Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network


BACKGROUND: Elucidating the genetic architecture of neuropsychiatric disorders remains a major scientific and medical challenge. Emerging genomic technologies now permit the analysis of somatic mosaicism in human tissues. The measured frequencies of single-nucleotide variants (SNVs), small insertion/deletion (indel) mutations, structural variants [including copy number variants (CNVs), inversions, translocations, and whole-chromosome gains or losses], and mobile genetic element insertions (MEIs) indicate that each neuron may harbor hundreds of somatic mutations. Given the long life span of neurons and their central role in neural circuits and behavior, somatic mosaicism represents a potential mechanism that may contribute to neuronal diversity and the etiology of numerous neuropsychiatric disorders.

ADVANCES: Somatic mutations that confer cellular proliferative or cellular survival phenotypes have been identified in patients with cortical malformations. These data have led to the hypothesis that somatic mutations may also confer phenotypes to subsets of neurons, which could increase the risk of developing certain neuropsychiatric disorders. Genomic technologies, including advances in long-read, next-generation DNA sequencing technologies, single-cell genomics, and cutting-edge bioinformatics, can now make it possible to determine the types and frequencies of somatic mutations within the human brain. However, a comprehensive understanding of the contribution of somatic mosaicism to neuropsychiatric disease requires a coordinated, multi-institutional effort. The National Institute of Mental Health (NIMH) has formed a network of 18 investigative teams representing 15 institutions called the Brain Somatic Mosaicism Network (BSMN) to generate a community resource of more than 10,000 DNA-sequencing data sets and will enable a cross-platform integrated analysis with other NIMH initiatives, such as the PsychENCODE project and the CommonMind Consortium.

OUTLOOK: A fundamental open question in neurodevelopmental genetics is whether and how somatic mosaicism may contribute to neuronal diversity within the neurotypical spectrum and in diseased brains. Healthy individuals may harbor known pathogenic somatic mutations at subclinical frequencies, and the local composition of neural cell types may be altered by mutations conferring prosurvival phenotypes in subsets of neurons. By extension, the neurotypical architecture of somatic mutations may confer circuit-level differences that would not be present if every neuron had an identical genome. Given the apparent abundance of somatic mutations within neurons, an in-depth understanding of how different types of somatic mosaicism affect neural function could yield mechanistic insight into the etiology of neurodevelopmental and neuropsychiatric disorders.

The BSMN will examine large collections of postmortem brain tissue from neurotypical individuals and patients with neuropsychiatric disorders. By sequencing brain DNA and single neuronal genomes directly, rather than genomic DNA derived from peripheral blood or other somatic tissues, the BSMN will test the hypothesis that brain somatic variants contribute to neuropsychiatric disease. Notably, it is also possible that some inherited germline variants confer susceptibility to disease, which is later exacerbated by somatic mutations. Confirming such a scenario could increase our understanding of the genetic risk architecture of neuropsychiatric disease and may, in part, explain discordant neuropsychiatric phenotypes between identical twins. Results from these studies may lead to the discovery of biomarkers and genetic targets to improve the treatment of neuropsychiatric disease and may offer hope for improving the lives of patients and their families.
Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network


Neuropsychiatric disorders have a complex genetic architecture. Human population-based studies have identified numerous heritable sequence and structural genomic variants associated with susceptibility to neuropsychiatric disease. However, these germline variants do not fully account for disease risk. During brain development, progenitor cells undergo billions of cell divisions to generate the ~80 billion neurons in the brain. The failure to accurately repair DNA damage arising during replication, transcription, and cellular metabolism amid this dramatic cellular expansion can lead to somatic mutations. Somatic mutations that alter subtile of neuronal transcriptomes and proteomes can, in turn, affect cell proliferation and survival and lead to neurodevelopmental disorders. The long life span of individual neurons and the direct relationship between neural circuits and behavior suggest that somatic mutations in small populations of neurons can significantly affect individual neurodevelopmental. The Brain Somatic Mosaicism Network has been founded to study somatic mosaicism both in inutephyal human brains and in the context of complex neuropsychiatric disorders.

The human body reaches a steady-state level of approximately $10^{14}$ cells in adulthood. Because DNA replication and DNA repair are imperfect processes (estimated at $\sim 0.27$ to 0.99 errors in $\sim 10^9$ nucleotides per cell division), somatic cells within an individual must differ in the presence of single-nucleotide variants (SNVs) and/or small insertion/deletion (inDEL) mutations (2–4). In addition to SNVs and inDEls, subsets of neurons also harbor structural variants [which include large (>1 Mb) copy number variants (CNVs), inversions, translocations, and whole-chromosome gains or losses (6–10)] and smaller mobile genetic element insertions (MEIs) (11–16). Here, we define somatic mosaicism as the existence of different genomes within the cells of a monozygotic individual. Well-known examples of somatic mosaicism include ichthyosis with confetti and lines of Blaschko (4).

Healthy neuronal development requires that neuronal stem cells and progenitor cells (NPCs) undergo tens of billions of cell divisions, both before birth and during the first years of life, to generate the ~80 billion neurons in the fully developed human brain (17). Because neurons are among the longest-lived cells in the body, the accumulation of somatic mutations (i.e., SNVs, indels, structural variants, and MEIs) within NPCs, or perhaps postmitotic neurons (18), could influence neuronal development, complexity, and function (19, 20). Indeed, mounting evidence indicates that somatic mutations in small populations of neurons contribute to various neurodevelopmental disorders (Table 1).

Genomic studies implicitly assume that every cell within an individual has the same genome. Family-based genetic studies, genome-wide association studies (GWAS), and exome sequencing analyses have identified numerous common, rare, and de novo germline SNVs and CNVs associated with an increased risk of autism spectrum disorder (ASD), schizophrenia, and bipolar disorder; but each variant only represents a minor component of population-level disease risk (21–24). In general, these approaches sequence the DNA from available clinical samples (e.g., peripheral blood) to interrogate an individual's germline genome; they do not account for any additional disease risk brought about by somatic mutations that occur during brain development. To address this knowledge gap, the National Institute of Mental Health (NIMH) supported the formation of the Brain Somatic Mosaicism Network (BSMN). Notably, several outstanding reviews have recently discussed how somatic mutations within the brain may contribute to neurologial disease (e.g., (2, 25, 26)). Here, we build on these discussions and highlight how somatic mutations within the brain may contribute to neuronal diversity. We also evaluate emerging genomic approaches to measure and validate somatic mosaicism and summarize BSMN efforts to generate a large publicly available resource to evaluate the contribution of somatic mosaicism to neuropsychiatric disease (Fig. 1).
Table 1. Mosaic mutations in genes and their associated signaling pathways and diseases. Disease abbreviations: CLOVES, Congenital lipomatous overgrowth, vascular malformations, and epidermal nevi; FCD, focal cortical dysplasia; GPCR, G protein–coupled receptor; HME, hemimegalencephaly; MCAP, megalencephaly-capillary malformation-polymicrogyria syndrome; MPPH2, megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome-2; NF, neurofibromatosis; RALD, Ras-associated autoimmune leukoproliferative disorder; TSC, tuberous sclerosis complex. Mosaicism abbreviations: G, germline; S, somatic; OS, obligatory somatic; MS, milder somatic; SHS, second-hit somatic.

<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Signaling pathway(s)</th>
<th>Disease(s)</th>
<th>Cellular function(s)</th>
<th>Cancer(s)</th>
<th>Cancer role</th>
<th>Mosaicism</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIK3CA</td>
<td>PI3K-AKT-mTOR</td>
<td>HME, mosaic overgrowth syndrome, type 2 segmental, CLOVES, MCAP</td>
<td>PI3K subunit, serine/threonine kinase</td>
<td>Cervical, various neoplasms, colorectal</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>AKT1</td>
<td>PI3K-AKT-mTOR</td>
<td>Proteus syndrome</td>
<td>Serine/threonine kinase</td>
<td>Breast, ovarian, colorectal</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>AKT2</td>
<td>PI3K-AKT-mTOR</td>
<td>Diabetes mellitus</td>
<td>Serine/threonine kinase</td>
<td>Ovarian, pancreatic, breast, colorectal, lung cancer</td>
<td>Oncogene</td>
<td>G/S</td>
</tr>
<tr>
<td>AKT3</td>
<td>PI3K-AKT-mTOR</td>
<td>HME, MCAP, MPPH2</td>
<td>Serine/threonine kinase</td>
<td>Melanoma, glioma, ovarian cancer</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>MTOR</td>
<td>PI3K-AKT-mTOR</td>
<td>FCD type II</td>
<td>Serine/threonine kinase</td>
<td>Carcinoma, glioblastoma, melanoma</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>DEPDC5</td>
<td>PI3K-AKT-mTOR</td>
<td>Epilepsy with FCD</td>
<td>mTORC1 repressor</td>
<td>Glioblastoma and ovarian tumors</td>
<td>Tumor suppressor</td>
<td>G/S</td>
</tr>
<tr>
<td>TSC1</td>
<td>PI3K-AKT-mTOR</td>
<td>TSC</td>
<td>Negative regulator of mTORC1</td>
<td>Renal angiomyolipomas</td>
<td>Tumor suppressor</td>
<td>SHS</td>
</tr>
<tr>
<td>TSC2</td>
<td>PI3K-AKT-mTOR</td>
<td>TSC</td>
<td>Negative regulator of mTORC1</td>
<td>Renal angiomyolipomas</td>
<td>Tumor suppressor</td>
<td>SHS</td>
</tr>
<tr>
<td>NRAS, BRAF, FGFR3, PIK3CA</td>
<td>PI3K-AKT-mTOR</td>
<td>Congenital melanocytic, other nevi; seborrheic keratosis</td>
<td>Cell cycle regulation</td>
<td>(FGFR3) bladder, cervical, urothelial</td>
<td>Oncogene</td>
<td>G/S</td>
</tr>
<tr>
<td>NF2</td>
<td>PI3K-AKT-mTOR</td>
<td>NF type 2</td>
<td>Negative regulator of Ras, mTOR pathways</td>
<td>Neurofibromas</td>
<td>Tumor suppressor</td>
<td>G/MS</td>
</tr>
<tr>
<td>NF1</td>
<td>PI3K-AKT-mTOR</td>
<td>NF type 1, Watson syndrome</td>
<td>Negative regulator of Ras pathway</td>
<td>Neurofibromas, leukemia</td>
<td>Tumor suppressor</td>
<td>SHS</td>
</tr>
<tr>
<td>BRAF, NRAS, KRAS</td>
<td>PI3K-AKT-mTOR</td>
<td>RAS</td>
<td>Pyogenic granuloma</td>
<td>(KRAS) breast, colorectal, other; (NRAS) thyroid, melanoma, other; (BRAF) melanoma, colorectal</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>HRAS, KRAS</td>
<td>PI3K-AKT-mTOR</td>
<td>RAS</td>
<td>Schimmelpenning-Feuerstein-Mims syndrome</td>
<td>Cell cycle regulation</td>
<td>(KRAS) bladder, breast, colorectal, pancreatic, other; (HRAS) Colorectal, bladder, kidney, other</td>
<td>Oncogene</td>
</tr>
<tr>
<td>KRAS</td>
<td>PI3K-AKT-mTOR</td>
<td>RALD</td>
<td>Cell cycle regulation</td>
<td>Breast, bladder, other</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>GNAQ</td>
<td>GPCR, MAPK</td>
<td>Sturge-Weber syndrome</td>
<td>G protein alpha subunit</td>
<td>Melanoma</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>GNAQ, GNAI</td>
<td>GPCR, MAPK</td>
<td>Dermal melanocytosis and phakomatosis pigmentovascularis</td>
<td>G protein alpha subunit</td>
<td>Melanoma</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>MAP3K3</td>
<td>MAPK</td>
<td>Verrucous venous malformation</td>
<td>Cell cycle regulation</td>
<td>Breast, colon, rectal cancers</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
</tbody>
</table>

continued on next page
<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Signaling pathway(s)</th>
<th>Disease(s)</th>
<th>Cellular function(s)</th>
<th>Cancer(s)</th>
<th>Cancer role</th>
<th>Mosaicism</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNAS (132, 133)</td>
<td>GPCR</td>
<td>McCune-Albright syndrome</td>
<td>G protein alpha subunit</td>
<td>Adenomas, carcinomas, ovarian neoplasms</td>
<td>Oncogene</td>
<td>OS</td>
</tr>
<tr>
<td>JAK2 (134, 135)</td>
<td>JAK-STAT</td>
<td>Myelofibrosis, polycythemia vera, and essential thrombocytopenia</td>
<td>Cell cycle regulation</td>
<td>Leukemia</td>
<td>Oncogene</td>
<td>SHS</td>
</tr>
<tr>
<td>SCN1A (136)</td>
<td>Sodium channel</td>
<td>Dravet syndrome</td>
<td>Neural excitation</td>
<td>–</td>
<td>–</td>
<td>G/MS</td>
</tr>
<tr>
<td>NLRP3 (137)</td>
<td>Caspase/ inflammasome</td>
<td>CINCA syndrome</td>
<td>Inflammasome subunit</td>
<td>–</td>
<td>–</td>
<td>G/MS</td>
</tr>
<tr>
<td>PORCN (138)</td>
<td>Wnt</td>
<td>Focal dermal hypoplasia</td>
<td>O-acyltransferase</td>
<td>–</td>
<td>–</td>
<td>G/MS</td>
</tr>
<tr>
<td>PIGA (139)</td>
<td>Hematopoiesis</td>
<td>Paroxysmal nocturnal hemoglobinuria</td>
<td>ER protein processing</td>
<td>Leukemia</td>
<td>–</td>
<td>OS</td>
</tr>
</tbody>
</table>

Box 1. Criteria used to prioritize somatic variants for functional characterization.

**Absence from the germ line**
We will focus on variants with a definitive somatic origin.

**Recurrence and frequency of somatic variation at the locus of interest**
We will prioritize loci at which somatic variations, across all types, recur in multiple disease samples but not in control samples.

**Mutation severity**
Highly deleterious variations will be prioritized for likely functional importance.

**Intersection with known disease loci and biochemical pathways**
Taking advantage of data on germline variations in brain disorders, we will prioritize loci that have been previously implicated in disease.

**Intersection with brain expression and epigenomic data**
Taking advantage of large, publicly funded consortia of human brain spatiotemporal expression data (e.g., BrainSpan) and epigenomic data (e.g., PsychENCODE and Roadmap Epigenomics), we will select genes that are expressed in brain regions associated with brain disorders and noncoding loci with potential regulatory function.

Mechanisms of somatic mosaicism
DNA damage occurs constantly in every cell in our bodies, and many components of the DNA damage response are essential for neurodevelopment. Single-strand and double-strand DNA breaks, as well as base mutations, arise as a consequence of DNA replication, transcription, epigenetic modification, cellular respiration, and environmental stressors. If the resultant damage is not accurately repaired, DNA mutations can occur that can lead to somatic variation among neurons and other cell types.

The nonhomologous end-joining (NHEJ) pathway of DNA repair is required for neurodevelopment. Mice deficient in NHEJ proteins exhibit extensive NPC apoptosis and often die prenatally (27). Intriguingly, the embryonic lethality and NPC apoptosis phenotypes are rescued in a p53-null mouse background, suggesting that genotoxic stress contributes to lethality (28). Consistent with these data, compound heterozygous mutations in DNA damage response genes (e.g., ataxia telangiectasia mutated (ATM), ataxia telangiectasia-related (ATR), and ATR-interacting protein (ATRIP)) can lead to increased mutational loads, neurodevelopmental brain defects, and neuronal degeneration (29–31). More broadly, deficits in other DNA repair pathways, such as transcription-coupled repair, homologous recombination, and nucleotide excision repair, also can lead to human neurodevelopmental phenotypes (32, 33).

Defects in different DNA repair pathways are associated with distinct somatic mutation profiles. For example, SNVs and indels can arise from errors during base excision repair, nucleotide excision repair, and transcription-coupled repair (33). Moreover, the action of the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like-3 (APOBEC3) family of cytosine deaminase proteins can lead to cytidine-to-uridine transition mutations on single-strand DNA that, upon replication, lead to guanosine-to-adenosine mutations on the opposing DNA strand (34). Errors made during DNA mismatch repair also can lead to either interspersed SNVs or indels within microsatellite repeat sequences, whereas errors made during double-strand break repair by homologous recombination, NHEJ, or alternative-NHEJ can lead to CNVs (35, 36).

Errors incurred during DNA replication or transcription also can lead to the formation of CNVs. Large, actively transcribed genes that undergo replication during late S-phase correspond to chromosomal fragile sites and are hot spots for the generation of genomic variants and translocations (37, 38). Because neuronal genes are overrepresented among the longest genes in the human genome, transcription may predispose these genes to somatic CNVs (39). Indeed, intragenic deletions within large, neuronally expressed genes (e.g., ATRIP, INPP5Q, NXXRN1, and CNTNAP2) are associated with ASD, intellectual disability, and other neurodevelopmental disorders (40, 41). Thus, if individuals harbor somatic CNVs at these loci in many neurons or in neurons within specific functional brain regions, they may be susceptible to neurological disease.

Long interspersed element-1s (LINE-1s or L1s) can mobilize (i.e., retrotranspose) within the brain, leading to another form of somatic variation (42). Active L1s encode two proteins, ORF1p and ORF2p, which are required for retrotransposition. ORF2p contains endonuclease and reverse transcriptase activities that are needed to “copy-and-paste” L1 sequences into a new genomic location by a mechanism termed target-site primed reverse transcription (TPRT) (42, 43). In addition to canonical TPRT, L1s occasionally can integrate into endogenous DNA lesions (44). Moreover, recombination events that arise either during (45, 46–47) or after L1 retrotransposition (48) can lead to the formation of structural variants.

Somatic mutations in human disease
**Mosaicism and structural brain abnormalities**
One of the most common causes of medically refractory pediatric epilepsy is focal dysplasia of the cerebral cortex. Until recently, the basis of
this disorder remained a medical mystery. Genetic studies of the most severe form of focal dysplasia, hemimegalencephaly, in which one entire cerebral hemisphere is enlarged in size, led to the identification of gain-of-function somatic mutations in the phosphatidylinositol-3-kinase (PI3K)–protein kinase B (Akt) and mammalian target of rapamycin (mTOR) signaling pathways (Table 1, Fig. 2). We now know that mutations in mTOR are the single largest contributor to focal dysplasia in pediatric epilepsy (49–53). Similarly, germline mutations in one allele of the TSC1 or TSC2 gene confer susceptibility to tuberous sclerosis, a disease characterized by facial and skin lesions, seizures, intellectual disability, cardiac and renal tumors, and cortical tubers (52). Because the Tsc1 and Tsc2 proteins are negative regulators of the mTOR-signaling pathway, a second somatically acquired mutation is required for disease onset. Somatic mutations that mildly activate the mTOR-signaling pathway also cause symmetrically overgrowth syndromes such as megalencephaly-capillary malformation syndrome, megalencephaly, and certain forms of polymicrogyria (49–51). Common to all of these phenotypes is the presence of hypertrophic neural-like “balloon” cells, which carry the somatic mutation yet fail to transform to a malignant cell type (52).

Somatic mutations that inappropriately activate Ras signaling or related signaling pathways can likewise confer proliferation and survival phenotypes to subsets of cells and cause neurological disease. For example, a gain-of-function somatic mutation in GNAQ, encoding G protein subunit alpha q, can lead to Sturge-Weber syndrome, a disease characterized by vascular anomaly in the brain, glaucoma, seizures, stroke, and intellectual disability (53). The same GNAQ mutation, occurring in a different somatic cell type later in development, can cause uveal melanoma (54). Because mutations in certain neurodevelopmental disorders (e.g., neurofibromatosis, tuberous sclerosis, Proteus syndrome, and other neurocognitive disorders) either activate proto-oncogenes or inactivate tumor suppressor genes, it is not surprising that similar mutations in non-neuronal cell types manifest as cancers. Intriguingly, postmitotic neurons are rarely the source of brain tumors, suggesting that postmitotic neurons may have safeguards that ensure against dedifferentiation and further proliferation.

Relative to germline mutations, somatic mutations can lead to milder cases of heritable neurodevelopmental disorders. For example, somatic mutations in genes involved in neuronal migration are estimated to represent 5 to 10% of de novo mutations and are detected more frequently in patients with unexplained brain malformations when studied with sensitive high-throughput sequencing methods (55). Moreover, somatic mutations within the LIS1 or DCX genes can lead to gross disruptions of neuronal migration, whereas germline mutations in LIS1 or DCX result in lissencephaly (56, 57). Results from several experiments also suggest that somatic mutations that lead to a reduction of gene copy number in migrating neurons can lead to cell-autonomous defects in neuronal migration, with severe epilepsy and intellectual disability as a consequence (56, 57).

**ASD and other common neuropsychiatric diseases**

Genetic approaches have not yet fully explained the etiology of ASD, bipolar disorder, schizophrenia, or Tourette syndrome. Although gene-by-gene and gene-by-environment interactions could, in principle, account for additional disease risk, somatic mosaicism is another potential mechanism that warrants exploration as a contributor to neuropsychiatric diseases (58). De novo SNVs and CNVs, particularly loss-of-function mutations, are significant contributors to ASD risk (21, 59–62). In addition to de novo germline mutations, a substantial number of de novo somatic mutations (i.e., ~5.4% of de novo events) are detected in the blood of ASD patients and are enriched in ASD probands (22). Somatic mosaic mutations also have been identified throughout postmortem ASD brains or, in some instances, in more localized areas in ASD brains (59). Evidence of continuous, widespread cortical mismigration, as seen in some mutant mice, has not been reported in the postmortem ASD brain (63, 64). However, NPCs from a subset of ASD patients with enlarged brain volumes are inherently more proliferative and display abnormal neurogenesis when compared to controls (65, 66). Other ASD patients have focal cortical abnormalities, including disorganized neurons and lamina, polymicrogyria, and other local surface malformations (67). Thus, in addition to specific mutations, additional cell cycles may further affect somatic mutational loads in patients.

Prenatal challenges to the immune system in animals (i.e., maternal immune activation) (68) can also lead to many features like those present in ASD brains. Maternal immune activation leads to increased cellular proliferation, brain size, and ASD-like behaviors in animal models (69–72). Intriguingly, an elevated prevalence of MEIs was observed in a primate model of maternal immune activation (73). Elevated MEI levels likewise are observed in schizophrenia (73) and Rett syndrome patients (74), suggesting that somatic MEI burden may play a role in the etiology of some neurodevelopmental and neuropsychiatric diseases.

**Methods to detect somatic mutations**

The difficulty in detecting a somatic mutation depends on its frequency within a cell population.
Whereas mutations affecting a large fraction (e.g., 50%) of cells are readily detected in bulk tissue sequencing experiments and generally result in high-confidence calls, mutations affecting one or a few cells are unlikely to be detected with bulk tissue sequencing approaches. The identification and validation of rare somatic mutations requires sequencing DNA derived from small pools of cells, single cells, or clonally reprogrammed cells followed by robust computational data analyses (Fig. 1).

**Bulk tissue approaches**
Whole-genome sequencing (WGS) or whole-exome sequencing (WES) of DNA derived from bulk brain tissue allows a straightforward approach to discovering somatic mosaicism (20). WGS and WES minimize sequencing artifacts that can confound downstream analyses and, in the case of WGS, provide an opportunity for identifying a wide range of structural rearrangements, including inversions and translocations. However, WGS and WES using standard sequencing depths have reduced statistical power to detect mutations that occur at low frequencies (i.e., <10% of cells in a population at 30 to 100x coverage). Although increasing sequence coverage allows detection of somatic variants at lower frequencies, it quickly becomes cost prohibitive. Moreover, WGS and WES do not provide information on how somatic variants are distributed across individual cell lineages within a bulk tissue sample.

**Sorted-pools approaches**
Fluorescence-activated cell or nuclei sorting (FACS/ FANS) can be used to isolate specific neural populations (e.g., NeuN+ neurons versus NeuN− cells or cortical inhibitory interneurons versus excitatory principal neurons). Analysis of sorted nuclei populations (e.g., 5000 or 500,000 cells) from specific brain regions increases the power to detect somatic mosaicism that arises in one lineage, because these genomes are no longer diluted by genomes derived from other lineages. Independent pools of sorted nuclei can then be subjected to RNA sequencing (RNA-seq) and quantitative reverse transcription polymerase chain reaction (qRT-PCR) to confirm cell type-specific gene expression profiles (75). In addition to increasing the power for detecting a somatic mutation, cell sorting before DNA extraction could yield information about the embryological origin and developmental trajectory of somatic variation across the brain. Large pools of sorted cells can yield enough DNA for the direct examination of somatic variants by WGS or WES. However, smaller pool sizes will only generate small amounts of DNA; thus, they are best suited for generating PCR amplicon libraries (e.g., as used in MEI detection and other targeted sequencing) or for subsequent whole-genome amplification (WGA).

**Single-cell approaches**
WGA can be used to analyze the genomes of single neurons (26). The spectrum of mutations identified from the genomes of single neurons can then be compared to germline variants in bulk tissue data derived from a non-neuronal control (e.g., brain dural fibroblasts or heart) to identify candidate somatic mutations (5). WGA approaches already are used in pre-implantation genetic screening of embryos (76, 77) and include (i) degenerate-oligonucleotide-primed PCR (DOP-PCR), (ii) multiple displacement amplification (MDA), and (iii) multiple annealing and looping-based amplification (MALBAC). Each method has its advantages and drawbacks. In general, DOP-PCR provides coverage evenly across the genome, which facilitates the detection of large CNVs and chromosomal aneuploidies. However, DOP-PCR has a higher read duplication rate, lower mapping rate, and lower recovery rate when compared with MDA and MALBAC (78) and is cost prohibitive for SNV, indel, and MEI detection. By comparison, MDA yields a high rate of artificial chimeric DNA molecules that can lead to false-positive calls in downstream analyses (79), whereas MALBAC exhibits reduced coverage of certain genomic regions (14, 16, 80), especially those rich in repetitive sequences (78). Considerable advances have recently been made in detecting SNVs (81, 82), CNVs (83), and MEIs (16) in WGA samples; however, best practices necessitate evaluating each WGA approach for the detection of specific types of somatic mosaicism.

Clonal expansion of single cells using human-induced pluripotent stem cell (hiPSC) technology or somatic cell nuclear transfer (SCNT) provides a biological alternative to WGA (80, 84). Any variant uniformly identified in the clonal line, but not in controls, represents a candidate somatic mutation that requires confirmation in the tissue of origin. In contrast, mutations introduced during cell culture will be present in a lower frequency of cells within a clonal cell line and can be discriminated from bona fide somatic mutations in downstream computational analyses. Although the clonal isolation and expansion of primary human neural stem and progenitor cells is possible, the analysis of human neuronal genomes using clonal reprogramming has several limitations. Foremost among these is the availability of live human neurons. Moreover, neither clonal reprogramming nor SCNT have been reported using human neurons; SCNT is further limited by the expense and availability of human oocytes. Finally, reprogramming approaches currently are only successful in ~10% of cells; thus, any neurons harboring highly aberrant genomes may be refractory to reprogramming. Despite these caveats, clonal reprogramming of human neurons is theoretically possible. In addition, it is noteworthy that mouse neurons reprogrammed by SCNT contain genomic rearrangements (e.g., kataegis and chromothripsis) that would be very challenging to validate using current WGA approaches (84).

**Computational methods for mutation detection**
WGS and WES have been used successfully to detect somatic SNVs in family-based studies of Mendelian disease and large-scale sequencing studies of human patient cohorts (2). To identify SNVs, most computational approaches compare call sets generated from an affected sample to those generated from a matched healthy/unaffected sample and/or a control population. These comparisons allow the identification and subsequent exclusion of germline polymorphisms from downstream analyses; however, care must be taken to...
ensure that any candidate somatic mutations are not germline variants that were missed in the matched control. In general, variant callers initially developed to detect mutations in cancer offer higher sensitivity for detecting mosaic SNVs when compared with standard approaches used to detect germline variants (85, 86).

Somatic CNVs can be detected by identifying deviations either from the expected depth of sequence or in the expected distances between paired-end sequencing reads. Similarly, inversions can be identified through differences in the orientations of paired-end sequencing reads. Numerous approaches have been developed to identify CNVs from WGS (7, 87–89), and most can be applied directly to identify somatic mutations. For example, recent studies using WGA in conjunction with WGS have identified megabase-scale de novo CNVs in human and mouse neurons based on differences in read-depth across genomic bins (6–9). CNVs are more difficult to identify using WES due to the biases encountered during the capture of target exons (90).

Somatic MEIs can be detected from bulk tissue, PCR amplicons generated from sorted-cell fractions, or single-cell WGA DNA using split-read and paired-end information (e.g., one paired-end read may map to the reference genome, whereas another may map to a MEI) (91, 92). Detecting low-frequency MEIs with fewer supporting reads requires careful bioinformatic analyses that can distinguish signal from noise, followed by experimental validation with orthogonal methods (14, 83). The analysis of single-cell data remains challenging due to the presence of chimeras generated during WGA (14, 16, 94); thus, care must be taken in calling MEIs.

Validation of somatic mutations

It is essential to validate all candidate somatic mutations. False-positive calls can arise from DNA sequencing errors, contamination with germline variants, chimeric molecules generated during single-cell WGA, PCR-induced nucleotide substitutions, and the failure to amplify certain genomic regions. False-negative calls are dependent on the allele frequency of the somatic mutation within the sample, the type of mutation, and the method of detection. Orthologous experimental methods are required to eliminate false-positives and to calibrate the confidence of detection for different types of somatic mutations. Validation experiments can then be performed on either the tissue of origin or amplified material used to discover the variant. The first approach represents a biological validation, which establishes the presence of a variant call in unamplified DNA from the source sample. The second approach represents a technical validation, which establishes the presence/absence of variant calls in the DNA source material used for discovery.

Biological/primary validation in the tissue of origin

Validation on unamplified DNA from the tissue of origin provides confirmation that a candidate call is a genuine somatic variant and rules out the possibility that it corresponds to a DNA amplification artifact or a mutation that occurred during clonal expansion. Biological validation requires a variant to be present in multiple cells in the tissue of origin at a frequency above experimental detection limits. As such, the failure to validate a variant in the tissue of origin does not necessarily represent a false call. For example, only ~50% of CNVs manifested in hiPSC clones could be directly confirmed in the primary fibroblast cells used to derive hiPSCs (80).

Somatic variants can be confirmed in unamplified cell source material by (i) targeted DNA capture followed by high-coverage (>100x) DNA resequencing, (ii) high-coverage sequencing of multiplexed PCR amplicons, and (iii) droplet digital PCR (ddPCR). These approaches vary in throughput and sensitivity. Targeted DNA capture and resequencing can require the creation of several thousand custom oligonucleotides designed to capture the genomic DNA either including or surrounding the putative variants. The captured DNA then is subjected to high-coverage paired-end DNA sequencing, yielding a typical sensitivity of variant detection in greater than 1% of cells. Amplicon sequencing involves PCR amplification of candidate loci followed by high-coverage paired-end DNA sequencing, yielding a typical sensitivity of variant detection in greater than 0.1% of cells. Finally, ddPCR involves partitioning a DNA sample into large numbers of individual droplets that generally contain one copy of template DNA. PCR takes place within these droplets, leading to the production of a fluorescent readout, either through the use of an intercalating dye or a fluorescent oligomer probe, to indicate the presence or absence of the PCR target of interest. Subsequent quantification of the fluorescent droplets allows a determination of the number of copies of the target locus present in the sample, yielding a typical sensitivity of variant detection in greater than 0.001% of cells (95). Although extremely sensitive, ddPCR requires the optimization of primers, probes, and amplification conditions, which is time-consuming and limits throughput.

The goal when employing biological validation procedures is to detect putative somatic variants and to assess, as precisely as possible, the frequency of each variant in that tissue of origin. Biological validation can (i) determine whether certain individuals in the population are more prone to somatic variation than others, (ii) investigate whether different areas of the brain and/or specific brain cell types have varying amounts and types of particular forms of somatic variation, (iii) assess whether developmental timing contributes to somatic variation, and (iv) reveal whether somatic variations increase as a function of the number of cell divisions and/or a function of age in postmitotic neurons.

Technical validation on source/amplified material

If a somatic variant is only present in a single cell, it will be impossible to validate in bulk tissue. Likewise, a variant present in very few cells may be difficult to validate in the tissue of origin. Thus, technical validation in the source DNA used to discover a putative variant can be used to determine whether a call is true or false. Technical validation typically employs PCR, qPCR, and Sanger sequencing of the locus in the DNA source material (e.g., WGA DNA or DNA from a clonal cell population). Multiple true/false verdicts form the basis for estimating false-discovery and false-negative rates in the resultant call sets.

Present understanding of the prevalence of somatic mutation in neurotypical individuals

Recent studies revealed that mosaic neuronal genomes are the rule, rather than the exception; every neuron probably has a different genome than the neurons with which it forms synapses. Not unexpectedly, SNVs are the most prevalent somatic mutations. A “triple calling” strategy was used to identify and validate clonal SNVs in...
MDA-amplified DNA from single neurons isolated from a neurotypical brain, leading to estimates of 1000 to 1500 SNVs per neuronal genome. By comparison to human cortical neurons, a SCNT experiment in reprogrammed mouse olfactory neurons detected hundreds of SNVs per neuron and a lower proportion of C-to-T transition mutations. Although the divergent SNV rates between these two studies may arise from technical differences (as discussed above), both approaches establish that SNVs represent an important form of somatic mutation in both human and mouse neurons.

Brain somatic CNVs initially were identified by comparing the sequences of bulk DNA derived from multicellular samples of different brain regions to the sequences of DNA derived from somatic tissues. The first single-cell study of neuronal CNVs analyzed 110 human frontal cortex neurons and found that 13 to 41% of the neurons contained at least one megabase-scale de novo CNV. Additional studies, which analyzed fewer neuronal genomes, confirmed that de novo CNVs occur in at least 10% of neurons. CNVs can be shared by multiple neurons and inherited in a clonal manner. Furthermore, megabase-scale CNVs typically alter the copy number of 10 or more genes in individual neurons. In addition to expression-level differences that can accompany gene copy number changes, mosaic neuronal CNVs also are expected to reveal or abate perinuclear alleles on a neuron-by-neuron basis in every individual.

L1 retrotransposition inserts the trans-scriptional regulation of genes in myriad ways. Initial studies used engineered L1s containing a retrotransposition indicator cassette to discover MEI activity in mouse brain and in human NPCs in vitro. Studies of MDA-amplified NeuN-positive nuclei isolated from a neurotypical human brain, followed by L1-transposon profiling or WGS, have since suggested that 0.2 to 1 L1 insertion occur per neuronal genome. Another report, which employed MALBAC WGA in conjunction with L1 capture technology (RC-seq), reported an average of 13 L1 insertions in every neuronal genome, although a subsequent study suggested a high false-positive rate in these data. By comparison, SCNT experiments in mouse olfactory neurons reported ≤1.3 MEIs per neuronal genome. An extrapolation of these data indicates that potentially billions of neurons in the neurotypical brain contain de novo MELS. Additional studies are required to determine whether L1s retrotranspose at varying rates in different brain regions, in different individuals, or preferentially insert into expressed genes, and whether other mobile elements (e.g., Alu retrotransposons) also contribute to individual neuronal genetic diversity.

**Generation of a community resource**

The BSMN will generate comprehensive maps of somatic genomic variation in neurotypical and diseased human brains, including a prioritized call set of confirmed somatic variants (Box 1) that may contribute to neuropsychiatric disease and epilepsy. Functional validation experiments will be performed using CRISPR/Cas9-mediated genome engineering, hiPSC-based neurogenesis, and mosaic mouse models generated by in utero electroporation (Fig. 3). The BSMN is initially determining concordance among disparate sequencing and bioinformatic approaches by performing a “common experiment” in which pulsed tissue from one neurotypical individual in the Lieber brain repository has been distributed to all of the working groups for independent assessment of mosaicism. The BSMN will generate an estimated 10,000 sequencing data sets that comprise >600 terabytes of data and facilitate data-sharing through the BSMN Knowledge Portal (www.synapse.org/bsmn) and the NIH MIMH Data Archive (https://data-archive. nimh.nih.gov). Coordinated analyses with data derived from some of the same brain samples by the CommonMind (www.synapse.org/cmn) and PsychENCODE (www.synapse.org/pec) initiatives may elucidate the effect of somatic mosaicism on tissue-wide gene expression. Data generated through the BSMN initiative will be released to the broader research community on an ongoing basis through a controlled-access mechanism that follows NIH policies and regulatory requirements.

**REFERENCES AND NOTES**


Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network

Science 356 (6336), . [doi: 10.1126/science.aal1641]

Editor's Summary

**Single-cell diversity in the brain**

The cells that make up an organism may all start from one genome, but somatic mutations mean that somewhere along the line of development, an organism's individual cellular genomes diverge. McConnell et al. review the implications and causes of single-cell genomic diversity for brain function. Somatic mutations caused by mobile genetic elements or errors in DNA repair may underlie certain neuropsychiatric disorders.

*Science*, this issue p. eaal1641

---

This copy is for your personal, non-commercial use only.

**Article Tools**

Visit the online version of this article to access the personalization and article tools:

http://science.sciencemag.org/content/356/6336/eaal1641

**Permissions**

Obtain information about reproducing this article:

http://www.sciencemag.org/about/permissions.dtl