science.sciencemag.org/cgi/content/full/science.aba7408/DC1

Supplementary Materials for

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, Shamil Sunyaev*

†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. Email: ssunyaev@rics.bwh.harvard.edu

Published 12 August 2021 on *Science* First Release
DOI: 10.1126/science.aba7408

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:
(available at science.sciencemag.org/cgi/content/full/science.aba7408/DC1)

Material and Methods.

Description of the population data

As a proxy for germline mutations, we used SNVs from TOPMed freeze 5 (WHI_GeneSTAR cohort was excluded, due to lack of access to it) (*9*) or gnomAD version r2.0.2 (*10*). The resulting number of individuals in TOPMed was 42,813, and in gnomAD - 15,063, and the corresponding numbers of SNVs are 292,382,053 and 182,057,341. These datasets have high average sequencing depth, above 30x. We did not have access to ethnicity of individuals in the TOPMed project, and did not apply any SNV stratification by ethnicity or phenotype of the carrier in both datasets. Only variants with allelic frequency below 10^{-4} were considered in further analysis, in order to minimize the effects of selection or biased gene conversion, in agreement with prior literature (*4*, *29*). 82% of TOPMed SNVs were below this threshold (Fig. S1B). The resulting set of SNVs included singletons, as no minimal frequency threshold was applied.

Preparation of the mutational matrix

To explore the uniformity of the base calling/sequencing quality, we examined the distribution of the number of mutations within 1 kb windows across the genome. This distribution was bimodal, with the first mode found at 0 SNVs per region (Fig. S3A). This mode clearly corresponded to regions of low quality. Therefore, we excluded 1kb loci with the abnormally low mutation counts (less than 50 mutations) from all the subsequent analyses.

Distribution of rare SNVs may provide a biased estimation for the mutation rates because some SNVs may have resulted from several independent recurrent mutational events. Such recurrent mutations would be more common among hypermutable sites, leading to underestimation of mutation rate. We developed a novel statistical approach to address this issue (see methods below) and adjusted the mutational matrix for potential recurrence.

For downstream analyses, the genome was binned into non-overlapping windows of 2, 5, 10, 30, 100 or 1000 kilobases in size, and mutation rate within each window was estimated as a ratio between the number of each of the 192 mutation types and the number of corresponding trinucleotide sites for each mutation type.

Accounting for recurrent mutations

We estimate the expected recurrence fractions of each mutation count by comparing the site frequency spectrum of different mutational contexts to a reference SFS in a designated low mutation rate context. This approach estimates properties of the distribution of mutation rates in order to predict recurrence fractions.

Conditional on the genealogy and mutation rate at a site, the number of mutations at count *i* in a sample is assumed to be Poisson.

$$Y_i|\mu, T_i \sim Poisson(\mu T_i),$$

where Y_i is the number of independent *i* count mutations in a sample of size n and T_i is the total length of coalescent tree branches subtending *i* sampled chromosomes. This will be a

good approximation if the considered count is sufficiently smaller than *n*, because mutations will be unlikely to occur on the same branch. In practice, what can be observed are composites of independent mutations C_j comprising all $Y = (Y_1, \ldots, Y_k)$ such that $\sum \quad Y_i \cdot i = j$. Let Ξ_j be a random variable giving the sample count of sites in class C_j.

The probability of being in a certain class, conditional on the mutation rate and branch lengths is

$$P(C_j|\mu, T_i) = \sum_{y:\Sigma i y_i = j} \quad \prod_{i}^{n} \quad P(Y_i = y_i|\mu, T_i)$$

$$= \sum_{y:\Sigma i y_i = j} \quad \prod_{i}^{n} \quad \frac{(\mu T_i)^{y_i}}{y_i!} e^{-\mu T_i}$$

$$= \sum_{y:\Sigma i y_i = j} \quad e^{-\mu \Sigma} \quad {}^{T_i} \prod_{i:y_i>0}^{n} \quad \frac{(\mu T_i)^{y_i}}{y_i!}$$

If L sites are observed, the overall distribution of counts in different classes is an overdispersed multinomial distribution, where probabilities differ among sites due to differences in genealogy and mutation rate. We approximate this distribution as Poisson

$$\Xi_j = Poisson\big(L \cdot P_j\big),$$

where

$$P_j = E_{\mu,T_i}[P(C_j|\mu, T_i)]$$

To obtain the likelihood of the observed counts of low-frequency alleles, we calculate the expected probability of observing each allele count. In theory, doing so involves integrating over the distribution of genealogies and mutation rates. In practice, we approximate the distribution of branch lengths as constant and only consider the first max moments of the mutation rate distribution. In the first equation, ignoring higher order moments of the mutation rate distribution is equivalent to ignoring observed alleles comprised of more than m_{max} independent mutations, or integer partitions of *j* with more than m_{max} components. We can break P_j into components corresponding to different levels of recurrence.

$$P_j = \sum_{r=1}^{j} \quad P_{r,j} \approx \sum_{r=1}^{m_{max}} \quad P_{r,j},$$

where

$$P_{r,j} = E_{\mu,T_i}\left[\sum_{y:\Sigma i y_i = j \wedge \Sigma y_i = r} \quad \prod_{i}^{n} \quad P(Y_i = y_i|\mu, T_i)\right]$$

When the mutation rate is low and recurrence can be ignored, the elements of the expected SFS are proportional to $E[T_j]$. We therefore use the SFS in the trinucleotide context with the lowest mutation rate, TTG>TAG, to estimate $E[T_j]$. We then estimate the five first moments of the mutation rate distribution in each trinucleotide context. We allow for a maximum of five independent mutations at each site (m_{max} = 5) and compute the likelihood of the first 70 entries of the SFS. We search for maximum likelihood parameter values of the

mutation rate moments using sequential least squares programming and the Basin-hopping algorithm as implemented in scipy. This provides a set of recurrence estimates, $\hat{P}_{r,j}$ which can be used to adjust each SNP in the observed SFS for the expected number of independent mutations it represents.

Normalization of mutation rates

A natural assumption is that number of mutations m_{ij} of mutation type j in a window i is drawn from Poisson distribution: $m_{ij} \sim Pois(\lambda_{ij} \cdot c_{ij})$, where λ_{ij} is mutation rate and $c_{ij}$is the number of contexts for mutation type j in a window i. Since Poisson process makes downstream statistical inference complicated, we use its normal approximation $m_{ij} \sim N(\lambda_{ij} \cdot c_{ij}, \sigma_{ij} = \sqrt{\lambda_{ij} \cdot c_{ij}})$. That way, observed mutation frequency m_{ij}/c_{ij} is linked to mutation rate λ_{ij}: $m_{ij}/c_{ij} \sim N(\lambda_{ij}, \sigma_{ij} = \sqrt{\lambda_{ij}/c_{ij}})$. To simplify downstream inference, we assume that mutation frequencies of a mutation type have shared across windows standard deviation $\hat{\sigma}_i$ shared across windows: $m_{ij}/c_{ij} \sim N(\lambda_{ij}, \sigma_{ij} = \hat{\sigma}_i)$, $\hat{\sigma}_i$ is an empirical standard deviation of m_{ij}/c_{ij} across genomic windows j. To account for different noise $\hat{\sigma}_i$ across mutation types, we further consider $x_{ij} = m_{ij}/(c_{ij} \cdot \hat{\sigma}_i)$ and $x_{ij} \sim N(\lambda_{ij}/\hat{\sigma}_{ij}, 1)$. An alternative approach could be to estimate window-specific and mutation type-specific σ_{ij} from the data and use it as observation weights in the objective function, but it is prone to be noisy in case of small number of mutations. In case $\hat{\sigma}_{ij} = \sigma_{ij}$, mutation types would have the same statistical variability. In the data, $\hat{\sigma}_{ij}$ includes statistical and biological variability across windows and generally larger than σ_{ij}. Throughout the paper spectra of mutational components are shown in variance normalized form, but are available for downloading in both variance normalized and standard forms (see https://github.com/hms-dbmi/spacemut) (*27*).

Volume-regularized NMF

Regional mutation frequencies $\underline{x_i}$ in a window i are assumed to be additive contribution of q mutational components:

$$\underline{x_i} = \sum_{p=1}^{q} \underline{s_p} \cdot I_{ip}, \quad (1)$$

where $\underline{s_p}$- spectrum of component $p$and I_{ip}- intensity of component p in window i. In particular, 10 kb non-intersecting genomic windows provide 263,870 vectors of regional mutation frequencies $\underline{x_i}$ of length 192.

NMF is a natural statistical framework of this biological model. It would seek to find non-negative matrices S of components spectra and I of components spatial intensities such that $X \approx I \cdot S^T$. However, NMF is not an identifiable problem in general cases (*5*, *30*): there are a potentially infinite number of pairs (I, S) that deliver the same quality of matrix X approximation. The assumptions under which NMF guarantees a unique solution are rather restrictive and likely do

not hold in the case of germline mutations (Fig. S1K). At the same time, lack of identifiability impedes unambiguous biological interpretation of NMF results.

In the noiseless case, NMF has a simple geometrical interpretation of finding a simplicial cone in the positive quadrant that contains all data points. Existence of multiple simplicial cones in the positive quadrant containing all data points is equivalent to the lack of NMF identifiability. Intuitively, there would exist many simplicial cones as soon as data points are distanced from facets of the positive quadrant. To overcome the issue, a common practice is to identify a simplicial cone of minimum volume that contains all of the data points (*31*). Importantly, recent theoretical advances reveal that under relatively mild assumptions on spread of column vectors of S (or I) the simplicial cone of minimum volume is unique and delivers the correct solution of the NMF problem (*32*).

In the case of datasets with large numbers of noisy observations, as happens in this study with 263,870 windows of regional germline mutation rates, an efficient approach would be a reformulation of NMF in "covariance domain" (*33*):

$$P = S \cdot R \cdot S^T, \quad (2)$$

where $S \geq 0, 1^T \cdot S = 1, R \geq 0, P = X^T \cdot X$ is co-occurrence of regional mutation type rates, $R = I^T \cdot I$ is co-occurrence of intensities of mutational components.

In the noiseless case, analogous to finding a simplicial cone of minimum volume, it was shown that under relatively mild assumptions on the spread of mutational spectra S, finding minimum volume of matrix R identifies unique and correct solution of (2) (*34*). Since real data are corrupted by noise and can be prone to model misspecifications, a more realistic approach is a two-step procedure. In the first step, matrix P is denoised using singular value decomposition by projecting vectors onto first q components: if $P = U\Lambda U^T$ than denoised $P \approx U_{1:m,1:q}\Lambda_{1:q,1:q}U^T_{1:m,1:q}$, where $m = 192$ mutation types. In the second step, quality of matrix P approximation and volume size of matrix R are balanced in the following optimization problem (*7*, *35*):

$$min_{R,S}|P - S \cdot R \cdot S^T|_F^2 + \lambda \cdot volume(R), S \geq 0, 1^T \cdot S = 1, R \geq 0, \quad (3)$$

where λ is volume regularization parameter and $|\ \ |_F$ is Frobenius norm. Volume of matrix R is proportional to its determinant. To alleviate computational issues of dealing with determinant, here we utilize a commonly used volume approximation $volume(R) = log(det(R) + \varepsilon)$, where ε is a small number to avoid zero logarithm (*7*).

However, the problem (3) includes a quadratic form $S \cdot R \cdot S^T$, which makes optimization of (3) non-trivial. To further simplify complexity and dimensionality of the optimization, we use square root decomposition $P = BB^T$ and decomposition $R = DD^T$. It can be shown that equality (2) entails $B \cdot U = S \cdot D$, where U is an orthonormal matrix, and $log(det(R) + \varepsilon) =$

$log(det(DD^T) + \varepsilon)$. Thus, optimization problem (3) is almost equivalent to the following optimization problem:

$$min_{S,D,U}|B \cdot U - S \cdot D|_F^2 + \lambda \cdot log(det(DD^T) + \varepsilon), S \geq 0, 1^T \cdot S = 1, D \geq 0, UU^T = I. \quad (4)$$

This problem can be solved using alternating optimization. After learning component spectra matrix S from (4), matrix I of component intensities can be identified via non-negative least squares using (1) and is described in the Methods section "Inference of spatial intensities". Non-convex optimization problem (4) is solved using iterative alternating optimization of matrices S, D and U. Optimization of individual matrices, largely following, is outlined below.

1) Optimization of non-negative D. At each iteration k, matrix D is optimized upon fixed $\{S_{k-1}, U_{k-1}\}$. Optimization problem (4) is majorated by a positive-definite quadratic form with respect to D using the following $logdet$ inequality from (*35*):

$$log(det(DD^T) + \varepsilon) \leq Tr(F_{k-1}DD^T) - log\ det\ F_{k-1} - K,$$

where $F_{k-1} = (D_{k-1}D_{k-1}^T + \varepsilon I)^{-1}$ and the equality holds when $D = D_{k-1}$. Thus, the original non-convex objective function (4) is majorated by the following convex function $g(D)$:

$$g(D) = |B \cdot U_{k-1} - S_{k-1} \cdot D|_F^2 + \lambda \cdot Tr(F_{k-1}DD^T).$$

Function $g(D)$ subject to non-negative constraints $D \geq 0$ is minimized using a local upper bound approximation:

$$g(D) \leq g(D_{k-1}) + \nabla g(D_{k-1}, S_{k-1}, U_{k-1})^T \cdot (D - D_{k-1}) + L_k \cdot |D - D_{k-1}|_F^2,$$

where L_k is Lipschitz constant $L_k = |(S_{k-1})^T S_{k-1} + \lambda \cdot (F_{k-1})^T \cdot F_{k-1}|_F$. It comes down to the following update:

$$D_k = max(D_{k-1} - 1/L_k \cdot \nabla g(D_{k-1}, S_{k-1}, U_{k-1}), 0) \quad (5)$$

Overall, optimization of D at each iteration consists of a predefined number of updates (5). To speed up convergence of convex upper bound (5) Nesterov acceleration is used at each update .

2) Optimization of non-negative S. At each iteration k, matrix S is optimized upon fixed $\{D_k, U_{k-1}\}$. Problem (4) is quadratic programming with respect to S. However, it requires simultaneous full matrix optimization, since rows of S are coupled in (4) due to column constraints $1^T \cdot S = 1$. To uncouple this dependence and provide small-scale per-column optimizations, an upper bound local approximation of S in (4) is used. Assuming $h(S) = |B \cdot U_{k-1} - S \cdot D_k|_F^2$, it follows that:

$$h(S) \leq h(S_{k-1}) + \nabla h(S_{k-1}) \cdot (S - S_{k-1}) + M_k \cdot |S - S_{k-1}|_F^2, \quad (6)$$

where Lipschitz constant $M_k = |(D_k)^T D_k|_F$. Solution of the upper boundary (6) subject to column simplex constraints $1^T \cdot S = 1, S \geq 0$, that can be calculated independently for each column, is:

$$S_k = Proj_{1^T \cdot S=1, S\geq 0}(S \quad - (S_{k-1} - 1/M_k \cdot G) \quad), \quad (7)$$

where $G(S_{k-1}, D_k, U_{k-1}) = (S_{k-1} D_k - B U_{k-1}) \cdot (D_k)^T$.Simplex projection (7)of each column vector can be solved efficiently (*36*). Overall, optimization of S at each iteration consists of a predefined number of updates (7). To speed up convergence, Nesterov acceleration is used at each update.

3) Optimization of orthonormal U. At each iteration k, matrix U is optimized upon fixed $\{D_k, S_k\}$. Problem (4) is orthogonal Procrustes problem with respect to $U$and has a closed-form solution.

The algorithm iteratively updates matrices D, S, U until convergence to a local optimum. Usually a small relative change of objective function or matrices is used as stopping criteria. For this TOPMed dataset, we observed that the algorithm often converges after 1000 iterations and always converges after 3000 iterations. Thus, the algorithm was run for 3000 iterations. Matrices D and S are updated 10 times inside each iteration to speed up convergence. Overall, vrnmf only guarantees convergence to a local optimum. However, 200 random initializations reveal convergence to a single stable optimum under a specifically chosen volume weight λ (see the Methods section "Selection of volume weight in vrnmf"). At each random initialization, $U$was set up as identity matrix and entries of matrices D, S were sampled uniformly from [0, 1] followed by normalization of columns S to unit sum.

Inference of spatial intensities

Vrnmf infers spectra of mutational components, but not intensities. The latter are estimated using non-negative least squares based on known mutation frequencies X and spectra of mutational components S:

$$\underline{x_i} = \sum_p \quad \underline{s_p} \cdot I_{ip} + \underline{s_{offset}}, \qquad (8)$$

where $I_{ip}$are regional intensities and $\underline{s_{offset}}$ is a residual spectrum of possibly unaccounted by vrnmf mutational forces, also estimated from (8).

The same procedure was applied to infer intensity of the components in *de novo* mutations.

Reflection correlation and test

Separating noise from meaningful components is a common challenge for NMF techniques. In addition to that, meaningful mutational processes can be classified as DNA strand-dependent, such as transcription-coupled NER, and strand-independent. We argue that the analysis of two DNA strands enables separation of noise from biological components and classification of the latter into strand-independent and strand-dependent processes. For technical purposes, the two

DNA strands are annotated as reference and non-reference strands, and mutation frequencies are estimated with respect to the reference strand. Since classification of reference/non-reference strands is artificial, sets of inferred mutational spectra using the reference and non-reference strands should be the same. This fact is used to separate biological components from noise components, which are not expected to be reproducible between inferences on reference/non-reference strands. The best Pearson correlation between the spectra of a component with components of another inferred set serves as a measure of reproducibility, called reflection correlation. Note that mutation frequencies with respect to the reference strand are identical to reverse complementary mutation frequencies with respect to the non-reference strand. Thus, a set of components inferred using the non-reference strand is identical to the reverse complementary set of components inferred using the reference strand. It indicates that reflection correlation can be estimated as the best Pearson correlation between spectra of components inferred using the reference strand and a set of reverse complementary components.

Components corresponding to a strand-independent process should have the same rates of reverse complementary mutation. As a consequence, reflection correlation should be the best with its own reverse complement. On the other hand, a strand-dependent process has unequal rates of at least some complementary mutation types; a component that reflects the strand-dependent process would be unequal to its own complement and correspond to a different reverse complementary component. Since sets of original and reverse complement components are expected to be identical, the strand-dependent process generates two different components that are the reverse complement of each other.

To summarize, reflection correlation between original and reverse complementary sets of components should partition them in three groups: noise with low reflection correlation, components reverse complementary to themselves corresponding to strand-independent processes, and pairs of reverse complementary components that reflect strand-dependent processes.
The reflection test is a core instrument for vrnmf parameter identification, model selection, and benchmarking of methods, as is shown below. Throughout the paper, average reflection correlation (ARC) of components of a vrnmf solution serves as a metric of the solution quality, and reflection correlation of 0.8 serves as a cutoff to separate meaningful components, which we call “reflected components”, from noise.

Selection of volume weight in vrnmf

Performance of vrnmf critically depends on volume weight λ in optimization problem (4): small weight would lead to insufficient reduction of volume making the problem similar to standard NMF, while large weight would lead to collapse of volume or, equivalently, zero values for at least some eigenvalues of matrix D. Vrnmf was run for 20 λ values sampled uniformly from 0 to 0.1 to infer 14 components. Average reflection correlation (ARC) of components was used as a criterion for overall quality of solution. Parameter $\lambda = 7.89 \cdot 10^{-3}$ with the highest ARC of more than 0.97 was selected. To ensure stability with respect to local optima, vrnmf with 50 random initializations was run for each λ and mean ARC across initializations was taken.

Selection of the number of components in vrnmf

ARC was used to choose a number N of components that together provide a full and sensitive representation of the underlying biology. For that, 20 random initializations of vrnmf were run for each N from 2 to 20 components using $\lambda = 7.89 \cdot 10^{-3}$, and the maximum number of reflected components across initializations was recorded for each N. The results show that all of the components are reflected for N up to 14 (except for $N = 9$) followed by a plateau at which almost no additional components have reflection property. We thus selected $N = 14$ components as the largest number at which all components are meaningful.

Selection of the window size

Size of the genomic window affects statistical power to detect mutational processes: small windows reduce power to detect larger-scale processes, while large windows dilute signals from smaller-scale processes. ARC was used as a criterion to identify the genomic window size that delivers the largest number of meaningful components. To do so, vrnmf was run for windows of size 2, 5, 10, 30, 100 and 1000 kilobases with $N = 14$ components, and reflection correlations were estimated. To account for different statistical properties of window-specific datasets, near-optimal volume weight λ was selected for each window following the procedure described in the section "Selection of volume weight in vrnmf": $\lambda = 8 \cdot 10^{-2}$ for 2 kb, $\lambda = 6 \cdot 10^{-2}$ for 5 kb, $\lambda = 3 \cdot 10^{-3}$ for 30 kb, $\lambda = 2 \cdot 10^{-3}$ for 100 kb, $\lambda = 5 \cdot 10^{-4}$ for 1000 kb. For each window, vrnmf was run for 10,000 iterations with 10 updates of matrices at each iteration. Window size of 10 kb demonstrates the best ARC across selected window sizes.

Spatial robustness

Robustness of each component spectrum was assessed using bootstrap of genomic windows. 14 components were inferred using 400 bootstrap sampling rounds. Maximum Pearson correlations between spectra of the original component and components in each bootstrap round were then calculated.

Comparison with alternative datasets

Stability of the inferred components was estimated with respect to mutation recurrence, SNP frequency, as well as their presence in another dataset (gnomAD). To estimate the effect of mutation recurrence, mutation frequencies without correction for recurrency (see the Methods section "Accounting for recurrent mutations") were used to infer 14 components with the same vrnmf parameters. To estimate the effect of SNP frequencies, singleton frequencies were used to infer 14 components using vrnmf with $\lambda = 1.4 \cdot 10^{-2}$ and the other parameters kept the same. To estimate potential bias of the TOPMed dataset, mutation frequencies in 100 kb windows were estimated using the gnomAD dataset (see Methods section "Preparation of mutational matrix") followed by inference of 14 components using vrnmf with $\lambda = 2.2 \cdot 10^{-2}$, and the other parameters kept the same. We used a window of 100 kb for the gnomAD dataset to account for different mutation densities in the datasets: 10 kb TOPMed windows contain on average 18 fold more mutations compared to 10 kb gnomAD windows, but only 1.8 fold more mutations compared to 100 kb gnomAD. In each case, parameter λ was selected as in the Methods section "Selection of volume weight in vrnmf" to enable near optimal performance.

Power analysis of the dataset

The dataset was subsampled up to the size of 1% of the original dataset. For each subsampled dataset vrnmf was run to infer $N = 2$ to 14 components with 5 random initializations for each N. We then selected a solution that had the maximum number of reflected components (reflection correlation > 0.8) across N initializations, and estimated 1) the number of reflected components, and 2) the number of components having Pearson linear correlation of at least 0.8 with the original mutational components.

Comparison of NMF methods.

Vrnmf was compared to standard NMF (*37*) and non-smooth NMF (*38*), which enforces sparseness of decomposition matrices, using *NMF* R package. Both standard (option method="brunet" in *NMF*) and non-smooth (option method="nsNMF" in *NMF*) versions of NMF were run using the original matrix X of standardized mutation frequencies with rank=14. To assess importance of denoising, performed in vrnmf, we additionally run methods with the same settings for denoised matrix X: if $X = U\Lambda U^T$ than denoised $X \approx U_{1:nrow(U),1:q}\Lambda_{1:q,1:q}V^T_{1:nrow(V),1:q}$, where $q = 14$ components. Additionally, to evaluate importance of volume regularization, vrnmf was performed for inference of 14 components with volume weight $\lambda = 0$ that effectively converts it to the standard NMF. Each method (standard NMF, standard NMF + denoising, nsNMF, nsNMF + denoising, vrnmf w/o volume regularization and vrnmf) was run 30 times to explore possible local optima. Quality of solutions was assessed based on the number of reflected components. Of note, volume-regularized NMF is the only algorithm that achieved a unique solution with 14 reflected components.

Statistical properties of mutational components

The scale of mutational components was defined using a linear autoregressive model. The spatial intensity of each mutational component was modeled as:

$$I_p = \sum_{k=1}^{M} a_k \cdot I_{p-k} + \xi_p,$$

where I_p is the intensity at position p, $a_k$are autoregressive coefficients and ξ_p is the residual noise. Order M of the model was chosen using Akaike Information Criterion. The *ar* R package was used to fit the autoregressive model. The scale of each process was defined as the half-life of the autoregressive model

$$hl = \frac{ln(0.5)}{ln(\sum_{k=1}^{M} a_k.)}$$

The contribution of each component was defined as the squared sum of intensities. Contributions of all components were then scaled to the unit sum.

Comparison with *de novo* data

To assess if the spatial distribution of *de novo* mutations is consistent with individual mutational processes, we pooled 309,287 *de novo* point mutations from two datasets (*11*, *12*)
For each component, we divided the genome in two bins with 10% of the genome with the highest activity of the component in TOPMed assigned to the first bin and the remaining 90% assigned to the second bin. We infer intensity of the components for *de novo* mutations in these two bins, as we did for individual windows in TOPMed (see Inference of spatial intensities, eq 8). We obtained confidence intervals for the intensities in each of the two bins by permuting *de novo* mutations 100 times.
A very similar approach was used to compare the prevalence of each component among *de novo* mutations of maternal and paternal origin. We calculated the ratio of component intensity in the 10% of the genome with the highest activity of the component in TOPMed, and the component intensity in the remaining genome. We obtained confidence intervals for the ratio of the intensities by permuting paternal and maternal *de novo* mutations 100 times.

Simulations to assess the limitations of the approach

The ability to infer mutational processes depends on their statistical properties, such as the spatial scale and magnitude of variation along the genome and specificity of the mutational spectrum. Limitations of vrnmf inference with respect to these statistical properties were analyzed by simulating a mutational processes and applying vrnmf to the resulting mutation rates.

Briefly, we simulated spectra and intensities of 14 mutational components corresponding to 4 strand-independent and 5 strand-dependent processes with diverse statistical properties. The mutational components were linearly combined to obtain regional mutation type rates along the genome that would reflect TOPMed genome-wide mutation spectrum and the number of mutations in TOPMed dataset (described in the paragraph below). Mutation counts along the genome were sampled from a Poisson process with regional mutation type rates and per-window nucleotide content identical to that in the TOPMed dataset. As with the TOPMed count matrix, the simulated matrix of mutation counts underwent pre-processing (estimation of regional normalized mutation type frequencies) followed by the vrnmf inference of 14 components. Vrnmf was run using $\lambda = 10^{-3}$ for 1,000 iterations. Vrnmf-inferred spectra of mutational components were compared to simulated ones to estimate the quality of recovery. Recovery quality of the simulated components was calculated as the maximum absolute Pearson correlation to one of the inferred components. Simulations were repeated 5000 times to assess each processes in a wide range of scales, fraction of mutations and spectra specificities.

In more detail, genome-scale intensities of components were assumed to follow a continuous Ornstein-Uhlenbeck (O-U) process. O-U process is a convenient framework that provides control of the scale and magnitude of spatial variation of the process, and at the same time enables tractable mathematical manipulations. The scale and magnitude of spatial variation were defined as the half-life and stationary variance of O-U process. Since spatial inference of components takes into account the signal of spatial variation but not the average of intensity across genomic loci, a critical parameter of O-U affecting inference power is the stationary variability v of intensities. We thus set up a constant mean level $m = 1/(\#windows \cdot window\_size 10^4)$ of an

O-U process I and used different values for the coefficient of variation $F = v/m$. That way, the overall number of generated mutations is constant $\int \quad I(p)dp = 1$, but the magnitude of variation along the genome reflected in F varies dramatically. To control the spatial scale, we vary the half-life hl of the O-U process. Parameters of mean m, the coefficient of variation F and half-life hl completely specify O-U process I:

$$dI = \lambda \cdot (m - I) \cdot dp + \sigma \cdot dW_p, \quad (1)$$

where λ is a reversion to the mean expressed as $\lambda = log2/hl$ and $\sigma = \sqrt{2 \cdot log(2) \cdot F/hl}$. To cover a wide range of parameters, For each process, the half-life hl is sampled from the log-uniform distribution in the interval 10^2-10^6 base pairs, $ln(hl) \sim U(ln(10^2), ln(10^6))$, and F is sampled from the log-uniform distribution in the interval 10^{-4}-$1.6 \cdot 10^{-1}$, $ln(F) \sim U(ln(10^{-4}), ln(1.6 \cdot 10^{-1}))$.

In practice, mutation counts are aggregated in non-intersecting windows. It was shown that the integral of the O-U process in non-intersecting windows, $J_w = \int_w \quad Idp$, is a strictly stationary Gaussian process with the following parameters (*39*):

$$E(J_w) = m, Var(J_w) = 1/\lambda^2 \cdot (\Delta - (1 - e^{-\lambda\Delta})/\lambda) \cdot \sigma^2, Cov(J_{w1}, J_{w2}) = r(w1 - w2, \lambda) \cdot \sigma^2,$$

where $r(k, \lambda) = 1/(2\lambda^3) e^{\lambda(1-|k|)\Delta} \cdot (e^{-\lambda\Delta} - 1)^2$ and $\Delta = 10^4$ is a window size. Thus, instead of base pair-resolution O-U process, we simulate window-resolution J_w process. To speed up simulations, J_{w+1} was simulated based on a previous iteration J_w:

$$P(J_{w+1}|J_w) \sim N(E(J_w), \sigma^2(J_w)),$$

where $E(J_w) = m + Cov(J_w, J_{w+1})/Var(J_w) \cdot (J_w - m)$, $\sigma^2(J_w) = Var(J_w) - Cov(J_w, J_{w+1})^2/Var(J_w)$.

Intensities are simulated sequentially for each window, based on the value of the previous window. To avoid negative values of intensity, it is assigned to zero if the sampled value is negative. The number of windows with initially negative sampled values is always less than 1000 (and frequently zero) out of 263,870.

Rate vector of the 192 mutation types was sampled using Dirichlet distribution $\vec{S} = Dir(\alpha \cdot \vec{1})$ with a concentration parameter α sampled log-uniformly in interval 0.01 to 10, $ln(\alpha) \sim U(ln(0.01), ln(10))$. Concentration parameter controls degeneracy of spectra. Spectra of components were then re-normalized to match the average observed genome-wide mutation frequencies in TOPMed. Rates $S_{i,j}$ of each spectra mutation type j were scaled by a factor r_j: $S_{i,j} \leftarrow S_{i,j} \cdot r_j$, where $r_j = \frac{m_j}{\sum_{w,i} \quad J_{wi} \cdot n_{wj} \cdot S_{ij}}$ with m_j being average genome-wide number of mutations rate of type j, n_{wj} being a number of context sites of mutation type j in window w.

Finally, the expected mutation rates of each type j in each window w is a linear combination of the components $v_{w,j} = \sum_i \quad J_{wi} \cdot S_{ij}$. Mutation counts of each type in each window $m_{w,j}$ were sampled from Poisson process with a rate $v_{w,j} \cdot n_{wj}$ proportional to mutation rate $v_{w,j}$. The inference procedure was then applied to matrix $m_{w,j}$ of simulated mutation counts: matrix $m_{w,j}$

was preprocessed, including normalization by the number of contexts and mutation type-specific standard deviations across windows, and inference of 14 components using vrnmf with volume weight $\lambda = 10^{-3}$and 1,000 iterations was performed.

Estimation of relative process activities in non-TOPMed datasets

In small datasets, such as *de novo* mutations from trio sequencing or mosaic mutations, it is impractical impossible to extract run mutational components using vrnmf, because mutation rate estimation in a window is dominated by sampling noise. At the same time, it is possible to estimate the contributions of already inferred components already inferred from large-scale datasets.

Assuming that observed mutations in a dataset are created by a mixture of the processes with predefined spatial intensities and spectra, we may formulate the generative model this inference problem as following:

$$\lambda_{ij} = \sum_{p=1}^{q} K_p \cdot \underline{s}_p \cdot I_{ip} + K_o \cdot \underline{s}_{offset},$$

where, λ_{ij} is a mutation rate of mutation type i in window j, $K_p$and $K_o$are prevalences of process *p* and the offset correspondingly, other notations are the same as in the Methods section "Volume-regularized NMF". In this formulation, any dataset may be deconvoluted into the processes that are already known with unknown contributions $\{K_p\}$ and K_o.
We optimized coefficients $\{K_p\}$ and K_o using Poisson regression: $m_{ij} \sim Pois(\lambda_{ij} \cdot c_{ij})$, where c_{ij} is the number of corresponding tri-nucleotide contexts.

In this instance, mutation counts were approximated by the Poisson distribution. We calculated parameter λ_{ij} for mutation type *j* in window *i* as:

$$\underline{\lambda}_{ji} = \sum_{p=1}^{q} K_p \cdot \underline{s}_{p,j} \cdot I_{ip} + K_o \cdot \underline{s}_{offset,j}$$

The optimization procedure starts with equal values of coefficients that, on average, generate the same amount of mutations as in the dataset. Then, we sequentially multiply each coefficient by a random factor a_p or $1/a_p$, with $a_p \in [0.77, 1.3]$. If substituting $K_p$for either $K_p \cdot a_p$ or $K_p \cdot 1/a_p$ increases likelihood, we assigned the corresponding max value for K_p, otherwise the coefficient remains the same. This procedure is repeated 10 times and usually converges to the final set of coefficients after 5-7 iterations. We find that different runs of the deconvolution converge to the same coefficients within 5% range.

We applied this deconvolution procedure to *de novo* mutations from trio sequencing data and to mosaic mutation. The ratios of coefficients in the two datasets are shown in Fig. 1F.

All processes except 12 and 13/14 significantly improves likelihood of the decomposition of *de novo* and mosaic mutations.

Associations with epigenetic tracks and DNA features

We relied on the analysis of correlations between mutational processes and epigenomic tracks to gain insight into biological mechanisms.

Coordinates of LINE elements were downloaded from UCSC (repeat masker track). In the absence of data from the relevant germline tissue, we used the track for Mcf7 cells. Replication fork direction was determined as in reference *6*.

Gene coordinates were obtained from the 'knownGenes' track downloaded from the UCSC genome browser. We measured gene bias within each window as the number of nucleotides transcribed on the reference strand minus the number of nucleotides transcribed on the strand complementary to the reference. Correlations between the direction of transcription and process 8/9 asymmetry were calculated for the top decile of the intensity of process 8/9.

Methylation levels for each CpG dinucleotide were obtained from Molaro et al, 2011 (*40*) and the methylation level per window was calculated as mean methylation value across all CpG sites within the window. Hydroxymethylation data was obtained from the dataset associated with reference (*41*). Because this track is very sparse, similarly to an earlier study (*25*), we considered any CpG site with the fraction of hydroxymethylated reads exceeding 0.1 as hydroxymethylated. The hydroxymethylation level of a window was calculated as the fraction of hydroxymethylated CpG dinucleotides among all CpG dinucleotides.

Histone modifications H3K4me3, H3K27ac and H3K4me1 were downloaded from the UCSC genome browser. These tracks were obtained for human embryonic stem cells as a potentially relevant cell type. Mappability was obtained from the UCSC genome browser.
Sex-specific recombination rates were obtained from the dataset associated with reference *3*.
CpG islands coordinates were downloaded from the UCSC genome browser.

Associations with the activity of nucleotide excision repair

Nucleotide excision repair (NER) effectively removes bulky lesions, and its activity is partly governed by chromatin structure. Kinetics of CPD and 6-4PP repair by NER was measured in (*15*). Repair of 6-4PP occurs within less than an hour, and thus it is unlikely to be relevant for the mutagenesis that operates in the germline, because divisions of spermatogonia take many days and the dictyate phase of oogenesis lasts for many years. Therefore, we focused on the repair of CPDs, a much slower process (*15*). The majority of UV-induced lesions occur in TT dinucleotides due to properties of UV radiation. To account for this bias, we normalized NER activity to TT dinucleotide content. Following this, we correlated NER efficiency with the intensity of each mutational process.

Correlations between NER activity and mutational processes are shown in Fig. S5A.

Clustered *de novo* mutations

In line with previous studies, we defined clustered *de novo* mutations as pairs of mutations observed in the same individual at distances less than 20,000 nucleotides (*11*, *42*). *De novo* mutations were obtained from the dataset associated with reference *3* and entire clusters were attributed to maternal or paternal origin if there was at least one phased mutation of this origin. Clusters that have mutations on both the paternal and maternal haplotype were excluded.

De novo and mosaic mutations

We aggregated 309,287 *de novo* mutations across two trio sequencing datasets. Clustered mutations were not filtered out. Note that phased mutations were available only for one dataset, and the analysis of maternal or paternal mutations were applied only to this subset. 2432 mosaic mutations were collected from three datasets: all mutations from study (*43*) and (*44*) and gonosomal mutations from study (*13*).

Alteration of mutation rate in gene bodies

To directly estimate the effect of transcription on mutation rate, we compared mutation rates for each of 12 mutation types on the non-transcribed strand of the gene to mutation rates 100 kbKB upstream and downstream of the gene (Fig. 3G and Suppl. Fig. S7). To reliably estimate the intensity of the process and the mutation rate within genes, only genes longer than 100 kbKB were considered. Differences in mutation rate between the gene and the flanking region were normalized to the genome-average mutation rate for each corresponding mutation type.

Maternal age effect in regions susceptible to process 8/9

“Maternal regions” were determined by the high intensity of the sum of components 8 and 9. We noticed a heavy tail of the process 8/9 intensity and considered the top decile of the process as “maternal regions” (Fig. S7).

Excess of maternal *de novo* C>G mutations in regions with high intensity of process 8/9

Mutational processes other than process 8/9 have similar prevalence among maternal and paternal mutations (Fig. S4F). Thus, we assumed that the ratio of maternal to paternal mutations should be similar across loci not mutated by process 8/9. We calculated maternal to paternal ratio for C>G mutations in regions with high intensity of process 8/9 and in the remaining genome. In dataset (28) in regions with high intensity of process 8/9 this ratio is equal to 622/701=0.89, in the remaining genome it is 826/5050=0.16. We expected 0.16*701=115 C>G mutations of maternal origin in regions of high intensity of process 8/9 to be contributed by processes other than process 8/9. The remaining 507 should be attributed to process 8/9.

Mutagenic effects of complex and simple crossovers

Observed number of paternal and maternal mutations within 100 kb window centered at simple and complex crossover was obtained from ref (28). The expected number of mutations

was calculated as the average mutation rate around the position of crossover that happened in another offspring.

Association of process 11 with enzymatic demethylation

TET1 and DNMT3B ChIP-seq peaks obtained from (*45*). Methylation levels here are an average of ENCODE whole genome bisulfite sequencing experiments ENCSR806NNG and ENCSR417YFD, which originate from testis and ovary samples, respectively, as potentially relevant tissue types. Mutation rates were calculated, using uncorrected TOPMED SNV data, for each of 192 mutation types in a trinucleotide context, centered on each peak, summed across each given set of peaks for each position, and then normalized to the genome average mutation rate per mutation type. The 192 mutational types were then aggregated into the given categories using a weighted average. Methylation levels were similarly averaged across the same set of peaks per position in the window.

Optimal scale of the processes

We showed that sampling noise decreases correlation coefficient between DNA features and the extracted processes. Many of the processes have a characteristic scale exceeding window size of 10 kb. Thus, it is possible that intensity of a process on a larger scale has a stronger association with epigenomic features because of reduced sampling noise. We explored properties of spatially smoothed intensities and asymmetries of the mutational processes using sliding windows of different sizes from 20 kb to 2 mb. Note that while smoothing reduces sampling noise, it reduces spatial precision. We argue that the optimal scale leads to the highest value of the explained variance by epigenomic features, and choose to determine the optimal windows sizes accordingly. In a few cases, we found that the explained variance behaved as a non-monotonic function of scale. In such cases, we always chose the first local maximum.

Supplementary Text, discussion of biological mechanisms behind the processes

Low fidelity of metC replication and deamination of metC are contributing to CpG>TpG mutations

Deamination of methylcytosine C (metC) leads to thymine and results in a T:G mismatch. This mismatch may be occasionally repaired into T:A instead of C:G by the Base Excision Repair

system (BER) (*53*). Alternatively, hypermutability of CpG transitions could be explained by methylated cytosines being inferior replication templates compared to unmethylated cytosines(*54*, *55*). The latter model is supported by the observation that CpG transitions demonstrate a remarkable asymmetry with respect to replication direction (Fig. S8A, correlation with asymmetry of process 3/4 is r=0.64 on 100 kb scale). It is thus surprising that process 10 appears to be strand symmetric (see spectra Fig. 4A). An appealing interpretation is that process 10 includes mutations arising due to spontaneous deamination, while CpG associated replication errors are absorbed into process 3/4 together with other replication errors. However, we cannot exclude that vrnmf simply lacks sensitivity to detect this asymmetry in the 192 dimensional space.

Enzymatic demethylation causes CpG transversions

Process 10 promotes transitions in methylated CpGs. By either model explained above, the frequency of this process would correlate with the level of methylation. In contrast, process 11 promotes CpG transversions, and correlates with low methylation levels. However, we believe process 11 does not act on unmethylated CpGs, but rather occurs during enzymatic demethylation. In support of this, we observe that CpG transitions correlate with methylation levels regardless of whether they are found inside or outside CpG islands, whereas CpG transversions only correlate with methylation levels inside of CpG islands (Fig. 4F), where active demethylation is highly coordinated (*59*, *60*). ChIP-seq peaks for TET1, a key enzyme involved in active demethylation (*56*), are accompanied by a large regional drop in methylation levels outside of CpG islands (Fig. S8C,E). These sites show both a proportionate increase in CpG transversions and a decrease in CpG transitions, in agreement with ref. (*58*). In contrast, sites of the methylation factor DNMT3B show only a small increase in methylation over the heavily-methylated background outside of CpG islands with a proportionate increase in CpG transitions, but no concomitant decrease in CpG transversions (Fig. S8D,F). Together, these results support the model that, rather than acting on unmethylated CpGs, process 11 occurs during enzymatic demethylation.

Processes 12/13 and 14

The 14 components are robust with respect to window size and are reproduced in the independent gnomAD dataset (*24*, *26*) (Fig. S2L). Natural selection and biased gene conversion can potentially bias the statistical properties of rare alleles relative to *de novo* mutations. Our results are robust to the allele frequency threshold and can be recapitulated on *de novo* mutations (*27–29*) from parent-child trio studies (Fig. S2L, Fig. S4D). We validated the results by excluding singletons, because singletons are enriched in sequencing errors (Fig. S2L).

Additional acknowledgments and phs numbers (In this section we acknowledge funding and participant of the individual cohorts comprising TOPMed freeze 5 NHLBI TOPMed)

In this section we acknowledge funding and participant of the individual cohorts comprising TOPMed freeze 5 *NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish*

The Amish studies upon which these data are based were supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119,
R01 HL121007, and P30 DK072488. See publication: PMID: 18440328

NHLBI TOPMed: Trans-Omics for Precision Medicine Whole Genome Sequencing Project: ARIC

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados

The Genetics and Epidemiology of Asthma in Barbados is supported by National Institutes of Health (NIH) National Heart, Lung, Blood Institute TOPMed (R01 HL104608-S1) and: R01 AI20059, K23 HL076322, and RC2 HL101651. For the specific cohort descriptions and descriptions regarding the collection of phenotype data can be found at: https://www.nhlbiwgs.org/group/bags-asthma. The authors wish to give special recognition to the individual study participants who provided biological samples and or data, without their support in research none of this would be possible.

NHLBI TOPMed: Cleveland Clinic Atrial Fibrillation Study

The research reported in this article was supported by grants from the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute grants R01 HL090620 and R01 HL111314, the NIH National Center for Research Resources for Case Western Reserve University and the Cleveland Clinic Clinical and Translational Science Award (CTSA) UL1-RR024989, the Department of Cardiovascular Medicine philanthropic research fund, Heart and Vascular Institute, Cleveland Clinic, the Fondation Leducq grant 07-CVD 03, and The Atrial Fibrillation Innovation Center, State of Ohio.

NHLBI TOPMed: The Cleveland Family Study (WGS)

Support for the Cleveland Family Study was provided by NHLBI grant numbers R01 HL46380, R01 HL113338 and R35 HL135818.

NHLBI TOPMed: Cardiovascular Health Study

This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01-HC85079, N01-HC-85080, N01-HC-85081, N01-HC-85082, N01-HC-85083, N01-HC-85084, N01-HC-85085, N01-HC-85086, N01-HC-35129, N01-HC-15103, N01-HC-55222, N01-HC-75150, N01-HC-45133, and N01-HC-85239; grant numbers U01 HL080295, U01

HL130114 and R01 HL059367 from the National Heart, Lung, and Blood Institute, and R01 AG023629 from the National Institute on Aging, with additional contributions from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at https://chs-nhlbi.org/pi. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program

This research used data generated by the COPDGene study, which was supported by NIH Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica

This study was supported by NHLBI grants R37 HL066289 and P01 HL132825. We wish to acknowledge the investigators at the Channing Division of Network Medicine at Brigham and Women's Hospital, the investigators at the Hospital Nacional de Niños in San
José, Costa Rica and the study subjects and their extended family members who contributed samples and genotypes to the study, and the NIH/NHLBI for its support in making this project possible.

NHLBI TOPMed: Diabetes Heart Study African American Coronary Artery Calcification (AA CAC)

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

NHLBI TOPMed: Boston Early-Onset COPD Study in the TOPMed Program

The Boston Early-Onset COPD Study (dbGaP accession number phs000946) was supported by the following NIH grants: R01 HL075478, U01 HL089856, and R01 HL113264.

NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I from the National Heart, Lung, and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

NHLBI TOPMed: Genes-environments and Admixture in Latino Asthmatics (GALA II) Study

Supported by NIH and NHLBI grant # R01HL117004; study enrollment supported by NIEHS grant # R01ES015794, the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos

Medical Faculty Development Program, Harry Wm. And Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II. All study collaborators: Shannon Thyne, UCSF; Harold J. Farber, Texas Children's Hospital; Denise Serebrisky, Jacobi Medical Center; Rajesh Kumar, Lurie Children's Hospital of Chicago; Emerita Brigino-Buenaventura, Kaiser Permanente; Michael A. LeNoir, Bay Area Pediatrics; Kelley Meade, Children's Hospital, Oakland; William Rodriguez-Cintron, VA Hospital, Puerto Rico; Pedro C. Avila, Northwestern University, Jose R. Rodriguez-Santana, Centro de Neumologia Pediatrica. The authors acknowledge the families and patients for their participation and thank the numerous health care providers and community clinics for their support and participation in GALA II. In particular, the authors thank study coordinator Sandra Salazar; the recruiters who obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez, Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez, Brenda Lopez, MD, Louis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jihan Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Ana Taveras. See publication: PMID: 23750510

NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy (GENOA)

Support for GENOA was provided by the National Heart, Lung, and Blood Institute (HL054457, HL054464, HL054481, and HL087660) of the National Institutes of Health.

NHLBI TOPMed: Genetic Epidemiology Network of Salt Sensitivity (GenSalt)

The Genetic Epidemiology Network of Salt-Sensitivity (GenSalt) was supported by research grants (U01HL072507, R01HL087263, and R01HL090682) from the National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD.

NHLBI TOPMed: Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

GOLDN biospecimens, baseline phenotype data, and intervention phenotype data were collected with funding from the National Heart, Lung and Blood Institute (NHLBI) grant U01 HL072524. Whole-genome sequencing in GOLDN was funded by NHLBI grant R01 HL104135 and supplement R01 HL104135-04S1.

NHLBI TOPMed: Heart and Vascular Health Study (HVH)

The research reported in this article was supported by grants HL068986, HL085251, HL095080, and HL073410 from the National Heart, Lung, and Blood Institute. NHLBI TOPMed: Hypertension Genetic Epidemiology Network (HyperGEN) The HyperGEN Study is part of the National Heart, Lung, and Blood Institute (NHLBI) Family Blood Pressure Program; collection of

the data represented here was supported by grants U01 HL054472 (MN Lab), U01 HL054473 (DCC), U01 HL054495 (AL FC), and U01 HL054509 (NC FC). The HyperGEN: Genetics of Left Ventricular Hypertrophy Study was supported by NHLBI grant R01 HL055673 with whole-genome sequencing made possible by supplement -18S1.

NHLBI TOPMed: The Jackson Heart Study

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts

from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staff and participants of the JHS.

NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National

Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and

genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, and by the National Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

NHLBI TOPMed: Whole Genome Sequencing of Venous Thromboembolism (WGS of VTE)

Funded in part by grants from the National Institutes of Health, National Heart, Lung, and Blood Institute (HL66216 and HL83141) and the National Human Genome Research Institute (HG04735).

NHLBI TOPMed: MGH Atrial Fibrillation Study

The research reported in this article was supported by NIH grants K23HL071632, K23HL114724, R21DA027021, R01HL092577,

R01HL092577S1, R01HL104156, K24HL105780, U01HL65962, R01HL128914, and American Heart Association, 18SFRN34110082. The research has also been supported by an Established Investigator Award from the American Heart Association (13EIA14220013) and by support from the Fondation Leducq (14CVD01).

NHLBI TOPMed: Partners HealthCare Biobank

We thank the Broad Institute for generating high-quality sequence data supported by the NHLBI grant 3R01HL092577-06S1 to Dr. Patrick Ellinor. The datasets used in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs001024.

NHLBI TOPMed: San Antonio Family Heart Study (WGS)
Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and
R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.
NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environment (SAGE) Study
Supported by NIH and NHLBI grant # R01HL117004; study enrollment supported by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. and Diana V. Hind Distinguished Professor
in Pharmaceutical Sciences II.
NHLBI TOPMed: Genome-wide Association Study of Adiposity in Samoans
Financial support from the U.S. National Institutes of Health Grant R01-HL093093. We acknowledge the assistance of the Samoa Ministry of Health and the Samoa Bureau of Statistics for their guidance and support in the conduct of this study. We thank the local village officials for their help and the participants for their generosity. The following publication describes the origin of the dataset: Hawley NL, Minster RL, Weeks DE, Viali S, Reupena MS, Sun G, Cheng H, Deka R, McGarvey ST. Prevalence of Adiposity and Associated Cardiometabolic Risk Factors in the Samoan Genome-Wide Association Study. Am J Human Biol 2014. 26: 491-501. DOI: 10.1002/jhb.22553. PMID: 24799123.
NHLBI TOPMed: The Vanderbilt AF Ablation Registry
The research reported in this article was supported by grants from the American Heart Association to Dr. Shoemaker (11CRP742009), Dr. Darbar (EIA 0940116N), and grants from the National Institutes of Health (NIH) to Dr. Darbar (R01 HL092217),
and Dr. Roden (U19 HL65962, and UL1 RR024975). The project was also supported by a CTSA award (UL1 TR00045) from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the NIH.
NHLBI TOPMed: The Vanderbilt Atrial Fibrillation Registry
The research reported in this article was supported by grants from the American Heart Association to Dr. Darbar (EIA 0940116N), and grants from the National Institutes of Health (NIH) to Dr. Darbar (HL092217), and Dr. Roden (U19 HL65962, and UL1 RR024975). This project was also supported by CTSA award (UL1TR000445) from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences of the NIH.
NHLBI TOPMed: Novel Risk Factors for the Development of Atrial Fibrillation in Women
The Women's Genome Health Study (WGHS) is supported by HL 043851 and HL099355 from the National Heart, Lung, and Blood Institute and CA 047988 from the National Cancer Institute, the Donald W. Reynolds Foundation with collaborative scientific support

and funding for genotyping provided by Amgen. AF endpoint confirmation was supported by HL-093613 and a grant from the Harris Family Foundation and Watkin's Foundation.

NHLBI TOPMed: Rare Variants for Hypertension in Taiwan Chinese (THRV)

The Rare Variants for Hypertension in Taiwan Chinese (THRV) is supported by the National Heart, Lung, and Blood Institute (NHLBI) grant (R01HL111249) and its participation in TOPMed is supported by an NHLBI supplement (R01HL111249-04S1). THRV is a collaborative study between Washington University in St. Louis, LA BioMed at Harbor UCLA, University of Texas in Houston, Taichung Veterans General Hospital, Taipei Veterans General Hospital, Tri-Service General Hospital, National Health Research

Institutes, National Taiwan University, and Baylor University. THRV is based (substantially) on the parent SAPPHIRe study, along with additional population-based and hospital-based cohorts. SAPPHIRe was supported by NHLBI grants (U01HL54527, U01HL54498) and Taiwan funds, and the other cohorts were supported by Taiwan funds.

NHLBI TOPMed: Genetics of Sarcoidosis in African Americans (Sarcoidosis)

National Institutes of Health (R01HL113326, P30 GM110766-01)

NHLBI TOPMed: Diabetes Heart Study African American Coronary Artery Calcification (AA CAC)

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

Supplementary Figures 1-10

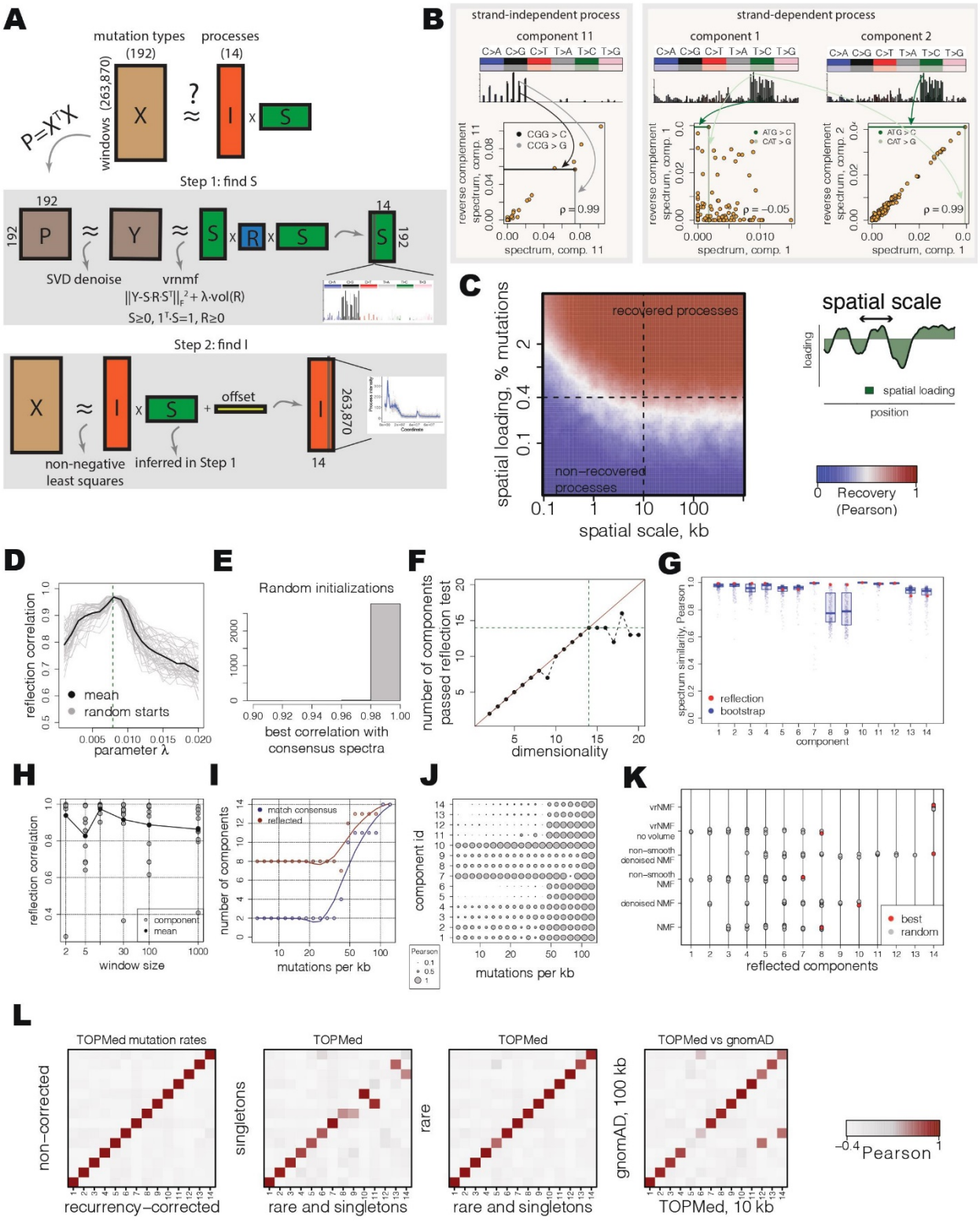

Figure S1. Statistical properties of inferred mutational processes.

A) Volume-regularized NMF pipeline is aimed at providing identifiable decomposition of the original matrix of regional mutation rates in a product of two non-negative matrices of intensities and spectra of underlying mutational components assuming sufficient spreading of mutational spectra. The problem is reformulated in terms of co-occurrence matrix *P* of mutation types followed by denoising of the matrix using SVD. Denoised matrix *Y* is then tri-factorized in non-negative spectra matrix *S* and intensity co-occurrence matrix *R* with regularization on *R* volume.

B) Visual interpretation of reflection correlation. For a strand-independent process, exemplified by component 11, mutation frequencies of complementary mutations are highly similar (left). Top: spectrum of component 11. Bottom: scatterplot of component 11 mutation frequencies (X axis) and its reverse complement (Y axis). For a strand-independent process, exemplified by components 1/2, mutation frequencies of complementary mutations of component 1 (as well as component 2) are different due to strand asymmetry (left scatterplot on the right). However, frequencies of mutations of component 1 and reverse complementary mutations of component 2 are highly similar (right scatterplot on the right). For each case, a pair of reverse complementary mutations are highlighted in spectra and their respective positions in scatterplots.

C) Theoretical scales and magnitudes of spatial variation of simulated mutational components that enable correct inference using vrnmf. Each dot of the heatmap shows mutational components with the given parameters of scale and magnitude of spatial variability. In practice, each dot represents an average of 200 simulated mutational components with the closest values of scale and magnitude of spatial variability. Color reflects the reconstruction quality of mutational components, measured as a Pearson correlation between simulated spectra and spectra inferred using vrnmf on simulated mutation counts. Overall, 70000 (= 5000 runs * 14 components) simulations generated independent sets of mutation counts. Each set of mutation counts reflected size and genome-wide spectrum of TOPMed dataset. To assess the performance of vrnmf, mutational processes inferred from simulated mutation counts were compared to the underlying simulated processes. Spatial intensity of a mutational process was simulated as Ornstein-Uhlenbeck (O-U) process (*46*) with reflection at zero intensity. Scale of spatial variation was defined as the half-life of O-U process. Scale and magnitude of spatial variation was defined as a product of half-life and stationary standard deviation of an O-U process and fraction of all mutations that the O-U process contributed. Informally, magnitude of spatial variation can be interpreted as a standard deviation of component's spatial intensity normalized to overall mutation counts. In simulations, scale and magnitude were sampled randomly in specific ranges to cover recoverable and non-recoverable components. The quality of recovery was assessed using maximum Pearson correlation between the spectrum of each simulated component and the components inferred using vrnmf on the simulated mutation counts. See the details in the Methods section "Simulations to assess the limitations of the approach".

D) Reflection test enables efficient selection of the volume regularization parameter λ (dashed vertical line, λ=7.89). Average Pearson reflection correlation of components (Y axis) was estimated for a range of λ (X axis). Grey curves show results for a set of random initializations of the vrnmf method, and the black curve shows the average reflection correlation across 50 random initializations.

E) Vrnmf optimization procedure empirically converges to a unique optimum. To assess existence of local optima of non-convex vrnmf optimization problem, vrnmf solution was estimated for 200 random initializations and compared to the average of solutions across initializations (consensus spectra) using λ=7.89. A histogram of Pearson correlation coefficients of vrnmf solutions with random initializations is shown.

F) Reflection test provides a strategy to choose an optimal number of components. The number of components with the reflection property (spectrum similar to the reverse complementary spectrum of any component among all extracted components) was calculated for a range of input numbers of components using λ=7.89. Dashed vertical/horizontal lines indicate 14 selected components.

G) Assessing quality of component spectra estimates through bootstrap sampling and the reflection property. For the bootstrap strategy, we assessed similarity of the extracted components with the components extracted using bootstraps of genomic windows. A total of 400 bootstrap samples were generated using λ=7.89. For each component, the most similar component in a bootstrap sample was recorded to estimate the confidence intervals (blue points). We assessed the reflection property using similarity of the spectrum of each extracted component to the spectrum of any reverse complementary component (red points).

H) Assessment of vrnmf performance using a range of windows shows near optimality of the 10 kb scale. Per-component (grey) and average (black) reflection correlations of vrnmf solutions were estimated for windows of 2, 5, 10, 30, 100 and 1000 kilobases using respective volume regularized λ=0.08, 0.06, 0.0078, 0.003, 0.002, 0.0005. Per-window parameters λ were selected to enable near optimal performance for a given window.

I) The number of detected components does not show saturation with an increasing size of the dataset. The number of components having the reflection property (red) or that match consensus components of the original dataset (blue) were estimated for a range of subsampling depths (X axis). Mutations of the original dataset were subsampled up to 1% of the data.

J) Estimates of the dataset size sufficient to detect individual components. Recovery of each component (Y axis), measured as the maximum absolute Pearson correlation of the original component with components inferred from subsampled datasets.

K) Volume-regularization and denoising of NMF significantly contribute to improved performance of vrnmf. Standard and non-smooth NMFs applied to the original and denoised matrices are compared to vrnmf, using the number of reflected components (Pearson reflection correlation of more than 0.8) as a quality metric. The best solutions (red) across 30 random initializations (grey) indicate that denoising improves performance of both NMF and non-smooth NMF, but only volume regularization predicts a single high-quality solution. Standard and non-smooth NMFs were run using *NMF* R package with 'brunet' and 'nsNMF' options respectively.

L) Recurrence corrected spectra of TOPMed 10 kb windows correlate with the spectra without recurrence correction (first panel from left), recurrence corrected spectra of TOPMed 10 kb windows correlate with the spectra calculated for singleton SNVs only (second panel from left) and with the spectra for rare variants excluding singleton SNVs (third panel from left),

Recurrence corrected spectra of TOPMed 10 kb windows correlate with the spectra calculated for gnomAD 100kb windows (right).

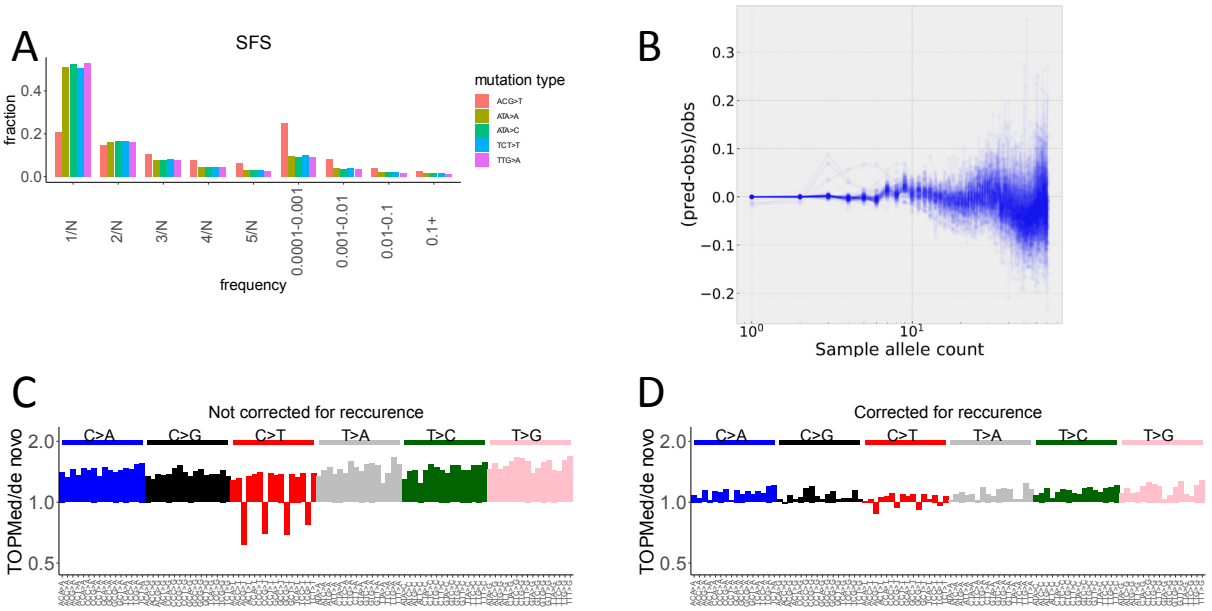

Figure S2. Recurrence adjusted mutational rates.

A) Site frequency spectra (SFS) for 5 different mutation types. Recurrence of mutations in CpG context lead to substantial depletion of singletons (see ACG>T mutations). Using differences in site frequency spectra between mutation types, our method estimates mutation recurrence for allele frequency class.

B) Comparison of the observed SFS with the expected SFS calculated using the recurrence model. Deviations between the observed and expected SFS are shown as function of allele frequency.

C,D) Mutation spectra of SNVs from TOPMed uncorrected for recurrence (**C**) and corrected for recurrence (**D**), divided by the spectra of *de novo* mutations. Values above 1 correspond to mutation types that are relatively more common among TOPMed SNVs; values below 1 correspond to mutation types depleted among TOPMed SNVs.

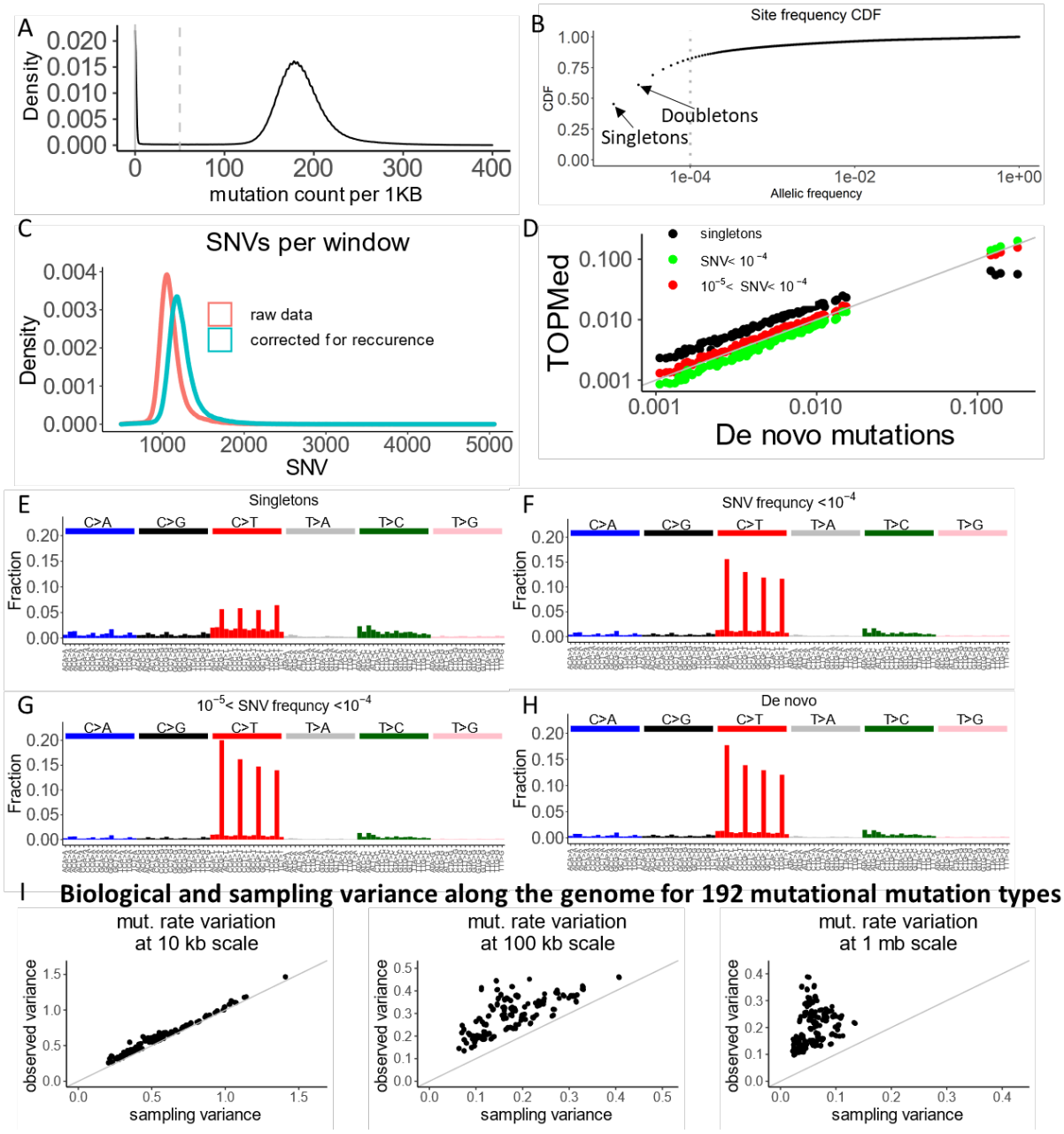

Figure S3. Properties of rare SNVs in TOPMed.

A) Distribution of the number of SNVs per 1 kb window. Windows with less than 50 SNVs (dashed grey line) were excluded.

B) Cumulative distribution of SNVs frequencies. 84% of the variants have allele frequency below 10^{-4}.

C) Distribution of the number of variants per 10 kb window, before and after adjustment for recurrent mutations.

D) Fractions of each of the 96 trinucleotide mutation types (unlike in the rest of this work, complementary mutation types are collapsed here) in the *de novo* mutation dataset and TOPMed SNVs at various allele frequency thresholds.

E-H) Spectra of SNVs in TOPMed with different allele frequency thresholds.

I) Standard deviation (SD) calculated for each of the 192 mutation types. On the X axis we show sampling variance (sampling noise), for uniform mutation rate along the genome for each mutation type. Y axis shows the observed deviance of the mutational frequency, resulting from both noise and biological variance.

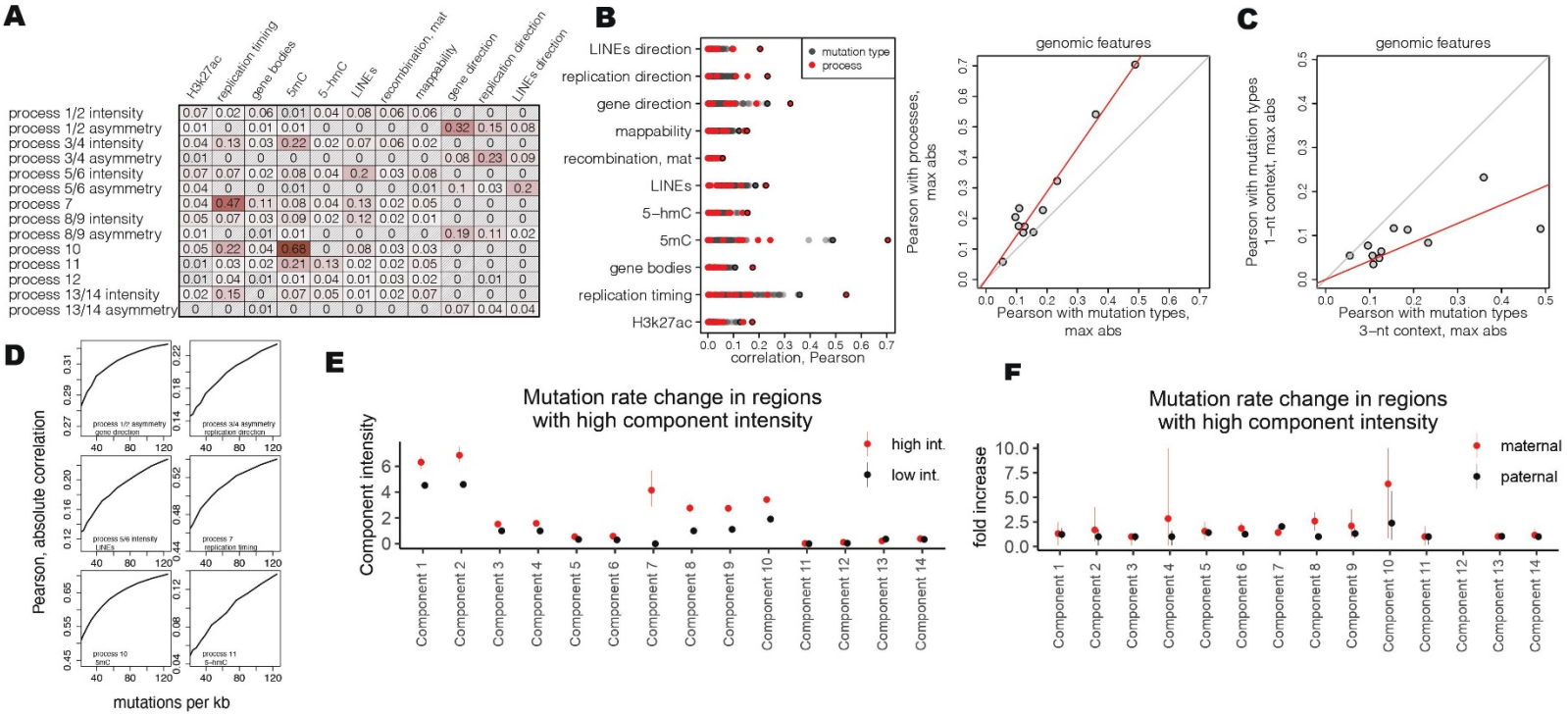

Figure S4. Germline processes correlate with genomic features.

A) Heatmap of ANOVA associations between intensities and genome features. Values indicate square root of relative increase of explained variance of a component by a genome feature after controlling for other genome features. As in Figure 1G, shaded heatmap elements indicate p-value > 0.001 (after Bonferroni correction) of the ANOVA model comparison of nested linear regressions of a component on genomic features with and without the selected genomic feature.

B) Intensities of mutational processes show significantly higher correlation with genomic features compared to raw mutation rates. Left: Pearson correlation of spatial tracks between each genomic feature (Y axis) and each process (red) or mutation type (grey) (X axis). Right: Maximum absolute Pearson correlation of each genomic feature (dots) with mutation processes (Y axis) and the 192 mutation type rates (X axis). For each genomic feature there is a mutational process that, on average, improves correlation by 40% (red linear fit compared to grey) compared to that of raw mutation type rates.

C) Consideration of tri-nucleotide contexts of mutation types enables improved correlation with genomic features. Comparison of the maximum absolute Pearson correlation of each of the eleven genomic features with mutation rates of the 192 tri-nucleotide mutation types (X-axis) and with the 14 standard point mutation types (12 point mutations and two strand-specific CpG>T) across genomic windows. Of note, the estimate of 192 tri-nucleotide mutation rates is statistically significantly nosier compared to 14 point mutation types due to sampling errors, indicative that the tri-nucleotide context confers unique biological information.

D) Correlations between the intensities of the mutational processes and genomic features are likely significantly underestimated due to insufficient sample size. Intensities estimated based on a subsampled TOPMed dataset (X axis: subsampling depth) show robust increase in

correlations with genomic features (Y axis: absolute Pearson correlation of intensity of a mutational process with a genomic feature) as sample size increases. This indicates that the true correlation is likely significantly higher than the correlation estimated based on the TOPMed dataset. A representative set of processes and their prominent genomic covariates are chosen.

E) To validate the spectra of the processes and their intensities we calculated their prevalence among *de novo* mutations. The genome was stratified by the intensity of each component in TOPMed (discovery dataset) and we measured the intensity of the corresponding component in the genomic bin that is expected to have a high intensity (top decile in discovery dataset) for *de novo* mutations (validation dataset). Error bars shows 95% bootstrap confidence intervals.

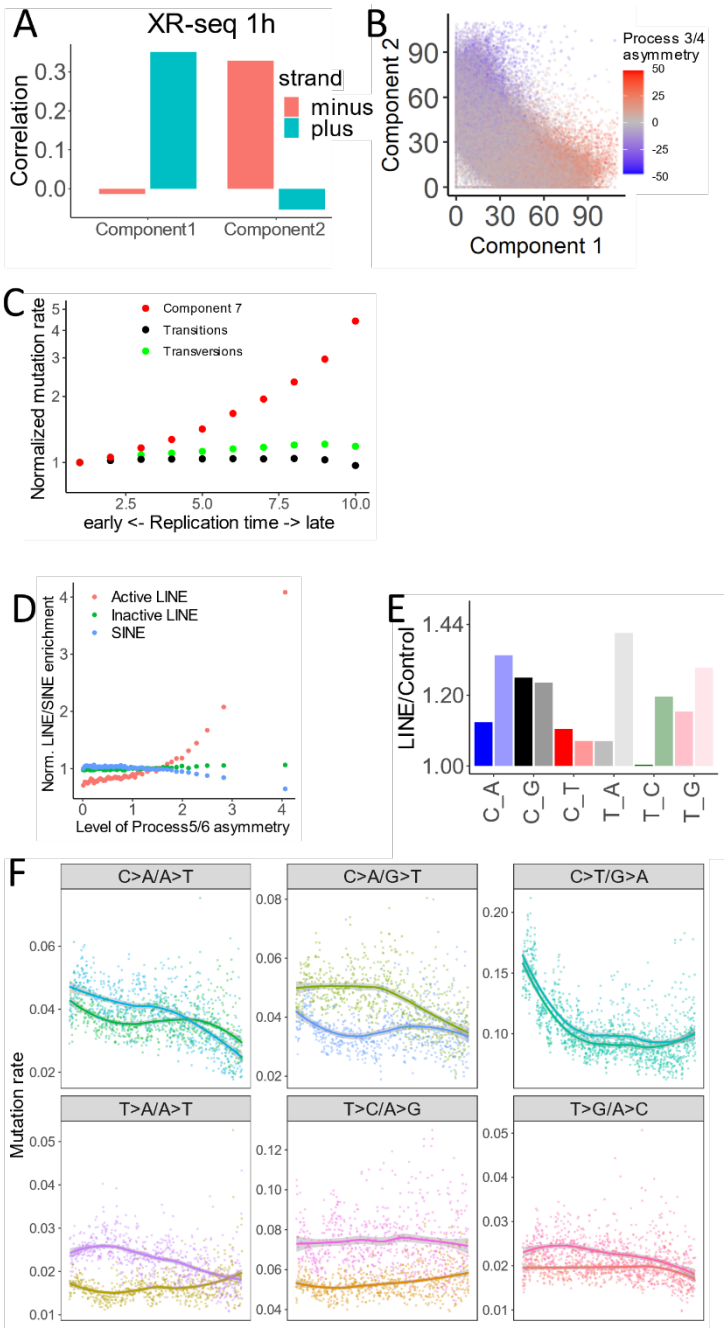

Figure S5. Germline processes correlate with genomic features.

A) Component 1 and component 2 show strand-specific correlation with XR-seq from Adar et al, 2016 (*28*). XR-seq measures activity of the NER system.

B) Intensities of process 1/2 on two DNA strands show a strong anticorrelation across genomic windows (r=-0.67). Asymmetry of process 1/2 is associated with the asymmetry of process 3/4.

C) Rates of transversions, transitions and intensity of process 7 across deciles of replication timing. Mutation rates in each decile were normalized to the mutation rate in the first decile.

D) Average fraction, normalized to the genome average, of nucleotides annotated as LINE repeats (pink: LINE from L1PA family, cyan: LINE from other families) across fifty genomic bins stratified by absolute value of the process 5/6 asymmetry.
E) Mutation rate on non-transcribed strand of LINE repeats normalized to the mutation rate in 10 kb flanks surrounding the LINEs.
F) Mutation rate in 10 nucleotide windows from start to the end of full L1PA LINE repeats. Complementary mutations are shown on the same panel.

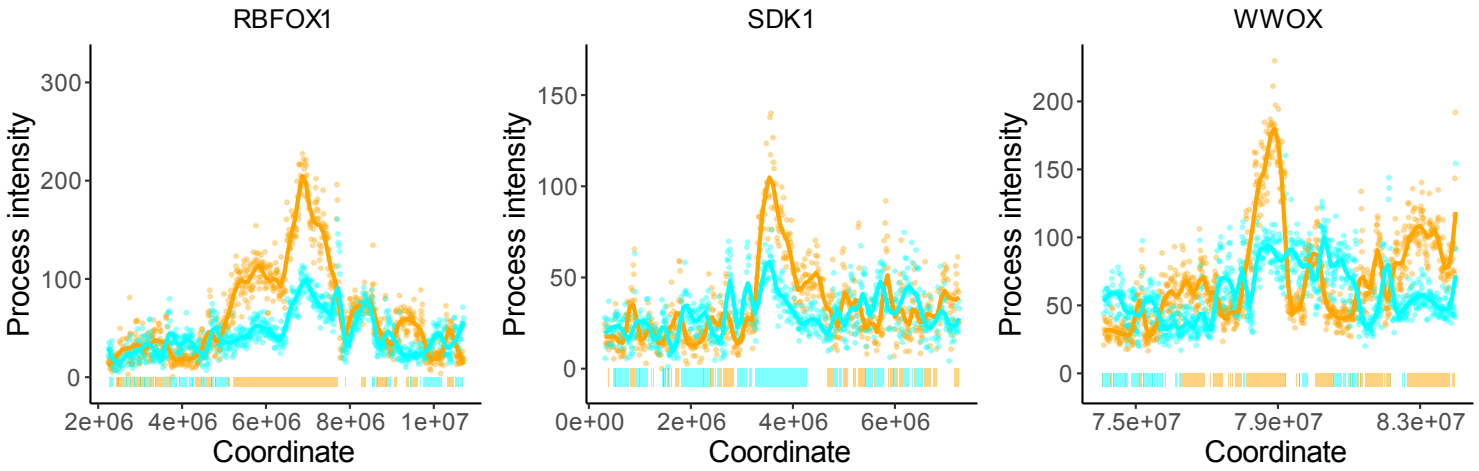

Figure S6. Spikes of the process 8/9.

Process 8/9 intensity spikes around genes *SDK1*, *WWOX* and *RBFOX1* on their non-transcribed strands. Bars at the bottom depict gene bodies (colors: cyan if transcribed strand is the reference strand and orange otherwise).

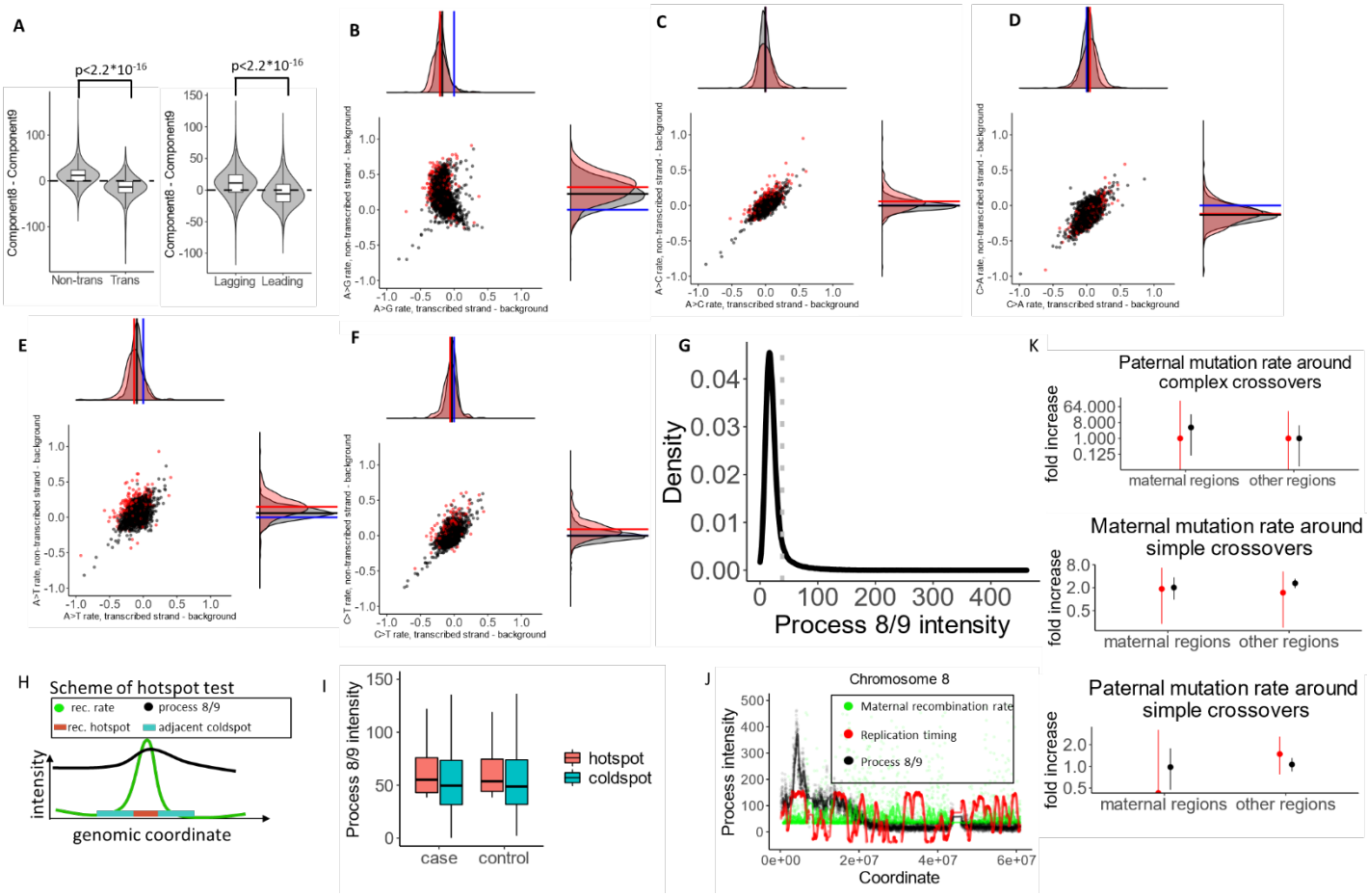

Figure S7. Properties of process 8/9.

A) Violin plots show asymmetry of process 8/9 with respect to transcription and replication.

B-F) Mutation type-specific comparison of mutation rates on the transcribed and non-transcribed strands of genes relative to 100 KB flanking regions. Red dots correspond to genes within maternal regions and black dots correspond to genes outside of maternal regions. Density plots on the right and at the top summarize the distributions on Y and X axes.

G) Heavy-tail distribution of the process 8/9 intensity. Maternal regions are defined as 10% of windows of highest process 8/9 intensity (grey line).

H) "Recombination hotspot test" to assess association of process 8/9 with maternal recombination within maternal regions. Genomic windows (10kb) are classified as hotspots or coldspots if window-averaged maternal recombination rate is in the highest decile or below the genome average respectively. The effect of recombination within maternal regions can be quantified by comparison of process 8/9 intensity between adjacent hoptspot and coldspot windows. We have chosen random windows within the maternal regions as a control.

I) Intensity of process 8/9 in recombination hotspots and in adjacent coldspots within the maternal regions. We have chosen random windows within the maternal regions as a control.

J) Example of a lack of strong association of process 8/9 intensity with the maternal recombination rate and replication timing on the left arm of chromosome 8.

K) Fold change in *de novo* mutation rate in 100kb windows around crossovers in spikes of process 8/9 and other regions.

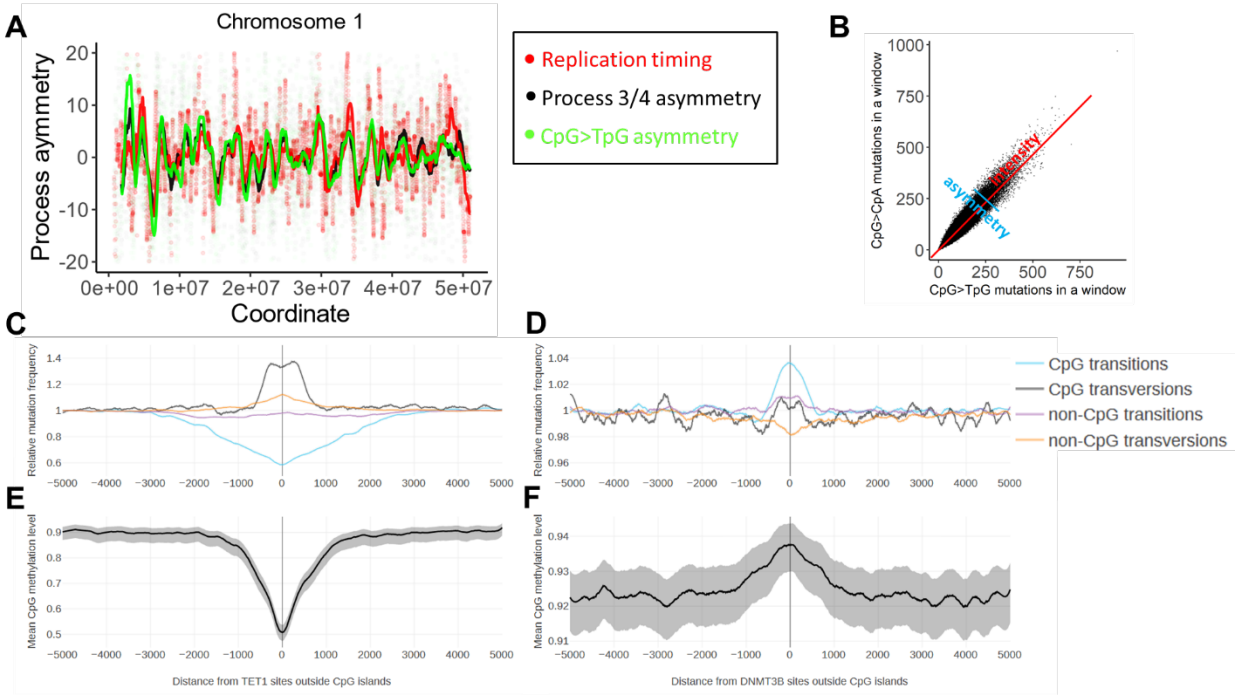

Figure S8. Weak asymmetry of CpG>TpG mutations associated with replication fork direction

A) Strong correlation of asymmetry of CpG>TpG/CpG>CpA mutations (green), asymmetry of process 3/4 (black) and derivative of replication timing (red). Role of the replication and deamination in CpG>TpG mutagenesis discussed in **supplementary text1.**

B) Reverse complementary CpG>TpG and CpG>CpA mutation types have strong positive correlation of mutation rates across genomic windows. Correlation of reverse complementary mutation rates is related to the balance of intensity and asymmetry of the underlying mutation processes. Strong positive correlation indicates that strand-independent factors of mutagenesis modulate CpG>TpG mutation rate is significantly stronger in comparison to strand-specific factors.

C) Mutation frequencies surrounding demethylation enzyme TET1 ChIP-seq sites outside of CpG islands.

D) Mutation frequencies surrounding methylation enzyme DNMT3B ChIP-seq sites outside of CpG islands.

E) Methylation levels surrounding TET1 sites outside of CpG islands. Shaded regions represent binomial proportion confidence intervals.

F) Methylation levels surrounding DNMT3B sites outside of CpG islands.

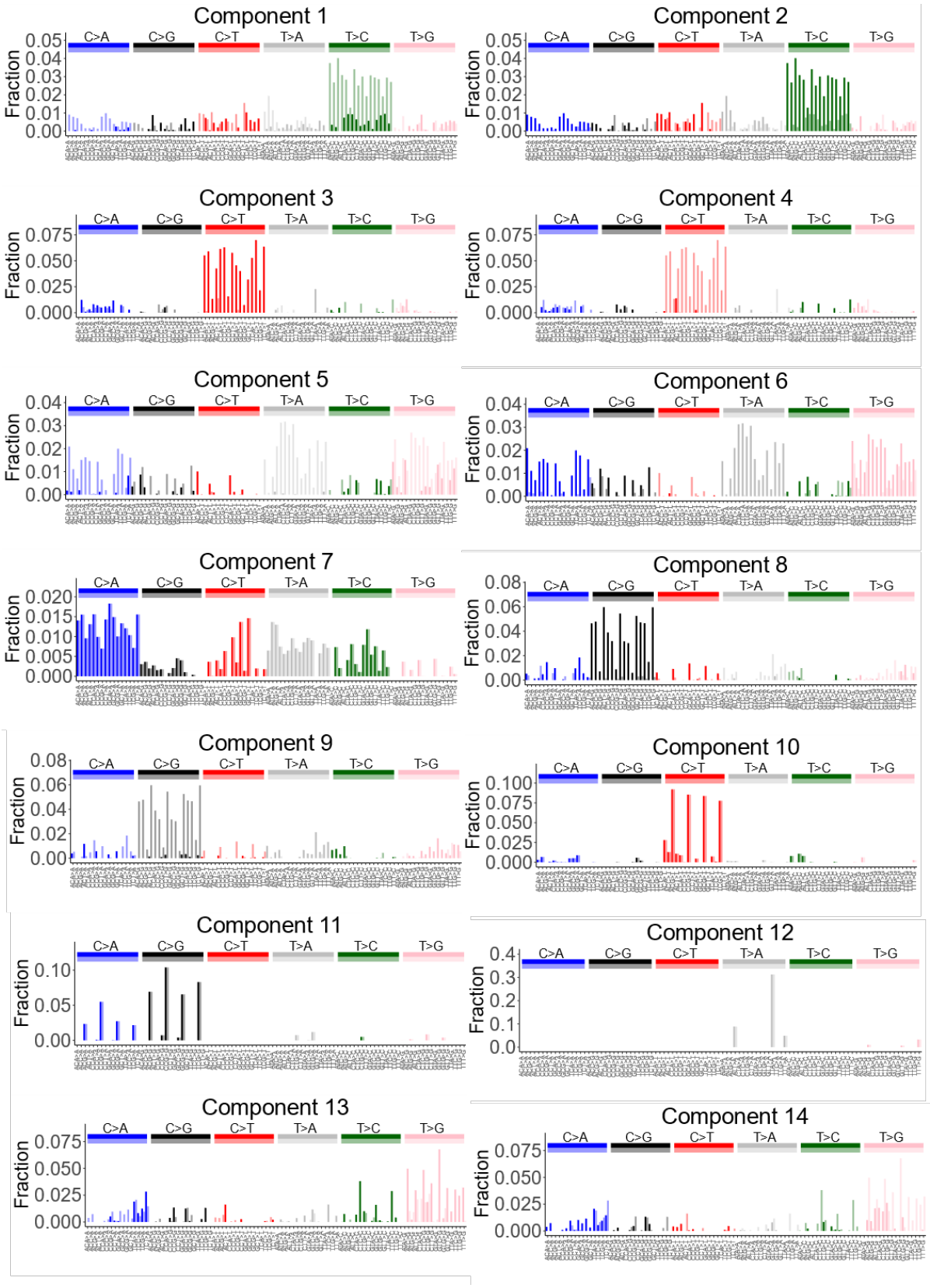

Figure S9. Spectra of the 14 components, where each mutation type is normalized to its standard deviation

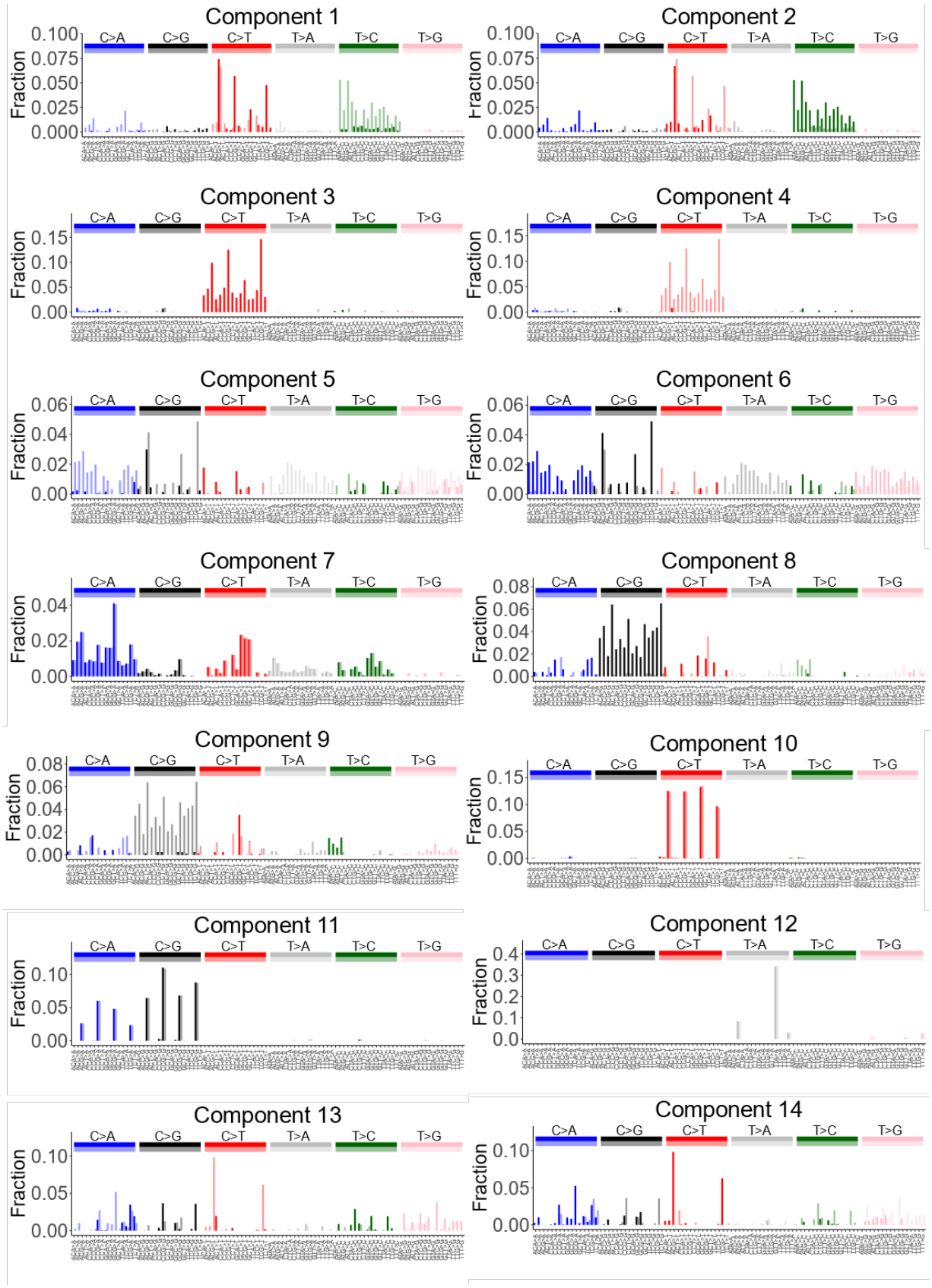

Figure S10. Unnormalized spectra of the 14 components.

Supplementary Tables 1-5

Table S1. ANOVA II associations of the process 1/2 asymmetry with the expression level of 54 tissues. The values show the increase of explained variance of the process 1/2 asymmetry by an expression level in a cell line, after controlling for the effect of expression in other cell lines. Tissues sorted by the value of explained variance.

ANOVA II: explained variance	Cell line
612.07	Testis
165.14	Brain.Frontal.Cortex.BA9.
146.83	Brain.Caudate.basal.ganglia.
128.38	Brain.Putamen.basal.ganglia.
97.49	Whole.Blood
73.25	Brain.Spinal.cord.cervical.c.1.
62.16	Pituitary
60.02	Fallopian.Tube
44.59	Adrenal.Gland
41.78	Brain.Substantia.nigra
39.90	Brain.Anterior.cingulate.cortex.BA24.
37.39	Brain.Cortex
33.13	Nerve.Tibial
31.80	Brain.Cerebellum
29.99	Artery.Tibial
29.01	Brain.Cerebellar.Hemisphere
26.69	Bladder
26.57	Esophagus.Gastroesophageal.Junction
24.58	Stomach
24.40	Cervix.Ectocervix
23.78	Small.Intestine.Terminal.Ileum
23.72	Ovary
22.78	Thyroid
22.64	Vagina
20.49	Artery.Coronary
18.19	Cells.EBV.transformed.lymphocytes
17.44	Brain.Nucleus.accumbens.basal.ganglia.

17.08	Esophagus.Muscularis
10.03	Uterus
9.63	Colon.Transverse
7.69	Muscle.Skeletal
7.50	Breast.Mammary.Tissue
7.08	Lung
6.73	Liver
6.57	Heart.Atrial.Appendage
6.14	Brain.Amygdala
4.51	Esophagus.Mucosa
4.06	Heart.Left.Ventricle
3.10	Kidney.Medulla
3.00	Adipose.Subcutaneous
2.46	Pancreas
2.14	Artery.Aorta
1.15	Prostate
1.12	Kidney.Cortex
0.78	Colon.Sigmoid
0.78	Skin.Sun.Exposed.Lower.leg.
0.71	Minor.Salivary.Gland
0.45	Brain.Hypothalamus
0.42	Cervix.Endocervix
0.28	Adipose.Visceral.Omentum.
0.19	Cells.Cultured.fibroblasts
0.05	Skin.Not.Sun.Exposed.Suprapubic.
0.04	Spleen
0.03	Brain.Hippocampus

Table S2. Optimal scale of inferred processes. Each process was smoothed with sliding windows 10 kb, 20 kb, 30 kb, 50 kb, 70 kb, 100 kb, 150 kb, 200 kb, 300 kb, 50 kb, 700 kb, 1000 kb, 1500 kb or 2000 kb in size. For each window size, the square root of the variance explained by the epigenetic features was calculated. Window size that maximizes explained variance considered as the optimal scale of the process.

Process	square root of explained variance without smoothing	optimal scale, KB	square root of explained variance with smoothing
Process 1/2 intensity	0.17	10	0.17
Process 1/2 asymmetry	0.36	20	0.38

Process 3/4 intensity	0.31	10	0.31
Process 3/4 asymmetry	0.26	100	0.36
Process 5/6 intensity	0.25	10	0.25
Process 5/6 asymmetry	0.23	10	0.23
Process 7	0.54	500	0.64
Process 8/9 asymmetry	0.18	10	0.18
Process 8/9 intensity	0.22	50	0.26
Process 10	0.75	10	0.75
Process 11	0.25	10	0.25
Process 12	0.07	500	0.18
Process 13/14 intensity	0.18	1000	0.56
Process 13/14 asymmetry	0.09	150	0.15

Table S3. Comparison of the maternal mutation rate in known and novel "maternal regions" and in the rest of the genome. Maternal mutations are from (*11*).

		Total size of regions	Number of all maternal mutations	Enrichment
Known (predicted maternal regions on chromosomes: 2,7,8,9,16)	Maternal regions	145 MB	2356	3.08
	All other regions	568 MB	3000	
Novel (predicted maternal regions on other chromosomes)	Maternal regions	119 MB	1200	1.91
	All other regions	1806 MB	9578	

Table S4. Comparison of clustered maternal mutation rate in known and novel "maternal regions" and in the rest of the genome. Maternal mutations are from (*11*).

		Total size of regions	Number of clustered maternal mutations	Enrichment
Known (predicted maternal regions on chromosomes: 2,7,8,9,16)	Maternal regions	145 MB	687	25.6
	All other regions	568 MB	105	
Novel (predicted maternal regions	Maternal regions	119 MB	147	8.13

on other chromosomes)	All other regions	1806 MB	275	

Table S5. Study Sequencing Acknowledgements.

TOPMed Accession #	Parent Study Short Name	Parent Study Full Name	TOPMed Phase	TOPMed Project	Omics Center	Omics Support Grant/Contract Number
phs000956	Amish	Genetics of Cardiometabolic Health in the Amish	CCDG co-funded	AFGen	BROAD	3R01HL121007-01S1
phs001211	ARIC	Atherosclerosis Risk in Communities Study	1	AFGen	BROAD	3R01HL092577-06S1
phs001211	ARIC	Atherosclerosis Risk in Communities Study VTE cohort	2	VTE	BAYLOR	3U54HG003273-12S2, HHSN268201500015C
phs001143	BAGS	New Approaches for Empowering Studies of Asthma in Populations of African Descent - Barbados Asthma Genetics Study	1	BAGS	ILLUMINA	3R01HL104608-04S1
phs001189	CCAF	Cleveland Clinic Atrial Fibrillation Study	1	AFGen	BROAD	3R01HL092577-06S1
phs000954	CFS	Cleveland Family Study - WGS Collaboration	1	CFS	UW NWGC	3R01HL098433-05S1
phs000954	CFS	Cleveland Family Study - WGS Collaboration	3.5	CFS	UW NWGC	HHSN268201600032I
phs001368	CHS	Cardiovascular Health Study	2	VTE	BAYLOR	3U54HG003273-12S2, HHSN268201500015C
phs001368	CHS	Cardiovascular Health Study	3	CHS	BAYLOR	HHSN268201600033I
phs000951	COPDGene	Genetic Epidemiology of COPD Study	1	COPD	UW NWGC	3R01HL089856-08S1
phs000951	COPDGene	Genetic Epidemiology of COPD Study	2	COPD	BROAD	HHSN268201500014C
phs000951	COPDGene	Genetic Epidemiology of COPD Study	2.5	COPD	BROAD	HHSN268201500014C
phs000988	CRA	The Genetic Epidemiology of Asthma in Costa Rica - Asthma in Costa Rica cohort	1	CRA_CAMP	UW NWGC	3R37HL066289-13S1
phs000988	CRA	The Genetic Epidemiology of Asthma in Costa Rica - Asthma in Costa Rica cohort	3	CRA_CAMP	UW NWGC	HHSN268201600032I
phs001412	DHS	Diabetes Heart Study	2	AA_CAC	BROAD	HHSN268201500014C
phs000974	FHS	Framingham Heart Study	1	AFGen	BROAD	3R01HL092577-06S1
phs000974	FHS	Framingham Heart Study	1	FHS	BROAD	3U54HG003067-12S2
phs000920	GALAII	Gene-Environment, Admixture and Latino Asthmatics Study	1	PGX_Asthma	NYGC	3R01HL117004-02S3

phs000920	GALAII	ATGC Gene-Environment, Admixture and Latino Asthmatics Study II Asthma	3	ATGC	UW NWGC	HHSN268201600032I
phs001345	GENOA	Genetic Epidemiology Network of Arteriopathy	2	AA_CAC	BROAD	HHSN268201500014C
phs001345	GENOA	Genetic Epidemiology Network of Arteriopathy	2	HyperGEN_GENOA	UW NWGC	3R01HL055673-18S1
phs001217	GenSalt	Genetic Epidemiology Network of Salt Sensitivity	2	GenSalt	BAYLOR	HHSN268201500015C
phs001359	GOLDN	Genetics of Lipid Lowering Drugs and Diet Network	2	GOLDN	UW NWGC	3R01HL104135-04S1
phs000993	HVH	Heart and Vascular Health Study	1	AFGen	BROAD	3R01HL092577-06S1
phs000993	HVH	Heart and Vascular Health Study	2	VTE	BAYLOR	3U54HG003273-12S2, HHSN268201500015C
phs001293	HyperGEN	Hypertension Genetic Epidemiology Network	2	HyperGEN_GENOA	UW NWGC	3R01HL055673-18S1
phs000964	JHS	Jackson Heart Study	1	JHS	UW NWGC	HHSN268201100037C
phs001402	Mayo_VTE	Mayo Clinic Venous Thromboembolism Study	2	VTE	BAYLOR	3U54HG003273-12S2, HHSN268201500015C
phs001416	MESA	Multi-Ethnic Study of Atherosclerosis	2	MESA	BROAD	3U54HG003067-13S1
phs001062	MGH_AF	Massachusetts General Hospital Atrial Fibrillation Study	1	AFGen	BROAD	3R01HL092577-06S1
phs001062	MGH_AF	Massachusetts General Hospital Atrial Fibrillation Study	1.4	AFGen	BROAD	3U54HG003067-12S2, 3U54HG003067-13S1
phs001062	MGH_AF	Massachusetts General Hospital Atrial Fibrillation Study	1.5	AFGen	BROAD	3U54HG003067-12S2, 3U54HG003067-13S1
phs001062	MGH_AF	Massachusetts General Hospital Atrial Fibrillation Study	CCDG co-funded	AFGen	BROAD	3UM1HG008895-01S2
phs001024	Partners	Partners Healthcare Biorepository	1	AFGen	BROAD	3R01HL092577-06S1
phs001215	SAFS	Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies	1	SAFS	ILLUMINA	3R01HL113323-03S1
phs001215	SAFS	Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies	legacy	SAFS	ILLUMINA	R01HL113322
phs000921	SAGE	Study of African Americans, Asthma, Genes and Environment	1	PGX_Asthma	NYGC	3R01HL117004-02S3
phs000921	SAGE	ATGC Study of African Americans, Asthma, Genes and Environment	3	ATGC	UW NWGC	HHSN268201600032I
phs001207	Sarcoidosis	Genetics of Sarcoidosis in African Americans	2	Sarcoidosis	BAYLOR	3R01HL113326-04S1
phs001207	Sarcoidosis	Genetics of Sarcoidosis in African Americans	3.5	Sarcoidosis	UW NWGC	HHSN268201600032I
phs000972	SAS / Samoans	Samoan Adiposity Study	1	SAS / Samoans	UW NWGC	HHSN268201100037C
phs000972	SAS / Samoans	Samoan Adiposity Study	2	SAS / Samoans	NYGC	HHSN268201500016C

phs001387	THRV	Taiwan Study of Hypertension using Rare Variants	2	THRV	BAYLOR	3R01HL111249-04S1, HHSN26820150015C
phs000997	VAFAR	Vanderbilt Atrial Fibrillation Ablation Registry	1	AFGen	BROAD	3R01HL092577-06S1
phs000997	VAFAR	Vanderbilt Atrial Fibrillation Ablation Registry	1.5	AFGen	BROAD	3U54HG003067-12S2, 3U54HG003067-13S1
phs000997	VAFAR	Vanderbilt Atrial Fibrillation Ablation Registry	CCDG co-funded	AFGen	BROAD	3UM1HG008895-01S2
phs000997	VAFAR	Vanderbilt Atrial Fibrillation Ablation Registry	CCDG co-funded year 2	AFGen	BROAD	3UM1HG008895-01S2
phs001032	VU_AF	Vanderbilt Genetic Basis of Atrial Fibrillation	1	AFGen	BROAD	3R01HL092577-06S1
phs001040	WGHS	Women's Genome Health Study	1	AFGen	BROAD	3R01HL092577-06S1

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. See supplementary materials and methods.

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