Multi-organ analysis of low-level somatic mosaicism reveals stage- and tissue-specific mutational features in human development

Hyeonju Son[1,†], Ja Hye Kim[2,†], Il Bin Kim[2,3], Myeong-Heui Kim[2 4], Nam Suk Sim[2,5], Dong-Seok Kim[6], Junehawk Lee[7], Jeong Ho Lee[2,4,*], and Sangwoo Kim[1,*]

[1]Department of Biomedical Systems Informatics, Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Republic of Korea.

[2]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea.

[3]Department of Psychiatry, Hanyang University Guri Hospital, Guri, Republic of Korea

[4]SoVarGen Inc., Daejeon, Republic of Korea

[5]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Republic of Korea.

[6]Department of Neurosurgery, Yonsei University College of Medicine, Seoul, Republic of Korea.

[7]Center for Supercomputing Applications, National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information, Daejeon, Republic of Korea

† These authors are contributed equally to work.

* Correspondence should be addressed to:

E-mail: jhlee4246@kaist.ac.kr and swkim@yuhs.ac

Most somatic mutations arising during normal development present as low-level in single or multiple tissues depending on the developmental stage and affected organs[1-4]. However, it remains unclear how the human developmental stages or mutation-carrying organs affect somatic mutations' features. Here, we performed a systemic and comprehensive analysis of low-level somatic mutations using deep whole-exome sequencing (WES; average read depth: ~500×) of 498 multiple organ tissues with matched controls from 190 individuals. We found that early-stage mutations shared between multiple organs are lower in number but showed higher allele frequencies than late-stage mutations [0.54 vs. 5.83 variants per individual: 6.17% vs. 1.5% variant allele frequency (VAF)] along with less nonsynonymous mutations and lower functional impacts. Additionally, early- and late-stage mutations had unique mutational signatures distinct from tumor-originate mutations. Compared with early-stage mutations presenting a clock-like signature across all studied organs or tissues, late-stage mutations show organ, tissue, and cell-type specificity in mutation count, VAFs, and mutational signatures. In particular, analysis of brain somatic mutations shows bimodal occurrence and temporal-lobe-specific mutational signatures. These findings provide new insight into the features of somatic mosaicisms dependent on developmental stages and brain regions.

Somatic mutations persistently occur in normal cells during the entire human lifetime[1]. Although unaccompanied with unregulated proliferation, as seen in cancer, these somatic mutations often present a degree of clonality depending on time and origin. For example, variants in the early stages of development tend to affect multiple organs of different germ layers and show high variant allele frequencies (VAFs), whereas those in later stages localize with low VAFs[5,6]. Somatic variants that occur after birth are theoretically transient and

restricted in a cellular level; however, mutations in stem or progenitor cells[7] or variants that confer clonal expansion[8] are persistent and accumulate during a lifetime and manifest a sufficient level of VAFs detectible in bulk-genome sequencing of tissues. Specifically, these tissue-level somatic mutations are crucial for the pathogenicity of non-cancerous or benign diseases, and the magnitude of aberrations is associated with their allele frequencies[9,10]. For example, mTOR-pathway-activating somatic mutations cause two types of intractable epilepsy (hemimegalencephaly and focal cortical dysplasia) depending on the time of mutation occurrence and VAFs (10–30% of VAFs in hemimegalencephaly, and 1–10% of VAFs in focal cortical dysplasia)[11-13]. Despite advances in the genetic identification of specific diseases, it still remains unclear how low-level but clone-forming somatic mosaicisms are generally characterized by the time and locations of their occurrence.

To address the questions, we performed a comprehensive analysis of low-level somatic mutations found in data from deep whole-exome sequencing (WES) of 498 tissues from 190 individuals (average read depth: ~500×) **(Fig. 1a)**. The cohort consisted of multiple organs, including brain (*n*=301), blood (*n*=100), liver (*n*=60), heart (*n*=13), and other peripheral tissues (*n*=24). The 190 individuals included patients with 'non-tumor' neurological disorders (*n*=133), brain tumors (glioblastoma and ganglioglioma, *n*=19), and non-diseased controls (*n*=38) **(Supplementary Table 1)**. This cohort enabled multi-dimensional analysis and specifically a direct comparison with cancer mutations identified from a same analysis procedure.

Regarding somatic mutations, we defined and used three different categories in the analysis: early-stage, late-stage, and tumor mutations **(Fig. 1b)**. Early-stage mutations were defined as mutations occurring during early embryonic development prior to gastrulation and

shared in multiple-organs, whereas late-stage mutations included late embryonic (post-gastrulation) and post-natal somatic mutations restricted in a single organ. Based on the definition, somatic mutation calling was conducted using an ensemble of robust variant callers: Mutect2[14], RePlow[15], and NeuSomatic[16] for 1,034 possible combinations of sample pairs. After strict filtration **(Fig. 1a)** and tests for organ specificity, we detected 103 early- and 997 late-stage mutations, as well as 583 tumor mutations. To validate the calls, 114 randomly selected single nucleotide variants (SNVs; ~10% of non-tumor mutations) were sequenced by targeted amplicon sequencing (TASeq) to ultra-high depth (average: 507,856×) and Sanger sequencing. Our call set achieved high precision in both the early-stage (89.47%, 17/19) and late-stage and tumor mutations (90.24%, 74/82) **(Fig. 1c and Supplementary Table 2)**. High concordance in VAFs across tissues (Pearson's correlation r=0.84; P=1.00×10^{-42}) and between WES and TASeq data (r=0.61; P=1.17×10^{-48}) confirmed the confidence of the calls **(Fig. 1d)**.

Additionally, we compared the quantitative traits of the mutations in terms of the number and allele frequency at different stages. On average, there were 0.54 early- and 5.83 late-stage somatic mutations per individual **(Fig. 2a)**. These numbers are roughly comparable to those of previous studies, which reported 0.53 shared and 3.15 non-shared somatic mutations in the brain (numbers were normalized to genomic size of 50 Mbp from whole-genome sequencing)[17,18]. It is possible that a slight increase in the number of late-stage (non-shared) might be due to the inclusion of blood samples, which are known to harbor ~3-fold more mutations than other peripheral tissues[6]. Apparently, the numbers of mutations in normal tissues were substantially lower than those of tumors (30.00 per individual). The overall numbers of the late-stage and tumor mutations positively correlated with age

(Pearson's *r:* late-stage, 0.44; and tumor, 0.4) (**Fig. 2b)**. Conversely, we found no correlation between early-stage mutations and age, confirming that these mutations are well confined to the designated period. Regarding indels, 0.047 and 0.68 somatic indels were found in the early- and late-stage per individual, respectively **(Fig. 2a)**. The proportion of indels in the early-stage (~8.7%) was slightly lower than that in the late-stage (~11.7%) and tumors (~10.7%). Because indels are more likely to be functionally damaging[19], these results might represent lower tolerance to damaging mutations in the early developmental phase. On the other hand, VAFs of the mutations were higher in the early-stage (6.17 ± 3.32%) relative to the late-stage (1.50 ± 3.29%), which is consistent with the general expectation that somatic mutations that arise earlier present higher VAFs **(Fig. 2c)**. VAFs of early-stage somatic mutations have been measured in several studies with different criteria for inclusion and presented a diverse range (0.3–55%)[6,18,20]. Because none of the studies directly observed multi-organ-shared mutations using matched tissue sets from the same individuals, our analysis provides a more realistic distribution of VAFs for mutations occurring before gastrulation. Notably, VAFs of somatic indels in the early-stage were lower than those of somatic SNVs (indels vs. SNVs: 4.00% vs. 6.40%) but higher in the late-stage and tumors (2.75% vs. 1.34% in the late-stage; and 18.47% vs. 14.78% in tumors). The lower VAFs of indels, which represents lower cellular proportion and later occurrence, might be also associated with lower tolerance to damaging mutations in the early-developmental phase.

We then conducted mutation-profile analysis to investigate the underlying mutagenic processes **(Fig. 3a–d)**. *De novo* signature extraction of the 1,494 somatic SNVs (94 early-stage, 880 late-stage, and 520 tumor SNVs) identified three novel signatures **(Fig. 3a)**: signatures A, B1, and B2, all of which exhibited C>T as the major base substitution while

showing additional T>C enrichment in signature A. Despite the overall similarity in mutational spectrum, especially between B1 and B2 (cosine similarity: 0.95), the clear distinction shown in the relative contribution to the sample groups confirmed the uniqueness of the signatures (*i.e.*, signatures A, B1, and B2 dominantly contributed to the early-, late-stage, and tumor SNVs, respectively) **(Fig. 3b)**. This also implies that somatic mutations from different stages have distinguishing contexts. Mapping of the three signatures to COSMIC Mutational Signatures (v3.1; June 2020)[21] identified clock-like SNV (SBS1, SBS5, and SBS40), and indel signatures (ID1, ID2, ID5, and ID8) as major components **(Fig. 3c)**. We noted that the relative contribution of the two well-known age-related signatures (SBS1 and SBS5) was altered from early- to late-stage SNVs (SBS1: 19% to 29%; and SBS5: 78% to 49%). The increased relative portion of SBS1 in late-stage somatic mutations appears to represent active proliferation and clonal expansion during late-embryonic and post-natal or aging periods[22,23]. Although the etiologies associated with most indel signatures remain unknown, the higher contributions of ID1 and ID2 in early-stage SNVs and ID5 and ID8 in late-stage SNVs were consistent with a previous finding[24].

Further assessment revealed differences between the early- and late-stage mutations in functional aspects. We found that early-stage mutations showed a lower ratio of non-synonymous to synonymous substitutions (dN/dS) (0.79) than did late-stage mutations (0.94), tumor mutations (0.94), and common germ-line coding variants (0.90; gnomAD Exome) **(Fig. 3d)**, indicating a stronger negative selection[25]. Additionally, early-stage mutations were less frequently (2.1%) located in trinucleotides with atypical mutability[26,27] than were late-stage mutations (8.0%), tumor mutations (8.2%), and common germ-line coding variants (9.9%) **(Fig. 3e)**. Sites with atypical mutability are more highly mutated in cancer than is expected to

occur randomly, indicating their functional significance and driverness in cancer[27,28]. Furthermore, genes that harbor early-stage mutations were lower in the probability of loss-of-function (LoF) intolerance (pLI score)[29] **(Fig. 3f)**, indicating that early-stage mutations are more enriched in LoF-tolerant genes. These results collectively implied the strong selective pressure in the early embryonic stage[30,31] that affects overall mutation characteristics that are less damaging possibly through the rejection of functionally-deleterious mutations.

We then investigated the characteristics of late-stage mutations, with a particular focus on diversity among organs and cell types. The numbers of mutations varied substantially by organ, with a smaller number in the brain (0.77 per individual) and higher number in the blood (9.24 per individual) relative to other peripheral organs (average: 1.13 per individual) **(Fig. 4a)**. However, the average number of VAFs was inversely proportional, with the highest number in the brain (7.32%) and the lowest in the blood (0.50%) **(Fig. 4b)**. Because VAFs generally decrease by the time of occurrence, we speculated that clonal somatic mutations in the brain occur relatively earlier but less frequently than those in the blood and other organs. The number of late-stage somatic mutations and the age of individuals showed a significant positive correlation (r=0.5; p=1.48×10^{-6}) in only the blood **(Fig. 4c)**, which has been well-documented by post-natal clonal hematopoiesis[32,33]. Moreover, unsupervised hierarchical clustering of the three signatures (A, B1, and B2) of the late-stage mutations identified that those of the brain primarily comprise signatures A (early-stage) and B2 (tumor), whereas blood mutations are closer to signature B1 (late-stage) and B2 (tumor) **(Fig. 4d)**. These results suggest that late-stage somatic mutations in the brain present a bimodal-like occurrence during the embryonic period shortly after gastrulation and the post-natal period accompanied by a tumor-originating mutational signature.

We then investigated the bimodal-like characteristics of the late-stage somatic mutations in the brain. First, we assessed the cell-type specificity of the somatic mutations in the brain by selecting two brain samples, which included one (NLE-P-0150) containing an early-stage mutation (5.47% VAFs) and the other (NLE-P-0225) five late-stage mutations (average: 8.00% VAF) **(Fig. 4e)**, each of which was sorted by fluorescence-activated nuclei sorting (FANS) to isolate three different cell types: neuronal ($NeuN^+$), oligogenic ($Olig2^+$), and others (negative). TASeq of the separated cell populations revealed that both early- and late-stage mutations are present in multiple cell lineages, but a large asymmetry of mutation frequencies among cell-types exists in the late-stage mutations **(Fig. 4e)**. These findings imply that the late-stage mutations in the brain occur later than the embryonic phase but relatively earlier during development in order to affect multiple lineages. We then subdivided the late-stage brain mutations into temporal and non-temporal areas and analyzed area-specific mutation signatures **(Fig. 4f)**. As previously reported, contributions to both areas were mainly from signatures A and B2; however, the degree of contribution of signature B2 was higher in the temporal lobe (70.3%) than non-temporal tissue (25.7%), revealing that the characteristics of somatic mutations in the temporal lobe are closer to those of tumor mutations. We speculated that the tumor-like mutational signatures in the temporal lobe might originate from neurogenesis activity (e.g., dentate gyrus) that confers clonal proliferation, as reported previously[34]. Furthermore, the strand specificity of the late-stage mutations in blood and tumor mutations showed enrichment of T>C mutations on transcribed strands **(Fig. 4g)**. Because transcription coupled repair occurs more frequently with higher transcription levels and this bias is increased in actively replicating templates[35,36], we again confirmed that clonal expansion-derived somatic mutations were included in the blood, similar to those in tumors[37].

In summary, based on a large scale of deep whole exome sequencing data using a total of 498 matched sample pairs from multiple organs in 190 individuals, we provided a more detailed picture of low-level but clone-forming somatic mutations, the counts, and characteristics of which are distinguished by time and space. We found that early-stage mutations, which arise prior to gastrulation and are shared in multiple organs, are lower in number and have lower functional impact than late-stage mutations restricted within a single organ. Moreover, we showed that late-stage mutations are associated with human mutational processes in the late-embryonic and post-natal developmental stages but that vary by organ, tissue, and cell lineages. In particular, late-stage mutations in the brain showed a bimodal-like occurrence over developmental stages and asymmetry of mutational features across brain-cell types and regions. Regarding the asymmetry of somatic mutations, asymmetric cell divisions resulting from early cellular bottlenecks of stochastic clonal selection contributed to an uneven variant fraction according to developmental timing[6,38]. These findings suggest that the VAFs of clone-forming somatic mutations reflect not only the timing of the mutation but also the cell fitness and cell-type specificity for given somatic mutations. Overall, the well-defined characteristics of each mutation group and target tissue according to their developmental period can confer an accurate representation of currently-observable somatic mutations and a better understanding of how they were generated.

Methods

Patient samples

The acquired freshly frozen brain and peripheral samples of 24 autism spectrum disorder (ASD) and five non-ASD cases from the National Institute of Child Health & Human Development (Bethesda, MD, USA) included various brain regions, such as the frontal, temporal, occipital, and cerebellar areas. Paired samples with other organs were derived from 13 ASD cases and five non-ASD case, and brain samples were obtained from 11 ASD cases. The Stanley Medical Research Institute (Rockville, MD, USA) supplied genomic DNA of brain tissue and other matched organs for 25 non-schizophrenia and 26 schizophrenia cases. Additionally, the Stanley Medical Research Institute provided genomic DNA for brain and matched liver tissues from patients with major depressive disorders. Fresh frozen brain samples of Alzheimer's disease (AD) were provided from the Netherlands Brain Bank (project number Lee-835) for 96 brain and matched blood samples for AD and non-demented control cases, and 15 samples of AD and non-demented control cases were obtained from the Human Brain and Spinal Fluid Resource Center (West Los Angeles Healthcare Center, Los Angeles, CA, USA), which is sponsored by NINDS/NIMH (Bethesda, MD, USA), the National Multiple Sclerosis Society (Raleigh, NC, USA), and the US Department of Veterans Affairs (Bethesda, MD, USA). Fresh frozen samples of lumbosacral lipoma were supplied from the Severance Children's Hospital of Yonsei University College of Medicine (Seoul, Republic of Korea). Bone tissues of non-syndromic craniosynostosis patients were provided from the Severance Hospital of Yonsei University College of Medicine. Subjects with refractory epilepsy, including focal cortical dysplasia and non-lesional epilepsy, and who had

undergone epilepsy surgery were enrolled through the Severance Children's Hospital of Yonsei University College of Medicine. Subjects with glioblastoma and ganglioglioma were enrolled from the Severance Hospital of Yonsei University College of Medicine and satisfied diagnostic criteria according to the 2016 World Health Organization Classification of Tumors of the Central Nervous System[39]. We were provided freshly-frozen samples of resected brain lesions.

Deep WES

Genomic DNA was extracted with either the QIAamp mini DNA kit (Qiagen, Hilden, Germany) from freshly frozen brain tissues or the Wizard genomic DNA purification kit (Promega, Madison, WI, USA) from blood according to manufacturer instructions. Each sample was prepared according to Agilent library preparation protocols (Human All Exon 50 Mb kit; Agilent Technologies, Santa Clara, CA, USA). Libraries were subjected to paired-end sequencing on an Illumina Hiseq 2000 and 2500 instrument (Illumina, San Diego, CA, USA) according to the manufacturer's instructions) with confidence-mapping quality (mapping quality score ≥ 20; base quality score ≥ 20).

Data processing and systematic variant calling

We checked the quality of the raw sequencing reads using FastQC[40] (v.0.11.7) software. The FASTQ-formatted sequencing reads of each sample that passed the quality check were aligned to the human reference genome (build 38; NCBI, Bethesda, MD, USA) using the BWA-MEM[41] algorithm and converted into a BAM file. The initial BAM file was updated with read groups, and duplicate information was excluded as it progressed through the steps

using Picard[42] and GATK[43]. Additionally, we performed local realignment and base-quality recalibration with GATK tools for each exome. BAM files that successfully underwent all of these steps were then used to measure contamination between samples, with the probability of swapping assessed using NGSCheckMate[44] software and cross-contamination tested using GATK tools. Vecuum[45] software was used to check for vector contamination during library construction, and Depth of Coverage (GATK) was used to measure sequencing depths. All processes not described in detail were performed based on the GATK best-practice pipeline.

Two or more tissue samples from each individual were paired using all-pairs testing. We performed the somatic mutation-detection pipeline (paired mode) with sample pairs as inputs using a three somatic variant caller; Mutect2[14] somatic variant-calling pipeline, excluding the panel of normal creation (SNVs and Indels), RePlow[15] (SNVs), and NeuSomatic[16] with the control of the false detection rate control performed by Varlociraptor[46] (Indels).

All mutations meeting the following conditions were removed from the initial mutation-detection results in the VCF format: oxoG-induced errors according to the method described by Costello et al.[47], common single-nucleotide polymorphisms by NCBI dbSNP[48] (build 153), segmental duplication and simple repeat regions according to the UCSC database[49], a mappability score >0.8 by Umap[50], and presence of an off-target region[51] whole genome without exome and the untranslated region.

Decisions regarding early and late mutations

After the removal of artefacts, somatic mutations with "PASS" results for both Mutect2 and other caller filters (RePlow/NeuSomatic) were classified as late-stage mutations. If the source of the sample was related to a brain tumor, it was separately regarded as a tumor mutation.

Early-stage mutations were initially categorized as such if the filter result of Mutect2 was “normal artifact” or RePlow (for only SNVs) was “normalFilter,” respectively. Additionally, these were assigned this category if they were called in Mutect2 only but not in RePlow. After confirming amino acid changes and genomic location, to confirm that the same mutation was detected from each individual, the validity of the mutation was statistically verified using the one-sample proportion test. The VAFs of each mutation were used as a criterion to determine whether the ratio of the ‘ref’ and ‘alt’ alleles of the other mutations satisfied the null hypothesis. Common mutations in different samples from each individual were tested, and mutations satisfying the criteria were classified as early-stage mutations.

Validation sequencing of candidate mutations using deep-targeted amplicon sequencing or Sanger sequencing

We then performed validation sequencing by randomly selecting mutations from each group. For validation, we used deep-targeted amplicon sequencing or Sanger sequencing of PCR-amplified DNA. Primers for PCR amplification were designed using Primer3 software (http://bioinfo.ut.ee/primer3-0.4.0/)[52]. Target regions were amplified by PCR using specific primer sets and high-fidelity PrimeSTAR GXL DNA polymerase (Takara, Shiga, Japan).

Sanger sequencing was performed using BigDye Terminator reactions and loaded onto a 3730xl DNA analyzer (Applied Biosystems, San Francisco, CA, USA).

Bioinformatics analysis

All somatic mutations excluded false positives by validation sequencing were annotated using VEP[53] (v.99.0) with "-everything -plugin ExACpLI" options. The results were evaluated using an in-house script to analyze the descriptive statistics of the properties of the basic mutations, effect of each gene, and possible correlations with patient demographics (age, disease, etc.). Non-negative matrix factorization-based novel signature extraction (200 iterations) and transcriptional strand-bias analysis were performed using the MutationalPatterns program[54]. The signature and 96-types variant contexts were fitted to clockwise Pan-Cancer Analysis of Whole Genomes (PCAWG) single-base substitution and small insertions and deletions signatures by deconstructSigs[55], Mutalisk[56] (date of use: March 2020), and YAPSA[57]. Mutability was calculated using NCBI MutaGene[26,27] (v.0.9.1.0) distributed as a Python package. The maximum-likelihood dN/dS method was applied by dNdScv (Wellcome Sanger Institute, Cambridge, UK)[25].

Nuclei extraction and FANS

Frozen brain samples were minced using pre-chilled razor blades and one or two drops of lysis buffer [0.2% Triton X-100, 1× protease inhibitor, and 1 mM DTT in 2% bovine serum albumin (BSA) in phosphate-buffered saline]. Lysis buffer (1 mL) was added to the homogenate and mixed by pipetting, after which the lysate was fixed in 1% paraformaldehyde at room temperature for 10 min, and the fixed lysate was quenched with

316 0.125 M glycine at room temperature for 5 min. The homogenate was then washed with
317 suspension buffer (1 mM EDTA and 2% BSA) and filtered with 40-µM cell strainer. The
318 sample was then incubated with anti-NeuN (mature neuronal marker; 1:1000) and anti-Olig2
319 (oligodendrocyte lineage marker; 1:500) overnight at 4°C, followed by washing with
320 suspension buffer and staining with the secondary antibody for 1 h at 4°C. After washing with
321 suspension buffer, nuclei were passed through a 40-µM cell strainer and stained with 1 µg
322 4′,6-diamidino-2-phenylindole. Nuclei used to isolate each cell type were analyzed and sorted
323 using a MoFlo Astrios EQ cell sorter (Beckman Coulter, Brea, CA, USA). Nuclei pellets
324 were centrifuged for 10 min at 1500*g* and processed immediately for gDNA extraction using
325 a QIAamp DNA micro kit (Qiagen) according to manufacturer instructions.
326
327

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2008050 to S. K.), the Suh Kyungbae Foundation (to J.H.L.), and a National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and Information and Communication Technology (ICT) (No. 2019R1A3B2066619 to J.H.L)

We thank the Netherlands Brain Bank (Lee-835) for Alzheimer and unaffected control cases., the National Institute of Child Health & Human Development for providing Autism and unaffected control cases, the Stanley Medical Research Institute for the brain and peripheral DNA provided of schizophrenia, major depressive disorders, and unaffected control cases, Seoul National University Hospital, Seoul National University College of Medicine for providing lumbosacral lipoma and non-syndromic craniosynostosis, and Severance Hospital, Yonsei University College of Medicine for providing samples of brain tumor and epilepsy, which were supplied to J.H.L.

Author contributions

S.K. and J.H.L designed and initiate the study. H.S. and J.H.K conducted main analysis. H.S. devised analysis pipeline and performed bioinformatics analysis. J.H.K. worked on sample organization, validation sequencing, and FANS. I.B.K, M-H.K., and N.S.S. prepped human tissue samples and performed whole-exome sequencing. D-S.K. performed the epilepsy surgeries, collected patient samples, and managed patient information. J.L. worked on analysis of sequencing data. H.S., J.H.K., J.H.L., and S.K. worked on data interpretation, and wrote the manuscript with input from coauthors. J.H.L and S.K. led the project.

352

353 **Competing interests**

354 J.H.L. is co-founder and CTO of SoVarGen Inc., which seeks to develop new diagnostics and

355 therapeutics for brain disorders. The other authors declare no competing interests.

356

357

References

1 Frank, S. A. Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 1**, 1725-1730, doi:10.1073/pnas.0909343106 (2010).

2 Freed, D., Stevens, E. L. & Pevsner, J. Somatic mosaicism in the human genome. *Genes (Basel)* **5**, 1064-1094, doi:10.3390/genes5041064 (2014).

3 Gajecka, M. Unrevealed mosaicism in the next-generation sequencing era. *Mol Genet Genomics* **291**, 513-530, doi:10.1007/s00438-015-1130-7 (2016).

4 Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting Somatic Mutations in Normal Cells. *Trends in genetics : TIG* **34**, 545-557, doi:10.1016/j.tig.2018.04.003 (2018).

5 Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422-425, doi:10.1038/nature13448 (2014).

6 Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718, doi:10.1038/nature21703 (2017).

7 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

8 Kakiuchi, N. & Ogawa, S. Clonal expansion in non-cancer tissues. *Nature Reviews Cancer* **21**, 239-256, doi:10.1038/s41568-021-00335-3 (2021).

9 D'Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci* **21**, 1504-1514, doi:10.1038/s41593-018-0257-3 (2018).

10 Mustjoki, S. & Young, N. S. Somatic Mutations in "Benign" Disease. *New England Journal of Medicine* **384**, 2039-2052, doi:10.1056/NEJMra2101920 (2021).

11 Sim, N. S. *et al.* Precise detection of low-level somatic mutation in resected epilepsy brain tissue. *Acta Neuropathol* **138**, 901-912, doi:10.1007/s00401-019-02052-6 (2019).

12 Baldassari, S. *et al.* Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. *Acta Neuropathol* **138**, 885-900, doi:10.1007/s00401-019-02061-5 (2019).

13 D'Gama, A. M. *et al.* Somatic Mutations Activating the mTOR Pathway in Dorsal Telencephalic Progenitors Cause a Continuum of Cortical Dysplasias. *Cell Rep* **21**, 3754-3766, doi:10.1016/j.celrep.2017.11.106 (2017).

14 Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054, doi:10.1101/861054 (2019).

15 Kim, J. *et al.* The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. *Nature communications* **10**, 1047, doi:10.1038/s41467-019-09026-y (2019).

16 Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nature communications* **10**, 1041, doi:10.1038/s41467-019-09027-x (2019).

17 Rodin, R. E. *et al.* The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nat Neurosci* **24**, 176-185, doi:10.1038/s41593-020-00765-6 (2021).

18 Breuss, M. W. *et al.* Somatic mosaicism in the mature brain reveals clonal cellular distributions during cortical development. *bioRxiv*, 2020.2008.2010.244814,

doi:10.1101/2020.08.10.244814 (2020).

19 Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics* **19**, R131-R136, doi:10.1093/hmg/ddq400 (2010).

20 Bae, T. *et al.* Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science (New York, N.Y.)* **359**, 550-555, doi:10.1126/science.aan8690 (2018).

21 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).

22 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

23 Lee, J. H. *et al.* Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243-247, doi:10.1038/s41586-018-0389-3 (2018).

24 Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *bioRxiv*, 2020.2011.2023.395244, doi:10.1101/2020.11.23.395244 (2020).

25 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e1021, doi:10.1016/j.cell.2017.09.042 (2017).

26 Goncearenco, A. *et al.* Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic acids research* **45**, W514-w522, doi:10.1093/nar/gkx367 (2017).

27 Brown, A. L., Li, M., Goncearenco, A. & Panchenko, A. R. Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLoS Comput Biol* **15**, e1006981, doi:10.1371/journal.pcbi.1006981 (2019).

28 Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context.

Nature Genetics **52**, 208-218, doi:10.1038/s41588-019-0572-y (2020).

29 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

30 Khokhlova, E. V., Fesenko, Z. S., Sopova, J. V. & Leonova, E. I. Features of DNA Repair in the Early Stages of Mammalian Embryonic Development. *Genes (Basel)* **11**, doi:10.3390/genes11101138 (2020).

31 Kermi, C., Aze, A. & Maiorano, D. Preserving Genome Integrity During the Early Embryonic DNA Replication Cycles. *Genes (Basel)* **10**, doi:10.3390/genes10050398 (2019).

32 de Haan, G. & Lazare, S. S. Aging of hematopoietic stem cells. *Blood* **131**, 479-487, doi:10.1182/blood-2017-06-746412 (2018).

33 Natarajan, P., Jaiswal, S. & Kathiresan, S. Clonal Hematopoiesis: Somatic Mutations in Blood Cells and Atherosclerosis. *Circ Genom Precis Med* **11**, e001926, doi:10.1161/circgen.118.001926 (2018).

34 Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (New York, N.Y.)* **359**, 555-559, doi:10.1126/science.aao4426 (2018).

35 Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-549, doi:10.1016/j.cell.2015.12.050 (2016).

36 Brachman, E. E. & Kmiec, E. B. DNA replication and transcription direct a DNA strand bias in the process of targeted gene repair in mammalian cells. *J Cell Sci* **117**, 3867-3874, doi:10.1242/jcs.01250 (2004).

37 Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related

Mutagenesis in Human Hematopoiesis. *Cell Reports* **25**, 2308-2316.e2304, doi:https://doi.org/10.1016/j.celrep.2018.11.014 (2018).

38 Bizzotto, S. *et al.* Landmarks of human embryonic development inscribed in somatic mutations. *Science (New York, N.Y.)* **371**, 1249-1253, doi:10.1126/science.abe1544 (2021).

39 Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803-820, doi:10.1007/s00401-016-1545-1 (2016).

40 Andrews, S. *FastQC: a quality control tool for high throughput sequence data.*, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).

41 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

42 *Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*, <http://broadinstitute.github.io/picard/> (2020).

43 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics* **43**, 11.10.11-33, doi:10.1002/0471250953.bi1110s43 (2013).

44 Lee, S. *et al.* NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. *Nucleic acids research* **45**, e103, doi:10.1093/nar/gkx193 (2017).

45 Kim, J. *et al.* Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics (Oxford, England)* **32**, 3072-3080,

doi:10.1093/bioinformatics/btw383 (2016).

46 Köster, J., Dijkstra, L. J., Marschall, T. & Schönhuth, A. Varlociraptor: enhancing sensitivity and controlling false discovery rate in somatic indel discovery. *Genome biology* **21**, 98, doi:10.1186/s13059-020-01993-6 (2020).

47 Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41**, e67-e67, doi:10.1093/nar/gks1443 (2013).

48 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **47**, D23-D28, doi:10.1093/nar/gky1069 (2018).

49 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

50 Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic acids research* **46**, e120-e120, doi:10.1093/nar/gky677 (2018).

51 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).

52 Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* **35**, W71-74, doi:10.1093/nar/gkm306 (2007).

53 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

54 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine* **10**, 33, doi:10.1186/s13073-018-0539-0 (2018).

55 Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome biology* **17**, 31, doi:10.1186/s13059-016-0893-4 (2016).

56 Lee, J. *et al.* Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic acids research* **46**, W102-W108, doi:10.1093/nar/gky406 (2018).

57 Hübschmann, D. *et al.* Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer*, doi:10.1002/gcc.22918 (2020).

513 **Figure legends**

514 **Figure 1.** Detection of early- and late-stage somatic variants in brain and matched peripheral
515 tissues. **a**, A schematic flow showing the bioinformatics pipelines of 301 brain tissues and
516 197 peripheral tissues from 190 individuals. To find somatic variants, Mutect2 and
517 RePlow/NeuSomatic were used for reciprocal mutation calling by all-pairs testing, followed
518 by post-call filtering. **b, c**, Early-stage, late-stage, and tumor mutations were classified with a
519 highly accurate precision rate (89.47%, early-stage; and 90.24% in late-stage and tumor
520 mutations). **d**, Correlation of VAFs from two matched tissues and WES and TASeq data.
521 VAFs were highly concordant between paired tissues ($r = 0.84$; $P < 0.0001$) and WES and
522 TASeq data ($r = 0.61$; $P < 0.0001$).

523

524 **Figure 2.** Basic descriptive statistics of somatic mutations. **a**, Number of somatic mutations
525 per individual in early-stage, late-stage, and tumor mutation groups. **b**, Age correlation with
526 somatic mutation counts in the groups. **c**, Average VAFs between the three mutation groups.

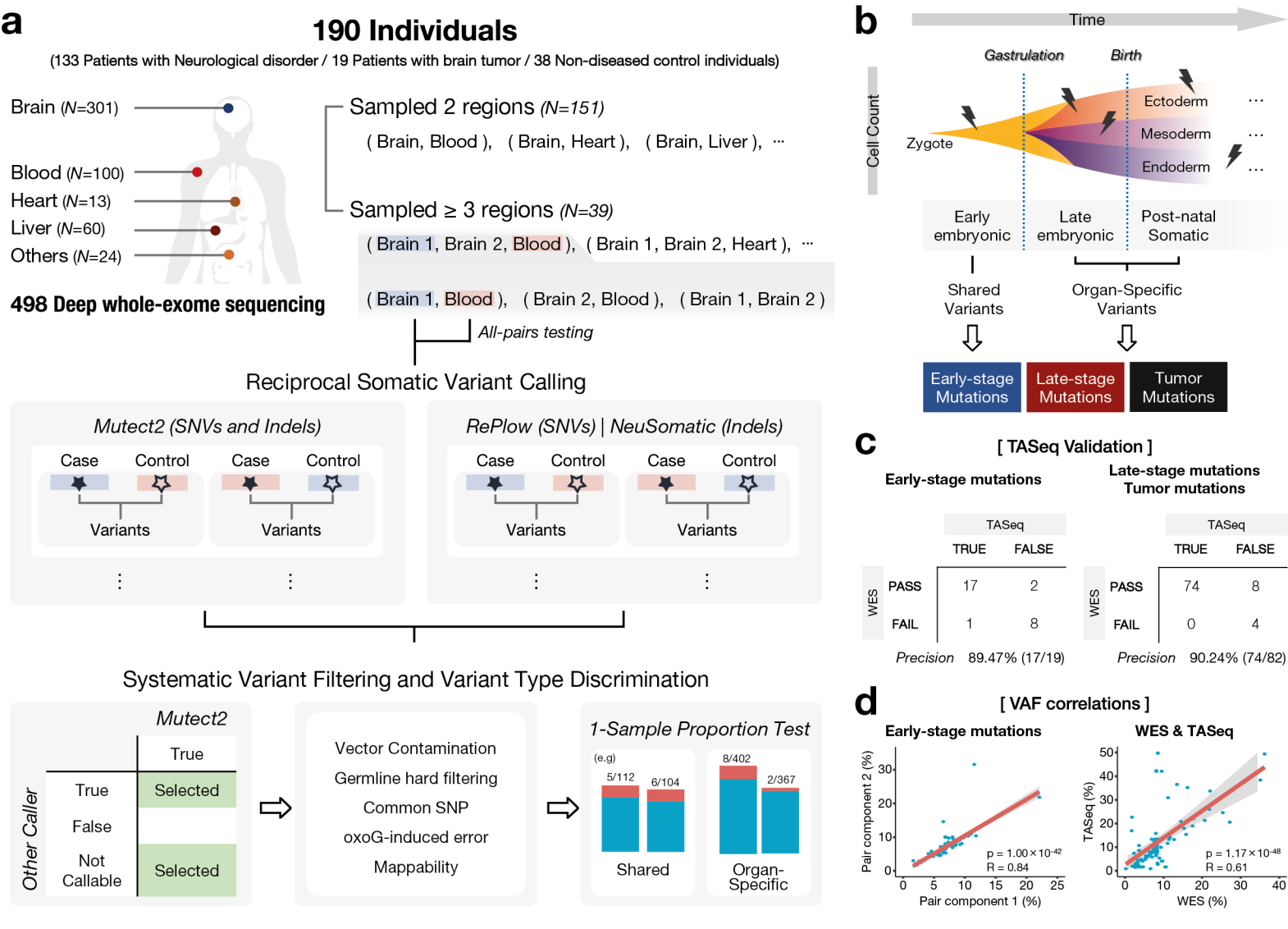
527

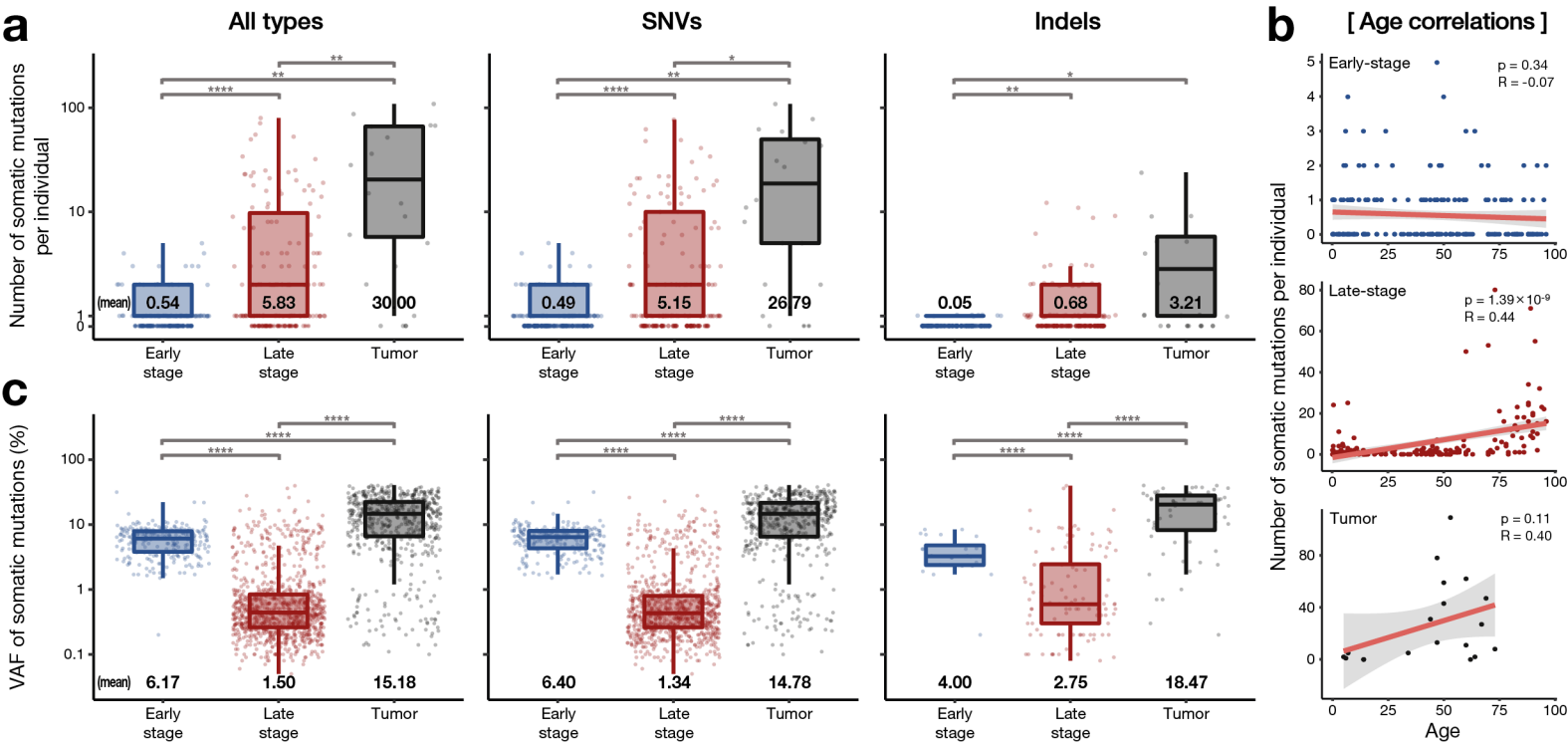
528 **Figure 3.** Mutational profile and functional analysis. **a**, *De novo* extraction of somatic
529 mutations by non-negative matrix factorization. **b**, Each group was classified according to
530 signature (A, B1, and B2). **c**, Relative contribution of the common clock-like signatures
531 (SBS1, SBS5, and SBS40 for single-base substitutions and ID1, ID2, ID5, and ID8 for indels)
532 from PCAWG signatures. **d**, The dN/dS score ratios, **e**, proportion of trinucleotides with
533 atypical mutability, and **f**, pLI score for gnomAD Exome and each group.

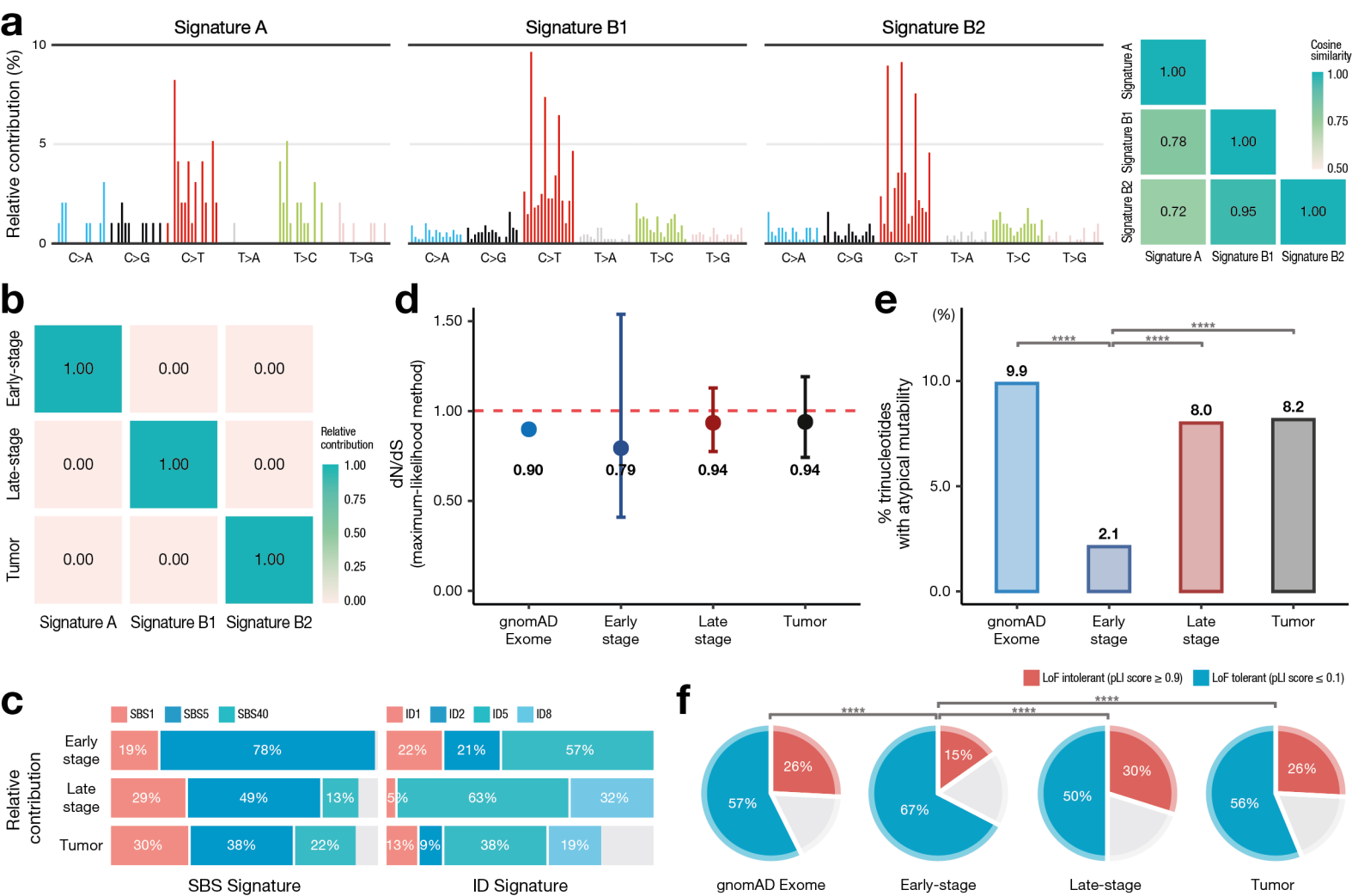
534

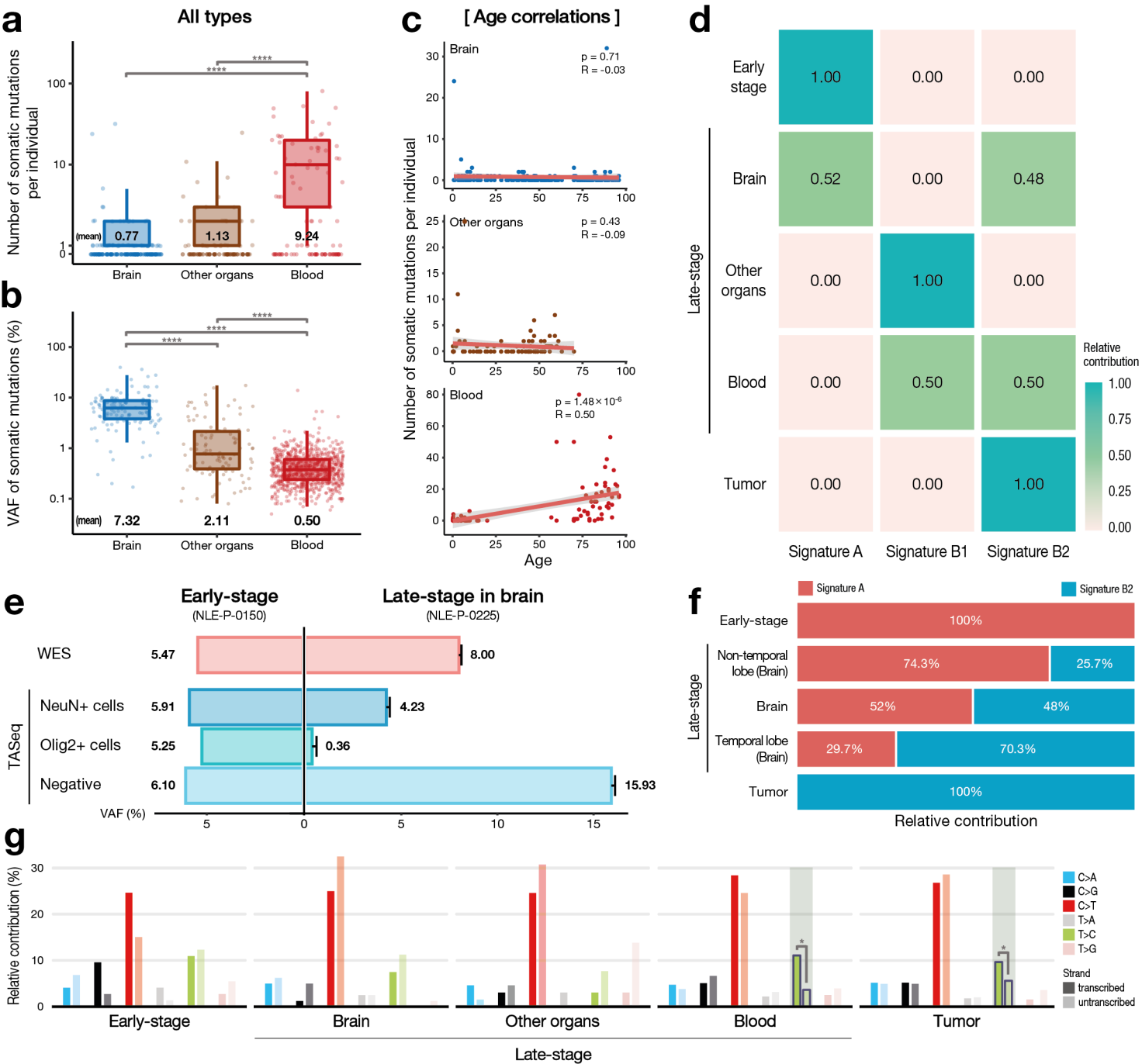
535 **Figure 4.** Analysis of late-stage mutations by mutation source for the organs (brain, blood,
536 and other organs). **a, b**, Number of mutations per individual and VAF distribution. **c**, Age

correlation with mutation counts. **d**, Unsupervised hierarchical clustering of late-stage mutations. Late brain somatic mutations were fit to signatures A and B2, whereas those in the blood were clustered to signatures B1 and B2. **e**, The VAFs of three different cell types [neuronal (NeuN+), oligogenic (Olig2+), and others (negative)] for early-stage and late-stage mutations in the brain. **f**, Signature distribution of late brain somatic variants divided among temporal and non-temporal areas or according to brain-disease status. **g**, Mutational-strand asymmetry. Late-onset blood and tumor mutations are noted as having strand-bias as T>C.


a
190 Individuals
(133 Patients with Neurological disorder / 19 Patients with brain tumor / 38 Non-diseased control individuals)
Brain (N=301)
Blood (N=100)
Heart (N=13)
Liver (N=60)
Others (N=24)
498 Deep whole-exome sequencing
Sampled 2 regions (N=151)
(Brain, Blood), (Brain, Heart), (Brain, Liver), ···
Sampled ≥ 3 regions (N=39)
(Brain 1, Brain 2, Blood), (Brain 1, Brain 2, Heart), ···
(Brain 1, Blood), (Brain 2, Blood), (Brain 1, Brain 2)
All-pairs testing
Reciprocal Somatic Variant Calling
Mutect2 (SNVs and Indels)
Case Control Case Control
Variants Variants
RePlow (SNVs) | NeuSomatic (Indels)
Case Control Case Control
Variants Variants
Systematic Variant Filtering and Variant Type Discrimination
Mutect2
Other Caller
True
True Selected
False
Not Callable Selected
Vector Contamination
Germline hard filtering
Common SNP
oxoG-induced error
Mappability
1-Sample Proportion Test
(e.g)
5/112 6/104
8/402 2/367
Shared
Organ-Specific
b
Time
Gastrulation
Birth
Cell Count
Zygote
Ectoderm
Mesoderm
Endoderm
Early embryonic
Late embryonic
Post-natal Somatic
Shared Variants
Organ-Specific Variants
Early-stage Mutations
Late-stage Mutations
Tumor Mutations
c
[TASeq Validation]
Early-stage mutations
TASeq
TRUE FALSE
WES PASS 17 2
WES FAIL 1 8
Precision 89.47% (17/19)
Late-stage mutations
Tumor mutations
TASeq
TRUE FALSE
WES PASS 74 8
WES FAIL 0 4
Precision 90.24% (74/82)
d
[VAF correlations]
Early-stage mutations
Pair component 2 (%)
Pair component 1 (%)
p = 1.00×10-42
R = 0.84
WES & TASeq
TASeq (%)
WES (%)
p = 1.17×10-48
R = 0.61


a
All types
SNVs
Indels
Number of somatic mutations per individual
(mean)
0.54
5.83
30.00
0.49
5.15
26.79
0.05
0.68
3.21
Early stage
Late stage
Tumor
b
[Age correlations]
Early-stage
p = 0.34
R = -0.07
Late-stage
p = 1.39×10-9
R = 0.44
Tumor
p = 0.11
R = 0.40
Number of somatic mutations per individual
Age
c
VAF of somatic mutations (%)
(mean)
6.17
1.50
15.18
6.40
1.34
14.78
4.00
2.75
18.47


a
Signature A
Signature B1
Signature B2
Relative contribution (%)
10
5
0
C>A
C>G
C>T
T>A
T>C
T>G
Cosine similarity
1.00
0.75
0.50
Signature A
Signature B1
Signature B2
1.00
0.78
1.00
0.72
0.95
1.00
b
Early-stage
Late-stage
Tumor
1.00
0.00
0.00
0.00
1.00
0.00
0.00
0.00
1.00
Relative contribution
1.00
0.75
0.50
0.25
0.00
Signature A
Signature B1
Signature B2
c
Relative contribution
SBS1
SBS5
SBS40
Early stage
19%
78%
Late stage
29%
49%
13%
Tumor
30%
38%
22%
SBS Signature
ID1
ID2
ID5
ID8
22%
21%
57%
5%
63%
32%
13%
9%
38%
19%
ID Signature
d
dN/dS (maximum-likelihood method)
1.50
1.00
0.50
0.00
0.90
0.79
0.94
0.94
gnomAD Exome
Early stage
Late stage
Tumor
e
(%)
% trinucleotides with atypical mutability
10.0
5.0
0.0
****
****
****
****
9.9
2.1
8.0
8.2
gnomAD Exome
Early stage
Late stage
Tumor
f
LoF intolerant (pLI score ≥ 0.9)
LoF tolerant (pLI score ≤ 0.1)
****
****
****
26%
57%
15%
67%
30%
50%
26%
56%
gnomAD Exome
Early-stage
Late-stage
Tumor


a
All types
Number of somatic mutations per individual
(mean) 0.77
1.13
9.24
Brain
Other organs
Blood
b
VAF of somatic mutations (%)
(mean) 7.32
2.11
0.50
c
[Age correlations]
Brain
p = 0.71
R = -0.03
Other organs
p = 0.43
R = -0.09
Blood
p = 1.48 × 10-6
R = 0.50
Number of somatic mutations per individual
Age
d
Early stage
Brain
Other organs
Blood
Tumor
Late-stage
Signature A
Signature B1
Signature B2
Relative contribution
e
Early-stage
(NLE-P-0150)
Late-stage in brain
(NLE-P-0225)
WES
NeuN+ cells
Olig2+ cells
Negative
TASeq
VAF (%)
f
Signature A
Signature B2
Early-stage
Non-temporal lobe (Brain)
Brain
Temporal lobe (Brain)
Tumor
Late-stage
Relative contribution
g
Relative contribution (%)
C>A
C>G
C>T
T>A
T>C
T>G
Strand
transcribed
untranscribed
Early-stage
Brain
Other organs
Blood
Tumor
Late-stage