Genome Biology

Comprehensive identification of somatic nucleotide variants in human brain tissue

Check for updates

Yifan Wang[1,2†], Taejeong Bae[3†], Jeremy Thorpe[4†], Maxwell A. Sherman[5,6†], Attila G. Jones[7,8†], Sean Cho[9,10,11], Kenneth Daily[12], Yanmei Dou[5], Javier Ganz[13,14,15], Alon Galor[5], Irene Lobon[16,17], Reenal Pattni[18,19], Chaggai Rosenbluh[7], Simone Tomasi[20], Livia Tomasini[20], Xiaoxu Yang[21,22], Bo Zhou[18,19], Schahram Akbarian[23,24], Laurel L. Ball[21,22], Sara Bizzotto[13,14,15], Sarah B. Emery[1], Ryan Doan[13,14,15], Liana Fasching[20], Yeongjun Jang[3], David Juan[16], Esther Lizano[16], Lovelace J. Luquette[5], John B. Moldovan[1], Rujuta Narurkar[25], Matthew T. Oetjens[1], Rachel E. Rodin[13,14,15], Shobana Sekar[3], Joo Heon Shin[25,26], Eduardo Soriano[17,27,28,29], Richard E. Straub[25], Weichen Zhou[2], Andrew Chess[7,8,23,30], Joseph G. Gleeson[21,22], Tomas Marquès-Bonet[16,31,32,33], Peter J. Park[5], Mette A. Peters[12], Jonathan Pevsner[9,10], Christopher A. Walsh[13,14,15], Daniel R. Weinberger[10,25,26,34,35], Brain Somatic Mosaicism Network, Flora M. Vaccarino[20,36], John V. Moran[1,37], Alexander E. Urban[18,19,38], Jeffrey M. Kidd[1,2], Ryan E. Mills[1,2] and Alexej Abyzov[3*]

* Correspondence: Abyzov.Alexej@mayo.edu
†Yifan Wang, Taejeong Bae, Jeremy Thorpe, Maxwell A. Sherman and Attila G. Jones contributed equally to this work.
3Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
Full list of author information is available at the end of the article

Abstract

Background: Post-zygotic mutations incurred during DNA replication, DNA repair, and other cellular processes lead to somatic mosaicism. Somatic mosaicism is an established cause of various diseases, including cancers. However, detecting mosaic variants in DNA from non-cancerous somatic tissues poses significant challenges, particularly if the variants only are present in a small fraction of cells.

Results: Here, the Brain Somatic Mosaicism Network conducts a coordinated, multi-institutional study to examine the ability of existing methods to detect simulated somatic single-nucleotide variants (SNVs) in DNA mixing experiments, generate multiple replicates of whole-genome sequencing data from the dorsolateral prefrontal cortex, other brain regions, dura mater, and dural fibroblasts of a single neurotypical individual, devise strategies to discover somatic SNVs, and apply various approaches to validate somatic SNVs. These efforts lead to the identification of 43 bona fide somatic SNVs that range in variant allele fractions from ~ 0.005 to ~ 0.28. Guided by these results, we devise best practices for calling mosaic SNVs from 250× whole-genome sequencing data in the accessible portion of the human genome that achieve 90% specificity and sensitivity. Finally, we demonstrate that analysis of multiple bulk DNA samples from a single individual allows the reconstruction of early developmental cell lineage trees.

Conclusions: This study provides a unified set of best practices to detect somatic SNVs in non-cancerous tissues. The data and methods are freely available to the scientific community and should serve as a guide to assess the contributions of somatic SNVs to neuropsychiatric diseases.

BMC

Introduction

Genomic sequence variants may be inherited vertically (i.e., transmitted through the germline) or generated after zygote formation (i.e., leading to somatic or gonadal mosaicism). It is well established that somatic mosaicism occurs in cells of phenotypically normal individuals [1–17] and can lead to various diseases [18]. However, the prevalence of somatic mosaicism and the extent to which it contributes to diseases outside of cancers requires elucidation [18].

Recent studies have estimated that each cell within the human brain contains hundreds to a few thousand somatic single-nucleotide variants (SNVs) and that a smaller fraction of cells harbor somatic copy number variations (CNVs) and mobile genetic element (i.e., retrotransposon) insertions [10, 15, 17, 19–22]. Dozens of somatic SNVs are present at high variant allele fractions (VAFs) across multiple tissues, indicating that they arose during early development [17, 23]. By comparison, some somatic SNVs are present at low VAFs and have limited tissue distributions, suggesting they arose later in development [15–17].

Single-cell DNA sequencing is the most direct approach to identify somatic variants. However, mutations introduced during the DNA amplification and/or generation of single-cell sequencing libraries, as well as non-uniform DNA amplification biases, make it difficult to discriminate bona fide mosaic SNVs from procedural artifacts [24]. Moreover, this approach for identifying mosaic SNVs requires sampling a large number of cells in a given individual and, consequently, is cost intensive.

Another approach to identify mosaic variants involves comparing bulk cell populations from two tissue samples derived from the same individual—the sample of interest and a control sample—as performed routinely during the analysis of cancer genomes. However, this approach is limited by the inability to define a proper control tissue because mosaic SNVs, particularly ones that arise during early development, are often present in multiple tissues across the body. Similarly, molecular barcoding approaches such as duplex sequencing can correct errors introduced by PCR amplification or sequencing and offer a > 10, 000-fold improvement of accuracy when compared to conventional WGS [25, 26]. However, the most accurate molecular consensus approaches require extremely high sequencing depth (1000× or higher) to ensure that each DNA molecule is sequenced multiple times, thereby effectively utilizing only a few percent of the generated reads for variant calling [27]. From a practical standpoint, this requirement restricts the main benefit of barcoding to targeted approaches. Thus, the development of a unified set of best practices to detect somatic SNVs from bulk whole-genome sequencing (WGS) datasets would provide an alternative, cost-effective approach to identify somatic SNVs.

In this study, members of the Brain Somatic Mosaicism Network (BSMN) conducted a coordinated, multi-institutional study that analyzed mosaicism in a single neurotypical brain sample and established unified standards for calling and validating mosaic SNVs from bulk WGS and WES data.
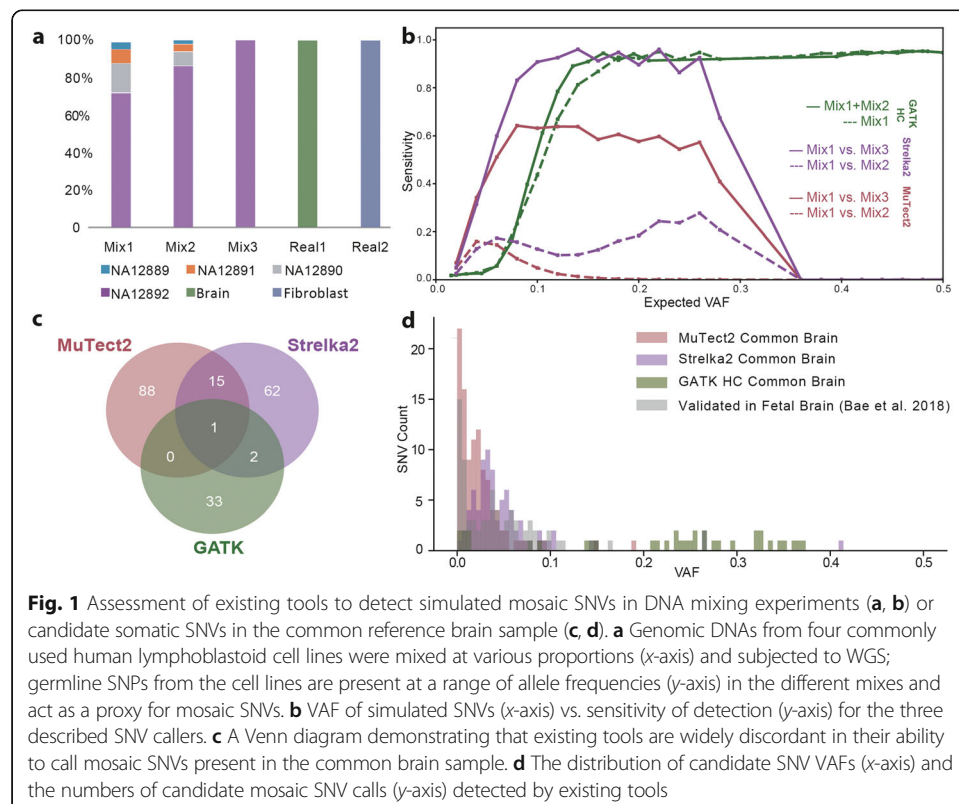
Results

The detection of simulated somatic SNVs in DNA mixing experiments

We first assessed the ability of three variant callers commonly used to detect germline single-nucleotide polymorphisms (SNPs) and somatic SNVs in cancers (i.e., the GATK

HaplotypeCaller [with a default value of 2 for ploidy] [28], MuTect2 [29], and Strelka2 [30]) to detect simulated somatic SNVs. We mixed genomic DNAs derived from transformed lymphoblastoid cell lines of four unrelated individuals at different proportions (Fig. 1a and Additional file 1: Fig. S1; see Mix 1 and Mix 2), sequenced the resultant mixtures to ~ 100× coverage, and assessed the ability of the callers to detect the simulated somatic SNVs (Fig. 1b and S2). In general, none of the callers were sensitive enough to reliably detect SNVs present at low (< 0.10) VAFs. Combining the Mix 1 and Mix 2 datasets to double sequencing coverage only marginally improved sensitivity (Fig. 1b and Additional file 1: Fig. S2). Moreover, MuTect2 and Strelka2, which are designed to detect somatic SNVs that are present in tumors but not matched normal samples, lacked the sensitivity to detect simulated mosaic SNVs shared between two matched samples (Fig. 1b and Additional file 1: Fig. S2). Indeed, given the mixing proportion and sequencing coverage, the p value of the difference in VAFs between Mix 1 and Mix 2 for a simulated somatic SNV is around 0.01. Thus, because there are ~ 4 million SNPs in an individual genome, the above p value is not sufficient to differentiate somatic SNVs from germline SNPs.

We next compared the performance of the three variant callers to detect putative somatic SNVs in ~ 250× coverage WGS data derived from postmortem dorsolateral prefrontal cortex (DLPFC; herein called the common reference brain; see below) and dural fibroblasts derived from a single neurotypical individual. We discarded variants having ~ 0.5 VAFs, as likely germline variants, and observed that the remaining putative somatic SNVs exhibited little overlap among the three callers (Fig. 1c). The VAF distributions identified using the variant callers also differed significantly from validated

Fig. 1 Assessment of existing tools to detect simulated mosaic SNVs in DNA mixing experiments (**a**, **b**) or candidate somatic SNVs in the common reference brain sample (**c**, **d**). **a** Genomic DNAs from four commonly used human lymphoblastoid cell lines were mixed at various proportions (x-axis) and subjected to WGS; germline SNPs from the cell lines are present at a range of allele frequencies (y-axis) in the different mixes and act as a proxy for mosaic SNVs. **b** VAF of simulated SNVs (x-axis) vs. sensitivity of detection (y-axis) for the three described SNV callers. **c** A Venn diagram demonstrating that existing tools are widely discordant in their ability to call mosaic SNVs present in the common brain sample. **d** The distribution of candidate SNV VAFs (x-axis) and the numbers of candidate mosaic SNV calls (y-axis) detected by existing tools

mosaic variants previously reported in fetal brain [17] (Fig. 1d). Thus, the naive application of existing callers appears to lack the sensitivity and precision necessary to robustly call somatic SNVs from bulk brain tissue, indicating a need for new approaches to address this challenge.
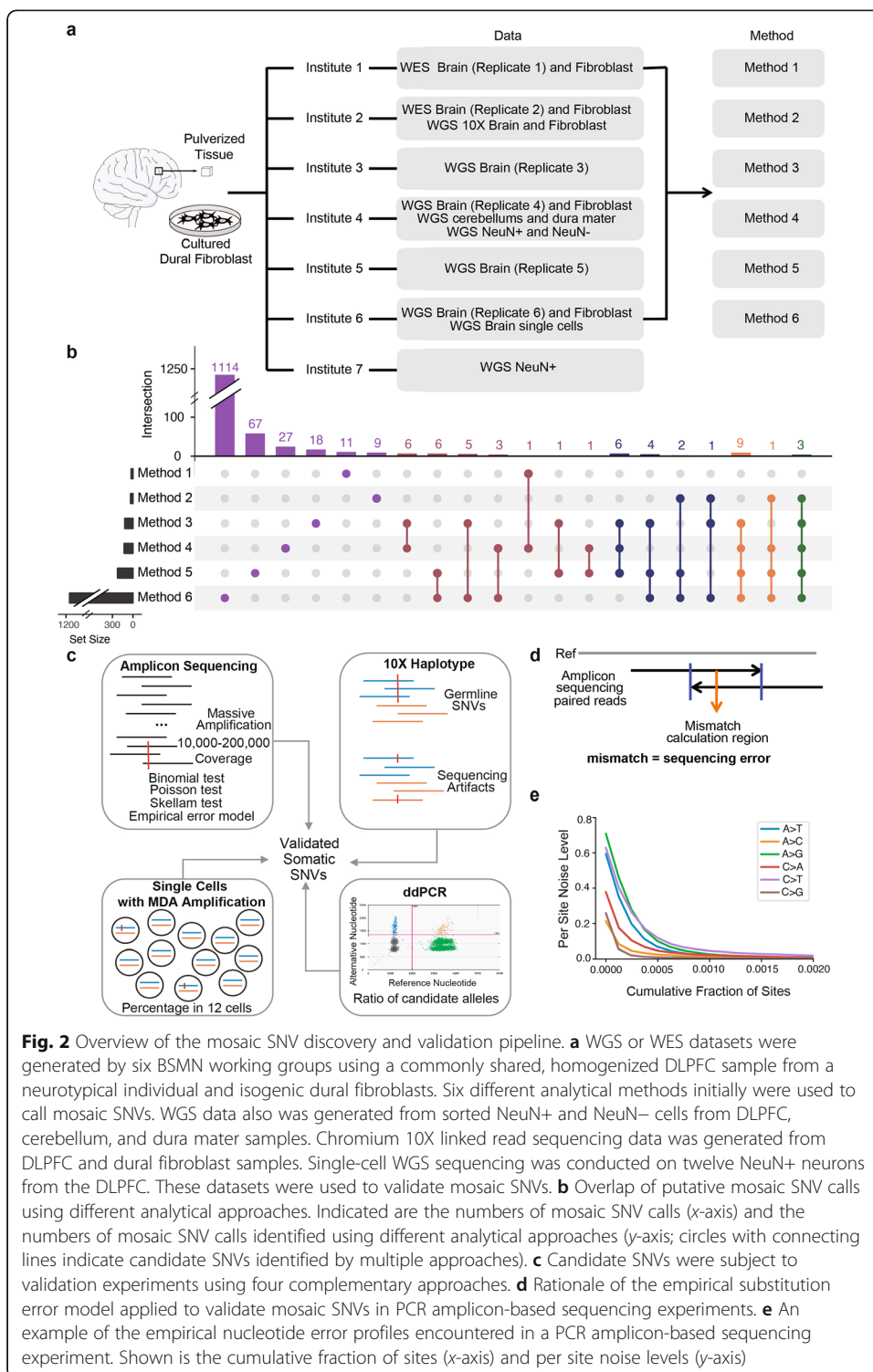
The generation of a somatic SNV call set from a neurotypical brain sample

To arrive at an unbiased approach to confidently identify somatic SNVs, we dissected a DLPFC sample from the reference brain of a neurotypical individual, pulverized it to enhance homogeneity, and distributed aliquots of this sample, as well as aliquots of a matched isogenic primary dural fibroblast cell line, to individual BSMN working groups. Each BSMN group independently performed DNA extractions and generated either whole-genome (four datasets ranging from ~ 85× to ~ 245× sequencing coverage) or whole-exome sequencing (two datasets of ~ 350× and ~ 435× sequencing coverage) data (Fig. 2a; replicates 1 through 6). The resultant reads were uniformly processed using a common workflow (see "Methods") and mapped to the human genome (version GRCh37d5) (Additional file 1: Fig. S3). Group-specific computational approaches and filtering strategies then were used to identify putative somatic SNVs (Fig. 2a; Methods 1 through 6).

The most widely used WGS technologies result in relatively short (~ 100–150 bps) sequence reads that are derived from the ends of small (typically of 350–450 bp) DNA fragments. As such, many reads do not unambiguously map to a single genomic locus. For example, it is difficult to confidently assign short DNA reads to evolutionarily young and/or human-specific retrotransposon-derived sequences (e.g., LINE-1 and Alu elements) [31, 32]. Similarly, segmental duplications sharing high sequence identity, as well as tandem repeat sequences present in centromeric, telomeric, and subtelomeric genomic regions, pose significant mapping challenges [33]. Thus, we applied the 1000 Genomes Project Strict Mask as a filter, which identifies positions in the human reference genome where sequence coverage does not significantly deviate from the expected average values across human populations. This mask defined ~ 73% of the human genome as accessible and each BSMN working group only called putative mosaic SNVs in the accessible fraction of the human genome.

The BSMN working groups used different analytical strategies to identify putative somatic SNVs (e.g., single sample calling, paired sample calling, or a combination of both strategies, as well as various filtering strategies; see "Methods"). Some groups opted for higher precision calling approaches, yielding fewer but presumably more precise mosaic SNV call sets. Other groups casted a wider net and opted to generate more inclusive but presumably less precise call sets. Consequently, the initial putative somatic SNV call sets were quite discordant and varied significantly among the working groups. For example, none of the putative SNV calls were identified by each of the six working groups and only three calls were identified by five of six groups (Fig. 2b and Additional file 1: Table S1).

We next assessed the sensitivity and specificity of the filters used by different BSMN working groups to generate individual call sets. First, we prioritized the 1114 putative mosaic SNV calls into four categories using the strategy outlined in Table 1. We reasoned that putative mosaic SNVs called by multiple groups using different approaches

Fig. 2 Overview of the mosaic SNV discovery and validation pipeline. **a** WGS or WES datasets were generated by six BSMN working groups using a commonly shared, homogenized DLPFC sample from a neurotypical individual and isogenic dural fibroblasts. Six different analytical methods initially were used to call mosaic SNVs. WGS data also was generated from sorted NeuN+ and NeuN– cells from DLPFC, cerebellum, and dura mater samples. Chromium 10X linked read sequencing data was generated from DLPFC and dural fibroblast samples. Single-cell WGS sequencing was conducted on twelve NeuN+ neurons from the DLPFC. These datasets were used to validate mosaic SNVs. **b** Overlap of putative mosaic SNV calls using different analytical approaches. Indicated are the numbers of mosaic SNV calls (x-axis) and the numbers of mosaic SNV calls identified using different analytical approaches (y-axis; circles with connecting lines indicate candidate SNVs identified by multiple approaches). **c** Candidate SNVs were subject to validation experiments using four complementary approaches. **d** Rationale of the empirical substitution error model applied to validate mosaic SNVs in PCR amplicon-based sequencing experiments. **e** An example of the empirical nucleotide error profiles encountered in a PCR amplicon-based sequencing experiment. Shown is the cumulative fraction of sites (x-axis) and per site noise levels (y-axis)

and from multiple data sources/replicates (Multi-call SNVs) would have the highest probability of yielding bona fide somatic SNVs. By comparison, SNVs identified using one approach from different data sources (Approach Singletons) and SNVs identified using multiple approaches from one data source (Data Source Singletons) might be

Table 1 Categories of validated mosaic SNVs

Categories	Definition	Number of mosaic SNVs selected for validation (out of total)	Number of validated mosaic SNVs
Multi-calls	Identified by multiple approaches, supportive evidence from multiple data sources	45 out of 45	33
Approach singletons	Identified by one approach, supportive evidence from multiple data sources	311 out of 1101	10
Data source singletons	Identified by multiple approaches, supportive evidence from one data source	4 out of 4	0
Absolute singletons	Identified by one approach, supportive evidence from one data source	40 out of 148	0

expected to have more false-positive calls when compared to Multi-call SNVs. Finally, SNVs identified using a single approach from a single data source (Absolute Singletons) might be expected to have the fewest number of true mosaic SNV calls.

We subjected 400 putative mosaic SNVs to validation experiments (Tables S1 & S2). The validation set included all the putative SNVs detected using two or more approaches (Data Source Singletons and Multi-calls; 49 calls), all of the calls detected individually using Methods 1–5 (Approach and Absolute Singletons; 124 calls), and a subset of randomly chosen calls detected individually by Method 6, which had the largest number of calls (Approach and Absolute Singletons; 227 calls). We divided the putative mosaic SNV calls into five sets of 100 variants; 20% of the SNV calls overlapped between the sets. The putative calls sets then were distributed to four BSMN working groups for PCR amplicon-based deep sequencing validation (Fig. 2c). Another group attempted to validate all 400 sites using a multiplex PCR-based targeted single-end resequencing assay (see "Methods"). Together, these strategies allowed testing of 272 putative SNVs by two independent working groups.

To aid in the validation of putative somatic SNVs, we conducted the following analyses: (i) whenever possible, we used the 3′ overlapping paired-end reads derived from each PCR amplicon to correct for sequencing errors (Fig. 2d), and (ii) we derived a nucleotide substitution error profile for each PCR amplicon-based sequencing library (Fig. 2e). To identify and eliminate systematic DNA amplification and sequencing artifacts, 4/5 groups also sequenced the same PCR validation amplicons from genomic DNA derived from the control NA12878 lymphoblastoid cell line [34]. This process yielded 233 concordant and 39 discordant putative validation calls. The majority (22/39) of discordant calls represented mosaic SNVs present at low VAFs (< 0.01).

We resolved the discordant mosaic SNV calls by prioritizing conclusions from amplicon-seq experiments with overlapping paired reads, which provided higher confidence base calls within mapping reads. Additionally, we considered orthogonal experimental evidence (i.e., whether the putative mosaic SNV was detected as an alternative read on a single haplotype in a ~ 70× coverage Chromium 10X linked read dataset and/or whether the putative mosaic SNV was detected in NeuN+ single-cell DNA sequencing datasets generated from the common reference sample) (Fig. 2c; see "Methods"). Finally, to verify the validation status, we performed droplet digital PCR (ddPCR) experiments for thirteen calls (Additional file 1: Table S3). These calls included a germline SNP within a heterozygous genomic duplication,

three false-positive calls, and nine previously validated mosaic SNVs. The ddPCR experiments demonstrated remarkable consistency with the VAFs determined using the PCR amplicon-based sequencing experiments, confirmed the predicted ~ 0.3 VAF for the germline SNP within a genomic duplication, invalidated the false-positive calls, and led to the re-validation of eight mosaic SNVs. One mosaic SNV validated by amplicon sequencing with a low VAF (0.0053) was not validated using ddPCR, highlighting potential limitations of PCR amplicon-based sequencing and/ or ddPCR to validate mosaic SNVs at very low VAFs.

In total, 43 mosaic SNVs were validated as bona fide somatic SNV calls. The validated mosaic SNVs were present at VAFs that ranged from ~ 0.005 to ~ 0.28. None of the mosaic SNVs occurred within coding exons, one (chr19:9,493,288 G>A) was present in a non-coding exon (i.e., 3′-UTR of *ZNF177*), and one intronic variant (chr8:103,281, 483 C>T) also had supportive reads in WES datasets, which only effectively survey 1 to 2% of the human genome. Moreover, consistent with previous data [17], the SNV mutation spectrum was dominated by C>T transitions (Additional file 1: Fig. S4). The filtering strategies, sources of false-positive calls, and caveats when conducting variant discovery and validation experiments are discussed in greater detail below.

Sources of false-positive SNV calls and filtering strategies

We sought to understand properties that distinguish bona fide somatic SNV calls from false-positive calls. Additional criteria used to identify and eliminate false-positive calls included (i) assessing the base quality score of reads supporting the candidate SNV; (ii) identifying imbalances in candidate SNV counts in forward vs. reverse sequencing reads; (iii) examining local assemblies of candidate somatic SNVs present within or near polymorphic insertion/deletion mutations (indels), homopolymeric tracts, and structural variants; and (iv) determining whether the candidate SNV arose on a single parental haplotype. Below we briefly summarize the sources of false-positive calls discovered in our study.

Germline SNPs

The PCR amplicon-based validation experiments coupled with the availability of Chromium 10X linked read data, single-cell DNA sequencing datasets, and germline copy number assessments in the common reference brain sample helped identify putative mosaic SNVs present at VAFs of < 0.5 that ultimately were determined to be germline SNPs. For example, our initial analyses identified 29 putative mosaic SNVs that appeared at < 0.5 VAFs; however, PCR amplicon-based sequencing validation demonstrated that these variants were present at ~ 0.5 VAFs. Consistently, Chromium 10X linked read data demonstrated that 18/29 variants were present on a single parental haplotype ("2 haplotype" calls) and 25/29 variants were present in at least six of twelve single-cell sequencing datasets (Fig. 3). Finally, 14/29 putative mosaic SNVs were located within regions of the genome containing germline copy number gains, leading to < 0.5 VAFs (Additional file 1: Fig. S5). This latter result highlights the importance of performing copy number characterization on experimental samples prior to calling mosaic SNVs.

Fig. 3 Summary of validation results for 400 candidate mosaic SNVs. Vertical lines represent candidate mosaic SNVs. Shaded rectangles to the right of the figure provide the keys to interpret the shading presented for each candidate SNV. There was concordance in true-positive mosaic SNV calls (PASS; green rectangle at bottom of figure) in multiple datasets and secondary validation experiments. Chromium linked read haplotype phasing and single-cell sequencing datasets also were effective in supporting a subset of bona fide mosaic SNV calls. By comparison, the VAFs of false-positive calls (red rectangle) are inconsistent across different datasets and often occur within or near insertion/deletion (indel) mutations, short tandem repeat sequences (STRs), homopolymeric nucleotide stretches, or copy number variants (CNVs). Importantly, the panel of normal (PON) filter, but not the comparison to WGS data from a control sample (i.e., to NA12878), was highly effective at identifying contaminating false-positive SNV calls (orange rectangle) and germline SNPs (gray rectangle). We lacked sufficient data to evaluate a subset of candidate SNVs (purple rectangle, NED—not enough data). The two green triangles at the top of the figure denote mosaic SNVs that validation experiments deemed to be false-positive calls; however, cell lineage analyses demonstrated that they are likely bona fide mosaic SNVs (see text and Fig. 4)

Sequencing and/or mapping artifacts

The largest proportion of false-positive calls (300/312) were represented by fewer than five supporting reads (i.e., low read count false-positive calls) in the 250× coverage replicate 6 WGS dataset, which had the highest coverage among all WGS datasets (Fig. 3). These false-positive calls likely represent DNA sequencing and/or mapping errors. Indeed, the use of Chromium 10X linked read datasets allowed us to unambiguously identify SNV calls that resulted from DNA sequencing artifacts. Twenty-four false-positive calls appeared as alternative reads on both parental haplotypes, leading to the appearance of four apparent haplotype states (Fig. 3; see "4 haplotype" calls) and are most simply explained by sequencing errors in reads originating from both parental haplotypes. By comparison, bona fide validated mosaic SNVs were detected as alternative reads on a single haplotype, giving the appearance of three haplotype states (i.e., the two parental haplotypes plus a "third" haplotype containing the somatic SNV on a single parental haplotype; Additional file 1: Fig. S6).

The effectiveness of applying Chromium 10X linked read datasets as a filter to evaluate bona fide somatic SNVs is limited by sequencing coverage. For example, at ~70× coverage, mosaic SNVs at < 0.03 VAFs were unlikely to have supporting alternative reads and therefore could not be assigned to a parental haplotype (Fig. 3; see "2 haplotype" for validated SNVs). Similarly, the relatively small number of single cells analyzed in this study limited the effectiveness in using these data to validate mosaic SNVs

present at low VAFs that were supported by independent PCR amplicon-based sequencing and/or ddPCR experiments (see below). Indeed, ~ 58% of validated and ~ 36% of false-positive SNVs lacked supporting alternative SNV reads in the single-cell datasets.

Genomic context

We noticed that candidate mosaic SNV reads located near or within polymorphic germline indels, short tandem repeats, or structural variants (which includes copy number variants and retrotransposon insertions) were the second largest source of false-positive calls. A systematic misalignment of these germline variants (Additional file 1: Fig. S7) and/or a systematic reduction in read counts near the germline variant (Additional file 1: Fig. S8) resulted in 75 false-positive calls (Additional file 2: Table S2). The systematic nature of these false-positive calls suggests they could be recurrently called as mosaic SNVs in unrelated samples. By extension, because sequencing errors and misalignments are known to be a problem in certain sequence contexts, some of the low read count false-positive calls and germline SNPs discussed above may also be due to systematic sequencing errors.

Unrelated samples and replicate samples

We next considered a filtering strategy that is based upon analyzing read support for a given mosaic SNV call in unrelated individuals. We searched for alternate reads containing the 400 candidate mosaic SNVs using ~ 30× coverage WGS data from the 2504 unrelated samples in the 1000 Genomes Project (ftp://ftp-trace.ncbi.nlm.nih.gov/1 000genomes/ftp/1000G_2504_high_coverage/). This analysis revealed that 183/400 calls were present at > 0.05 VAFs in more than five individuals and likely represent systematic false-positive calls or germline variants. Indeed, 17/183 were germline, 153/183 were false positives, and 13/183 require additional data to classify (i.e., not enough data [NED] Fig. 3). This panel of normal (PON) filter provided discriminative power to detect false-positive calls that likely arise from systematic technical errors encountered during Illumina sequencing and/or read misalignments in certain genomic contexts across the VAF spectrum. For example, we found instances where sequencing errors occurred within short tandem repeats (e.g., an A>T at chrX:150,725,216 in the sequence of 5′-CTC [A>T]CTCTCTCTCT; also see below). Given its importance, we calculated the PON mask across the entire human genome and have made these data available to the scientific community (https://bsmn.synapse.org).

Next, we searched for alternate reads in sequencing data derived from additional brain regions of the common reference sample. These datasets consisted of 75-240X WGS data derived from the cerebellum, neuronal, and non-neuronal cell fractions of the DLPFC, dura matter, and dural fibroblasts. Genotyping the 400 candidate variants using high stringency mapping quality (q30) and base quality (Q30) sequencing reads revealed consistent support for both germline SNPs and bona fide mosaic SNV assignments across multiple datasets (Fig. 3). Namely, all the germline and 35/43 validated somatic SNVs were present at > 0.01 VAFs in at least three replicate datasets, which is consistent with previous studies [17]. However, we observed that 9/12 false-positive calls with substantial read support (≥ 5) and 12/18 false-positive calls with a VAF > 0.02 passed this filtering criterion, but likely arose from systematic sequencing errors. Thus, sequencing multiple brain regions and different tissues from the common reference

sample is powered to confirm bona fide mosaic SNV calls, but does not replace the requirement of using the PON filter to remove systematic false-positive calls.

Best practices for calling mosaic SNVs

Consolidation of the above validation approaches led to the identification of 43 bona fide somatic SNVs. We next sought to derive a unified set of best practices to call somatic SNVs in WGS data. We found the input "discovery" call set has a significant impact on final sensitivity and determined that a GATK ploidy option of 50 for the 250× coverage WGS replicate 6 dataset, which corresponds to approximately five sequencing reads per haplotype, generally allowed the best approach to identify validated mosaic SNVs (Fig. 4d and Additional file 1: Fig. S9). We applied a number of filters to remove false-positive calls (see below). Another effective strategy for mosaic SNV detection is to apply a machine learning approach that unifies multiple read and genomic features, as well as read-based phasing information, into a predictive model such as MosaicForecast [35]. For our best practices framework, we incorporated MosaicForecast as an additional step to obtain the highest accuracy call sets (Additional file 1: Fig. S9).

Based on the above data, the BSMN has the following recommendations for calling somatic SNVs from WGS data derived from non-cancerous cells (Fig. 4a, see "Methods"): (i) call variants with GATK using a ploidy setting that corresponds to 20% of the overall sequencing coverage (e.g., a ploidy value of 50 for 250X WGS coverage); (ii) eliminate inaccessible genomic regions using the 1000 Genomes Strict Mask; (iii) discard germline variants that have a population allele frequency of > 0.001 (e.g., by comparing the candidate mosaic SNV calls to established catalogs of population
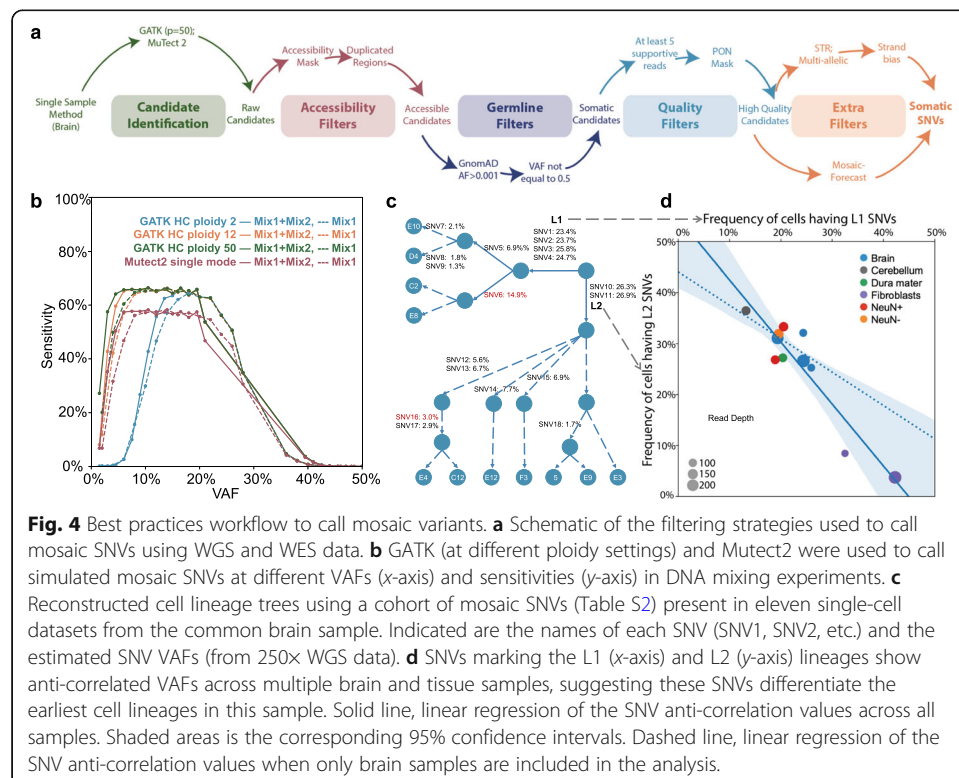
Fig. 4 Best practices workflow to call mosaic variants. **a** Schematic of the filtering strategies used to call mosaic SNVs using WGS and WES data. **b** GATK (at different ploidy settings) and Mutect2 were used to call simulated mosaic SNVs at different VAFs (*x*-axis) and sensitivities (*y*-axis) in DNA mixing experiments. **c** Reconstructed cell lineage trees using a cohort of mosaic SNVs (Table S2) present in eleven single-cell datasets from the common brain sample. Indicated are the names of each SNV (SNV1, SNV2, etc.) and the estimated SNV VAFs (from 250× WGS data). **d** SNVs marking the L1 (*x*-axis) and L2 (*y*-axis) lineages show anti-correlated VAFs across multiple brain and tissue samples, suggesting these SNVs differentiate the earliest cell lineages in this sample. Solid line, linear regression of the SNV anti-correlation values across all samples. Shaded areas is the corresponding 95% confidence intervals. Dashed line, linear regression of the SNV anti-correlation values when only brain samples are included in the analysis.

variations such as the Genome Aggregation Database [36]; (iv) eliminate calls having ~ 0.5 VAFs by binomial test significance with a p value cutoff above 10^{-6}; (v) carefully scrutinize calls in genomic regions exhibiting copy number gains and other structural variants; (vi) mandate candidate mosaic SNVs have at least five independent non-duplicated supporting reads that have minimum values of 20 for mapping and base quality; (vii) identify and eliminate false-positive calls using the PON filter based on the data from the 1000 Genomes Project; (viii) filter calls using available machine learning algorithms; (ix) optionally interrogate calls by applying filters that identify forward-reverse read imbalances and/or haplotype information (e.g., 10X Genomics linked read haplotype datasets); and (x) optionally validate candidate mosaic SNVs using additional wet-bench experimental approaches (e.g., PCR amplicon-based sequencing and/or ddPCR). This best practices workflow is freely available at: https://doi.org/10.5281/zenodo.4321679 and https://github.com/bsmn/bsmn-pipeline.

Applying the first seven steps of this best practices workflow to 250X WGS replicate 6 data from the common DFPLC brain sample resulted in the identification of 40 candidate mosaic SNVs with > 0.025 VAFs; 34 of these SNVs also were present in our original validation set. Two additional SNVs, which are absent from the set of 43 validated mosaic SNVs, also are likely true variants (discussed below). As an additional benchmarking metric, we used the best practices workflow to detect simulated SNVs in the DNA mixing experiment; however, because the simulated mosaic variants actually represent germline SNPs, we omitted the PON filter for this analysis. The sensitivity to detect simulated SNVs across the entire genome at VAFs ranging from ~ 0.02 to ~ 0.25 was ~ 65% (Fig. 4b) but increased to ~ 90% when we only considered simulated SNVs present in the accessible portion of the genome. The latter estimate is comparable to what we observed for a subset of 35 validated mosaic SNVs with > 0.02 VAF (34 validated out of 35 [~ 97%]).

Using somatic SNVs to recreate cell lineage trees

We next asked whether the identified mosaic SNVs could be used to infer developmental cell lineage trees. We first examined whether the union set of 49 SNVs identified in our validation call set and by the best practices workflow were shared among the twelve NeuN+ neuron single-cell datasets. One neuron (Additional file 1: Fig. S11, cell E2) was excluded from this analysis, as it lacked evidence for any of the 49 SNVs, perhaps because of inefficient whole-genome DNA amplification. The VAFs of the 49 SNVs also were estimated from the following WGS datasets: DLPFC (4 samples), NeuN+ (2 datasets), and NeuN− (1 dataset) DLPFC cell fractions, cerebellum (1 dataset), dura mater (1 dataset), and dural fibroblasts (2 datasets). Hierarchical clustering using the single-cell genotype and SNV VAF data allowed the reconstruction of a cell lineage tree (Fig. 4c; see "Methods"). Two lineages (L1 and L2) were clearly separable by diagnostic mosaic SNVs in the eleven single cells (Fig. 4c and Additional file 1: Fig. S11; compare SNVs 1–4, which define the L1 lineage, with SNV10 and SNV11, which define the L2 lineage). Moreover, a subset of SNVs could be used to infer L1 (e.g., SNV5, SNV6, SNV7) or L2 (e.g., SNV12, SNV14, SNV18) sub-lineages. Notably, the L1 and L2 lineages were evident in ectodermally derived brain tissue, ectodermally and mesodermally derived dura mater, and mesodermally derived dural fibroblasts, suggesting the L1 and

L2 lineages originated prior to gastrulation and diverged during early embryonic development, possibly after the first post-zygotic cleavage.

We next tested whether we could discriminate the presumed sister L1 and L2 lineages without relying upon the use of single-cell sequencing data. We reasoned that frequencies of SNVs defining the L1 and L2 lineages would covary across DNA samples derived from different tissues of the common reference sample. For example, if the fraction of cells from the L1 lineage predominates in one tissue sample, we should observe a corresponding decrease in the fraction of cells from the L2 lineage in that tissue sample and vice versa. Moreover, the sum of allelic frequencies of SNVs defining the L1 and L2 lineages should approach 0.5 VAF (which corresponds to a total cell frequency of 100%). We compared the VAFs of all 49 mosaic SNV pairs in nine brain region and two dural fibroblast samples sequenced to at least 75× coverage and then calculated how well the VAF sums of each SNV pair matches 0.5 across the WGS bulk samples (see "Methods"). The eight best scoring SNV pairs consisted of a SNV (SNVs 1–4) from the L1 lineage and a SNV from L2 lineage (SNV10 and SNV11; Additional file 1: Fig. S12). Although the VAFs of SNVs defining the L1 and L2 lineages differed among tissue samples, their sums consistently approached 0.5, which correspond to a cell fraction value of ~ 100% (Fig. 4c, d). Using a similar strategy, we also were able to identify anti-correlated mosaic SNV VAFs that defined L1 and L2 sub-lineages that likely arose later in development (Additional file 1: Fig. S12). Thus, this analysis suggested the possibility of discovering and analyzing mosaic SNVs across multiple bulk tissues to reconstruct early developmental cell lineages.

The recreation of cell lineage trees served as an independent means to gain support for bona fide mosaic SNV calls. For example, a somatic SNV defining a L1 lineage sub-branch (SNV6: chr7:112,461,481 G>T) was not present in our call set of 43 validated SNVs (even though it "passed" the Chromium 10X linked read haplotype test) because it had a very low VAF (< 0.0002) in PCR amplicon-based sequencing experiments. However, SNV6 exhibited a consistent VAF across multiple brain region datasets and no apparent alignment or sequencing artifacts surround the site. In addition, and consistent with the reconstructed lineage tree (Fig. 4c and Additional file 1: Fig. S11), the sum of the SNV5 and SNV6 VAFs for L1 sub-lineages was in general agreement with the VAF of the SNVs defining the L1 lineage across samples. Based on these analyses, SNV6 is likely a bona fide mosaic SNV. Using a similar strategy, we also concluded that a mosaic SNV defining a L2 sub-lineage (SNV16: chr6: 96,086,198 A>C) that was not in our original validation set also likely is a bona fide somatic SNV (see "Methods"). Thus, in total, we were able to validate 45 somatic mosaic SNVs in the common reference brain.

Discussion

The BSMN has described the successful implementation of a coordinated, multi-institutional approach to discover somatic SNVs in a neurotypical brain sample. These efforts led to the discovery of a catalog of high-confidence mosaic SNVs that are present at > 0.02 VAFs at ~ 65% genome-wide detection sensitivity; 42/43 validated somatic SNVs are present at > 0.02 VAFs in at least one WGS replicate. When adjusted for discovery sensitivity, this number of somatic SNVs is consistent with previous estimates that ~ 1.3 SNVs arise per cell division per cell during early development [17].

A major factor currently affecting the sensitivity of our approach towards discovering mosaic variants at a > 0.02 VAF is the necessity of restricting variant calling to the portion of the genome that can be assessed using short-read DNA sequencing technologies. The refinement of single-molecule long-read sequencing approaches (e.g., Pacific Biosciences and Oxford Nanopore technologies) should enable SNV calling in currently inaccessible genomic regions; however, the sequencing depth required for mosaic SNV discovery currently makes the application of these approaches prohibitively expensive.

In principle, the same workflow described above also can be applied to WES data. However, due to uneven coverage of sequence data across the genome, the detection sensitivity of mosaic variants can vary in WES data. Moreover, filtering out germline variants in duplicated genomic regions is difficult due to the lack of precise copy number estimates from WES data; thus, we mainly focused our analyses on WGS data.

The use of orthogonal approaches to interrogate somatic SNV calls was instrumental in our validation analyses. In general, these approaches (e.g., PCR amplicon-based sequencing, Chromium 10X linked read sequencing, the analysis of single-cell sequencing datasets, and ddPCR) were effective at identifying false-positive calls and providing validation support for bona fide somatic SNVs, but likely are not suitable for genome-wide de novo mosaic SNVs discovery. Moreover, each approach has limitations. For example, PCR amplicon-based sequencing initially led to the mis-assignment of a small number of bona fide somatic SNV calls as false positives. Similarly, the Chromium 10X linked read and single-cell sequencing data filters generally were only effective in supporting a small number of somatic SNV calls present at higher VAFs. Finally, although ddPCR proved to be a highly reliable validation approach, the effort and cost required for developing robust assays for each individual predicted mosaic SNV only allowed it to be applied to assess a small number of variants. It is noteworthy that distinguishing somatic SNVs present at low VAFs from false-positive calls (e.g., sequencing artifacts or calls arising from read misalignment errors) remains a formidable challenge.

Our analysis provided tantalizing suggestive evidence that the use of somatic SNVs discovered in bulk WGS sequencing data may ultimately provide a cost-effective approach to reconstruct early embryonic cell lineage trees when compared to more expensive single-cell sequencing and clonal analysis approaches. However, because we only analyzed a limited number of bulk DNA samples from a single individual, additional data will be needed to explore the potential advantages and limitations of tracing cell lineages from mosaic SNVs discovered in bulk WGS datasets.

In sum, the BSMN has developed an efficient approach to identify and validate somatic SNVs that should serve as a guide to ultimately assess the contributions of somatic SNVs to neuropsychiatric diseases. This study generated ~ 7 Tb genomic sequencing data that have been deposited into the NIMH Data Archive, which is freely available to the greater research community.

Methods

Data generation

Mixing experiment (Mount Sinai School of Medicine, contributed by Attila G. Jones)

Lymphoblastoid cell lines (LCLs) of four grandparents from CEPH/Utah pedigree 1463 were ordered from the Coriell Institute (catalog numbers: GM12889, GM12890,

GM12891, GM12892). Genomic DNAs (gDNA) from the cells were prepared, mixed in different proportions (as indicated in Fig. 1a), and subjected to WGS. Briefly, DNA libraries were prepared from 1 μg of mixed gDNA per sample using the Illumina PCR-free TruSeq DNA Library Prep kit according to the protocol provided by the manufacturer. The constructed libraries were quantified using the KAPA Library Quantification Kit and real-time PCR. Then, 150 bp paired-end reads were generated using the Illumina HiSeq X platform. The sequencing experiments were designed to yield ~ 200× coverage on each sample and were carried out at a GeneWiz facility (South Plainfield, NJ).

Sample processing (Lieber Institute for Brain Development, contributed by Daniel Weinberger)

The samples from the common reference brain numbered 5154 were dissected from deep frozen tissue from various brain regions in a standardized routine that uses a dental drill to minimize tissue injury and RNA degradation. To ensure that all distributed samples came from a singular source, the brain samples were first uniformly homogenized and then aliquoted for distribution to all but one BSMN node (Yale University and Institut de Biologia Evolutiva), which received a piece of frozen tissue. The disbursed tissue and related data were de-identified, delinked, and coded prior to sharing with the BSMN nodes. To obtain viable fibroblasts we collected a biopsy from the meninges (dura mater) of the same individual, suspended it in appropriate media (including DMEM (Gibco), HiFBS (Gibco), GlutaMAX (Gibco), Antibiotic/Antimycotic Solution (Gibco), MEM NEAA (Gibco), 2 βME (Gibco)), and then cultured it into a fibroblast line. The resultant fibroblast cell lines were serially expanded in T-25 and T-75 tissue culture flasks, harvested, re-suspended in media as a large homogenous culture, and subsequently were aliquoted into smaller tubes for storage/redistribution.

Whole-exome sequencing replicate 1 (University of Michigan, contributed by Sarah Emery and Yifan Wang)

Genomic DNA was extracted from 5154 brain tissue using the MagAttract HMW DNA Kit (Qiagen, Germantown, MD). The length of gDNA was determined by either standard gel electrophoresis in 0.4% agarose or pulse field gel electrophoresis in 1% agarose and $0.5 \times$ TBE for 16 h at 6 V/cm and a 1200 angle with an initial switch time of one second and a final switch time of six seconds. Whole-exome sequencing was performed on all extracted gDNAs. Duplicate libraries were made for each sample by shearing 75–200 ng of gDNA to 350 bp. Libraries were purified with 0.65× SPRIselect beads (Beckman Coulter) and quantitated using a Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific, Carlsbad, CA). A 50 ng aliquot of sheared DNA was saved for later use and the remaining 400–800 ng was used for exome target enrichment. Target enrichment was performed using SeqCap EZ Exome Probes v3.0 (Roche Sequencing Solutions, Pleasanton, CA) according to the protocol provided by the manufacturer with a 72 h incubation for hybridization and 12–16 cycles of post-capture ligation-mediated PCR (LM-PCR) to amplify the captured DNA. The quantity of the captured DNA was measured using a Qubit™ dsDNA HS Assay Kit and target enrichment was determined by calculating the abundance of control targets in post-capture

libraries relative to the abundance of these targets in the pre-capture libraries as outlined in SeqCap_EZ_UGuide_v5.4 (Roche Sequencing Solutions). Each library was sequenced on an individual HiSeq lane on the HiSeq X series. Library QC and sequencing was performed at Novogene Corporation (Davis, CA).

Whole-exome sequencing replicate 2 (University of California San Diego, contributed by Laurel Ball)

Pulverized brain cortex (0.98 g) and fibroblasts were provided by the Lieber Institute for Brain Development, (Baltimore, MD). Fibroblasts were cultured in fibroblast culture media containing MEM (Gibco), 20% FBS (Gibco), and 1× Pen/Strep (Gibco). Confluent fibroblasts were harvested and gDNA was extracted from pulverized brain and fibroblast samples using Qiagen Maxiprep kits according to the protocols provided by the manufacturer. Genomic DNA samples were prepared for whole-exome sequencing using the Agilent SureSelect XT Human All Exon v.5 kit. Then, 125 bp paired-end reads (median insert size ~ 210 bp) were generated using the Illumina HiSeq X 2500 platform. The sequencing experiments were designed to yield three datasets of ~ 100× coverage on each sample, sequencing completed at the New York Genome Center, NY.

Whole-genome sequencing replicate 3 (Yale University and Institut de Biologia Evolutiva, contributed by Irene Lobon)

A frozen piece of DLPFC from the 5154 common reference brain sample was provided by the Lieber Institute for Brain Development (Baltimore, MD). A scalpel was used to scrape off its surface and subdivide it into smaller pieces. Genomic DNA was extracted using the DNeasy Blood and Tissue kit (QIAGEN). The NEBNext Ultra II DNA Library Prep Kit for Illumina was used for library preparation. Then, 125 bp paired-end reads were generated using the Illumina HiSeq4000 platform. The sequencing experiments were designed to yield 90× coverage. Sequencing was carried out by Macrogen.

Whole-genome sequencing replicate 4 (Lieber Institute for Brain Development, contributed by Rujuta Narurkar and Joo Heon Shin)

Nuclei were isolated from pulverized postmortem brain tissue (0.7 g) using sucrose gradient centrifugation (4 °C, 25000 rpm, 1 h) and incubated with anti-NeuN-488 antibody (Millipore MAB377X, 1:1000) in a phosphate buffered saline (PBS) solution containing bovine serum albumin (BSA at 0.1%). Nuclei were sorted into NeuN+ and NeuN− fractions by FACS (BD FACSAria™III). Fibroblasts were cultured in fibroblast culture media (DMEM high glucose, FBS, L-glutamine, N.E. amino acids, Pen/Strep; all Invitrogen) and confluent fibroblasts were harvested for gDNA isolation. DNA was extracted from: (i) 0.3 g post-mortem bulk brain tissue; (ii) NeuN+ and NeuN− nuclear fractions; and (iii) fibroblast cell lines using the DNeasy Blood and Tissue Kit (QIAGEN) that included an RNAse A treatment step (QIAGEN) according to protocols provided by the manufacturer. For all samples, the libraries were prepared using the Illumina TruSeq DNA PCR-Free Library Prep Protocol with 1.5 μg of DNA from NeuN+ and NeuN− nuclear fractions and 3 μg of DNA from bulk cortex and fibroblast samples. Then, 150 bp paired-end reads were generated using the Illumina HiSeq X platform. The sequencing experiments were designed to yield 90× coverage. Library construction and

sequencing were conducted at Psomagen, previously known as Macrogen (Rockville, MD, USA).

Whole-genome sequencing replicate 5 (Harvard University, contributed by Javier Ganz)

Pulverized brain tissue was provided by Lieber Institute for Brain Development (Baltimore, MD). DNA was isolated from bulk brain tissue using lysis buffer from the QIAamp DNA Mini kit (Qiagen) followed by phenol chloroform extraction and isopropanol cleanup. DNA was then processed at Macrogen using the TruSeq DNA PCR-Free library preparation (Illumina) followed by a minimum of 30X sequencing of seven separate libraries on the Illumina HiSeq X Ten, for a total minimum coverage of 210×.

Whole-genome sequencing replicate 6 (Yale University, contributed by Liana Fasching)

Pulverized brain cortex (1.02 g) and fibroblasts were provided by the Lieber Institute for Brain Development (Baltimore, MD). Nuclei were isolated from pulverized postmortem brain tissue (0.7 g) using sucrose gradient centrifugation (4 °C, 25,000 rpm, 1 h) and incubated with anti-NeuN-488 antibody (Millipore MAB377X, 1:1000) in PBS and BSA. Nuclei were sorted into NeuN+ and NeuN− fractions by FACS (BD FACSAria™ III). Fibroblasts were cultured in fibroblast culture media (DMEM high glucose, FBS, L-glutamine, N.E. amino acids, Pen/Strep; all Invitrogen) and confluent fibroblasts were harvested for gDNA isolation. DNA was extracted from (i) 0.3 g post-mortem bulk brain tissue; (ii) NeuN+ and NeuN− nuclear fractions; and (iii) fibroblast cell lines using the DNeasy Blood and Tissue Kit (QIAGEN) that included an RNAse A treatment step (QIAGEN) according to protocols provided by the manufacturer. Illumina Truseq DNA PCR-free (350 bp insert) libraries were prepared for all samples using 1.5 μg of DNA from NeuN+ and NeuN− nuclear fractions and 3 μg of DNA from bulk cortex and fibroblasts. Then, 150 bp paired-end reads were generated using the Illumina platform. The experiments were designed to yield 30× sequence coverage for NeuN+ and NeuN− nuclear fractions and 210× sequence coverage for bulk cortex and fibroblast samples. Library construction and sequencing were conducted at Macrogen (Rockville, MD, USA).

Whole genome sequencing of single cells (Yale University and Stanford University, contributed by Liana Fasching, Livia Tomasini, and Bo Zhou)

Nuclei were isolated from fresh frozen pulverized postmortem brain tissue using sucrose gradient centrifugation (4 °C, 25000 rpm, 1 h) and incubated with anti-NeuN-488 antibody (Millipore MAB377X, 1:1000) in PBS and BSA. Ninety-five NeuN+ single nuclei were sorted into a 96-well plate by FACS (BD FACSAria™ III). Whole genome amplification was performed by multiple displacement amplification (MDA) using the REPLI-g Single Cell Kit (QIAGEN) according to protocols provided by the manufacturer. After amplification, each genomic DNA was purified using the DNeasy Blood & Tissue Kit (QIAGEN). Quality control was performed using a multiplex PCR for four arbitrary loci from different human chromosomes [12]. Amplified single nuclei were subsequently excluded if less than four loci were amplified. Sixty-seven out of ninety-five amplified nuclei (70.5%) passed the 4-locus multiplex PCR quality control. Low coverage sequencing was performed as a second quality control to assess locus dropout

rate. Paired-end barcoded WGS libraries were prepared for the 67 MDA reactions and sequenced on one lane of HiSeq 4000 (2 × 100 bp). Twelve out of the sixty-seven MDA reactions (18%) were selected for further sequencing. Illumina Truseq DNA PCR-free libraries were prepared for seven of the cells and sequenced on a HiSeq X (2X150 bp) at 30X coverage.

Samples of MDA amplified DNA from five cells were size selected for fragments > 10 kb on the BluePippin instrument (Sage Science, Beverly, MA, USA) using the S1 Maker 10 kb High Pass protocol provided by the manufacturer and then diluted to 1 ng/μl and used as input for the Chromium reagent delivery system [37, 38] from 10X Genomics (Pleasanton, CA, USA), where high molecular weight (HMW) DNA fragments are partitioned into > 1 million droplets, uniquely barcoded (16 bp) within each droplet, and subjected to random priming and isothermal amplification following standard manufacturer's protocol. Afterwards, the emulsion was broken, and the barcoded DNA molecules were released and converted to a Chromium 10X linked read library in which each library molecule retains its "HMW fragment barcode". Read-pairs generated in this manner (i.e., linked reads) that come from the same HMW DNA fragment can be identified by their "HMW fragment barcode" and subsequently used to phase variants onto megabase scale haplotypes. The final libraries (8 cycles of PCR amplification) were diluted to 5 nM and sent to Macrogen (Rockville, MD, USA) for sequencing (2 × 151 bp) on the Illumina HiSeq X.

Chromium linked read sequencing (University of Michigan, contributed by Sarah Emery and Yifan Wang)

Pulverized, frozen brain tissue and dural fibroblasts (FIBRO) from a deceased, neurotypical, male individual (5154) were received from the Lieber Institute for Brain Development (Baltimore, MD). Dural fibroblast were cultured in DMEM (Gibco/ Life Technologies) supplemented with 10% FBS (Gibco/Life Technologies), 2% Glutamax (Gibco/ Life Technologies), and 1% Antibiotic, Antimycotic (Gibco/Life Technologies). Cells were cultured for 3–10 weeks and passaged when they reached 85–95% confluence.

Genomic DNA (gDNA) was extracted from 5154 brain tissue and dural fibroblasts using the MagAttract HMW DNA Kit (Qiagen, Germantown, MD). The length of gDNA was determined by either standard gel electrophoresis in 0.4% agarose or pulse field gel electrophoresis in 1% agarose and 0.5 x TBE for 16 h at 6 V/cm and a 1200 angle with an initial switch time 1 s and a final switch time of 6 s. For 5154 brain gDNA, 5154 FIBRO, an aliquot containing 1–5 μg of gDNA was sent to HudsonAlpha Discovery (Huntsville, AL) for linked read sequencing using 10X Genomics technology (Pleasanton, CA). Long Ranger v2.2 (10X Genomics) was used to align reads and then call and phase SNPs to obtain haplotype information for each read.

Data uniform processing at Amazon Web Services (Mayo Clinic, contributed by Taejeong Bae)

All sequencing data generated by each BSMN node were uploaded to the shared S3 bucket of the Amazon Web Services (AWS). We uniformly processed the data in the AWS system to prepare aligned BAM files. We implemented a mapping workflow

similar to the best practices protocol suggested by GATK. The original FASTQ files, or those obtained from the conversion of BAM files, were split by flowcell lanes using an in-house awk script so that reads in each lane form the same read group in alignment. For each sample, the reads from FASTQ files were aligned to the human reference genome GRCh37d5 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz) using bwa (version 3.7.16a), sorted per each read group, and merged into a single BAM file with sambamba (version 0.6.7). The merged BAM files were marked for duplicate reads using PICARD (v2.12.1). Then, we performed indel realignment and base quality recalibration using GATK (v3.7–0), resulting in the final uniformed processed BAM files. This mapping workflow was installed and run in an AWS cluster prepared by AWS ParallelCluster (https://github.com/aws/aws-parallelcluster).

Calling variants in the reference brain

Method 1 (University of California in San Diego, contributed by Xiaoxu Yang)

Both tissue-specific and tissue-shared mosaic variants were called from the WES sequencing data. Brain- and fibroblast-specific variants were called using Mutect2 (GATK3.8) [28] and Strelka2 [30]; the BAM files from the brain sample (combined and non-combined from independent sequencing libraries) and fibroblast sample were treated as "tumor-normal" and "normal-tumor" pairs separately and cross-compared between each other. Variants called by both callers were listed. Mosaic variants shared between the brain and fibroblast samples were called using the single mode of MosaicHunter [39] by either combining all brain replicates or calling each separate sample. Variants that passed all the MosaicHunter filters also were listed. Mosaic candidates from the combined lists were further filtered using the following criteria: (i) the variant had more than 3 reads for the alternative allele; (ii) the variant was not present in UCSC repeat masker or segmental duplications; (iii) the variant was at least 2 bp away from a homopolymeric tract; and (iv) the variant exhibited a gnomAD allele frequency lower than 0.05. Variants that exist in the 1000 genome project (phase 3) also were excluded from the analysis. Variants from both exome data sources were tested and a combination of tissue-specific mosaic variants and tissue-shared mosaic variants were collected and the credible interval of VAFs were calculated using a Bayesian based method described previously [40].

Method 2 (Yale University and Institut de Biologia Evolutiva, contributed by Irene Lobon)

Somatic SNVs were called for each single sample using VarScan 2 [41] with relaxed parameters (–min-coverage 1 –min-reads2 1 –p value 1 and –min-var-freq 0.000001) to increase sensitivity. Calls were excluded if they overlapped the following somatic non-callable regions: (i) the 1000 Genomes strict mask non-pass positions; (ii) the WGAC segmental duplications track; (iii) the dbSNP common database; or (iv) positions with a mappability < 1 for the corresponding read length. A Fisher's exact test was used to filter out false-positive calls caused by random sequencing errors, and a binomial test across the samples was used to remove germline heterozygous variants. Regions with higher coverage were determined using a chi-squared test. Strand biased candidates were identified with a Poisson and a Fisher's exact test. Variants within 5 bp

of indels, near homopolymers, or within 300 bp of each other were also filtered during the analysis. Similarly, calls with a biased position in the reads for one of the alleles were also excluded from the analysis. Additionally, somatic SNVs were called with GATK 3.8 [28] Haplotype Caller (–ploidy 4) and filtered with the somatic non-callable, sequencing error, and binomial tests. The final call set consisted of mosaic SNVs passing all VarScan2 filters in at least three samples and all Haplotype Caller filters in at least one sample.

Method 3 (Harvard University, contributed by Alon Galor and Max Sherman)

To generate a list of high-confidence brain somatic variants for the common experiment sample, we considered three high-quality, high-depth Mixed-Brain samples generated from the replicate 6, replicate 4, and replicate 3 datasets. We produced putative somatic variant calls for each of these three samples with Mutect2 (v2.3.5) [29], using the highest-depth common experiment fibroblast sample available, generated by replicate 6, as a matched normal sample. Only variants flagged as "PASS" were retained for further analysis. We then leveraged Mutect2 to generate a Panel of Normals, using 75 200× WGS Brain samples. To construct this panel, we ran Mutect2 in its tumor-only mode on each of these 75 samples, making use of non-default parameters to produce more lenient calls. If a putative call from the common experiment sample was present at greater than a 1% allele fraction in two or more of the panel of normal samples, it was filtered out. Subsequently, we removed putative calls coinciding with Mutect2 tumor-only calls (again, using the same non-default parameters described above) of the remaining Mixed-Brain fibroblast samples. Next, we excluded putative somatic variants in the 1000G, ExAC, and ESP5600 common variant databases, as well as those located in inaccessible regions, according to criteria outlined by the 1000 Genome Project Strict Mask. We also removed putative somatic variants in indel, copy number variant, segmental duplication, and structural variation genomic regions. To generate indel calls, we employed the GATK Best Practices Germline Variant Calling Workflow. Variants in structural variant regions were defined as those in a 200-bp region, centered at the SNV, where the number of clipped reads was in the ≥ 95th percentile of 200-bp regions genome-wide. Variants were deemed to lie in a copy number variant or segmental duplication region if a 200-bp region, centered at the SNV, had GC-normalized depth in the ≥ 95th percentile of 200-bp regions genome-wide.

In order to prioritize variants, we further tiered them using evidence available for the calls in additional bulk brain tissue samples and single brain cells. In particular, we annotated putative SNVs observed on exactly one haplotype in phased bulk tissue, deemed significantly present in other Mixed-Brain samples by a joint-genotyping likelihood method, or present in single cells, as determined by an in-house method for analyzing single cells for evidence of mosaic events.

Method 4 (Kennedy Krieger Institute, contributed by Jeremy Thorpe)

Candidate somatic variants were identified leveraging high-depth whole-genome sequencing of four brain prefrontal cortex and cultured dural fibroblast samples. Paired brain fibroblast variant calling was performed with MuTect2 (v4.0.1.0) [29], Strelka2 (v2.8.4) [30], and MosaicHunter (v1.0) [39]. Additionally, single brain samples were

analyzed with the GATK (v3.8-0) HaplotypeCaller [28] with ploidy 5 parameter. Germline InDels and CNVs were called with the GATK Best Practices Germline Variant Calling Workflow and CNVkit (v1.1), respectively.

Quality filters were applied to mosaic SNV candidates. Variants were excluded if they (1) occurred in error-prone genomic intervals, as defined by the 1000 Genomes Strict Mask and/or UCSC RepeatMasker; (2) clustered within 1 kbp; (3) were within the 50 bp adjacent to InDels; (4) had a read depth > 2 standard deviations of mean coverage; (5) were likely germline events occurring as GnomAD single-nucleotide polymorphisms > 0.1% allele frequency; (6) occurred within a CNV; (7) exhibited a read strand bias using binomial test; (8) were heterozygous sites using binomial test; (9) failed Fisher's exact test for allele strand bias; (10) had minor allele frequency < 0.01 to exclude likely false-positive, low-frequency events. Following removal of low-quality sites, candidates were tiered using additional bulk and single-cell datasets.

We tiered candidates by leveraging biological replicates, bulk 10X sequencing, and single-cell data. Candidates were tiered on supporting evidence of the mosaic allele in multiple independent replicates, single-cell allele support, and concordance of mosaic alleles within a single haplotype by haplotype phasing of 10X data.

Method 5 (Mayo Clinic, contributed by Taejeong Bae)

For calling somatic variants, we applied a single sample approach using the GATK (3.7–0) Haplotype Caller [28] to each high-depth WGS dataset. To increase sensitivity towards calling somatic variants present at < 10% variant allele frequencies, we used ploidy options of 2 to 10 for GATK Haplotype Caller. We took only "PASS" calls from each raw call set. Those "PASS" SNVs were filtered out using common variant databases from the 1000 Genomes Project, ExAC, GnomAD, ESP5600, and Kaviar to exclude known germline variants. Then, we filtered out the SNV sites residing in the genomic region of non-P bases of the 1000 Genomes Project Strict Accessibility Mask to filter out calls in problematic genomic regions. Remaining heterozygous germline variants were excluded by applying binomial test ($P < 0.00001$), which identifies variants with 0.5 VAFs given the numbers of total reads and supporting reads. For checking strand bias of each candidate site, we used binomial test ($P < 0.05$) for evenness of the counts of both strand reads and Fisher's exact test ($P < 0.05$) to identify imbalance of strand ratios between reference (REF) and alternative (ALT) bases. We filtered out the sites having either bias. We excluded the sites having more than two alleles (multi-allelic sites). Finally, the germline CNV status for the +/− 1 kb region of each candidate site was checked using the genotype function of CNVnator. We removed the sites with > 2.5 estimated copy number as likely duplicated regions.

Method 6 (University of Michigan, contributed by Yifan Wang)

Candidate variants from paired brain and dural fibroblast samples were called using MuTect [29] and Strelka [30] with the default parameters. At the same time, candidate variants from single brain samples were called using the GATK Haplotype Caller with parameter of ploidy set to 5. Candidate mosaic SNV sites then were filtered out using multiple quality filters. We first filtered out variants that overlapped with repetitive regions or low mappability regions, including regions covered by UCSC RepeatMasker

simple repeats, Segmental Duplications, Simple Repeat tracks, regions not covered by 1000 Genomes Project Strict Accessibility Mask, and any variants not within ± 3 standard deviations of mean sequencing coverage. We excluded common variants in GnomAD with a population allele frequency larger than 0.1%. During the process of counting alleles at candidate positions, reads with mapping quality scores lower than the 90 percentile of control sites (sites with high confidence), with more than 3% mismatches, as well as candidate sites exhibiting a base quality score lower than 20 were excluded from further analysis. After accounting for different alleles at candidate positions, we excluded sites with a candidate allele frequency larger than 0.01 in the control sample NA12878 (from the Genome in a Bottle Project). We applied Fisher's exact test to exclude the sites whose alternative alleles are enriched on one strand compared to the other. We then filtered out sites with a known indel within ± 5 base pairs of the variant. In the end, we had an allele frequency cutoff at 0.03 to exclude the extremely low-frequency sites.

After removing the low-quality sites, we applied a binomial test with false-discovery protection using the Benjamini-Hochberg procedure [42] to filter out the heterozygous sites. We also used the haplotype information from the 10X linked read common reference brain sequencing dataset to further filter out false-positive sites. We additionally removed candidate sites located within 100 bp of each other.

Validation with targeted PCR amplicon-based sequencing approaches

Different PCR amplicon-based sequencing approaches were used to validate mosaic SNV candidates identified using the different computational pipelines. Each of the validation procedures is described below.

Amplicon 1 (Yale University, contributed by Liana Fasching and Simone Tomasi)

We validated the called SNVs by PCR amplicon-seq using following samples: (i) reference tissue—pulverized cortex DNA; (ii) reference tissue—fibroblast DNA; and (iii) control lymphoblastoid cell line DNA—NA12878. We determined the genomic location of each SNV using the Genome Browser. We designed each primer pair by selecting an 800-nucleotide-long DNA template surrounding the SNV (400 nucleotides up- and 399 nucleotides down-stream of each SNV) using Primer3 [43]. The specificity of primers was confirmed by UCSC in silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr) (Additional file 3:Table S5).

The PCR amplification was performed using the Phusion High-Fidelity DNA polymerase (Thermo Fisher). The optimal annealing temperature was predicted using the T_m calculator tool provided by the Thermo Fisher website. The presence and size of each amplicon was confirmed by gel electrophoresis (2% agarose gel). In samples where we detected multiple PCR products, DNA bands with the correct amplicon sizes were extracted via gel extraction and purification using 2% agarose precast gels (E-Gel EX, Invitrogen) and gel imager (E-Gel Safe Imager connected to E-Gel iBase, Invitrogen), according to the recommendations provided by the manufacturer. Amplified DNA fragments were purified using the QIAquick PCR Purification Kit (Qiagen). Samples were multiplexed and sequenced under following conditions: MiSeq paired-end, 300 bp, 3 samples/lane.

Amplicon 2 (University of Michigan, contributed by Sarah Emery and Yifan Wang)

Putative mosaic SNVs were validated by high-throughput sequencing of amplicons that contain the SNV and then calculating the relative abundance of reads containing the alternate allele. Primers were designed using Primer 3 software [43] with 300–400 bp of genomic sequencing surrounding the SNV as the input. Since the read length of the amplification product is ~ 300 bp, we were able to gain an overlapped region between the paired reads. With the overlapped regions containing the somatic SNV candidates, we were able to sequence each candidate site twice and increase the accuracy of sequencing result by excluding reads with non-concordant bases from the pair ended reads at the candidate SNV positions. If possible, SNPs known to be heterozygous in our samples were included in the amplified sequence used as input. Primers were tested in silico [44] to confirm they uniquely target the correct region of the genome. Phusion High-Fidelity DNA Polymerase (New England Biolabs) was used according to instructions provided by the manufacturer for amplification, and primers were cycled under varying conditions to determine optimal PCR mix and annealing temperature. To generate amplicons for sequencing, either NA12878 or gDNA from the 5154 common reference brain tissue was used as a template. The PCR product was purified with 0.7× SPRIselect beads (Beckman Coulter) and 10% of product was visualized on an agarose gel to confirm that only one amplicon of the correct size was present. If the size of the amplicon agreed with the size predicted from the primer design, we then sequenced the targeted amplified genomic fragment using an Illumina MiSeq sequencer. If the size was incorrect, we designed a second set of primers to obtain unique amplification. If none of the primers worked, the candidate was flagged with "primer not designed." Protocols and reagents from NEBNext Ultra™ DNA Library Prep Kit for Illumina (New England Biolabs) were used for end repair and dA-tailing and to ligate NextFlex adapters (Perkin Elmer, Waltham, MA) onto amplicons. After ligation, reactions were purified with 0.7× SPRIselect beads (Beckman Coulter) and PCR enrichment of adapter-ligated DNA was performed for 10 cycles using NEBNext Ultra™ DNA Library Prep Kit (New England Biolabs). Amplified libraries were purified with 0.7× SPRIselect beads and sequenced with MiSeq Reagent Kit v3, 600 cycle PE on MiSeq sequencer (Illumina, San Diego, CA).

Amplicon 3 (Mount Sinai School of Medicine, contributed by Chaggai Rosenbluh)

We sent all 400 target sites to Paragon Genomics and obtained a CleanPlex Custom NGS Panel allowing us to perform two multiplex PCR-based targeted resequencing assays for a total of 382 target sites. Eighteen of 400 sites were deemed not reliably targetable. The panels were designed using a proprietary primer design algorithm (sequences are available upon request) and proprietary background cleaning and molecular barcoding technologies to amplify the regions of interest and attach adapters for Illumina sequencing. The panel was iteratively optimized in silico prior to amplifying DNA from the common sample. The amplicons ranged in size from 250 to 425 bp (including the adapter sequences). A single MiSeq run provided ample coverage for all successfully targeted sites.

Amplicon 4 (Harvard University, contributed by Javier Ganz)

Targeted validation was attempted on 100 candidate mosaic variants and primers for 98 of those were successfully generated using BatchPrimer3 [45]. PCR primers were synthesized with flanking Ion Torrent adapters P and A. Each primer pair was unique and their sequence served as a barcode for variant identification. Phusion HotStart II DNA Polymerase (Thermo) CR was used for DNA amplification over 25 cycles following the manufacturer guidelines. PCR products were pooled and purified using AMPure XP technology (Agencourt) to be sequenced on the Ion Torrent S5 platform (Ion 530 chip). After demultiplexing and trimming, reads were mapped using BWA-MEM and locally realigned with GATK. Average coverage of mappable reads was 96,350× per variant. Mosaic fractions were identified using custom scripts based on SAMtools mpileup.

Amplicon 5 (Kennedy Krieger Institute, contributed by Jeremy Thorpe)

We validated candidate mosaic SNVs by targeted, high-coverage amplicon sequencing. PCR primers were designed with Primer3 [43] to generate amplicons from 300 to 500 bp and in silico verified to minimize the possibility off-target products. Amplicon sizes and purity were confirmed by PCR followed by agarose gel electrophoresis. Amplicons were quantitated, purified, and pooled for each of the NA12878 and 5154 brain tissue samples. NGS libraries were prepared using Illumina TruSeq Nano DNA kit and following standard Illumina NGS library preparation protocols. Libraries were sequenced with MiSeq Reagent Kit v3, 600 cycle PE on a MiSeq. Following demultiplexing and trimming, paired-end reads were merged with FLASH [46] and aligned with BWA-MEM to hs37d5 human reference genome. The average depth of sequencing was 103,335× per candidate. Target mosaicism was assessed by SAMtools mpileup and custom scripts.

Amplicon sequencing data analysis pipeline (University of Michigan, contributed by Yifan Wang)

We constructed a framework for uniform processing of PCR amplicon validation data. The PCR amplicon paired-end sequencing data was assembled to a single read using PEAR [47] with the non-concordant bases between the two reads set to N with a base quality of zero. We then aligned the reads to GRCh37d5 using BWA-MEM. Indel realignment was performed using the Genome Analysis Toolkit [28]. After preprocessing, we applied a series of filters to evaluate the putative mosaic SNV calls. We required a minimum of 200 reads covering the candidate mosaic SNV prior to making a "call" decision. Sites with fewer than 200 reads covering the candidate SNV were assigned the flag of "read not enough." We then compared the allele fractions of the candidate mosaic SNV in the amplicon data from common reference brain sample and from NA12878. Since the same mosaic SNV event is unlikely to occur in two different individuals, we reasoned that bona fide mosaic SNVs should only be present in the brain sample. We applied a hard cutoff that removed any site with a candidate allele fraction > 0.01 in NA12878 and additionally applied a Skellam test to compare the putative mosaic SNV allele frequencies in the two samples. Together, these criteria excluded likely false-positive mosaic SNV calls that arise from sequencing errors in

certain genomic contexts. We additionally established an empirical error model to exclude false-positive mosaic SNVs that likely arose during DNA amplification, library preparation, and/or sequence processing. We further evaluated PCR amplicon sequencing errors by assessing the mismatch rate (second allele frequency) in the overlapping region of the paired-end reads. We then used the 95th percentile of the mismatch rate distribution found for each type of base change as the sequencing error cutoff for base changes at the candidate positions. Finally, we excluded candidate mosaic SNVs with > 0.40 VAFs, as they likely represent germline SNPs.

We performed some additional final checks on the validated mosaic SNV events. First, we tested whether there was a positive correlation between the putative mosaic SNV VAFs in the WGS and PCR amplicon-based sequencing experiments (Fig. S13). Second, we tested whether there was a positive correlation between the ratios of the second highest allele frequency and the third highest allele frequency with the putative somatic SNV VAF in the WGS dataset (Fig. S13); the third highest allele frequency did not correlate with the variant allele frequency (Fig. S13).

Variant validation using single cells (Harvard University, contributed by Alon Galor and Max Sherman)

To distinguish a true mosaic variant from a multiple displacement amplification artifact in single-cell sequencing, we adapted the method of Dong et al. [24]. First, we augmented their kernel smoothing method to utilize phased genotypes [48] as this improves estimation of allelic imbalance at a given locus. Phasing was performed on the GATK HaplotypeCaller germline variants from replicate 6 bulk brain tissue using Eagle [49] and the 1000 Genomes Phase 3 reference panel [50]. Second, we determined the optimal kernel size to maximize accuracy using PaSD-qc [51], which was determined to be 10 kb. Third, we instituted a likelihood ratio test to distinguish a first-round amplification artifact from a true mosaic variant; let $f(X;p)$ be the probability of observing reads X where p is the expected proportion of alternate reads. Under the null hypothesis that the reads are generated due to amplification artifact, $p = 0.125$. Under the alternate hypothesis that the reads are generated due to a true mosaic variant, $p = 0.125 + \theta$ where p is the allelic balance estimated by kernel smoothing [24]. Since the models of the two hypotheses are nested, $-2 \times (\log f(X; 0.125 + \theta)/f(X; 0.125)) \sim X_1^2$. We applied this model at all 400 sites at which we attempted amplicon validation across the 12 deeply sequenced single cells to obtain high-quality single-cell genotypes. Scripts for running this model are available at https://github.com/parklab/SCGenotyper_mini.

Variant validation using 10X linked read data (University of Michigan, contributed by Yifan Wang and Jeffrey Kidd)

We used haplotype information provided by 10X Genomics linked read sequencing data to eliminate false-positive mosaic SNV calls and to provide support for bona fide mosaic SNVs. Briefly, the 10X Genomics linked read approach is a barcoded short-read sequencing technology, where individual HMW genomic DNA molecules (~ 50 kb) initially are attached to a bead then incorporated into droplets. Within each droplet, the HMW DNA

fragments are fragmented into small pieces, assigned a unique barcode, and then undergo amplification. The resultant DNA libraries from each individual droplet then were subject to standard Illumina short-read sequencing. Short reads containing the same barcodes were aligned and re-assembled into larger linked reads using the LongRanger pipeline (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger). Overlapping individual HMW fragments that shared informative germline SNP alleles used reconstruct longer, continuous haplotype segments. This results in the assignment of sequencing reads derived from the initial HMW fragments to individual haplotypes. Using this information, we identified and removed candidate mosaic SNVs that were found on both parental haplotypes, which likely arose from systematic sequencing errors. Moreover, the haplotype information could be used to identify candidate mosaic SNVs that were represented > 90% of the reads on a single parental haplotype and are likely germline SNPs. By comparison, bona fide somatic SNVs are detected as a minor allele on a single parental haplotype. Tools for analyzing linked read BAMs are available at https://github.com/KiddLab/hapfilter-10X.

Variant validation using ddPCR (Stanford University, contributed by Reenal Pattni)

We performed ddPCR using the primers and probe listed in Additional file 4: Table S6. We initially designed 29 ddPCR assays to assess candidate mosaic SNVs present at < 3% VAFs and/or near know SVs (Additional file 1: Table S3). This initial set included (i) 4 mosaic SNVs at higher VAFs (12–28% in amplicon sequencing); (ii) 5 mosaic variants at lower VAFs (1–2.5% in amplicon sequencing); (iii) 3 SNVs deemed to be germline SNVs; and (iv) 17 false-positive calls (including 6 that were present at < 0.5% VAFs). The ddPCR reaction mixture (20 µL) contained 2 µL of bulk gDNA or NA12878 (25 ng) template, 250 nM of each primer (reference and assay specific pairs), and 250 nM TaqMan probe (Ref-HEX and assay specific-FAM dye labeled) in a 1× Bio-Rad Supermix (Bio-Rad, Hercules, CA, USA). The 20 µl PCR reaction was mixed with Bio-Rad droplet generator oil and partitioned into 15,000–20,000 droplets using the Bio-Rad QX-100 droplet generator (Bio-Rad). The generated droplets were transferred to a 96-well PCR reaction plate and sealed with a pierceable sealing foil. PCR conditions were 10 min at 95 °C, 40 cycles of denaturation for 30 s at 94 °C, assay specific anneal/ extension temperatures 50–57.5 °C (as noted in table) for 60 s with ramp rate of 2.5 °C s^{-1}, followed by 10 min at 98 °C and a hold at 4 °C. Post-PCR amplification each droplet was checked for fluorescence to count the number of droplets that yielded positive/ negative results, using the Bio-Rad QX-100 droplet reader (Bio-Rad) and frequencies calculated based on total droplet counts per well. In the end, we were able to design informative ddPCR validation tests for 13 candidate mosaic SNVs. For the remaining 16 variants, we could not design effective probes because the variants were within AT-rich or repetitive regions or the assay failed to separate clusters at various annealing temperatures.

Validation result consolidation (University of Michigan and Mayo Clinic, contributed by Yifan Wang and Alexej Abyzov)

We tested 383 sites using PCR amplicon-based sequencing (Fig. 3). Of these, 280 were determined to be false-positive calls, 64 had read support above background levels

(eventually called as 9 germline SNVs, 17 false-positive calls, and 38 true mosaic SNVs), and 39 had a discordant validation status between two different BSMN groups (eventually called as 2 germline SNVs, 32 false-positive calls, and 5 true mosaic SNVs).

To further evaluate the discordant calls, we considered experiments yielding overlapping end sequence reads as more reliable because they resulted in higher-quality sequencing data (i.e., higher confidence base calls, reduced indels, and better mappability); these analyses allowed us to resolve 11/39 candidates. Other information that allowed the interrogation of the remaining calls included (1) determining the number of haplotypes observed in 10X genomics linked read sequencing datasets (i.e., "3 haplotypes" were observed for true-positive calls and "4 haplotypes" were observed for false-positive calls); (2) eliminating SNVs at a < 0.005 VAF in the PCR amplicon-based sequencing experiments (as such variants are unlikely be called from WGS data); (3) eliminating calls with fewer than 3 reads per WGS replicate; (4) eliminating variants within homopolymer tracts, CNVs, within 3 bps of indels, and near a germline Alu insertion (Fig. S8). We also applied the above criteria to invalidate 17/64 candidates with read support above background levels. Furthermore, 9/64 and 2/39 discordant candidates were germline SNVs, as their VAF approached 0.5 and only "2 haplotypes" were observed for the candidate calls in 10X linked read sequencing data. Following ddPCR, one SNV candidate initially assigned as a false positive was reassigned as a true SNV. Finally, based on the examination of 10X linked read sequencing datasets, one of the 17 candidates that lacked enough data from the PCR amplicon-based sequencing experiments ("NED"; not enough data) was determined to be a germline SNV. By consolidating all validation results, we arrived at a confident set of 43 mosaic SNVs.

Panel of normal (PON) mask using 1000 genomes 2504 whole-genome sequencing libraries (University of Michigan, contributed by Yifan Wang and Jeffrey Kidd)

We applied a PON mask for all the candidate somatic SNV sites to exclude the possible sequencing artifacts using the 1000 Genomes Project whole-genome sequencing libraries. We collected the counts for reference alleles and alternative alleles for each candidate site in each of the 2504 samples with a read quality cutoff at 20 and a base quality cutoff at 20. Candidate sites that have greater than 0.05 candidate allele frequency in more than 5 samples were identified as false positives. Thus, the PON mask filter out possible recurrent artifacts that are induced during sequencing or library preparation.

Classifying false-positive calls (Harvard University, contributed by Max Sherman)

We consolidated the above filtering strategies into a naïve Bayes classifier and assessed our ability to characterize the 400 putative candidate mosaic SNVs using a leave-one-out cross-validation method. Briefly, we iteratively trained the model on validation results from 399 mosaic SNVs and predicted the validation status of the held-out SNV. This classifier uses features such as the allele count of a candidate somatic SNV across the 2504 samples in the 1000 Genomes Project as well as Chromium 10X linked read and single-cell properties to aid in the discrimination of true somatic SNVs from false-positive calls (see below). Given that trade-offs between maximizing the number of true-positive calls and excluding false-negative calls are inevitable in statistical modeling, we ultimately arrived at conservative

calling parameters, where the classifier achieved 90% precision, 42% sensitivity, and 99% specificity. While this model suffers with respect to sensitivity, the chosen values maximize the rejection of false-positive calls at the expense of removing some false-negative calls (Additional file 1: Table S4). For example, false-negative calls were mostly present at low VAFs (median = 0.013) and lacked haplotype support in Chromium 10X linked read (i.e., 15/25 false-negative calls lacked haplotype support, whereas only 4/18 true-positive calls lacked haplotype support) and/or single-cell sequencing datasets (i.e., 22/25 false-negative calls lacked support in single-cell datasets as compared to 9/18 true-positive calls). Overall, these results were consistent with our prior observations and indicated that integrating various sequence characteristics can distinguish between bona fide mosaic SNVs and false-positive signals. Indeed, these criteria, as well as additional filtering strategies, have been consolidated for use in a recently developed machine learning algorithm, MosaicForecast [35].

Naive Bayes classifier combining validation approaches (Harvard University, contributed by Max Sherman)

We trained a naïve Bayes classifier to provide the probability that a given site will pass or fail amplicon validation based on the properties of the site in the 75 WGS panel of normal samples, 10X sequencing from the common brain tissue, and single-cell whole-genome sequencing. Given a set of training sites, we calculated the conditional likelihood of a site having the following properties given its validation status (pass or fail):

- Panel of normal properties:
  - ○ Presence of alternate reads in ≤5 of the 75 panel of normal samples
  - ○ Presence of alternate reads in >5 of the 75 panel of normal samples
- 10X properties:
  - ○ PASS 10X filters (see above Methods)
  - ○ FAIL 10X filters for any reason
  - ○ No 10X information available due to lack of alternate reads
- Single cell properties:
  - ○ No alternate reads present in any single cells
  - ○ Alternate reads present in <4 single cells
  - ○ Alternate reads present in ≥4 single cells
  - ○ Locus flagged as PASS by single cell genotyping method
  - ○ Locus flagged as LOWQUAL by single cell genotyping method
  - ○ Locus flagged as ARTIF by single cell genotyping method
  - ○ Locus has no flag assigned by single cell genotyping algorithm

The conditional likelihood was calculated as the number of sites exhibiting that property and having a certain validation status divided by the number of sites with that validation status. To prevent events of probability zero, each category was given a prior pseudocount of 1. The prior probability of passing (failing) was calculated as the number of test sites in the training data which passed (failed) amplicon validation divided by the number of test sites in the training data.

Let B be a binary random variable which takes on the value 0 if the site fails validation and 1 if it passes validation. Let S be the set of properties given above for some new locus:

$$\Pr(B \mid S) = \frac{\prod_{s \in S} \Pr(s \mid B) \, \Pr(B)}{\prod_{s \in S} \Pr(s \mid B = 0) \, \Pr(B = 0) + \prod_{s \in S} \Pr(s \mid B = 1) \, \Pr(B = 1)} \, .$$

The sensitivity and specificity of this approach was calculated using leave-one-out cross-validation on the 400 sites.

Implementation of the best practices workflow (Mayo Clinic, contributed by Taejeong Bae)

The best practices workflow was implemented as a computational pipeline composed of a series of bash job scripts compatible with the job scheduler and Sun Grid Engine. The software includes the steps of uniform data processing, variant calling with GATK at various ploidies, and variant filtering. It is accessible at https://github.com/bsmn/bsmn-pipeline. Although the pipeline is executable at any cluster system using Sun Grid Engine as a job scheduler, we used AWS ParallelCluster (https://github.com/aws/aws-parallelcluster) as the running environment for the BSMN project. The configuration used for setting up the ParallelCluster is available at https://github.com/bintriz/bsmn-aws-setup.

Cell lineage tree reconstruction (Mayo Clinic, contributed by Taejeong Bae)

To reconstruct a cell lineage tree from single-cell data, we examined whether the union set of 49 SNVs identified in our validation call set and called by the application of the best practices workflow to DLPFC replicate 6 WGS were shared among twelve NeuN+ neuron single-cell datasets. An SNV was considered "present" if we could find at least one supporting read in a single cell. The VAFs of the 49 SNVs also were estimated from the following WGS datasets: DLPFC (4 samples), NeuN+ (2 datasets), and NeuN− (1 dataset) DLPFC cell fractions, cerebellum (1 dataset), dura mater (1 dataset), and dural fibroblasts (2 datasets). We conducted hierarchical clustering of SNVs based upon their VAFs in different tissues using average linkage with Canberra distance. Clustering of single cells was performed by assessing shared genotypes using Jaccard distance. The final lineage tree was constructed by manual inspection of shared genotypes in each cell and the SNV VAF values across tissues. SNVs defining earlier developmental lineages have higher VAFs than SNVs defining later developmental sub-lineages. Six SNVs with highest VAFs clustered together and were present in two distinct groups of single cells (the L1 and L2 lineages). We found that 4/6 SNVs (SNVs 1–4) denote the L1 lineage and 2/6 SNVs (SNVs 10 and 11) denote the L2 lineage. This analysis was conducted using the SciPy package of python. Notably, one single-cell dataset lacked the 49 SNVs and was excluded from the analysis.

Two calls (SNV6 and SNV16) were not validated in our PCR amplicon-based sequencing experiments, but passed the 10X Genomics linked read haplotype test. Further analyses indicated that SNV6 and SNV16 exhibited consistent VAFs across multiple sequencing replicates, that no apparent alignment or sequencing artifacts surround the SNV site, and that these SNVs defined sub-branches of L1 and L2 lineages. Moreover, SNV16 appeared together with SNV17 in a single cell and had similar VAFs, suggesting

they represent markers for a sub-branch of the L2 lineage and, consequently, SNV16 is a bona fide mosaic SNV.

Predicting early cell lineages from bulk DNA (Mayo Clinic, contributed by Taejeong Bae)

To identify mosaic SNV pairs that putatively were marking different cell lineages in WGS bulk DNA, we calculated all possible pairwise VAF anti-correlation values for the 49 somatic mosaic SNVs identified in our validation set and best practices pipeline. For each pairwise comparison, we calculated the mean distance of VAFs of the SNV pair across multiple tissue samples according to the diagonal line: $VAF_1 + VAF_2 = 0.5$. We then used that distance as a score to rank SNV pairs by how well they fit the diagonal line (the lower score is better). The eight best scores corresponded to SNV pairs, where one represented the L1 lineage and the other one represented the L2 lineage. For seven pairs, the scores were significantly different (< 0.05) from the other tested pairs (Fig. S12).

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02285-3.

Additional file 1:. Figure S1-S13, Table S1,S3,S4.

Additional file 2:. Table S2. List of 400 candidate SNVs chosen for validation experiments.

Additional file 3:. Table S5. List of amplicon sequencing primers used to validate the 400 candidate mosaic SNVs.

Additional file 4:. Table S6. List of primers and probes used in ddPCR validation experiments.

Additional file 5. BSMN_Membership_List.

Additional file 6. Review history.

Review history
The review history is available as Additional file 6.

Peer review information
Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
The Brain Somatic Mosaicism Network consortium designed the research. Contributors to each different data set, analytical approach, and variant validation are marked in "Methods." FMV, JVM, AEU, JMK, REM, AA, AC, JGG, TMB, PJP, MAP, JP, CAW, and DRW provided project oversight. JG, IL, RP, CR, ST, LT, BZ, LLB, SBE, LF, RN, JHS generated the data. YW, TB, JT, MAS, AG, IL, XY, YJ, JMK, and AA analyzed the data. TB, YJ, and YW performed uniform processing and workflow implementation. YW, TB, JT, MAS, and SC contributed to display item preparation. YW, TB, MAS, FMV, JVM, AEU, JMK, REM, and AA drafted the manuscript. Multiple authors contributed sections in "Methods," edits, and comments. All authors read and approved the final manuscript.

Availability of data and materials
Data and call sets have been deposited in the NIMH Data Archive (NDA Study ID 792: https://doi.org/10.15154/1504248) [52] and can be accessed as part of the NIMH Data Archive permission groups: https://nda.nih.gov/user/dashboard/data_permissions.html. The PON mask for GRCh38 can be accessed from https://www.synapse.org/#!Synapse:syn22024464. The PON mask was generated using 1000 Genomes Project high-coverage whole-genome sequencing of 2504 individuals available at ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_

coverage. The BSMN workflow is available on both GitHub [53] (https://github.com/bsmn/bsmn-pipeline) and Zenodo (doi: https://doi.org/10.5281/zenodo.4321679).

Ethics approval and consent to participate
The analyzed samples were de-identified and derived from an individual postmortem, and IRB approval was not required by any of the participating institutions. The external data we utilized was generated as part of the 1000 Genomes Project and has unrestricted access.
The experimental methods comply with the Helsinki Declaration.

Competing interests
J.V.M. is an inventor on patent US6150160, is a paid consultant for Gilead Sciences, serves on the scientific advisory board of Tessera Therapeutics Inc. (where he is paid as a consultant and has equity options), and currently serves on the American Society of Human Genetics Board of Directors. The other authors do not declare competing interests.

Author details
[1]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA. [3]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA. [4]Program in Biochemistry, Cellular and Molecular Biology, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA. [5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [6]MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA. [7]Department of Cell, Developmental and Regenerative Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [8]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [9]Department of Neurology, Kennedy Krieger Institute, Baltimore, MD 21205, USA. [10]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA. [11]Present Address: Arcus Biosciences, Hayward, CA 94545, USA. [12]Sage Bionetworks, Seattle, WA, USA. [13]Division of Genetics and Genomics, Manton Center for Orphan Disease, and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA. [14]Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA 02115, USA. [15]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [16]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), PRBB, 08003 Barcelona, Catalonia, Spain. [17]Department of Cell Biology, Physiology and Immunology, and Institute of Neurosciences, University of Barcelona, 08028 Barcelona, Spain. [18]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA 94305, USA. [19]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. [20]Child Study Center, Yale University, New Haven, CT 06520, USA. [21]Department of Neurosciences, University of California San Diego, La Jolla, CA, USA. [22]Rady Children's Institute for Genomic Medicine, San Diego, CA, USA. [23]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [24]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [25]Lieber Institute for Brain Development, Baltimore, MD 21205, USA. [26]Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD, USA. [27]Vall d'Hebron Institut de Recerca, 08035 Barcelona, Spain. [28]Centro de Investigación en Red sobre Enfermedades Neurodegenerativas (CIBERNED), 28031 Madrid, Spain. [29]ICREA Academia, 08010 Barcelona, Spain. [30]Icahn Institute for Data Science and Genomic Technologies, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [31]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain. [32]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036 Barcelona, Spain. [33]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. [34]Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [35]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, USA. [36]Department of Neuroscience, Yale University, New Haven 06520, CT, USA. [37]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA. [38]Tashia and John Morgridge Faculty Scholar, Stanford Child Health Research Institute, Stanford, CA 94305, USA.

References
1. Abyzov A, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature. 2012;492:438–42. https://doi.org/10.1038/nature11629.
2. Laurie CC, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet. 2012;44: 642–50. https://doi.org/10.1038/nature11629.
3. Jacobs KB, et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet. 2012;44:651–8. https://doi.org/10.1038/ng.2271.
4. Loh PR, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature. 2018;559:350–5. https://doi.org/10.1038/ng.2270.
5. Gore A, et al. Somatic coding mutations in human induced pluripotent stem cells. Nature. 2011;471:63–7. https://doi.org/10.1038/s41586-018-0321-x.
6. Ji J, et al. Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem Cells. 30:435–40. https://doi.org/10.1002/stem.1011(2012.
7. Piotrowski A, et al. Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat. 2008;29: 1118–24. https://doi.org/10.1002/humu.20815.
8. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci U S A. 2012;109:18018–23. https://doi.org/10.1073/pnas.1213736109.
9. Bruder CE, et al. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet. 2008;82:763–71. https://doi.org/10.1073/pnas.1213736109.

10. McConnell MJ, et al. Mosaic copy number variation in human neurons. Science. 2013;342:632–7. https://doi.org/10.1016/j.ajhg.2007.12.011.

11. Baillie JK, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7. https://doi.org/10.1126/science.1243472.

12. Evrony GD, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151:483–96. https://doi.org/10.1038/nature10531.

13. Saini N, et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. Plos Genet. 2016;12:e1006385. https://doi.org/10.1016/j.cell.2012.09.035.

14. Abyzov A, et al. One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. 2017;27:512–23. https://doi.org/10.1371/journal.pgen.1006385.

15. Lodato MA, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science. 2018;359:555–9. https://doi.org/10.1101/gr.215517.116.

16. Blokzijl F, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4. https://doi.org/10.1126/science.aao4426.

17. Bae T, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science. 2018;359:550–5. https://doi.org/10.1038/nature19768.

18. McConnell MJ, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science. 2017;356 https://doi.org/10.1126/science.aan8690.

19. Cai X, et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep. 2014;8:1280–9. https://doi.org/10.1126/science.aal1641.

20. Evrony GD, et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron. 2015;85:49–59. https://doi.org/10.1016/j.celrep.2014.07.043.

21. Erwin JA, et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci. 2016;19:1583–91. https://doi.org/10.1016/j.neuron.2014.12.028.

22. Zhu, X. et al. Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. Nat Neurosci. 2021;24:186-96. https://doi.org/10.1101/660779

23. Lodato MA, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science. 2015;350:94–8. https://doi.org/10.1038/nn.4388.

24. Dong X, et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14:491–3. https://doi.org/10.1126/science.aab1785.

25. Schmitt MW, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012;109:14508–13. https://doi.org/10.1073/pnas.1208715109.

26. Kennedy SR, et al. Detecting ultralow-frequency mutations by duplex sequencing. Nat Protoc. 2014;9:2586–606. https://doi.org/10.1038/nprot.2014.170.

27. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet. 2018;19:269–85. https://doi.org/10.1038/nrg.2017.117.

28. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. https://doi.org/10.1038/nmeth.4227.

29. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9. https://doi.org/10.1101/gr.107524.110.

30. Kim S, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4. https://doi.org/10.1038/s41592-018-0051-x.

31. Zhou W, et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. 2020;48:1146–63. https://doi.org/10.1093/nar/gkz1173.

32. Goerner-Potvin P, et al. Computational tools to unmask transposable elements. Nat Rev Genet. 2018;19:688–704. https://doi.org/10.1038/s41576-018-0050-x.

33. Chaisson MJP, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784. https://doi.org/10.1038/s41467-018-08148-z.

34. Zook JM, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32:246–51. https://doi.org/10.1038/nbt.2835.

35. Dou Y, et al. Accurate detection of mosaic variants in sequencing data without matched controls. Nat Biotechnol. 2020;38:314–9. https://doi.org/10.1038/s41587-019-0368-8.

36. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. https://doi.org/10.1038/s41587-019-0368-8.

37. Marks P, et al. Resolving the full spectrum of human genome variation using linked-reads. Genome Res. 2019;29:635–45. https://doi.org/10.1038/s41586-020-2308-7.

38. Zheng GX, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34:303–11. https://doi.org/10.1101/gr.234443.118.

39. Huang AY, et al. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. Nucleic Acids Res. 2017;45:e76. https://doi.org/10.1038/nbt.3432.

40. Yang X, et al. Genomic mosaicism in paternal sperm and multiple parental tissues in a Dravet syndrome cohort. Sci Rep. 2017;7:15677. https://doi.org/10.1093/nar/gkx024.

41. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76. https://doi.org/10.1038/s41598-017-15814-7.

42. Freed D, Pevsner J. The contribution of mosaic variants to autism spectrum disorder. Plos Genet. 2016;12:e1006245. https://doi.org/10.1101/gr.129684.111.

43. Untergasser A, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40:e115. https://doi.org/10.1371/journal.pgen.1006245.

44. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64. https://doi.org/10.1093/nar/gks596.

45. You FM, et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics. 2008;9:253. https://doi.org/10.1101/gr.229202.

46. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63. https://doi.org/10.1186/1471-2105-9-253.

47. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30:614–20. https://doi.org/10.1093/bioinformatics/btr507.

48. Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat Commun. 2019;10:3908. https://doi.org/10.1093/bioinformatics/btt593.

49. Loh PR, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443–8. https://doi.org/10.1038/s41467-019-11857-8.

50. Auton A, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. https://doi.org/10.1038/ng.3679.

51. Sherman MA, et al. PaSD-qc: quality control for single cell whole-genome sequencing data using power spectral density estimation. Nucleic Acids Res. 2018;46:e20. https://doi.org/10.1038/nature15393.

52. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human brain tissue. NIMH Data Archive, Study ID 792. 2020. https://doi.org/10.15154/1504248

53. Wang, Y. et al. Comprehensive identification of somatic nucleotide variants in human brain tissue. Github. 2020. https://github.com/bsmn/bsmn-pipeline. Access date: 01/01/2018.

Publisher's Note