500x. FASTQ files were
431 aligned to the GRCh37d5 human reference genome with BWA (v0.7.17) *mem* command. Aligned
432 data were processed by GATK (v3.8.1) and Picard (v2.18.27) for marking duplicate, sorting,
433 INDEL realignment, base quality recalibration, and germline variant calling. The up- and down-
434 sampling expanded this dataset into a pool of 990,000 different variants. Depth ratios were
435 calculated as defined. To avoid the situation that randomly generated mutations falls on a
436 common SNP position in the genome, which would bias the training and benchmarking, gnomAD
437 allele frequencies were randomly assigned from 0 to 0.001 for simulated mosaic positive and
438 from 0 to 1 for simulated negative variants, which were established as homozygous or
439 heterozygous.

440

441 **BioData1:**

442 Variant information and raw sequencing reads from 80-120x PCR-amplified PE-150 WGS data
443 of 29 samples from 6 normal individuals were extracted from published data^{16, 17} on SRA
444 (SRP028833, SRP100797, and SRP136305). 921 variants identified from WGS of samples from
445 different organs of the donors and validated by orthogonal experiments were selected and
446 labeled as mosaic positive. 492 genomic positions from the control samples validated with 0%
447 AF were selected and labeled as negative. 162 variants with known sequencing artifacts were
448 first filtered by MosaicHunter, manually selected and labeled as negative. The 1575 genomic
449 positions were also down-sampled and up-sampled for a targeted depth of 30, 50, 100, 150, 200,
450 250, 300, 400, and 500x, to expand this dataset into a pool of 14,175 different conditions. Depth
451 ratios were calculated accordingly, gnomAD allele frequencies, segmental duplication, and
452 repeat masker information was annotated.

453 The entire BioData1 and random subsampling from SimData1 were combined to generate a
454 training and validation dataset with approximately 200,000 variants from the 1,000,000 training
455 variants. 180,000 variants were selected for model training, 45% from SimData1 and 55% from
456 resampling of BioData1. This dataset was used for the model training and evaluation of the
457 sensitivity and specificity of the selected model, and their features including AF distribution and
458 biological appearances were very similar to published biological data (Supplementary Fig. 1).

459

460 **BioData2:**

461 To estimate the performance of the pre-trained models and select the model with the best
462 performance for DeepMosaic-CM, we introduced an independent gold-standard dataset.
463 Variants were computationally detected from replicated sequencing experiments generated from
464 6 distinct sequencing centers and validated in 5 different centers, known as the common
465 reference tissue project from the Brain Somatic Mosaicism Network²³. 400 variants underwent
466 multiple levels of computational validations including haplotype phasing, CNV exclusion,
467 population shared exclusion, as well as experimental validation such as whole-genome single
468 cell sequencing, Chromium Linked-read sequencing (10X Genomics), PCR amplicon
469 sequencing, and droplet digital PCR. After validation, 43 true positive MVs and 357 false positive
470 variants were determined as gold-standard evaluation set for low-fraction single nucleotide MVs
471 from the 250x WGS data²³. We extracted deep whole-genome sequences for those variants,
472 labeled them accordingly and used them as gold standard validation set for model selection
473 (Supplementary Fig. 2).

474

475 BioData3:

476 To evaluate the performance of DeepMosaic-CM trained on a different portion of biological
477 variants, we included another large-scale validation experiment we recently generated. Variant
478 information and raw sequencing reads of 300x PCR-free PE150-only WGS of 18 samples from
479 9 different brain regions, cerebellum, heart, liver, and both kidneys of one individual were
480 extracted from the capstone project of the Brain Somatic Mosaicism Network¹⁸. 1400 genomic
481 positions with variants identified from WGS sample and reference homozygous/heterozygous
482 controls validated by orthogonal experiments were selected and labeled as positive and negative
483 according to the experimental validation result. The 1400 genomic positions were also down-
484 and up-sampled for a targeted depth of 30, 50, 100, 150, 200, 250, 300, 400, and 500x. Depth
485 ratios were calculated accordingly, gnomAD allele frequencies, segmental duplication, and
486 repeat masker information were annotated.

487

488 SimData2:

489 To compare the performance of DeepMosaic and other software to detect mosaicism on
490 simulated data, we randomly generated another simulation dataset, with the following
491 modifications: 1] only 7610 variants on non-repetitive region of chromosome 22 were considered
492 true positive genomic positions; 2] random errors were generated from the Illumina NovaSeq
493 sequencer error model. 3] Data was randomly down-sampled and up-sampled for a targeted
494 depth of 50, 100, 200, 300, 400, and 500x. A total of 439,200 different variants were generated.
495 FASTQ files were aligned and processed with BWA (v0.7.17), SAMtools (v1.9), and Picard
496 (v2.18.27). The data was subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both
497 paired mode and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast
498 (v8-13-2019) with different models trained for different read depth (250x model for depth \geq 300x).

499

500 SimData3:

501 We further generated another simulation dataset in a way that was fundamentally different from

502 the training data with a positive:negative ratio similar to real data¹⁸ to compare the performance
503 of DeepMosaic and other software for the detection of mosaic variants. We selected 30,090
504 genomic positions with reference homozygous genotype from a different genomic region (the
505 entire Chromosome 1) of the whole-genome deep sequences from the ‘Genome In a Bottle’
506 sample HG002 (NA24345)²⁶. The genomic positions from the 30,090 positions were genotyped
507 as homozygous and fulfilled additional criteria 1] zero alternative bases in the raw sequencing
508 data; 2] no detectable insertions/deletions in the position of interest; 3] have a genomic distance
509 of at least 1000 bases between each other. On this clear background, 15,471 of them were
510 labeled as “true negative” with reference homozygous genotype, 6868 were labeled as “true
511 positive” mosaic variants with expected alternative AF 0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15,
512 0.20, and 0.25 (on average 763 variants for each genotype); 7751 were labeled as “true negative”
513 heterozygous variants with alternative AF 0.50; the latest version of a different software
514 BAMSurgeon (updated 24 Dec 2020) was used to generate this simulation dataset and retain
515 the sequencing errors from the original biological samples. The original bam file was first up-
516 sampled, alternative reads were replaced to generate the expected AF, mapped back to the
517 genome and merged back to the bam file, according to the software manual²⁵. Bam files with
518 and without simulated data were downsampled to 500x, 400x, 300x, 200x, 100x, and 50x. The
519 data were subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both paired mode and
520 single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with
521 different models trained for different read depth (250x model for depth \geq 300x), the performance
522 of the 180,540 points were evaluated.

523

524 **BioData4:**

525 This additional dataset was used to compare the performance of DeepMosaic and other mosaic
526 variant callers on biological samples. 16 WGS samples from blood and sperm of 8 individuals
527 were sequenced at 200x²⁷ (PRJNA588332). WGS was performed using an Illumina TrueSeq
528 PCR-free kit with 350bp insertion size and sequenced on an Illumina HiSeq sequencer. Reads
529 were aligned to the GRCh37 genome with BWA (v0.7.15) *mem* and duplicates were removed
530 with sambamba (v0.6.6) and base quality recalibrated by GATK (v3.5.0). Processed BAM files
531 were subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both paired mode and single
532 mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with 200x
533 models trained for the specific depth. Data from one of the individuals (F02) was down-sampled
534 to 150x, 100x, 50x, and 30x with the SAMtools (v1.9) *view* command for the further benchmark
535 of DeepMosaic.

536 **Neural network building and model training**

537 For the 10 neural network architectures, Inception-v3, Resnet and Densenet were imported from
538 PyTorch’s (v1.4.0) built-in library, while the 7 different builds of EfficientNet were imported from
539 the efficientnet_pytorch (v0.6.1) Python (v3.7.1) package. The final fully connected layer of each
540 model was replaced to be fed into 3 output units representing intermediate results instead of the
541 default 1,000 output units for the 1,000 ImageNet classes to significantly reduce the total images

542 required to extract basic features such as edges, stripes from raw images. A transfer-learning
543 method was adopted for model training. Each model's initial pre-trained weights provided by
544 Pytorch and efficientnet_pytorch packages were trained on the ImageNet dataset. Before model
545 training, we randomly divided the entire training dataset (including down-sampling and up-
546 sampling of SimData1 and BioData1) to 80% "training" and 20% "evaluation" sets and fixed the
547 split during model training while shuffling the order within the training set and evaluation set for
548 each training epoch to form mini-batches for gradient descent. Each network architecture was
549 trained using a batch size of 20 with a stochastic gradient descent (SGD) optimizer with learning
550 rate of 0.01, and momentum of 0.9. The training was terminated until the training losses
551 plateaued and evaluation accuracy reached 90% for each model architecture. The training was
552 conducted on NVIDIA Kepler K80 GPU Nodes on San Diego Supercomputer Centre's Comet
553 computational clusters.

554 **Network selection**

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

562 **Independent model training and evaluation for DeepMosaic-CM**

563 To evaluate the performance of DeepMosaic-CM when trained on a different portion of biological
564 variants, 15 epochs were trained for the EfficientNet-b4 architecture on 5 different training sets
565 consisting of 122,424 genomic positions. EfficientNet was imported from the efficientnet_pytorch
566 (v0.6.1) Python (v3.7.1) package. The 5 different training sets were generated based on
567 SimData1, BioData1, SimData2, and BioData3. 1] BioData only: 40,808 variants from the entire
568 of BioData1 and BioData3 were pooled. The overall positive:negative ratio was 26.8%:73.2%. 2]
569 SimData only for SimData1: 40,808 variants were selected from SimData1 with the matched
570 number of positive and negative labels as BioData only. 3] SimData only for SimData2: 40,808
571 variants that were agreed by both MuTect2 and Strelka2 as "positive" or agreed by both methods
572 as "negative" were selected from SimData2 with the matched number of positive and negative
573 labels compared to BioData only. 4] BioData+SimData for SimData1: 40,808 variants half from
574 BioData half from SimData only for SimData1 were selected with the matched number of positive
575 and negative labels compared to BioData only. 5] BioData+SimData for SimData2: 40,808
576 variants half from BioData half from SimData only for SimData2 were selected with the matched
577 number of positive and negative labels compared to BioData only. Each network architecture
578 was trained using a batch size of 4 with a stochastic gradient descent (SGD) optimizer with a
579 learning rate of 0.01, and momentum of 0.9. Fifteen different epochs were trained on each of the
580 5 training sets described above, and the model after each epoch is saved for performance

581 evaluation. The training was conducted on NVIDIA GTX 980 GPU Nodes on San Diego
582 Supercomputer Center's Triton Shared Computing Cluster (TSCC). The training performance of
583 the models was further evaluated on BioData2, which has not been used for any of the training
584 procedures.

585 **Usage of DeepMosaic**

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

588 **Orthogonal validation with deep amplicon sequencing**

589 Deep amplicon sequencing analysis was applied to 241 variants from the 1355 candidates
590 detected by all 5 mosaic variant callers from the 200× WGS of 16 samples²⁷ to experimentally
591 confirm the validation rate of DeepMosaic as well as other methods. PCR products for
592 sequencing were designed with a target length of 160-190 bp with primers being at least 60 bp
593 from the base of interest. Primers were designed using the command-line tool of Primer3³³ with
594 a Python (v3.7.3) wrapper. PCR was performed according to standard procedures using GoTaq
595 Colorless Master Mix (Promega, M7832) on sperm, blood, and an unrelated control. Amplicons
596 were enzymatically cleaned with ExoI (NEB, M0293S) and SAP (NEB, M0371S) treatment.
597 Following normalization with the Qubit HS Kit (ThermoFisher Scientific, Q33231), amplification
598 products were processed according to the manufacturer's protocol with AMPure XP Beads
599 (Beckman Coulter, A63882) at a ratio of 1.2x. Library preparation was performed according to
600 the manufacturer's protocol using a Kapa Hyper Prep Kit (Kapa Biosystems, KK8501) and
601 barcoded independently with unique dual indexes (IDT for Illumina, 20022370). The libraries
602 were sequenced on a NovaSeq platform with 100 bp paired-end reads. Reads from deep
603 amplicon sequencing were mapped to the GRCH37d5 reference genome by BWA mem and
604 processed according to GATK (v3.8.2) best practices without removing PCR duplicates. Putative
605 mosaic sites were retrieved using SAMtools (v1.9) mpileup and pileup filtering scripts described
606 in previous TAS pipelines²⁷. Variants were considered positively validated for mosaicism if 1]
607 their lower 95% exact binomial CI boundary was above the upper 95% CI boundary of the control;
608 2] their AF was >0.5%. The number of reference and alternative alleles calculated from the
609 Amplicon validation was provided in Supplementary Table 1.

610 **Analysis of different categories of variants overlap with different genomic features**

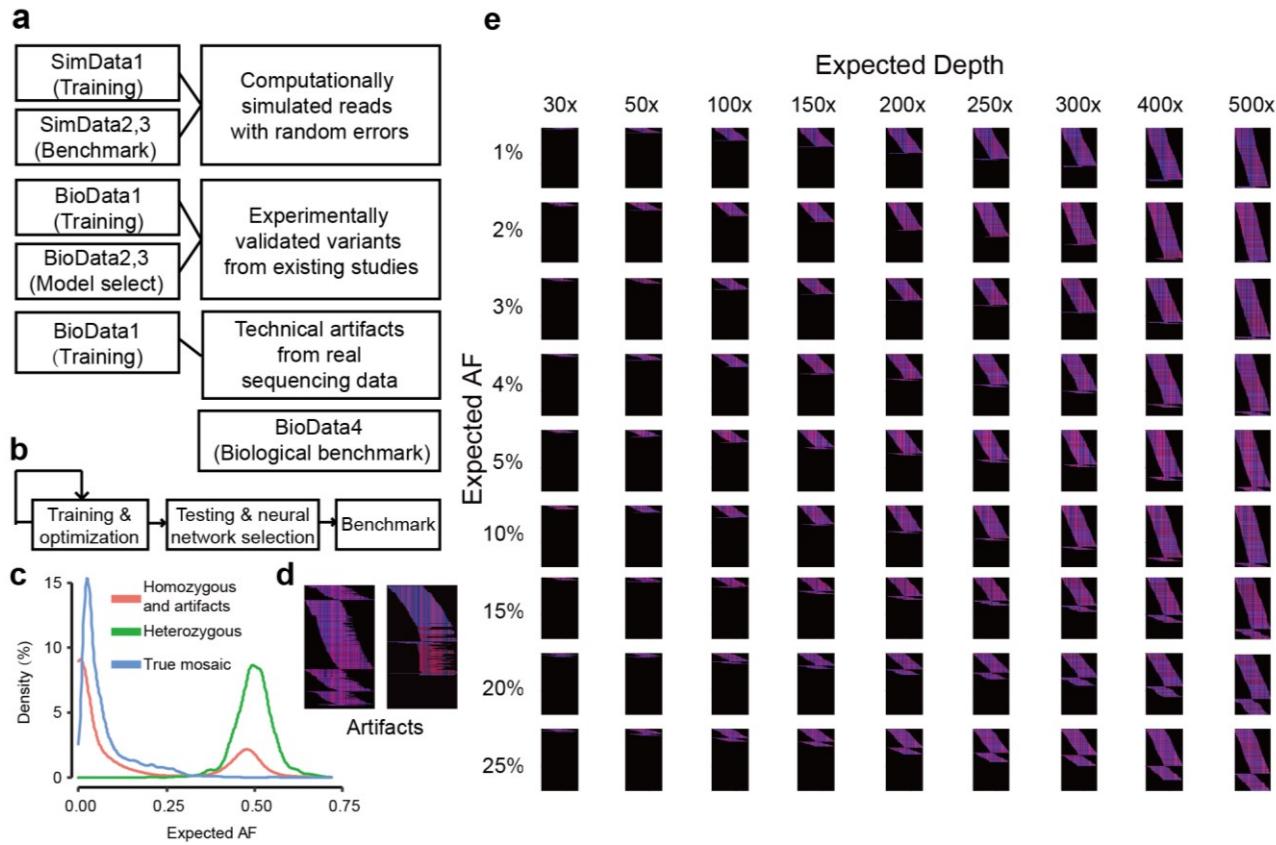
611 In order to assess the distribution of MVs and their overlap with genomic features, an equal
612 number of variants (mSNVs/INDELs as in group G1-G7 in Supplementary Fig. 6) was randomly
613 generated with the BEDtools (v2.27.1) shuffle command within the region from Strelka2 without
614 the subtracted regions (e.g. repeat regions). This process was repeated 10,000 times to generate
615 a distribution and their 95% CI. Observed and randomly subsampled variants were annotated
616 with whole-genome histone modifications data for H3k27ac, H3k27me3, H3k4me1, and
617 H3k4me3 from ENCODE v3 downloaded from the UCSC genome browser
(<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/>)—specifically for the overlap with

619 peaks called from the H1 human embryonic cell line (H1), as well as peaks merged from 10
620 different cell lines (Mrg; Gm12878, H1, Hmec, Hsmm, Huvec, K562, Nha, Nhek, and Nhlf). Gene
621 region, intronic, and exonic regions from NCBI RefSeqGene
622 (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz>); 10
623 Topoisomerase 2A/2B (Top2a/b) sensitive regions from ChIP-seq data (Samples: GSM2635602,
624 GSM2635603, GSM2635606, and GSM2635607); CpG islands: data from the UCSC genome
625 browser (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/>); genomic regions with
626 annotated early and late replication timing³⁴; high nucleosome occupancy tendency (>0.7 as
627 defined in the source, all values were extracted and merged) from GM12878; enhancer genomic
628 regions from the VISTA Enhancer Browser (<https://enhancer.lbl.gov/>); and DNase I
629 hypersensitive regions and transcription factor binding sites from Encode v3 tracks from the
630 UCSC genome browser (wgEncodeRegDnaseClusteredV3 and wgEncodeRegTfbsClusteredV3,
631 respectively).

632

633

Supplementary Information:



634

635

Supplementary Figure 1

Training strategies and examples of training data for DeepMosaic.

636

637

(a) More than 200,000 training and validation variants were generated for DeepMosaic, including computational simulations (SimData1), 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 biological data (BioData3 and BioData4) 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.

638

639

640

641

642

643

644

645

646

647

648

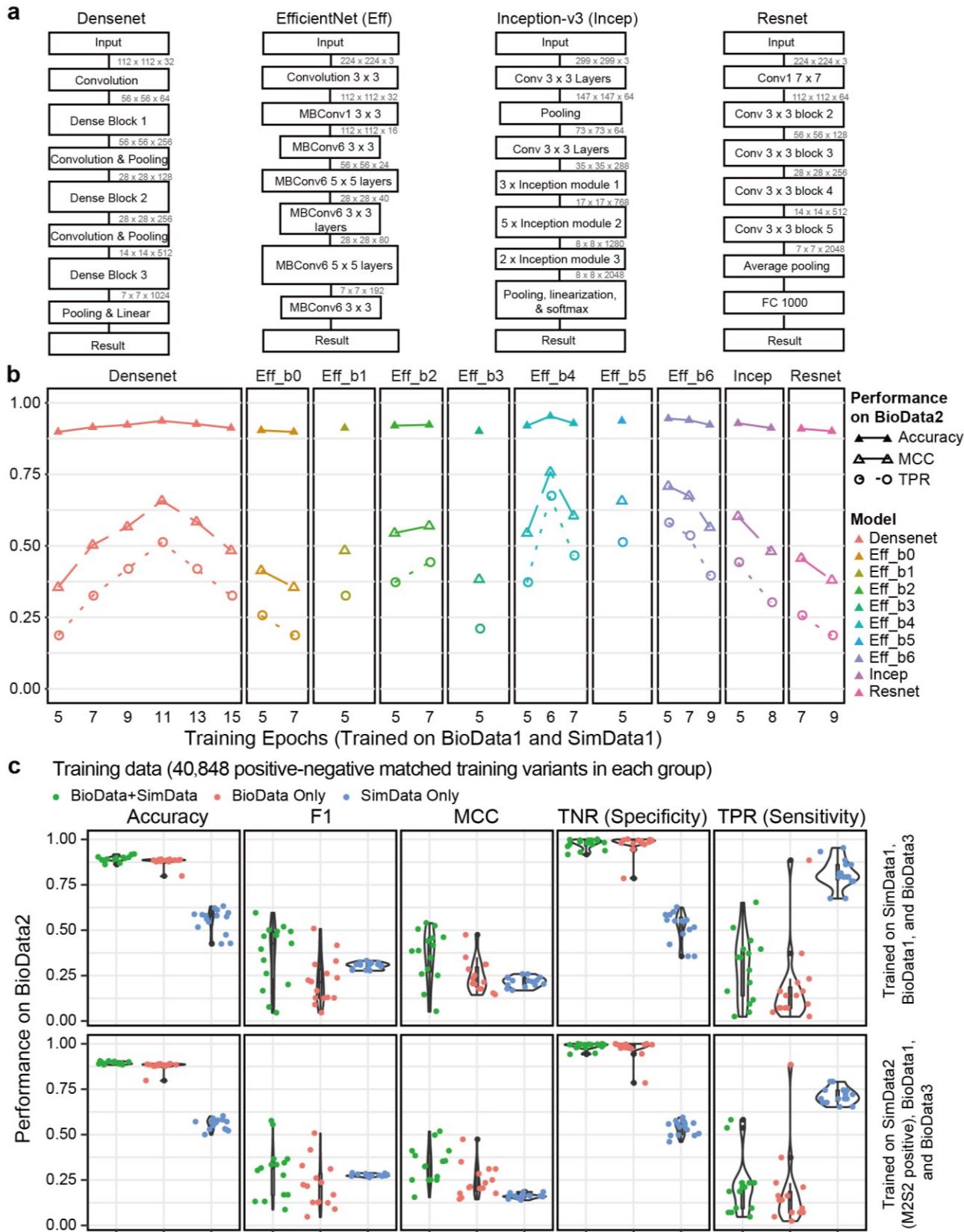
649

650

651

(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 30x, 50x, 100x, 150x, 200x, 250x, 300x, 400x and 500x, mutant allelic fractions (AFs) from

652 the simulated data that were set as 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25% and shown.

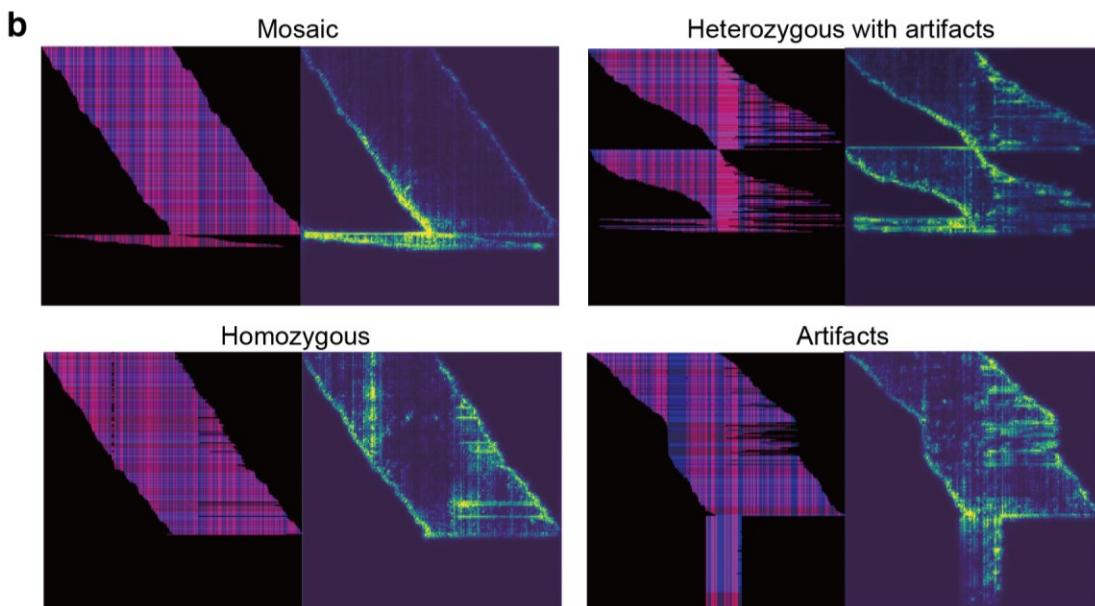
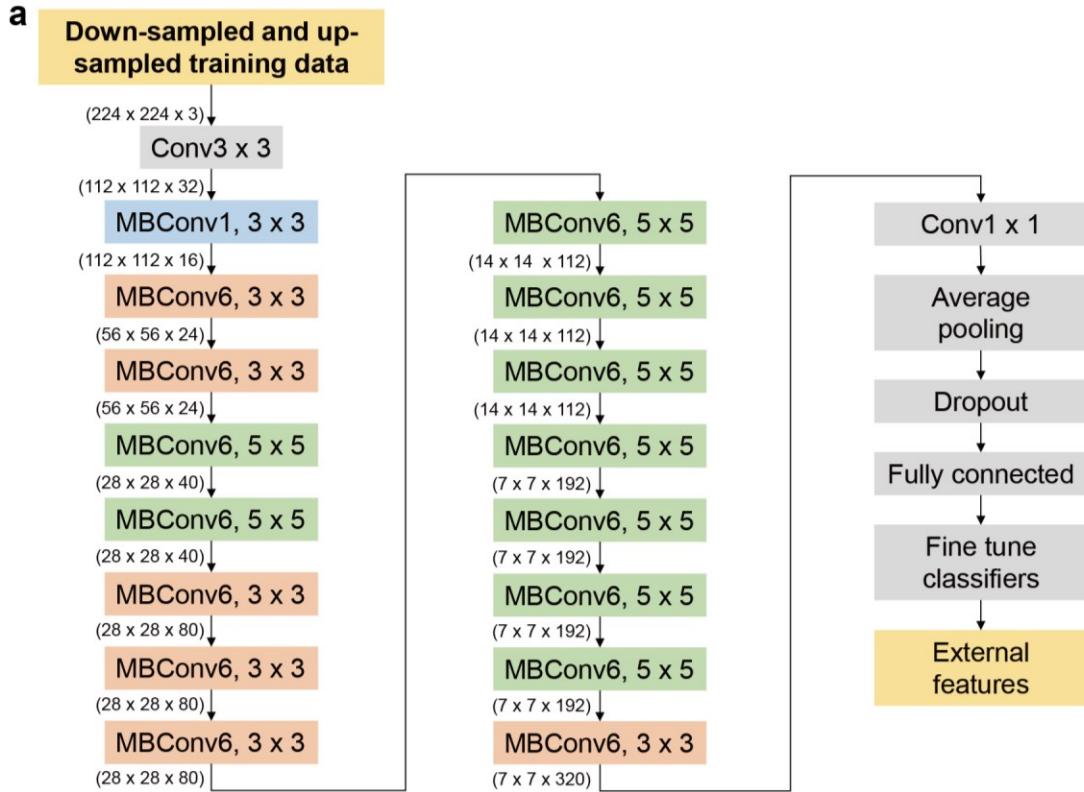


654 **Supplementary Figure 2**

655 **Network model selection based on independent gold-standard testing set.**

656 (a) Comparison of network structures implementing a variety of classification algorithms. For
657 different build versions of EfficientNet, only a general structure is shown. Inception v3 was used
658 in DeepVariant, and Resnet was used in NeuSomatic. (b) All models were trained on 180,000
659 training variants from BioData1 and SimData1 until the models reach training accuracy > 0.9.
660 Accuracy, Matthews's correlation coefficient (MCC), and Sensitivity of different network
661 structures trained with the same data with different epochs. EfficientNet-b4 trained at 6 epochs
662 demonstrated the highest Accuracy, MCC, and Sensitivity on the gold standard validation set²³
663 (BioData2); thus it was used as the default core model for DeepMosaic. We additionally provide
664 an option for experienced users to train their own models with self-labeled training data. (c)
665 EfficientNet-b4 models were trained on 5 additional datasets, each for 15 epochs. The training
666 datasets were generated with different compositions of biologically validated data and simulated
667 data. Models trained only on simulated data showed overall higher sensitivity but much lower
668 specificity on the gold standard evaluation set (BioData2) due to the high fraction of false-positive
669 calls. Models trained only on biological data showed similar overall performance compared with
670 models trained on a mixture of biological and simulated data. All three training sets are generated
671 with the same number of positive and negative data points as the biological data and with the
672 same number of total variants. M2S2 Positive: training variants were labeled positive by both
673 MuTect2 and Strelka2.

674



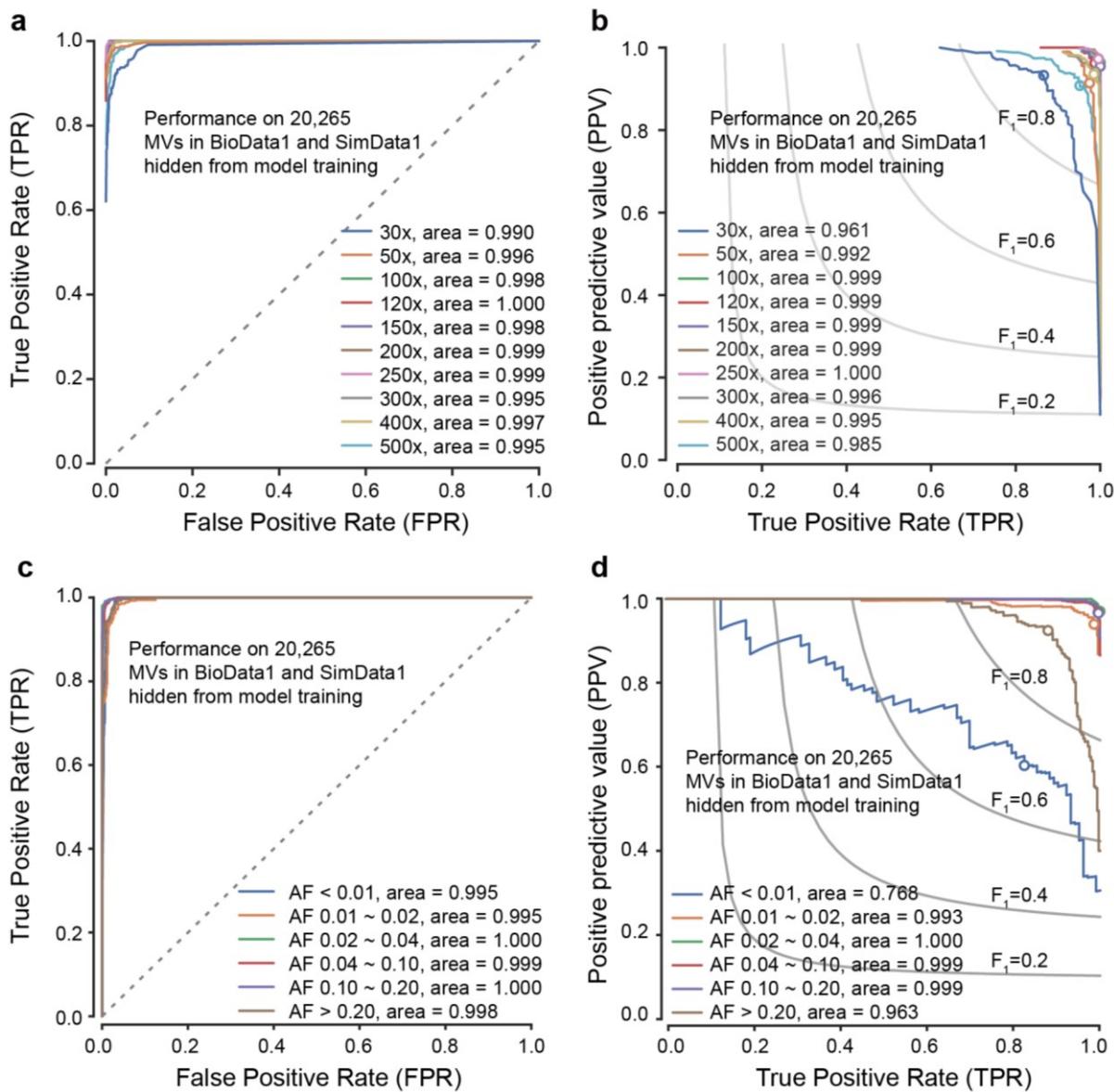
675

676 **Supplementary Figure 3**

677 The convolutional neural network of the DeepMosaic default model and gradient
678 visualization with guided backpropagation for the DeepMosaic default model
679 (**EfficientNet-b4**).

680 (a) Down- sampled and up-sampled image files coded from original BAM files were used as
681 input. 16 mobile convolutional layers were adapted from EfficientNet-b4, with optimized
682 parameter size and structures. Numbers represent the dimensions of trained hyperparameters.
683 (b) A mosaic, a homozygous, and a heterozygous variant with artifacts, as well as a technical
684 artifact, are shown here for the gradient visualization with guided backpropagation method²³
685 implemented for the DeepMosaic core model, EfficientNet-b4 trained at epoch 6, left: image
686 coding, right: gradient heatmap. The edges of bases, the sequence information, as well as
687 other high-dimensional information, are highlighted by the model.

688

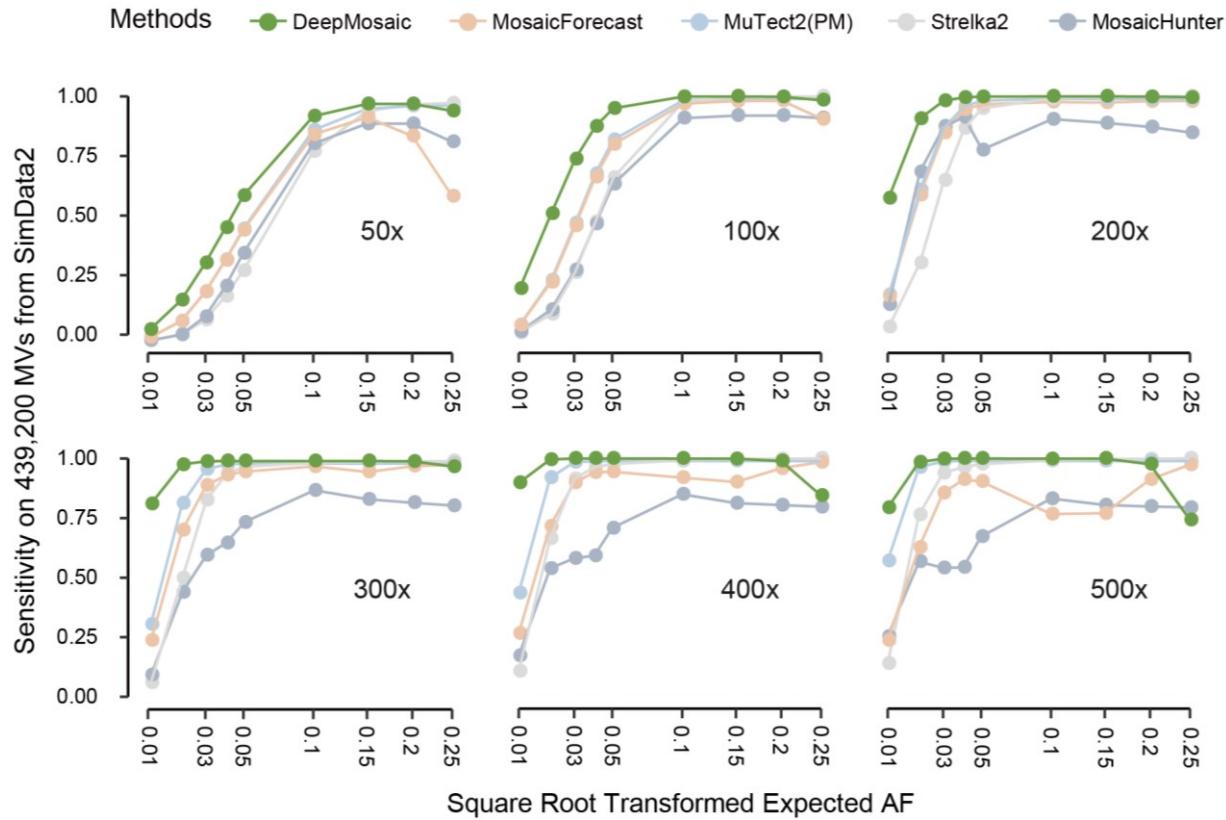


689

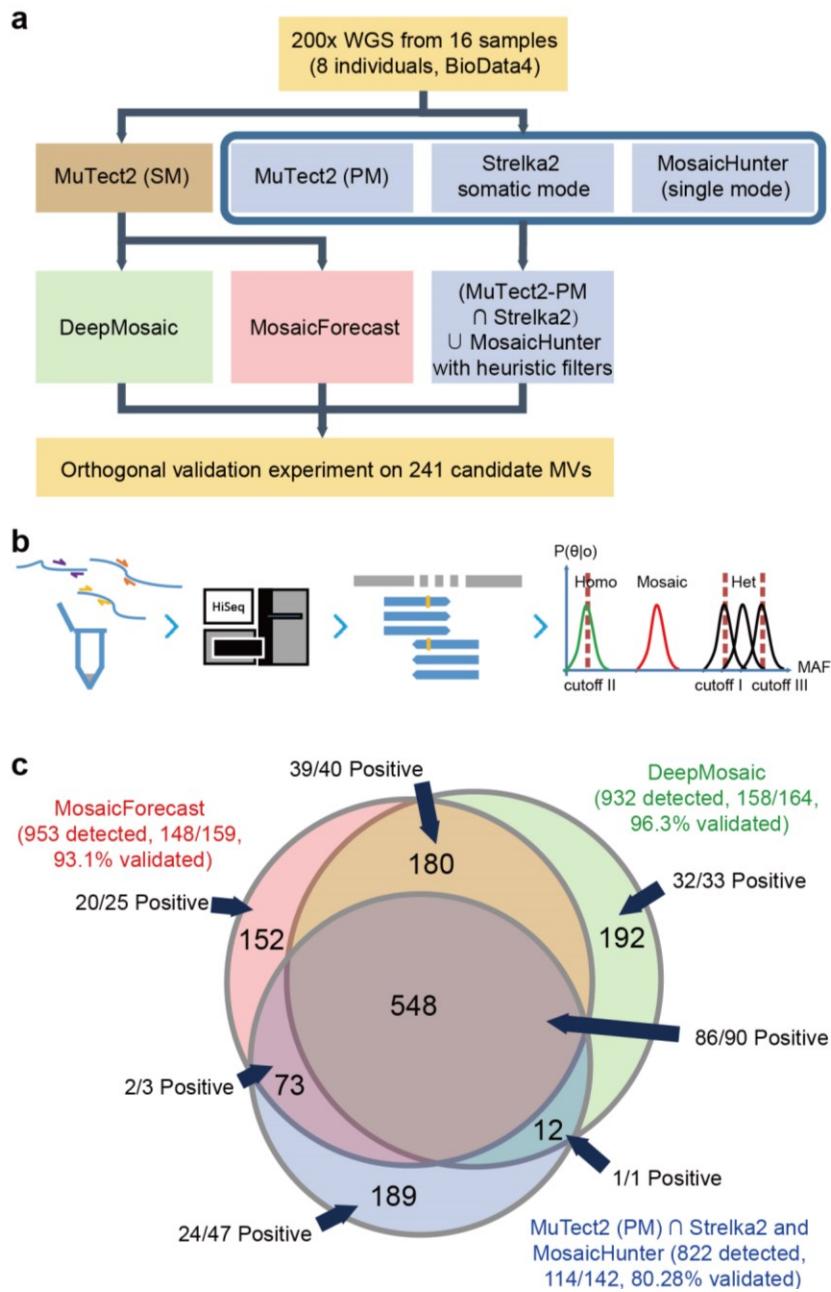
Supplementary Figure 4

Performance of DeepMosaic default model (EfficientNet-b4) on data hidden from training.

690 (a) Receiver operating characteristic (ROC) curve for DeepMosaic. True positive rates (TPR)
 691 and false-positive rates (FPR) were evaluated from 20,265 variants (BioData1 and SimData1)
 692 hidden from model training and model selection. Colors show groups of intended read depth. (b)
 693 Precision-recall curves for DeepMosaic, evaluated from the 20,265 hidden variants, dots showed
 694 the performance of the default parameters for DeepMosaic-CM. (c) ROC curve for DeepMosaic.
 695 TPR and FPR were evaluated from 20,265 variants (BioData1 and SimData1) hidden from model
 696 training and model selection. Colors show groups of bins of different expected AFs. (d) Precision-
 697 recall curves for DeepMosaic, evaluated from the 20,265 hidden variants, dots showed the
 698 performance of the default parameters for DeepMosaic-CM for different AF bins. Iso-F1 curves
 699 were shown for each precision-recall pairs with identical F1 scores labeled in (b) and (d).
 700



701
702 **Supplementary Figure 5**
703 **Performance of DeepMosaic and other mosaic variant callers on SimData2.**
704 Sensitivity of DeepMosaic and other mosaic callers on 439,200 independently simulated
705 benchmark variants (SimData2) at simulated read depths and AFs. DeepMosaic performed
706 equally well or better than other tested methods, especially at lower expected AFs.
707



708

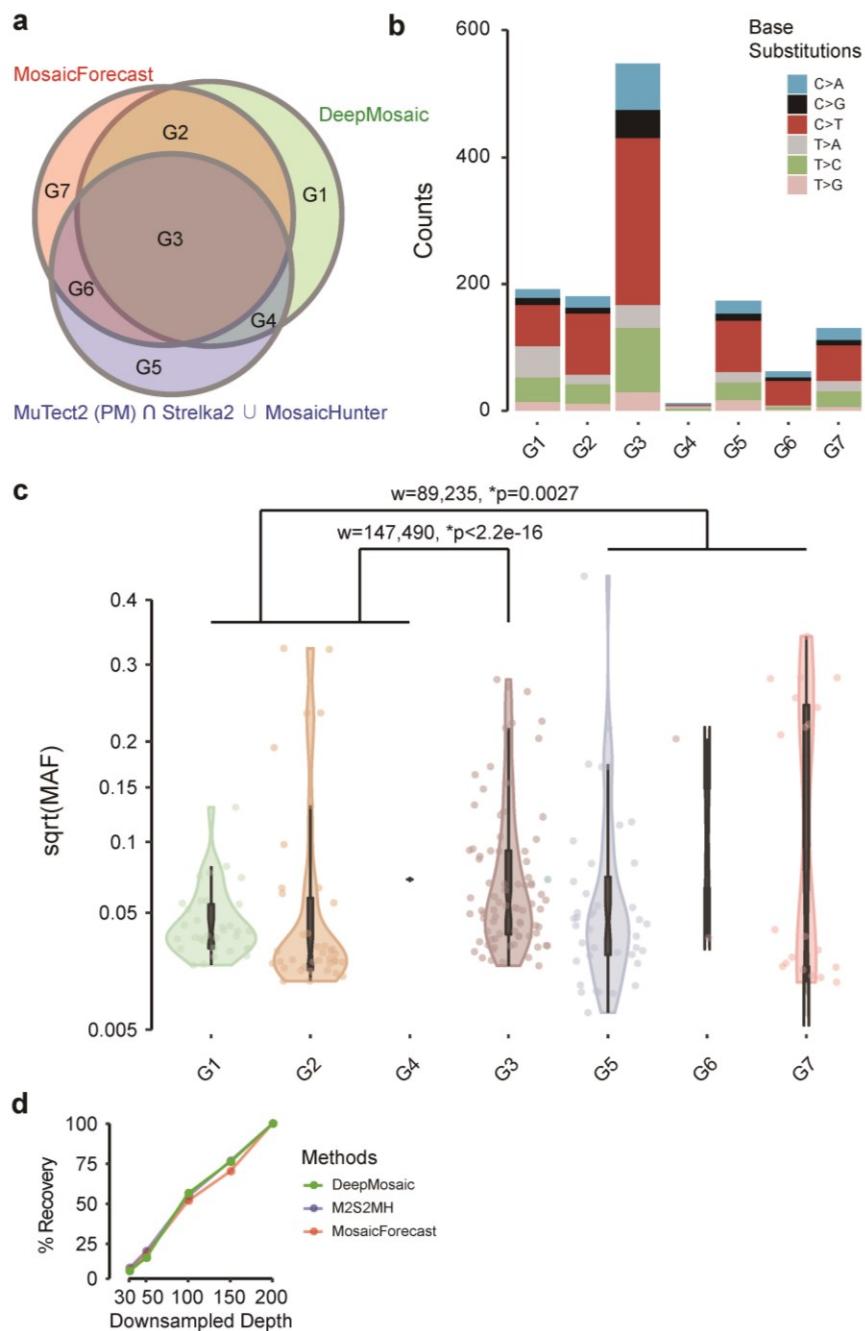
709 **Supplementary Figure 6**
710 **Comparison of DeepMosaic and traditional mosaic variant calling strategies on a**
711 **biological dataset (BioData4)**

712 (a) Compared with the mosaic variant calling strategy (M2S2MH) used in a previous publication,
713 DeepMosaic and MosaicForecast strategies are also listed. (b) Schematics for amplicon
714 validation. Primers were designed for different candidates and amplicons were collected for
715 Illumina sequencing. Information from aligned reads were calculated and genotypes were
716 determined. (c) Venn diagram of the experimentally validated results and the portions of variants
717 from different study strategies. DeepMosaic demonstrated a 96.3% (158/164) validation rate. Of

718 all the 932 variants identified by DeepMosaic, 39.91% (372/932) were missed by the MuTect2
719 Strelka2 MosaicHunter pipeline²⁷ with validation rate 97.26 (71/73) and 21.89% (204/932) were
720 missed by the MosaicForecast⁴ pipeline with validation rate 97.06 (33/34).

721

722



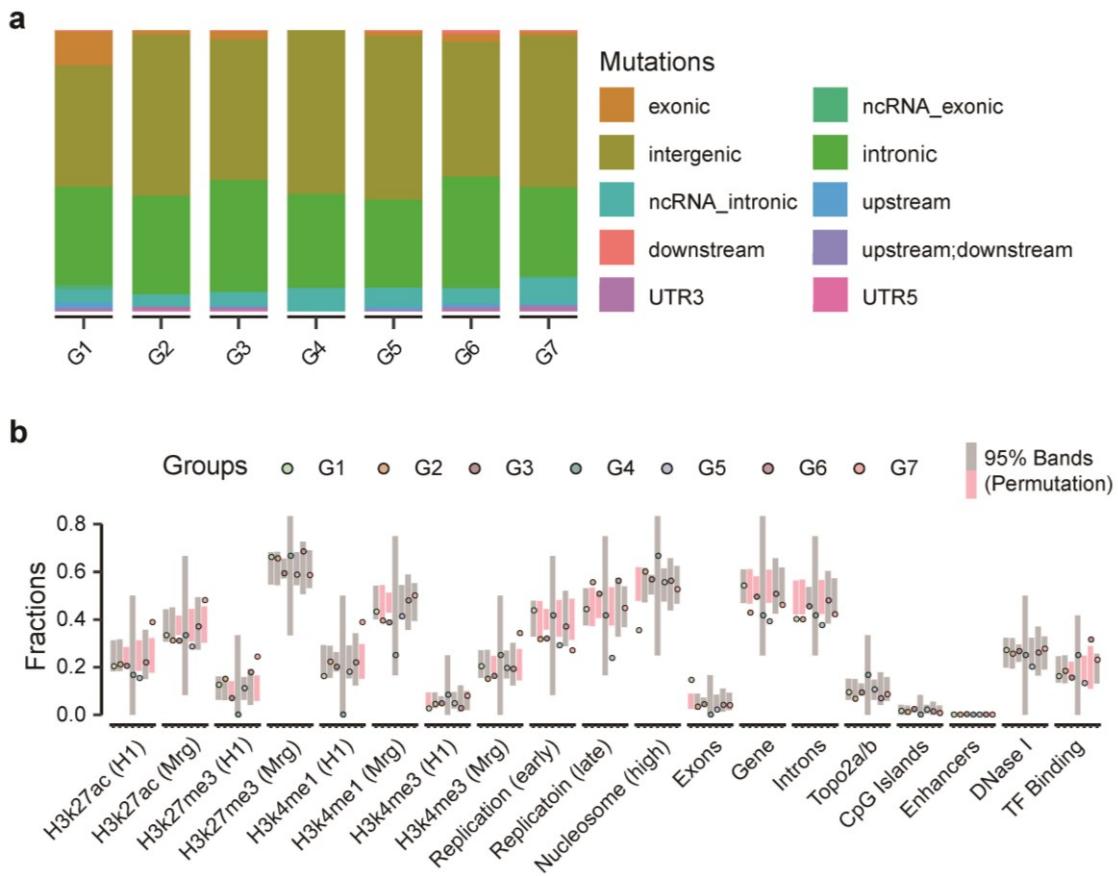
723

724 **Supplementary Figure 7**

725 **Comparison of features of variants called by DeepMosaic and other pipelines.**

726 **(a)** Different overlapping groups of variants detected by the 3 pipelines were separated into 7
 727 groups. **(b)** DeepMosaic-specific (G1) variants present similar base-substitution features
 728 compared with variants detected by the MuTect2-Strelka2-MosaicHunter combined pipeline as
 729 well as MosaicForecast pipeline (G2-G7). **(c)** Allelic fractions of the variants detected in the
 730 original WGS sample showed that DeepMosaic-specific variants (G1, G2, and G4) showed a

731 significantly lower average AF than variants detectable by all 3 pipelines (G3, $p<2.2e-16$ by a
732 two-tailed Wilcoxon rank sum test with continuity correction) and lower than variants detectable
733 only in other pipelines (G5, G6, and G7, $p=0.0027$ by a two-tailed Wilcoxon rank sum test with
734 continuity correction). **(d)** Recovery rate of DeepMosaic, M2S2MH, and MosaicForecast at
735 different depths from down sampling of BioData3. DeepMosaic showed a similar variant recovery
736 rate compared with M2S2MH and MosaicForecast, even when considering the lower AF variants
737 detected by DeepMosaic.
738



739

740

Supplementary Figure 8

741

Enrichment of genomic features for variants called by DeepMosaic and conventional 742 methods.

743

744 Variants called from different pipelines shared similar variant types and contributions. The
745 groups are defined the same as Supplementary Fig. 6a. The relative contribution of different
746 types of MVs is stable between different variant groups. **(b)** Enrichment analysis of variants in
747 different genomic features. Unlike the variants shared with other callers, DeepMosaic-specific
748 (G1) variants present depletion in high nucleosome occupancy regions. 10,000 permutation was
749 carried out on randomly selected gnomAD variants, significant comparisons are shown in pink.
750 Overall DeepMosaic-specific variants (G1) do not show significantly different genomic features
751 compared with permutation intervals.

752 **Supplementary Text: Members of the Brain Somatic Mosaicism Network**

753 Boston Children's Hospital: August Yue Huang, Alissa D'Gama, Caroline Dias, Christopher A.
754 Walsh, Javier Ganz, Michael Lodato, Michael Miller, Pengpeng Li, Rachel Rodin, Robert Hill,
755 Sara Bizzotto, Sattar Khoshkhoo, Zinan Zhou
756
757 Harvard University: Alice Lee, Alison Barton, Alon Galor, Chong Chu, Craig Bohrson, Doga
758 Gulhan, Eduardo Maury, Elaine Lim, Euncheon Lim, Giorgio Melloni, Isidro Cortes, Jake Lee,
759 Joe Luquette, Lixing Yang, Maxwell Sherman, Michael Coulter, Minseok Kwon, Peter J. Park,
760 Rebeca Borges-Monroy, Semin Lee, Sonia Kim, Soo Lee, Vinary Viswanadham, Yanmei Dou
761 Icahn School of Medicine at Mt. Sinai: Andrew J. Chess, Attila Jones, Chaggai Rosenbluh,
762 Schahram Akbarian
763
764 Kennedy Krieger Institute: Ben Langmead, Jeremy Thorpe, Jonathan Pevsner, Sean Cho
765
766 Lieber Institute for Brain Development: Andrew Jaffe, Apua Paquola, Daniel Weinberger,
767 Jennifer Erwin, Jooheon Shin, Michael McConnell, Richard Straub, Rujuta Narurkar
768
769 Mayo Clinic: Alexej Abyzov, Taejeong Bae, Yeongjun Jang, Yifan Wang
770
771 Sage Bionetworks: Cindy Molitor, Mette Peters
772
773 Salk Institute for Biological Studies: Fred H. Gage, Meiyang Wang, Patrick Reed, Sara Linker
774
775 Stanford University: Alexander Urban, Bo Zhou, Xiaowei Zhu
776
777 Universitat Pompeu Fabra: Aitor Serres Amero, David Juan, Inna Povolotskaya, Irene Lobon,
778 Manuel Solis Moruno, Raquel Garcia Perez, Tomas Marques-Bonet
779
780 University of Barcelona: Eduardo Soriano
781
782 University of California, Los Angeles: Gary Matherne
783
784 University of California, San Diego: Danny Antaki, Dan Averbuj, Eric Courchesne, Joseph
785 Gleeson, Laurel Ball, Martin Breuss, Subhojit Roy, Xiaoxu Yang
786
787 University of Michigan: Diane Flasch, Trenton Frisbie, Huiru Kopera, Jeffrey Kidd, John
788 Moldovan, John V. Moran, Kenneth Kwan, Ryan Mills, Sarah Emery, Weichen Zhou, Xuefang
789 Zhao
790
791 University of Virginia: Aakrosh Ratan

792

793 Yale University: Alexandre Jourdon, Flora M. Vaccarino, Liana Fasching, Nenad Sestan, Sirisha
794 Pochareddy, Soraya Scuderi

795

796 **Data availability**

797 WGS data used to generate the training set are available at the Sequence Read Archive (SRA),
798 Accession No. SRP028833 and SRP100797). The gold standard WGS data and validated
799 capstone project data are available at the National Institute of Mental Health Data Archive
800 (NIMH NDA Study ID 792 and 919: <https://dx.doi.org/10.15154/1504248>) and the Brain
801 Somatic Mosaicism Consortium Data Portal. Simulated data generated from NA24385 (HG002)
802 are available at <https://humanpangenome.org/hg002/>. The independent sperm and blood deep
803 WGS data are available at SRA (Accession No. PRJNA588332).

804 **Code availability**

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

807 **Acknowledgment**

808 The authors thank Dr. Yanmei Dou for helping to set up the MosaicForecast pipeline. We thank
809 Prof. Michael K. Gilson for the help in computational resources. We thank Profs. Peter J. Park,
810 Garrison W. Cottrell, John V. Moran, Melissa Gymrek, Drs. Patrick J. Reed, August Y. Huang,
811 and Si-Jin Cheng for their valuable comments and suggestions. This work was supported by the
812 National Institute of Mental Health (U01MH108898), Rady Children's Institute for Genomic
813 Medicine and Howard Hughes Medical Institute. The authors thank San Diego Supercomputer
814 Center (SDSC) (TG-IBN190021). This publication includes data generated at the UC San Diego
815 IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from
816 a National Institutes of Health SIG grant (#S10 OD026929).

817 **Contributions:**

818 X.Y., X.X., and J.G.G. conceived this project with input from M.B. and D.A.; X.Y. designed the
819 study workflow and managed the project. X.X. implemented the image representation and neural
820 network classifier under supervision and instruction by X.Y.; X.Y., C.L. and X.X. generated the
821 training data with the help from D.A., R.D.G., and L.W.; X.X. performed the training and model
822 selection under supervision by X.Y.; independent dataset were processed by M.B., D.A., and
823 R.D.G. under supervision by J.S. and J.G.G.; X.Y. and M.B. performed the validation
824 experiments with help from L.L.B. and C.C.; X.Y. and X.X. wrote the original and revised
825 manuscript with input from all listed authors; X.Y. and J.G.G. edited the manuscript. DeepMosaic

826 is benchmarked on part of the Brain Somatic Mosaicism Network (BSMN) common brain
827 experiment and common analysis pipeline for SNVs contributed by Y.W., T.B. under supervision
828 by A.A. and the BSMN capstone project contributed by M.B., X.Y., D.A., and X.X. under
829 supervision by J.G.G.; J.G.G. supervised this project. All authors discussed the results and
830 contributed to the final manuscript.

831 **Competing interests:**

832 The authors declare no competing interests.

833