Science

REPORTS

Cite as: V. B. Seplyarskiy*et al*., *Science* 10.1126/science.aba7408 (2021).

Population sequencing data reveal a compendium of mutational processes in the human germ line

Vladimir B. Seplyarskiy[1,2]†, Ruslan A. Soldatov[2]†, Evan Koch[1,2], Ryan J. McGinty[1,2], Jakob M. Goldmann[3], Ryan D. Hernandez[4,9], Kathleen Barnes[5], Adolfo Correa[6,7,8], Esteban G. Burchard[9,10], Patrick T. Ellinor[11], Stephen T. McGarvey[12,13,14], Braxton D. Mitchell[15,16,17], Ramachandran S. Vasan[18,19], Susan Redline[20,21], Edwin Silverman[22], Scott T. Weiss[20,21,22], Donna K. Arnett[23], John Blangero[24,25], Eric Boerwinkle[26,27], Jiang He[28,29], Courtney Montgomery[30], D.C. Rao[31], Jerome I. Rotter[32], Kent D. Taylor[32], Jennifer A Brody[33], Yii-Der Ida Chen[34], Lisa de las Fuentes[31,35], Chii-Min Hwu[36], Stephen S. Rich[37], Ani W. Manichaikul[37], Josyf C. Mychaleckyj[37], Nicholette D. Palmer[38], Jennifer A. Smith[39,40], Sharon L.R. Kardia[40], Patricia A. Peyser[40], Lawrence F. Bielak[40], Timothy D. O'Connor[41,42,43], Leslie S. Emery[44], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium‡, TOPMed Population Genetics Working Group, Christian Gilissen[3], Wendy S. W. Wong[45], Peter V. Kharchenko[2], Shamil Sunyaev[1,2]*

[1]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands. [4]Quantitative Life Sciences, McGill University, Montreal, QC, Canada. [5]Department of Medicine, University of Colorado Denver, Aurora, CO 80045, USA. [6] Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [7] Department of Pediatrics, University of Mississippi Medical Center, Jackson, MS, USA. [8]Department of Population Health Science, University of Mississippi Medical Center, Jackson, MS, USA. [9]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. [10]Department of Medicine, University of California, San Francisco, CA, USA. [11]Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [12]International Health Institute, Brown University, Providence, RI, USA. [13]Department of Epidemiology, Brown University, Providence, RI, USA. [14]Department of Anthropology, Brown University, Providence, RI, USA. [15]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [16]Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [17]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA. [18]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [19]Framingham Heart Study, Framingham, MA, USA. [20]Department of Medicine, Harvard Medical School, Boston, MA, USA. [21]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [22]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [23]Department of Epidemiology, University of Kentucky, Lexington, KY, USA. [24]Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [25]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [26]University of Texas Health Science Center at Houston, Houston, TX, USA. [27]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA. [28]Department of Epidemiology, Tulane University, New Orleans, LA, USA. [29]Tulane University Translational Science Institute, Tulane University, New Orleans, LA , USA. [30]Division of Genomics and Data Science, Department of Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [31]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA. [32]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [33]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [34]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [35]Department of Medicine, Cardiovascular Division, Washington University School of Medicine, St. Louis, MO, USA. [36]National Yang-Ming University School of Medicine, Taipei, Taiwan. [37]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [38]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [39]Department of Epidemiology, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA. [40]Survey Research Center, Institute for Social Research, University of Michigan 426 Thompson St, Room Ann Arbor, MI 48104, USA. [41]Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [42]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [43]University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center, Baltimore, MD, USA. [44]University of Washington Department of Biostatistics, Seattle, WA 98195, USA. [45]Inova Translational Medicine Institute (ITMI), Inova Health Systems, Falls Church, VA, USA.

†These authors contributed equally to this work. ‡https://www.nhlbiwgs.org/topmed-banner-authorship ; see "Additional Authors from the Trans-Omics for Precision Medicine Program" for full banner author list (excluding primary authors above).*Corresponding author. E-mail: ssunyaev@rics.bwh.harvard.edu

Biological mechanisms underlying human germline mutations remain largely unknown. We statistically decompose variation in the rate and spectra of mutations along the genome using volume-regularized nonnegative matrix factorization. The analysis of a sequencing dataset (TOPMed) reveals nine processes that explain the variation in mutation properties between loci. We provide a biological interpretation for seven of these processes. We associate one process with bulky DNA lesions that resolve asymmetrically with respect to transcription and replication. Two processes track direction of replication fork and replication timing, respectively. We identify a mutagenic effect of active demethylation primarily acting in regulatory regions and a mutagenic effect of LINE repeats. We localize a mutagenic process specific to oocytes from population sequencing data. This process appears transcriptionally asymmetric.

Decades of experimental research in mutagenesis revealed the various error modes of DNA replication and repair (*1*, *2*) but did not elucidate which mechanisms are primarily responsible for human germline mutations. Statistical analysis of sequencing

datasets can quantify contributions of relevant mechanisms.

Cancer genomics has been propelled by the analysis of "mutation signatures" (*3*). Signature extraction relies on differential mutagen exposure of tumor samples and is not transferable to germline mutation, beyond insights from comparing human populations (*4*).

Here, we use variation of mutation rate along the genome to model germline mutagenesis. Our model assumes that several mechanisms generate mutations. We approximate these mechanisms by processes characterized by spectra of 192 tri-nucleotide mutation types and variable intensities along the genome (Fig. 1A). Inference of mutational processes from variability of mutational spectra across genomic loci represents a classic nonnegative matrix factorization (NMF) problem.

NMF separates a set of nonnegative source signals (here, mutational processes) from observed nonnegative signal mixtures (here, mutation frequencies). However, NMF can have many solutions with the same quality of approximation (*5*). To find identifiable solutions, applications in cancer assume that tumors are exposed to few mutagenic forces characterized by unique mutation types (*6*).

Classic NMF performs poorly in our case (fig. S1K), likely because mutagenic forces with complex spectra act in most loci. Recently developed "volume-regularized" NMF (vrnmf) (*6*, *7*) guarantees a unique solution under mild assumptions of sufficiently spread source signals. Vrnmf finds the most distinct (positively defined) mutational spectra (Fig. 1B and fig. S1A). For germline mutations, our implementation of vrnmf (*8*) delivers a unique interpretable solution, outperforming standard NMF in simulations and data (fig. S1, K and C).

A powerful way to assess the biological relevance of the inferred processes is provided by the symmetry between antiparallel strands of DNA. Strand-specific footprints of molecular machineries such as transcription and replication break this symmetry. Mutational mechanisms coupled with these machineries are strand-dependent. For example, A>G mutations are depleted within genes on the transcribed strand due to the action of transcription-coupled repair (TCR) (*2*). Mutation processes uncoupled from the action of strand-specific machineries are strand-independent (Fig. 1C).

We assign mutation types with respect to the genome reference independently of the direction of transcription and replication. For some genes the reference strand is transcribed, while for others it is non-transcribed. As a consequence of TCR, some genic regions display depletion of A>G mutations and others display depletion of the complementary T>C mutations (Fig. 1C).

For a strand-dependent mutational mechanism, our statistical procedure would infer two independent components with reverse-complimentary spectra (Fig. 1C and fig. S1B). Following the example above, the intensity of A>G mutations in one component would be identical to the intensity of T>C in the other. In contrast, a strand-independent mechanism would correspond to a single self-complementary component (the intensity of A>G would be identical to the intensity of T>C). All biologically meaningful components would either be self-complementary or arise in mutually complementary pairs. We disregard spurious processes not conforming to this complementarity rule using the "reflection test" (Fig. 1, C and D) (*8*).

We applied our method to 292 million very rare (allele frequency below 10^{-4}) single nucleotide variants (SNVs) from 42,813 individuals in the TOPMed freeze 5 (*9*). We applied a statistical correction (*8*) to account for multiple independent mutations that occurred in the same site (*10*). This correction increased the estimated rate f CpG>TpG mutations by 1.8-fold (fig. S2). To capture the regional variation, we binned the genome into non-overlapping windows of 10 kb (fig. S3C), the scale that maximizes the number of reflected components (fig. S1H).

Application of vrnmf to this dataset identifies 14 components passing the "reflection test" and robust to resampling. They are reproduced in gnomAD (*10*) and recapitulated in de novo mutation data (figs. S1, G and L, and S4D) (*11*, *12*). The 14 components correspond to 9 processes, 5 strand-dependent represented by two components) and 4 strand-independent (Fig. 1D and fig. S1L). Eight processes correlate exclusively with one or two genomic features including gene bodies, replication timing, direction of replication, and chromatin accessibility (Fig. 1E and fig. S4A). Processes have, on average, 40% higher correlation with genomic features than any individual trinucleotide mutation type (fig. S4B). This is remarkable given that the inference was solely based on mutation density.

Up to 5% of heritable mutations arise during early mosaic divisions (*13*). Intensities of several processes differ between de novo germline mutations and early mosaic mutations (Fig. 1F).

Broadly, mutations arise from replication errors or from DNA damage. Bulky damage is resolved in a strand-specific manner within gene bodies due to the action of TCR (*2*) and due to the preferential error-prone damage bypass on the lagging strand during replication (*14*).

Strand-dependent Process 1/2 is represented by mutually complementary components 1 and 2 (Figs. 1D and 2A). The strand asymmetry, measured as the difference between intensities of components 1 and 2, correlates with the directions of both transcription (r=0.32) and replication (r=-0.15), and with gene expression (Table S1). Process 1/2 correlates with the experimentally obtained TCR activity (*15*) (fig. S5A). We conclude that this process is a footprint of the asymmetric resolution of bulky DNA damage. Process 1/2 has reduced intensity in early development, and transcriptional asymmetry of mosaic A>G/T>C mutations has the opposite direction (Figs. 1F and 2A).

Strand-dependent process 3/4 likely captures asymmetric replication errors. Its asymmetry correlates with the direction of replication (Fig. 2B, r=0.34 at the optimal 100kb scale, Table S2).

This process is unlikely to be mediated by bulky DNA damage, because the correlation with direction of transcription is small. We hypothesize that process 3/4 reflects differential replication fidelity between leading and lagging strands (*1*), offering the first probable footprint of replicative errors.

Process 5/6 has an elevated intensity on non-transcribed strands of L1PA LINE repeats (Fig. 2C and fig. S5E). The two percent of the genome with the strongest asymmetry of process 5/6 are four-fold enriched with L1PA repeats, but not with other LINEs (fig. S5D).

Strand-independent process 7 closely tracks replication timing (RT) (r=0.64 at the optimal 500 kb scale) (Fig. 2D). The modest association of mutation rate with RT has been long known (*16*, *17*). It is stronger for transversions (*17*), especially C>A (*18*). The intensity of process 7 increases by 4.4-fold from the earliest to latest RT decile, while the rate of transversions increases by just 20% (fig. S5C). Process 7 is substantially more active in early development (Figs. 1F and 2D).

Strand-dependent process 8/9 is dominated by C>G transversions. It is characterized by local spikes totaling 264 Mb (Fig. 3, A to C); just 10% of the genome harbors 67% of clustered de novo mutations of maternal origin and includes all known (*19*) and many new regions of accelerated maternal mutagenesis (Fig. 3D and tables S3 and S4).

Process 8/9 displays a 50-200% increase in the rate of C>G mutations on the non-transcribed strand compared to gene flanks (Fig. 3, E to G, and figs. S6 and S7A;). This effect is especially pronounced for long, fragile genes (*WWOX, RBFOX1, CSMD1, FHIT, SDK1*) covered by spikes of the process. We interpret this as transcription-associated mutagenesis in oocytes that is possibly induced by localized susceptibility to DNA damage (*20*). Mutations in spikes of process 8/9 show a much stronger effect of maternal age than the remaining genome (Fig. 3H).

Accumulation of maternal mutations with age in non-dividing oocytes cannot be mediated by replication. Literature favors resolution of double strand breaks (DSB) as a likely mechanism (*11*, *19*, *21*). Complex crossovers in spikes of process 8/9 have 399-fold elevated C>G mutation rates (Fig. 3I) in line with (*11*). However, all complex crossovers contribute only 10 out of 507 additional C>G mutations in spikes of process 8/9 in the de novo mutation dataset (*11*) suggesting that this is an important but not a major mechanism.

Process 10 characterized by CpG transitions is known to be mediated by methylcytosine deamination or by erroneous replication over methylcytosine (*8*, *22*) (Fig. 4, A and B, and fig. S8A).

Process 11 is represented by CpG transversions (Fig. 4C). It is likely a footprint of enzymatic demethylation, which proceeds through hydroxymethylated cytosines and abasic sites as intermediates (*23*). Unfinished repair of abasic sites results in CpG transversions (*24*). Process 11 negatively correlates with cytosine methylation (*25*) only in demethylated CpG islands and positively correlates with hydroxymethylation (*26*) (Fig. 4, D to G, and fig. S8). Process 11 has an increased activity in early mosaic mutations (Figs. 1F and 4H), likely driven by the demethylation wave in early zygote (*26*).

The remaining unexplained processes 12 and 13/14 (figs. S9 and S10) are responsible for a small fraction of mutations (Fig. 1E).

Subsampling suggests no statistical signs of saturation for the number of detectable processes with respect to sample size (figs. S1, I and J, and S4D). Due to limited power, the contributions of known mechanisms such as recombination are not observed in our analyses.

REFERENCES AND NOTES

1. T. A. Kunkel, D. A. Erie, Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu. Rev. Genet.* **49**, 291–313 (2015). doi:10.1146/annurev-genet-112414-054722 Medline
2. J. A. Marteijn, H. Lans, W. Vermeulen, J. H. J. Hoeijmakers, Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014). doi:10.1038/nrm3822 Medline
3. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen, M. R. Stratton; PCAWG Mutational Signatures Working Group; PCAWG Consortium, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). doi:10.1038/s41586-020-1943-3 Medline
4. K. Harris, J. K. Pritchard, Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017). doi:10.7554/eLife.24284 Medline
5. H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, S. H. Jensen, Theorems on positive data: On the uniqueness of NMF. *Comput. Intell. Neurosci.* **2008**, 764206 (2008). doi:10.1155/2008/764206 Medline
6. X. Fu, K. Huang, N. D. Sidiropoulos, W.-K. Ma, Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications. *IEEE Signal Process. Mag.* **36**, 59–80 (2019). doi:10.1109/MSP.2018.2877582
7. A. M. S. Ang, N. Gillis, Algorithms and Comparisons of Nonnegative Matrix Factorizations With Volume Regularization for Hyperspectral Unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**, 4843–4853 (2019). doi:10.1109/JSTARS.2019.2925098
8. Seplyarskiy, Soldatov, Materials and methods online.
9. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S. B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B.

M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021). doi:10.1038/s41586-021-03205-y Medline

10. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, C. A. Aguilar Salinas, T. Ahmad, C. M. Albert, D. Ardissino, G. Atzmon, J. Barnard, L. Beaugerie, E. J. Benjamin, M. Boehnke, L. L. Bonnycastle, E. P. Bottinger, D. W. Bowden, M. J. Bown, J. C. Chambers, J. C. Chan, D. Chasman, J. Cho, M. K. Chung, B. Cohen, A. Correa, D. Dabelea, M. J. Daly, D. Darbar, R. Duggirala, J. Dupuis, P. T. Ellinor, R. Elosua, J. Erdmann, T. Esko, M. Färkkilä, J. Florez, A. Franke, G. Getz, B. Glaser, S. J. Glatt, D. Goldstein, C. Gonzalez, L. Groop, C. Haiman, C. Hanis, M. Harms, M. Hiltunen, M. M. Holi, C. M. Hultman, M. Kallela, J. Kaprio, S. Kathiresan, B.-J. Kim, Y. J. Kim, G. Kirov, J. Kooner, S. Koskinen, H. M. Krumholz, S. Kugathasan, S. H. Kwak, M. Laakso, T. Lehtimäki, R. J. F. Loos, S. A. Lubitz, R. C. W. Ma, D. G. MacArthur, J. Marrugat, K. M. Mattila, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, J. B. Meigs, O. Melander, A. Metspalu, B. M. Neale, P. M. Nilsson, M. C. O'Donovan, D. Ongur, L. Orozco, M. J. Owen, C. N. A. Palmer, A. Palotie, K. S. Park, C. Pato, A. E. Pulver, N. Rahman, A. M. Remes, J. D. Rioux, S. Ripatti, D. M. Roden, D. Saleheen, V. Salomaa, N. J. Samani, J. Scharf, H. Schunkert, M. B. Shoemaker, P. Sklar, H. Soininen, H. Sokol, T. Spector, P. F. Sullivan, J. Suvisaari, E. S. Tai, Y. Y. Teo, T. Tiinamaija, M. Tsuang, D. Turner, T. Tusie-Luna, E. Vartiainen, M. P. Vawter, J. S. Ware, H. Watkins, R. K. Weersma, M. Wessman, J. G. Wilson, R. J. Xavier, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi:10.1038/s41586-020-2308-7 Medline
11. B. V. Halldorsson, G. Palsson, O. A. Stefansson, H. Jonsson, M. T. Hardarson, H. P. Eggertsson, B. Gunnarsson, A. Oddsson, G. H. Halldorsson, F. Zink, S. A. Gudjonsson, M. L. Frigge, G. Thorleifsson, A. Sigurdsson, S. N. Stacey, P. Sulem, G. Masson, A. Helgason, D. F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019). doi:10.1126/science.aau1043 Medline
12. J.-Y. An, K. Lin, L. Zhu, D. M. Werling, S. Dong, H. Brand, H. Z. Wang, X. Zhao, G. B. Schwartz, R. L. Collins, B. B. Currall, C. Dastmalchi, J. Dea, C. Duhn, M. C. Gilson, L. Klei, L. Liang, E. Markenscoff-Papadimitriou, S. Pochareddy, N. Ahituv, J. D. Buxbaum, H. Coon, M. J. Daly, Y. S. Kim, G. T. Marth, B. M. Neale, A. R. Quinlan, J. L. Rubenstein, N. Sestan, M. W. State, A. J. Willsey, M. E. Talkowski, B. Devlin, K. Roeder, S. J. Sanders, Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018). doi:10.1126/science.aat6576 Medline
13. T. A. Sasani, B. S. Pedersen, Z. Gao, L. Baird, M. Przeworski, L. B. Jorde, A. R. Quinlan, Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife* **8**, e46922 (2019). doi:10.7554/eLife.46922 Medline
14. V. B. Seplyarskiy, E. E. Akkuratov, N. Akkuratova, M. A. Andrianova, S. I. Nikolaev, G. A. Bazykin, I. Adameyko, S. R. Sunyaev, Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nat. Genet.* **51**, 36–41 (2019). doi:10.1038/s41588-018-0285-7 Medline
15. S. Adar, J. Hu, J. D. Lieb, A. Sancar, Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2124–E2133 (2016). doi:10.1073/pnas.1603388113 Medline
16. J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, S. R. Sunyaev, Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009). doi:10.1038/ng.363 Medline
17. A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat, S. R. Sunyaev, S. A. McCarroll, Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012). doi:10.1016/j.ajhg.2012.10.018 Medline
18. I. Agarwal, M. Przeworski, Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 17916–17924 (2019). doi:10.1073/pnas.1900714116 Medline
19. J. M. Goldmann, V. B. Seplyarskiy, W. S. W. Wong, T. Vilboux, P. B. Neerincx, D. L. Bodian, B. D. Solomon, J. A. Veltman, J. F. Deeken, C. Gilissen, J. E. Niederhuber, Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018). doi:10.1038/s41588-018-0071-6 Medline
20. S. Jinks-Robertson, A. S. Bhagwat, Transcription-associated mutagenesis. *Annu. Rev. Genet.* **48**, 341–359 (2014). doi:10.1146/annurev-genet-120213-092015 Medline
21. Z. Gao, P. Moorjani, T. A. Sasani, B. S. Pedersen, A. R. Quinlan, L. B. Jorde, G. Amster, M. Przeworski, Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9491–9500 (2019). doi:10.1073/pnas.1901259116 Medline
22. R. C. Poulos, J. Olivier, J. W. H. Wong, The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* **45**, 7786–7795 (2017). doi:10.1093/nar/gkx463 Medline
23. X. Wu, Y. Zhang, TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017). doi:10.1038/nrg.2017.33 Medline
24. K. Chan, M. A. Resnick, D. A. Gordenin, The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair (Amst.)* **12**, 878–889 (2013). doi:10.1016/j.dnarep.2013.07.008 Medline
25. F. Supek, B. Lehner, P. Hajkova, T. Warnecke, Hydroxymethylated cytosines are associated with elevated C to G transversion rates. *PLOS Genet.* **10**, e1004585 (2014). doi:10.1371/journal.pgen.1004585 Medline
26. H. Bagci, A. G. Fisher, DNA demethylation in pluripotency and reprogramming: The role of tet proteins and cell division. *Cell Stem Cell* **13**, 265–269 (2013). doi:10.1016/j.stem.2013.08.005 Medline
27. solrust, pkharchenko, *hms-dbmi/spacemut: Description of the data and code*, Zenodo (2021); https://doi.org/10.5281/zenodo.4494404.
28. solrust, pkharchenko, *kharchenkolab/vrnmf: Volume-regularize NMF*, Zenodo (2021); https://doi.org/10.5281/zenodo.4495386.
29. J. Carlson, W. S. DeWitt, K. Harris, Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Curr. Opin. Genet. Dev.* **62**, 50–57 (2020). doi:10.1016/j.gde.2020.05.024 Medline
30. D. L. Donoho, V. C. Stodden, When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts? (2004); https://doi.org/10.7916/D88D05N7.
31. M. D. Craig, Minimum-volume transforms for remotely sensed data. *IEEE Trans. Geosci. Remote Sens.* **32**, 542–552 (1994). doi:10.1109/36.297973
32. D. Ciuonzo, On Time-Reversal Imaging by Statistical Testing. *IEEE Signal Process. Lett.* **24**, 1024–1028 (2017). doi:10.1109/LSP.2017.2704612
33. R. Da Ponte Barbosa, A. Ene, H. L. Nguyen, J. Ward, in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* (2016), pp. 645–654.
34. X. Fu, K. Huang, N. D. Sidiropoulos, Q. Shi, M. Hong, Anchor-Free Correlated

Topic Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1056–1071 (2019). doi:10.1109/TPAMI.2018.2827377 Medline
35. X. Fu, K. Huang, B. Yang, W.-K. Ma, N. D. Sidiropoulos, Robust Volume Minimization-Based Matrix Factorization for Remote Sensing and Document Clustering. *IEEE Trans. Signal Process.* **64**, 6254–6268 (2016). doi:10.1109/TSP.2016.2602800
36. W. Wang, M. Á. Carreira-Perpiñán, Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv:1309.1541 [cs, math, stat]* (2013) (available at https://arxiv.org/abs/1309.1541).
37. J.-P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4164–4169 (2004). doi:10.1073/pnas.0308531101 Medline
38. A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, R. D. Pascual-Marqui, Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 403–415 (2006). doi:10.1109/TPAMI.2006.60 Medline
39. A. Gloter, Parameter estimation for a discrete sampling of an intergrated Ornstein-Uhlenbeck process. *Statistics* **35**, 225–243 (2001). doi:10.1080/02331880108802733
40. A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, A. D. Smith, Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029–1041 (2011). doi:10.1016/j.cell.2011.08.016 Medline
41. M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012). doi:10.1016/j.cell.2012.04.027 Medline
42. H. Jónsson, P. Sulem, G. A. Arnadottir, G. Pálsson, H. P. Eggertsson, S. Kristmundsdottir, F. Zink, B. Kehr, K. E. Hjorleifsson, B. Ö. Jensson, I. Jonsdottir, S. E. Marelsson, S. A. Gudjonsson, A. Gylfason, A. Jonasdottir, A. Jonasdottir, S. N. Stacey, O. T. Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, K. Stefansson, Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018). doi:10.1038/s41588-018-0259-9 Medline
43. Y. S. Ju, I. Martincorena, M. Gerstung, M. Petljak, L. B. Alexandrov, R. Rahbari, D. C. Wedge, H. R. Davies, M. Ramakrishna, A. Fullam, S. Martin, C. Alder, N. Patel, S. Gamble, S. O'Meara, D. D. Giri, T. Sauer, S. E. Pinder, C. A. Purdie, Å. Borg, H. Stunnenberg, M. van de Vijver, B. K. T. Tan, C. Caldas, A. Tutt, N. T. Ueno, L. J. van 't Veer, J. W. M. Martens, C. Sotiriou, S. Knappskog, P. N. Span, S. R. Lakhani, J. E. Eyfjörd, A.-L. Børresen-Dale, A. Richardson, A. M. Thompson, A. Viari, M. E. Hurles, S. Nik-Zainal, P. J. Campbell, M. R. Stratton, Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017). doi:10.1038/nature21703 Medline
44. R. E. Rodin, Y. Dou, M. Kwon, M. A. Sherman, A. M. D'Gama, R. N. Doan, L. M. Rento, K. M. Girskis, C. L. Bohrson, S. N. Kim, L. J. Luquette, D. C. Gulhan, P. J. Park, C. A. Walsh, The Landscape of Mutational Mosaicism in Autistic and Normal Human Cerebral Cortex. *bioRxiv* 2020.02.11.944413 [Preprint]. 12 February 2020. https://doi.org/10.1101/2020.02.11.944413.
45. J. Chèneby, M. Gheorghe, M. Artufel, A. Mathelier, B. Ballester, ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018). doi:10.1093/nar/gkx1092 Medline
46. G. E. Uhlenbeck, L. S. Ornstein, On the Theory of the Brownian Motion. *Phys. Rev.* **36**, 823–841 (1930). doi:10.1103/PhysRev.36.823

ACKNOWLEDGMENTS

Funding: This work was supported by National Institutes of Health (NIH) grants R35GM127131, R01MH101244, U01HG009088 and R01 HG010372. R.A.S and P.V.K. were supported by NHLBI (R01HL131768). SR was supported by NIH R35HL135818. Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Specific funding sources for each study and genomic center and authors contribution into dataset creation are given in Supplementary Text 1 and Table S5. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. **Authors contribution**: V.S.B. and R.A.S. idea of the project, data analysis. S.S. and P.V.K. supervision of the study, E.M.K. development of the correction for recurrent mutations, R.J.M analysis of TET and DNMT3B binding. V.S.B., R.A.S., R.J.M., S.S wrote the manuscript. Contribution of other authors listed in Supplementary Text 1 and Table S5. **Competing interests:** we declare no conflicts of interests. Dr. Emery is currently an employee of Bristol Myers Squibb. Bristol Myers Squibb had no role in the funding, design, conduct, and interpretation of this study." **Data and materials availability:** There are no big dataset associated with the paper, all the code and necessary data available are available at (*27*) (https://doi.org/10.5281/zenodo.4494404): http://pklab.med.harvard.edu/ruslan/spacemut. Vrnmf R package is available at (*28*) (https://doi.org/10.5281/zenodo.4495386): https://github.com/kharchenkolab/vrnmf.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/cgi/content/full/science.aba7408/DC1
Materials and Methods
Supplementary Text
Additional acknowledgments and phs numbers
Figs. S1 to S10
Tables S1 to S5
References (*29*–*46*)
MDAR Reproducibility Checklist

31 December 2019; accepted 14 July 2021
Published online 12 August 2021
10.1126/science.aba7408

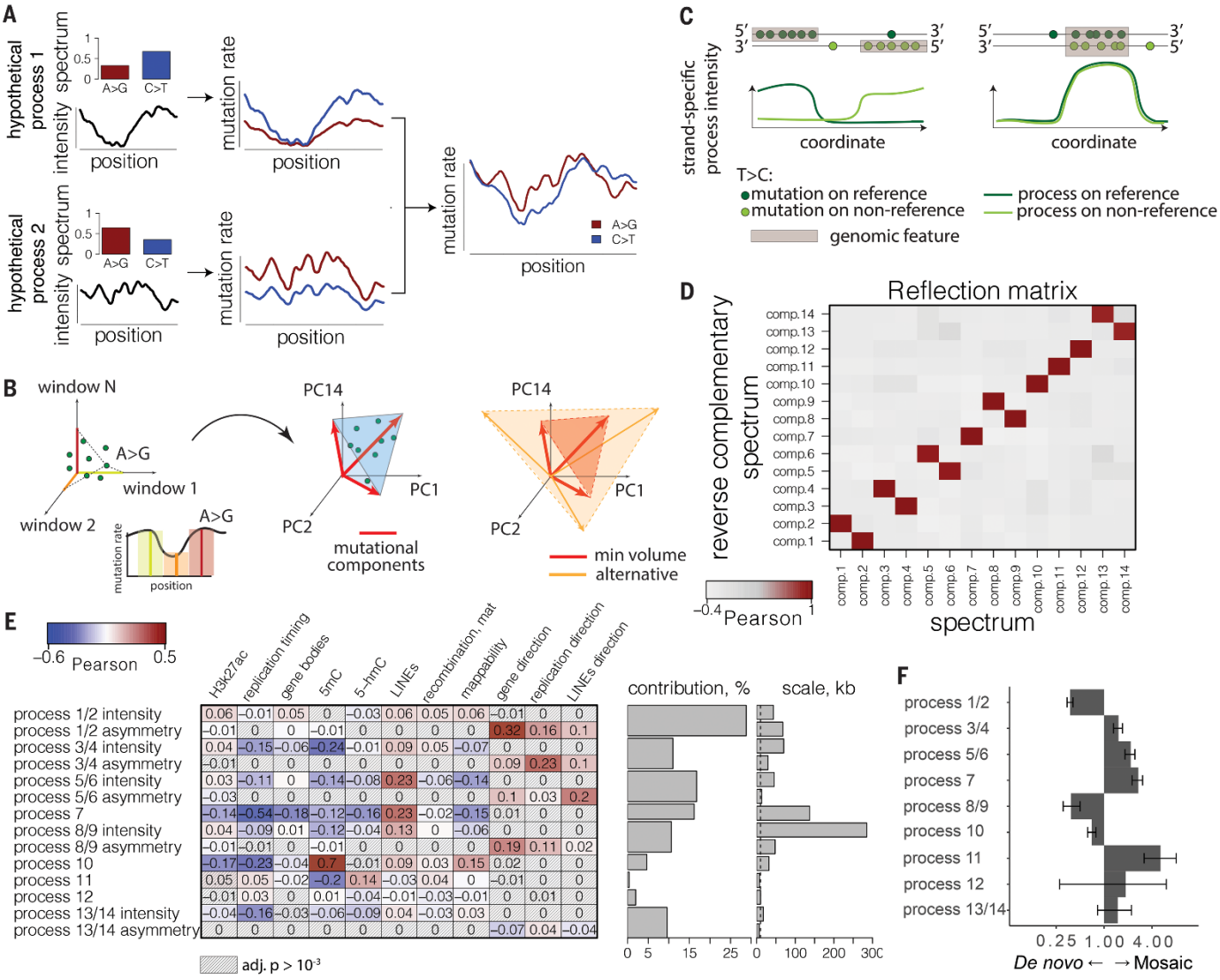

Fig. 1. Inference of spatially-varying mutational processes in germ line. (**A**) Mutational data are modeled as the sum of processes defined by spectra and positional intensities. Two hypothetical processes (left) produce rates of the two mutation types (middle) that together generate the data (right). (**B**) Inference steps for volume-regularized NMF (vrnmf): rates of mutation types in each locus (left) are represented in the low-dimensional space (PCs, middle). In that space, vrnmf searches for the cone of minimum volume containing all the data; standard NMF identifies any cone containing the data (right). (**C**) A strand-independent process (left) has equal rates of mutations on reference strand and complementary mutations on non-reference strand at each locus. A strand-dependent mutational process (right) has unequal complementary mutation rates. (**D**) Reflection matrix reveals strand-dependency of processes. Correlation of spectrum of one mutational component with reverse complementary spectrum of another component separates the components into self-correlated and mutually correlated pairs. (**E**) Left: correlations of process intensities with genomic features. For strand-dependent processes, intensity is the sum of the two components, and asymmetry is the difference. Shaded correlations have Bonferroni-corrected *p*-value > 0.001. High values of the replication timing track correspond to early replicating regions. Middle: fraction of mutations contributed by the process. Right: spatial scale of intensities. (**F**) Ratios between contributions of the processes to *de novo vs* early zygotic mutations.

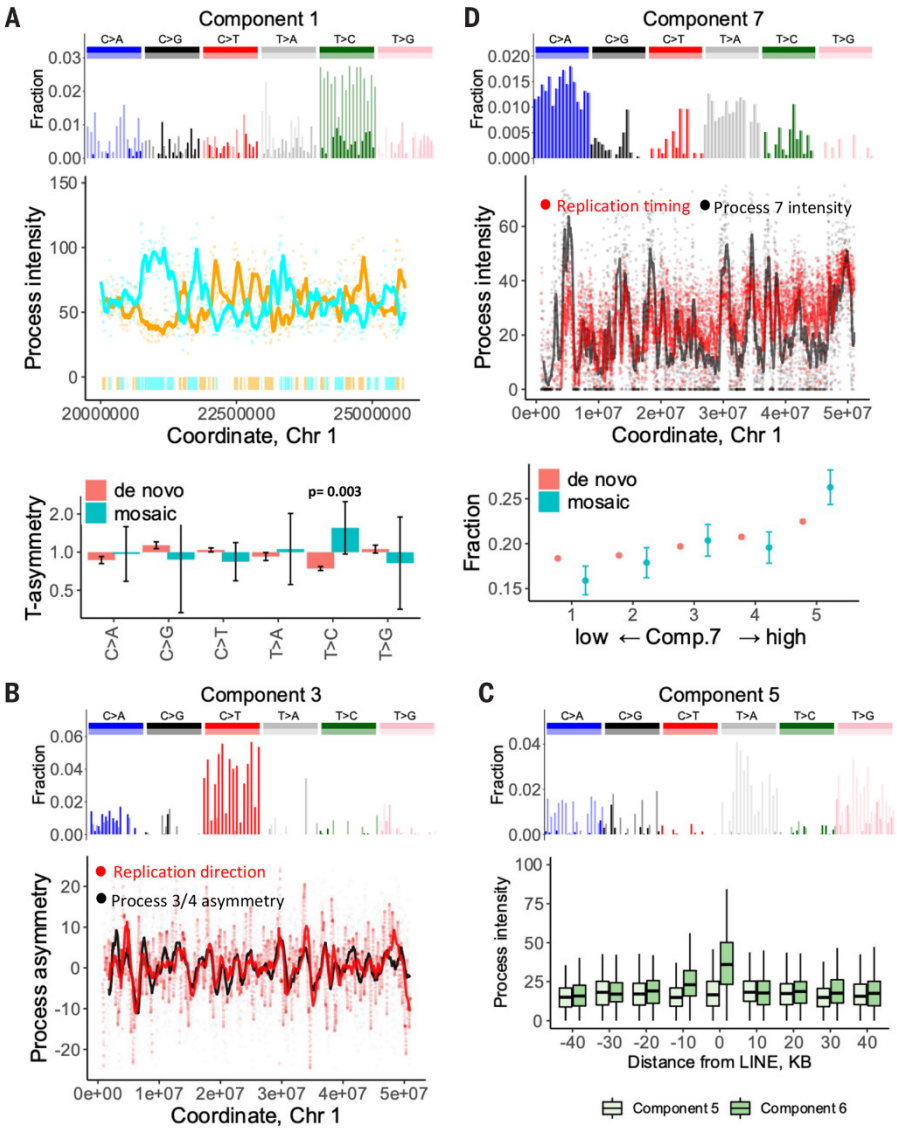

Fig. 2. Mutational processes are associated with distinct genomic features. (**A**) Top: The spectrum of component 1 (the reference strand is bright, and the non-reference strand is translucent). Mutation types are normalized to standard deviations; Middle: Example of Intensities of components 1 and 2. The bars depict gene bodies colored by direction of transcription. Bottom: Transcriptional asymmetry of de novo and mosaic mutations. (**B**) Top: The spectrum of component 3. Bottom: The association between asymmetry of process 3/4 and direction of replication. (**C**) Top: The spectrum of component 5. Bottom: intensities of component 5 and component 6 near LINEs (template strand). (**D**) The spectrum of component 7 (top), and its association with replication timing (middle). (bottom) Fractions of de novo and mosaic non-CpG mutations as function of activity of process 7

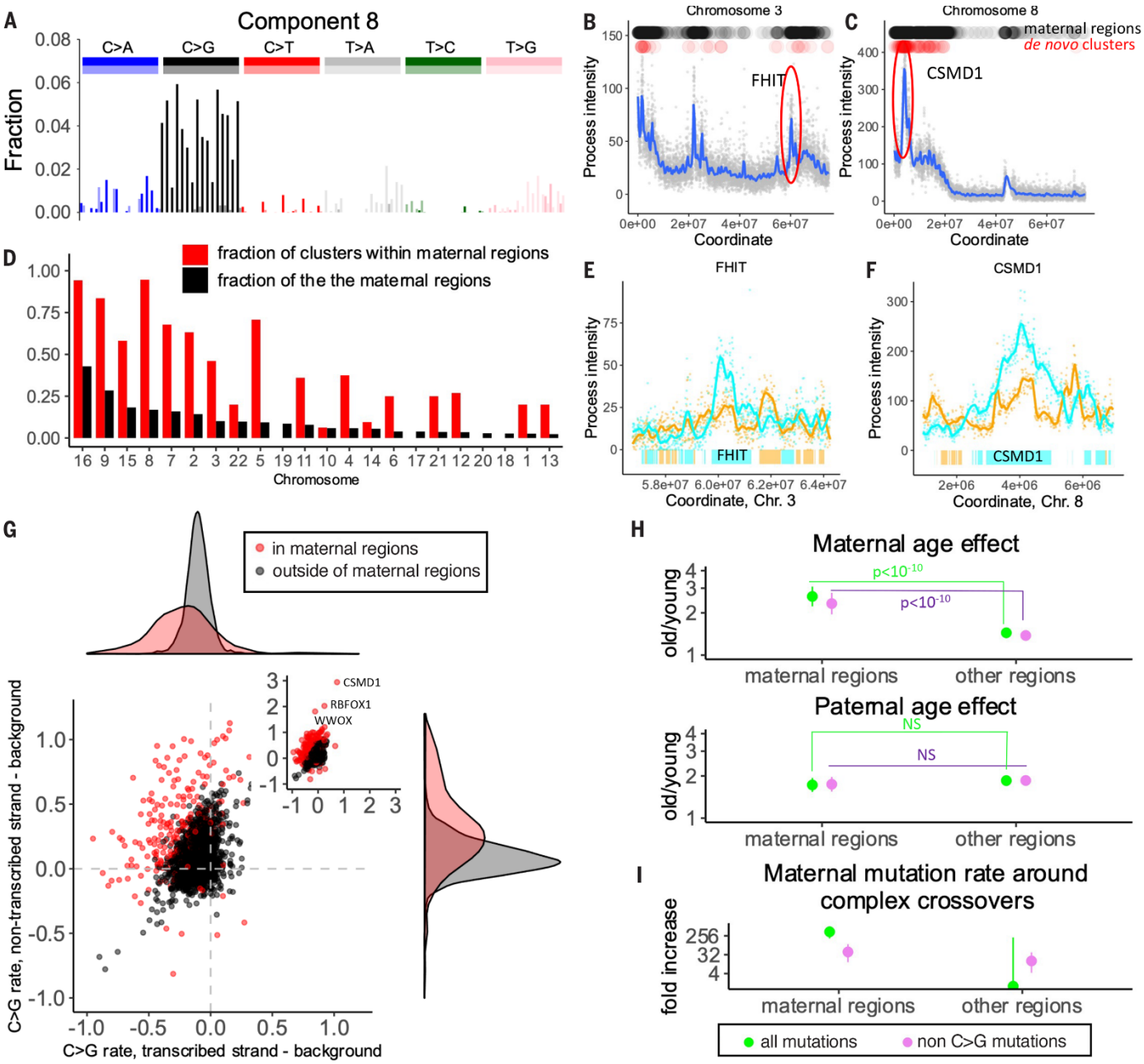

Fig. 3. Oocyte-specific mutational process. (**A**) The spectrum of component 8. (**B and C**) Examples of spikes of process 8/9 (black dots) alongside de novo maternal clustered mutations (*11*) (red dots). (**D**) Enrichment of maternal clustered de novo mutations (*11*) in spikes of process 8/9. (**E** and **F**) Spikes of process 8/9 around *FHIT* and *CSMD1* on non-transcribed strands. The bars depict gene bodies colored by direction of transcription. (**G**) C>G mutation rate on transcribed or non-transcribed strands compared to 100 kb flanks. (**H**) Ratio of parent-specific de novo mutation rates between the first and the last parental age quartiles. (**I**) Fold change in maternal de novo mutation rate in 100kb windows around complex crossovers.

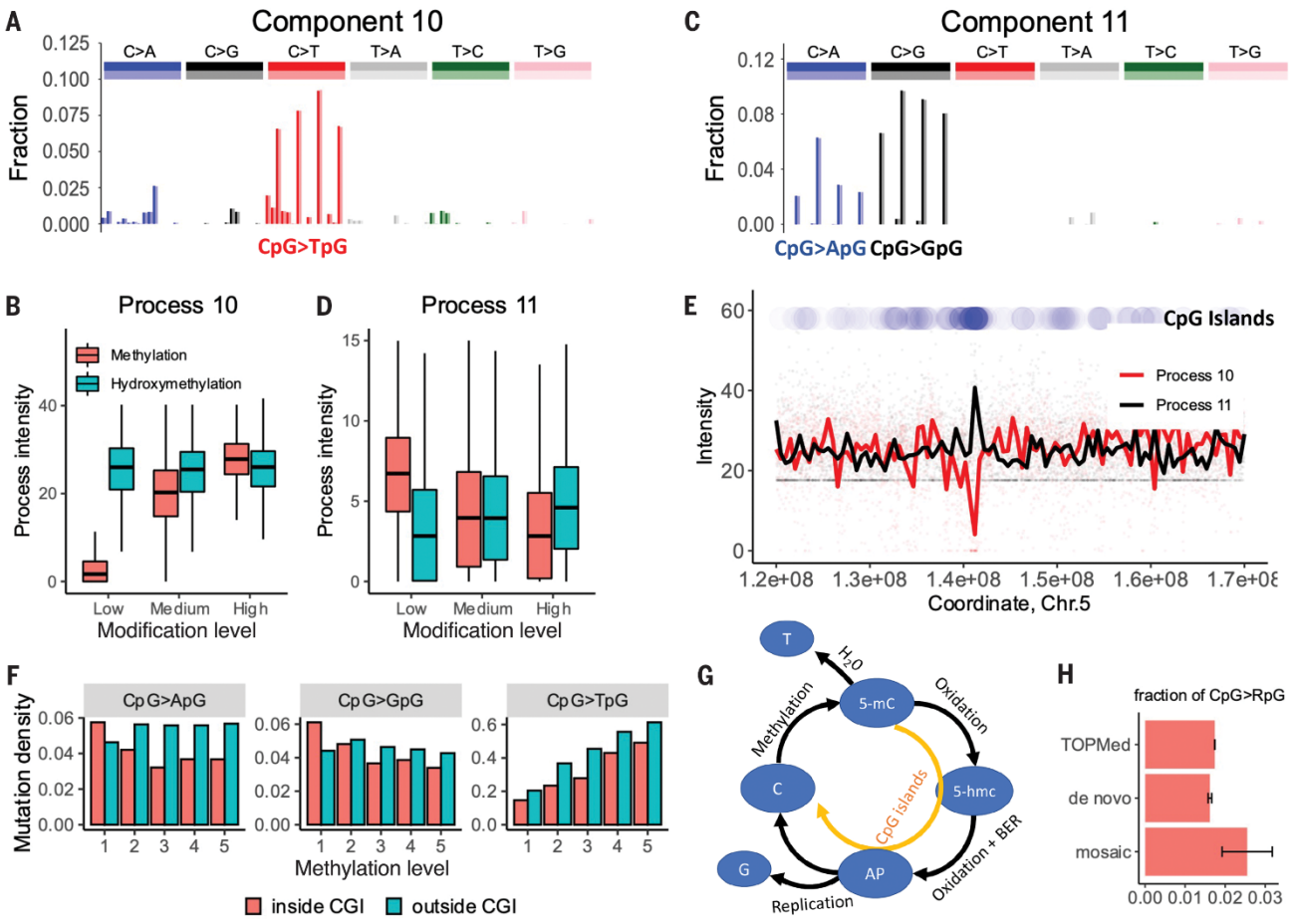

Fig. 4. Cytosine deamination and cytosine demethylation. (**A** and **C**) Spectra of components 10 and 11. (**B**, **D**) The intensity or processes 10 (**C**) and 11 (**D**) as function of methylation and hydroxymethylation. (**E**) Process 10 increases and process 11 decreases in CpG islands (CGI). (**F**) Dependency of CpG mutations on methylation within and outside CGI. (**G**) Mechanisms suggested for processes 10 and 11. Oxidation of methylcytosine (5-mC) leads to hydroxymethylcytosine (5-hmC), which is removed by glycosylase, leaving an abasic site (AP). If not repaired prior to replication, AP sites are causing CpG>GpG or CpG>ApG mutations (**H**) Fraction of CpG transversions among mosaic mutations, de novo mutations and rare polymorphisms.

Science

Population sequencing data reveal a compendium of mutational processes in the human germ line

Vladimir B. Seplyarskiy, Ruslan A. Soldatov, Evan Koch, Ryan J. McGinty, Jakob M. Goldmann, Ryan D. Hernandez, Kathleen Barnes, Adolfo Correa, Esteban G. Burchard, Patrick T. Ellinor, Stephen T. McGarvey, Braxton D. Mitchell, Ramachandran S. Vasan, Susan Redline, Edwin Silverman, Scott T. Weiss, Donna K. Arnett, John Blangero, Eric Boerwinkle, Jiang He, Courtney Montgomery, D.C. Rao, Jerome I. Rotter, Kent D. Taylor, Jennifer A Brody, Yii-Der Ida Chen, Lisa de las Fuentes, Chii-Min Hwu, Stephen S. Rich, Ani W. Manichaikul, Josyf C. Mychaleckyj, Nicholette D. Palmer, Jennifer A. Smith, Sharon L.R. Kardia, Patricia A. Peyser, Lawrence F. Bielak, Timothy D. O'Connor, Leslie S. Emery, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, Christian Gilissen, Wendy S. W. Wong, Peter V. Kharchenko and Shamil Sunyaev

published online August 12, 2021

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.