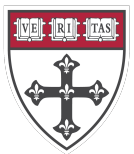
Poisson process factorization for mutational signature analysis with genomic covariates

Jeffrey W. Miller

Department of Biostatistics, Harvard University

Nonparametric Bayesian Inference - Computational Issues
Jan 12, 2026 || ICERM || Brown University



HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Dana-Farber
Cancer Institute

Giovanni Parmigiani

Alessandro Zito

A blast from the past...

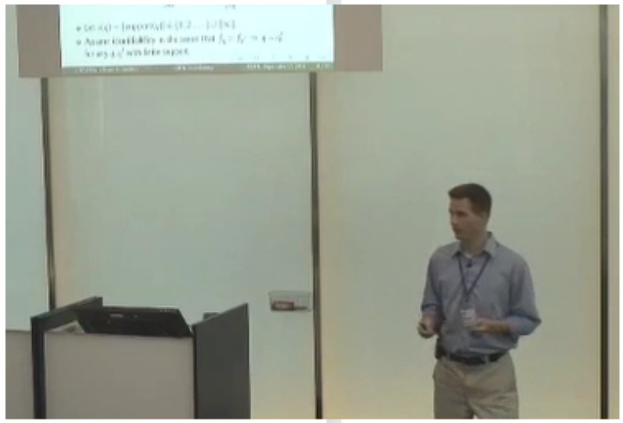
My first introduction to the Bayesian community was speaking at the ICERM workshop on Bayesian Nonparametrics in 2012 as a PhD student at Brown.

Dirichlet process mixtures are inconsistent for the number of components in a finite mixture

Jeffrey W. Miller
and
Matthew T. Harrison

Division of Applied Mathematics
182 George Street
Providence, RI 02912

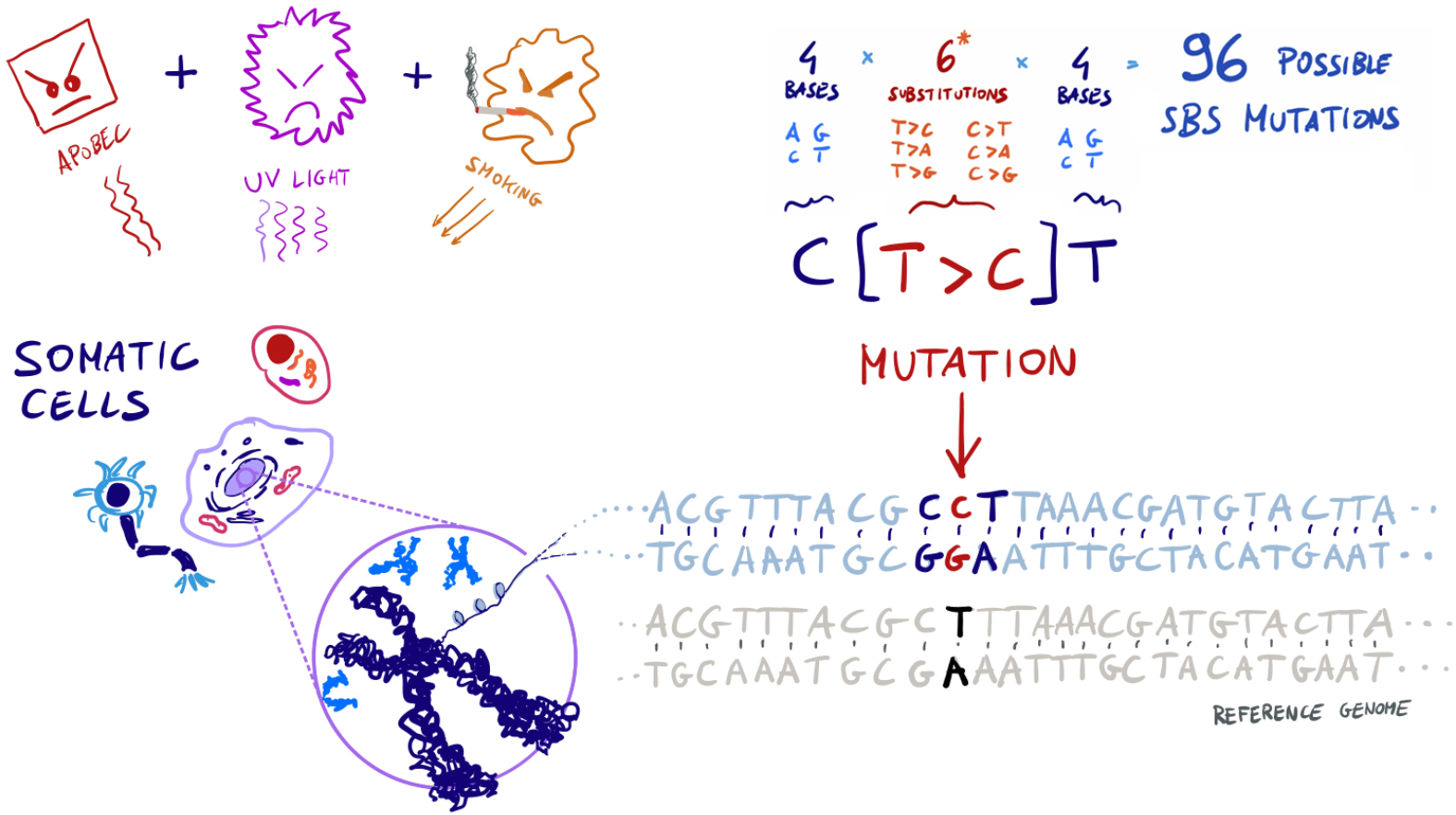
ICERM, September 17, 2012



Cancer cells have DNA mutations due to many processes

- Various processes cause mutations, such as environmental exposures and cellular dysregulation.

- Each mutational process has been found to consistently produce each mutation type at a relatively constant rate.



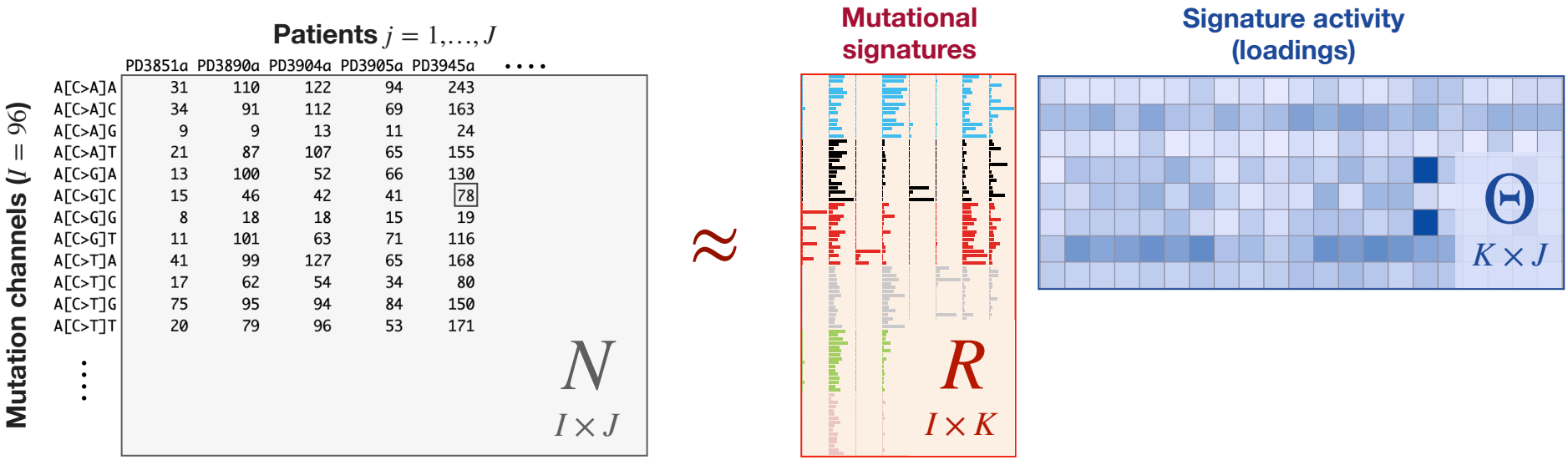
Mutational signatures analysis

- Non-negative matrix factorization (NMF) is used to recover these rates (referred to as "mutational signatures") and patient-specific exposures (Alexandrov et al., 2013).

- For mutation type $i = 1, \ldots, I$ and patient $j = 1, \ldots, J$, the usual model to let

$$N_{ij} \sim \text{Poisson}\left(\sum_{k=1}^{K} r_{ik}\theta_{kj}\right)$$

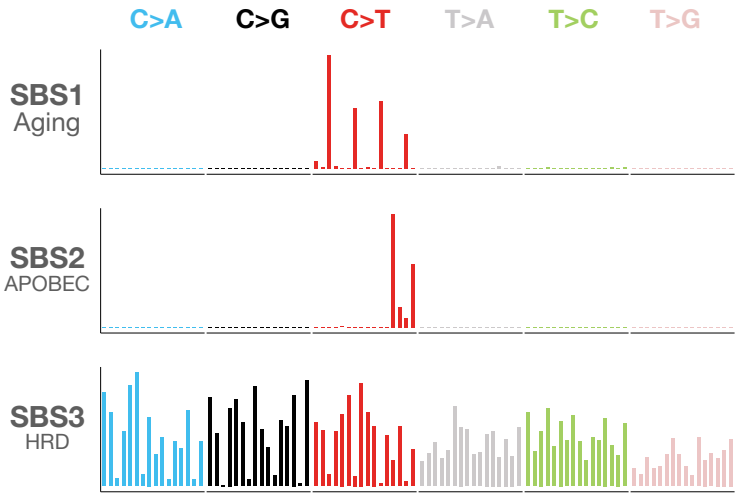
be the number of mutations observed, where $r_{ik}, \theta_{kj} \geq 0$ and $\sum_i r_{ik} = 1$.



Existing methods assume homogeneity across the genome

- Numerous mutational signatures have been discovered with the NMF approach, opening up exciting new directions in cancer research and treatment.
- Existing methods implicitly assume that mutational process activity is homogeneous across the genome.

COSMIC
Catalogue Of Somatic Mutations In Cancer



Mutational signatures improve precision therapies

RESEARCH ARTICLES | AUTHOR CHOICE | SEPTEMBER 04 2018

Real-time Genomic Characterization of Advanced Pancreatic Cancer to Enable Precision Medicine FREE

Andrew J. Aguirre; Jonathan A. Nowak; Nicholas D. Camarda; Richard A. Moffitt; Arezou A. Ghazani; Mehlika Hazar-Rethinam; Srivatsan Raghavan; Jaegil Kim; Lauren K. Brais; Dorisanne Ragon; Marisa W. Welch; Emma Reilly; Devin McCabe; Lori Marini; Kristin Anderka; Karla Helvie; Nelly Oliver; Ana Babic; Annacarolina Da Silva; Brandon Nadres; Emily E. Van Seventer; Heather A. Sha Nadine J. McCl Douglas A. Rut Deborah Knoer Bruce E. Johns William C. Hahn

TECHNICAL REPORT
https://doi.org/10.1038/s41588-019-0390-2
nature genetics

Check f
+ Author & Art
Cancer Discov
https://doi.org/1

Detecting the mutational signature of homologous recombination deficiency in clinical samples

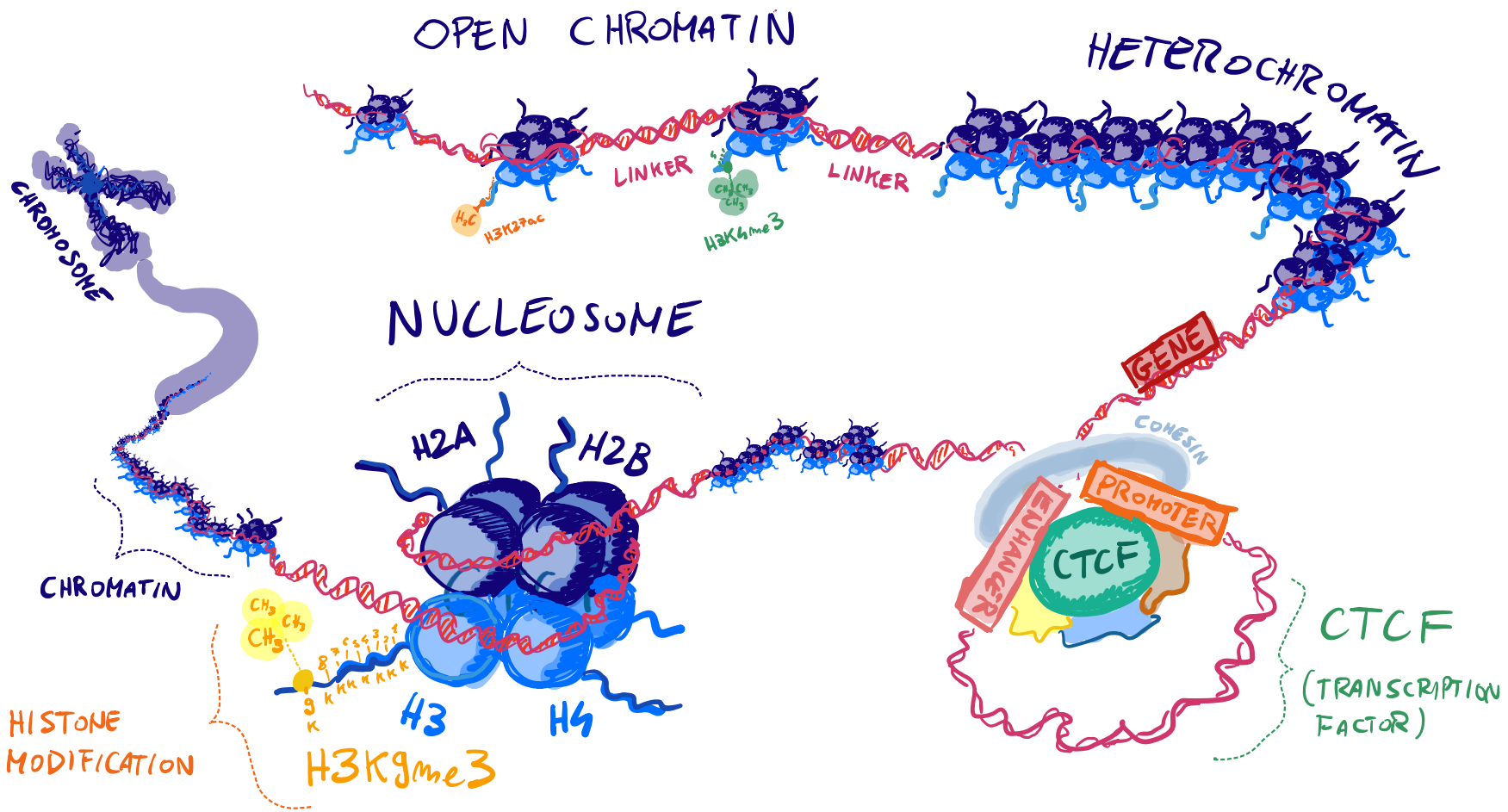
Doga C. Gulhan[1], Jake June-Koo Lee[1], Giorgio E. M. Melloni[1], Isidro Cortés-Ciriano[1,2] and Peter J. Park[1*]

Mutations in *BRCA1* and/or *BRCA2* (*BRCA1/2*) are the most common indication of deficiency in the homologous recombination (HR) DNA repair pathway. However, recent genome-wide analyses have shown that the same pattern of mutations found in *BRCA1/2*-mutant tumors is also present in several other tumors. Here, we present a new computational tool called Signature Multivariate Analysis (SigMA), which can be used to accurately detect the mutational signature associated with HR deficiency from targeted gene panels. Whereas previous methods require whole-genome or whole-exome data, our method detects the HR-deficiency signature even from low mutation counts, by using a likelihood-based measure combined with machine-learning techniques. Cell lines that we identify as HR deficient show a significant response to poly (ADP-ribose) polymerase (PARP) inhibitors; patients with ovarian cancer whom we found to be HR deficient show a significantly longer overall survival with platinum regimens. By enabling panel-based identification of mutational signatures, our method substantially increases the number of patients that may be considered for treatments targeting HR deficiency.

Heterogeneity of the genome

- However, the genome is highly heterogeneous. The structure and cellular processes of the genome are affected by a range of features that are position-specific.

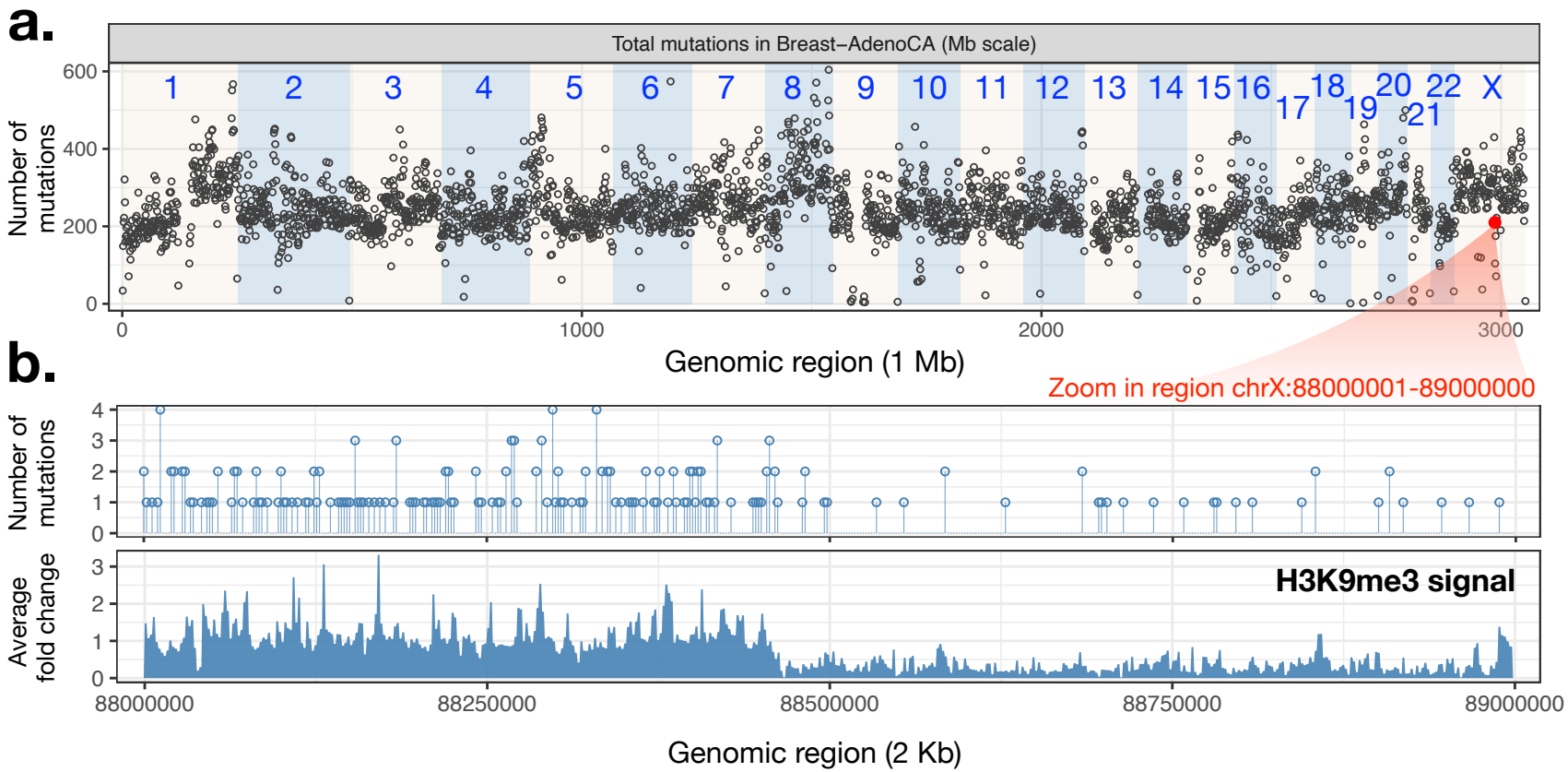
- Features such as GC content, methylation, DNA accessibility, and epigenetic modifications like histones marks strongly affect genomic processes.



Signature activity correlates with genomic features

- Mutational burden varies across the genome, and recent work has found that mutational signature activity correlates with epigenetic marks (Otlu et al., 2023).

- However, Otlu et al. (2023) is based on *post hoc* correlation with epigenetic marks, rather than modeling the joint effect of genomic features.



Objective and Challenges

Objective

- A Bayesian framework that **links genomic features with mutational signatures**.

Challenges

- Position-specific modeling of mutational signature activity for each patient is needed.
- Copy number affects exposure in a patient-specific and position-specific manner.
- Selecting K, the rank of the factorization, is important to avoid over- or under-fitting.
- Computation is challenging since there are around 3×10^9 positions in the genome.

Our contribution: Poisson process factorization (PPF)

- We introduce a new Bayesian modeling framework that incorporates:
  1. a **Poisson process** model for spatial count data,
  2. **non-negative matrix factorization** (NMF) of the intensity function, and
  3. **log-linear model** for the NMF weights.

- For mutational signatures analysis, this enables:
  1. attribution of individual mutations to signatures,
  2. improved accuracy of signature estimation, and
  3. inference of the effect of genomic features on mutational signature activity,

- We develop **computationally efficient** estimation and posterior inference algorithms.

Poisson process factorization (PPF)

- An inhomogeneous Poisson process Z on the real line is a completely random measure defined via an intensity function $\lambda : [0, T) \to \mathbb{R}_+$ such that

$$Z(A) \sim \text{Poisson}\Big(\int_A \lambda(t)\, \mathrm{d}t \Big), \quad A \subset [0, T).$$

Definition: Poisson process factorization

A *Poisson process factorization* model is a multivariate Poisson process (Z_{ij}) where the intensity functions $\lambda_{ij} : [0, T) \to \mathbb{R}_+$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$ factor as

$$\lambda_{ij}(t) = \sum_{k=1}^{K} \underbrace{r_{ik}}_{\substack{\text{Mutational} \\ \text{signatures}}} \times \underbrace{\vartheta_{kj}(t)}_{\substack{\text{Position-specific} \\ \text{exposures}}}, \qquad t \in [0, T).$$

- If $T = 1$ and $\vartheta_{kj}(t) = \theta_{kj}$ for all $t \in [0, T)$ and all i, j, k, then this reduces to the usual Poisson NMF model used in previous work.

Log-linear model of position-specific exposures

- For each signature k and patient j, we model the **position-specific exposures** as

$$\vartheta_{kj}(t) = \frac{1}{2}\,\theta_{kj}\,c_j(t)\,e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)}, \qquad t \in [0, T).$$

- **Baseline exposures:** $\theta_{kj} \geq 0$ is the baseline exposure of patient j to signature k.
- **Copy number:** $c_j : [0, T) \to \mathbb{R}_+$ with $c_j(t) = 2$ under normal conditions.
- **Genomic covariates:** $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))^\top \in \mathbb{R}^p$, with $x_\ell(t)$ denoting the value of covariate ℓ at position t.
- **Regression coefficients:** $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^\top \in \mathbb{R}^p$.
- Although genomic position t is technically discrete, we make a continuous approximation by using $t \in [0, T) \subset \mathbb{R}_+$ where $T \approx 3 \times 10^9$ nucleotides.

Prior distributions

Under this model, the number of mutations of type i in region A for patient j is

$$Z_{ij}(A) \mid r, \theta, \beta \sim \text{Poisson}\left(\sum_{k=1}^{K} r_{ik}\theta_{kj} \int_A \frac{1}{2}\, c_j(t)\, e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)} \mathrm{d}t \right).$$

For the priors, we generalize the **compressive NMF** approach (Zito and Miller, 2024):

- Signatures: $r_k = (r_{1k}, \ldots, r_{Ik}) \sim \text{Dirichlet}(\alpha_{1k}, \ldots, \alpha_{Ik})$
- Baseline exposures: $\theta_{kj} \mid \mu_k \sim \text{Ga}\left(a,\ \frac{a}{\mu_k} \int_0^T \frac{1}{2} c_j(t)\, \mathrm{d}t \right)$
- Relevance weights: $\mu_k \sim \text{InvGa}(aJ + 1,\ \varepsilon aJ)$
- Regression coefficients: $\boldsymbol{\beta}_k \mid \sigma_k^2 \sim \text{N}(\mathbf{0}, \sigma_k^2 I_p), \quad \sigma_k^2 \sim \text{InvGa}(100, 1)$

The compressive hyperprior on μ_k shrinks the weights of any unneeded factors to zero, similarly to overfitted mixture models (Rousseau and Mengersen, 2011).

Posterior of the model

- The likelihood of the intensity function λ_{ij} for mutation type i, patient j, is

$$\mathcal{L}(\lambda_{ij}; t_1, \ldots, t_{N_{ij}}) = \exp\left(-\int_0^T \lambda_{ij}(t)\,\mathrm{d}t \right) \prod_{n=1}^{N_{ij}} \lambda_{ij}(t_n)$$

where $t_1, \cdots, t_{N_{ij}}$ are the positions of mutations of type i (Daley and Vere-Jones, 2003).

Log-posterior of the Poisson process factorization (PPF) model

$$\log \pi\left(r, \theta, \beta, \mu, \sigma^2 \,\middle|\, \{t_1, \ldots, t_{N_{ij}}\}_{ij}\right) =$$

$$= -\sum_{jk} \theta_{kj} \int_0^T \frac{1}{2}\, c_j(t)\, e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)} \mathrm{d}t + \sum_{ij} \sum_{n=1}^{N_{ij}} \log \left(\sum_{k=1}^{K} r_{ik} \theta_{kj} \frac{1}{2}\, c_j(t_n)\, e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)} \right)$$

$$+ \log \pi(r)\, \pi(\theta \mid \mu)\, \pi(\mu)\, \pi(\beta \mid \sigma^2)\, \pi(\sigma^2) + \text{const},$$

subject to $\sum_{i=1}^{I} r_{ik} = 1$ for all $k = 1, \ldots, K$.

- We use *maximum a posteriori* (MAP) estimation and MCMC for Bayesian inference.

Attribution of individual mutations to signatures

- For computation, we employ data augmentation with multinomial random variables

$$W_{ij}(t_n) = \big(W_{ij1}(t_n), \ldots, W_{ijK}(t_n)\big) \sim \text{Mult}\big(1;\, p_{ij1}(t_n), \ldots, p_{ijK}(t_n)\big),$$

  where

$$p_{ijk}(t_n) = \mathbb{P}(W_{ijk}(t_n) = 1 \mid r, \theta, \beta) = \frac{r_{ik}\theta_{kj}e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)}}{\sum_{s=1}^{K} r_{is}\theta_{sj}e^{\boldsymbol{\beta}_s^\top \mathbf{x}(t_n)}}.$$

- Conditional on W, this yields **conjugate updates** for all parameters except β_k, for which we use elliptical slice sampling.

- $p_{ijk}(t_n)$ is the probability that a mutation of type i at position t_n in patient j was generated by signature k, under the model.

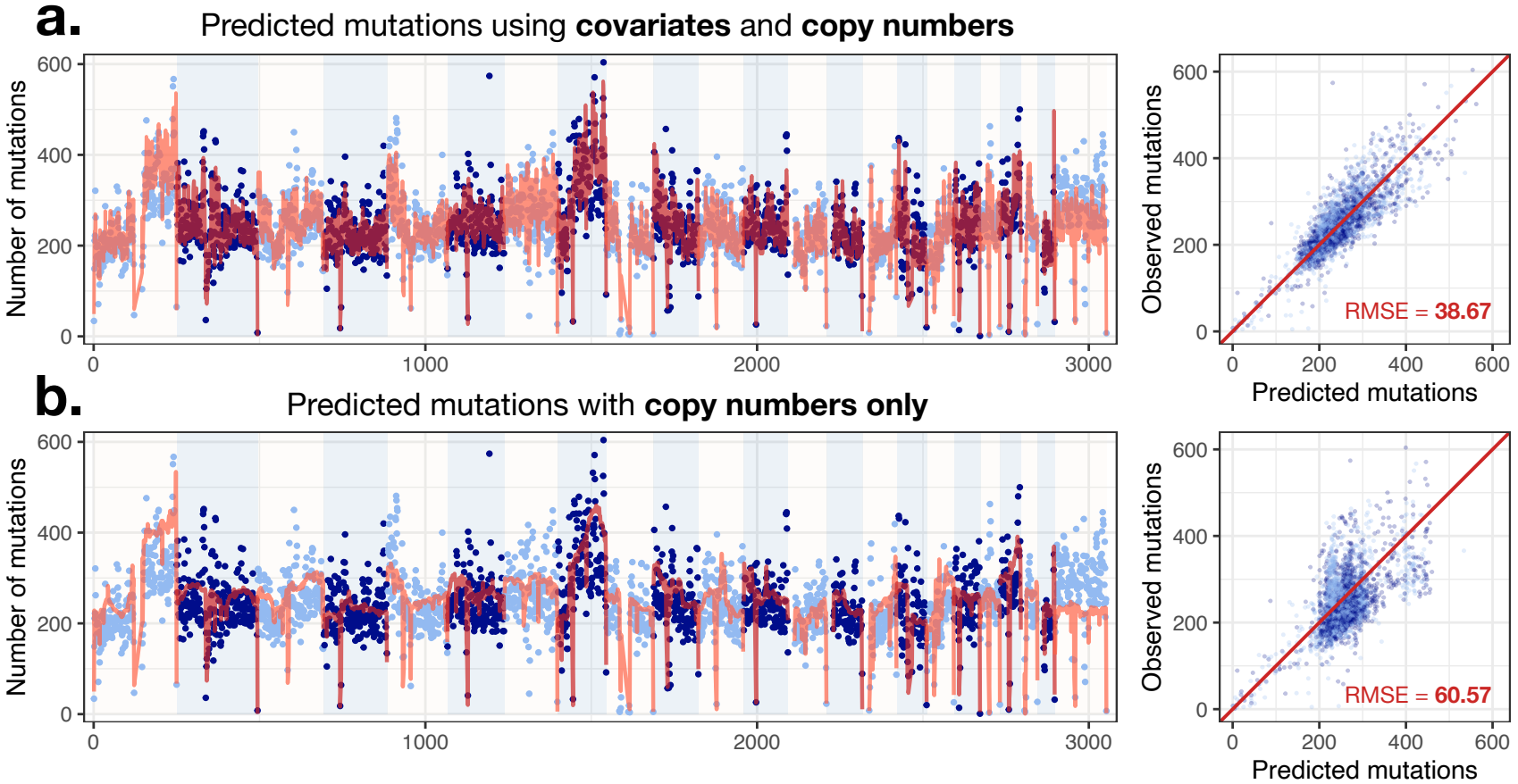
- Unlike usual NMF, **each mutation has its own signature assignment probabilities**.

- The scientific and medical implication of this is that we can infer and quantify uncertainty in the mechanism that generated each mutation.

Analysis of ICGC breast adenocarcinoma cohort

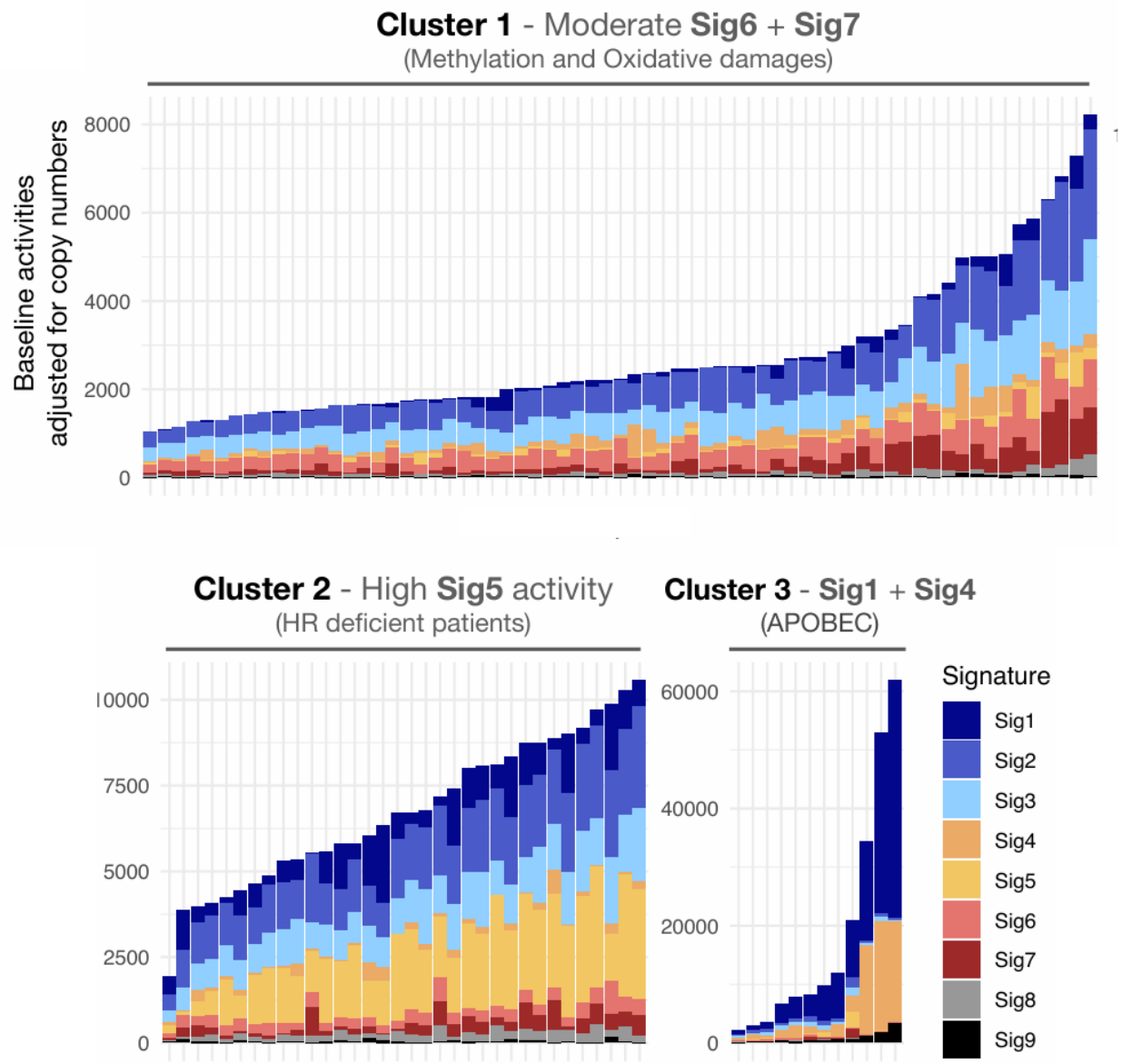
- We analyze whole-genome sequencing data from 113 women with breast cancer from the `Breast-AdenoCA` ICGC cohort.

- The data consist of 707,104 total mutations altogether, for which we have the genomic location (e.g., `chrX:77364730-77364827`) and the mutation type (e.g., A[C>T]C).



Clusters of patients

- Clustering the normalized baseline exposures yields interpretable groups of patients.

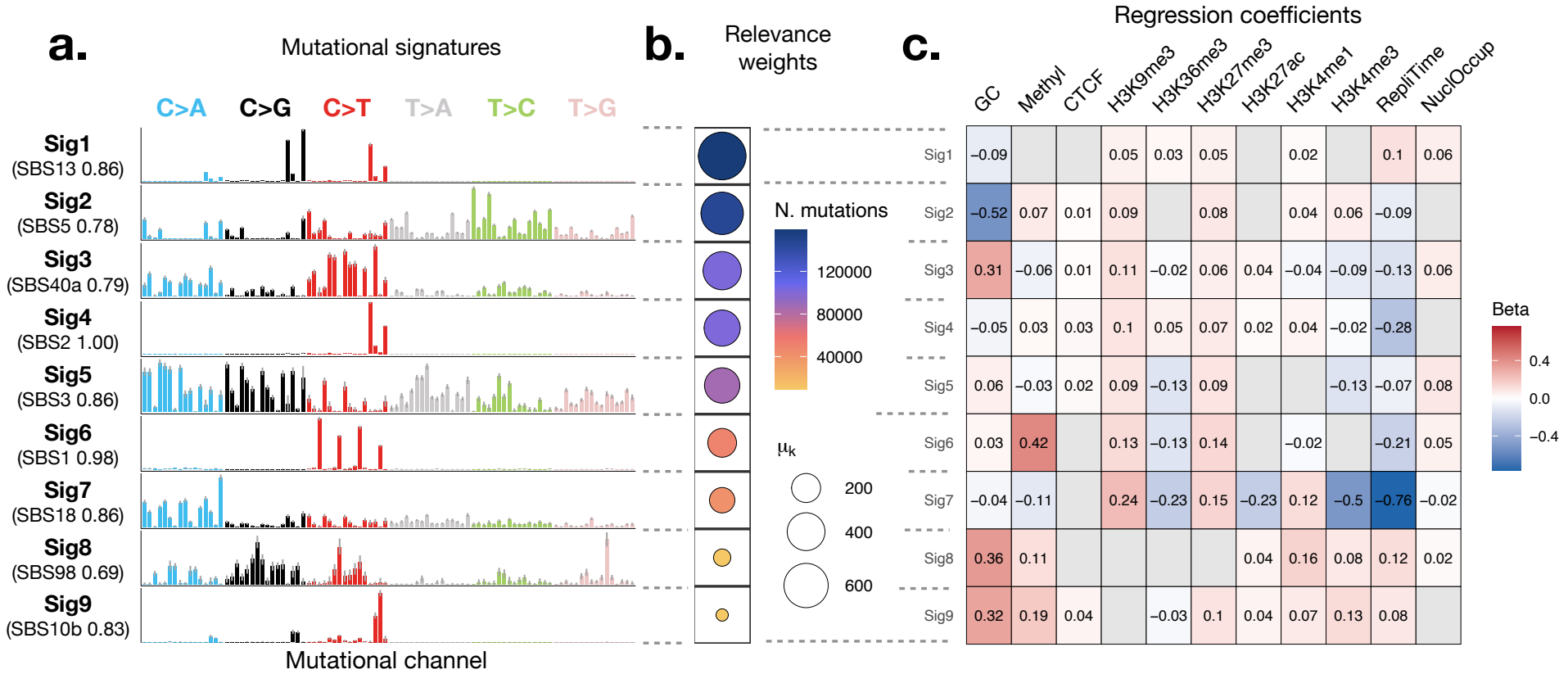


Genomic covariates

- For these data, we have a range of genomic covariates that have previously reported effects on mutation rate.

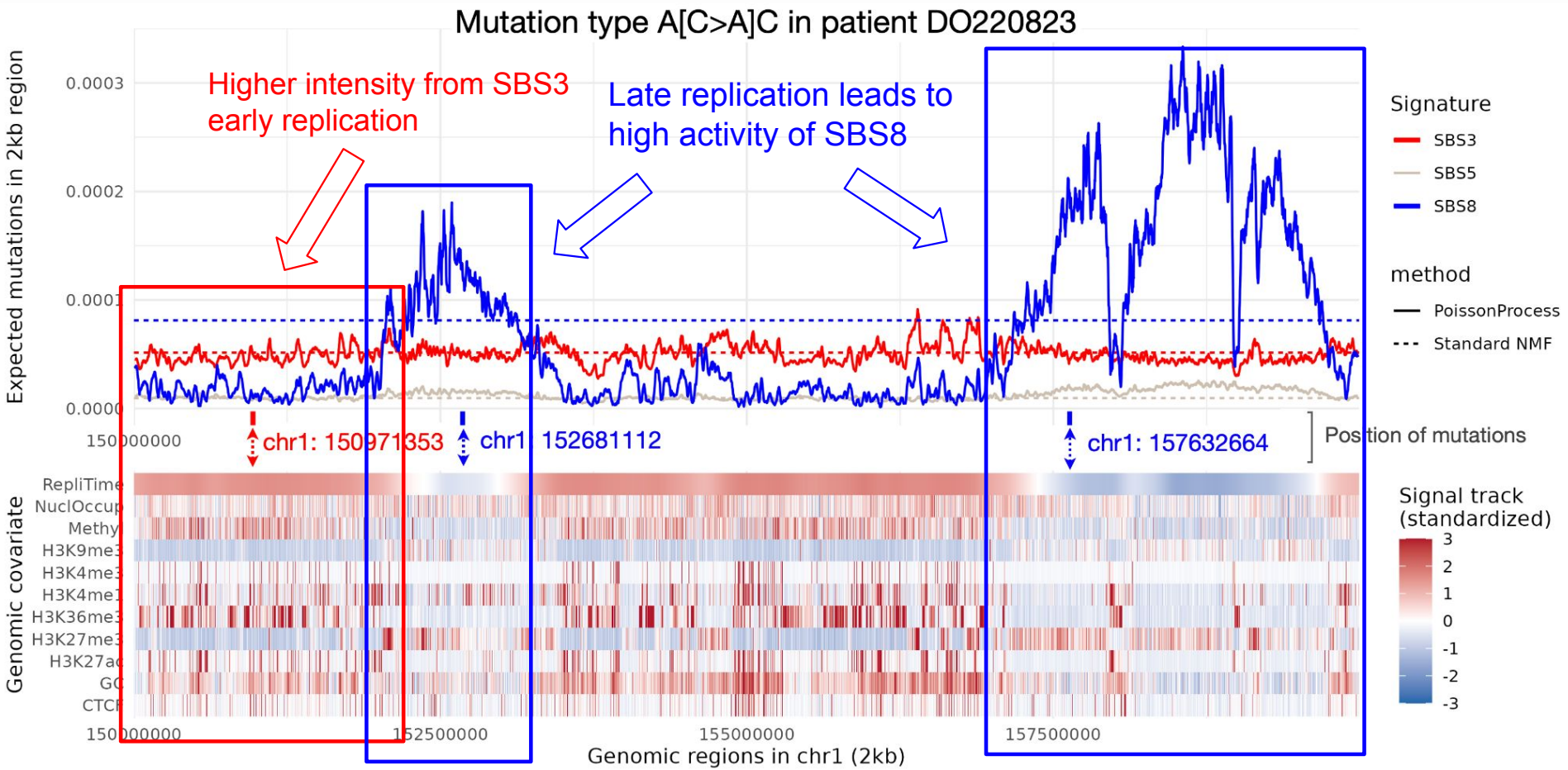
Covariate	Assay	Description and role	Mutations	Reference
Nucleosome occupancy	MNase-seq	Degree to which DNA is wrapped around nucleosomes; higher values indicate packed chromatin.	Periodic patterns	Pich et al. (2018)
H3K27ac	Histone ChIP-seq	Marker of active enhancers and promoters; associated with gene activation.	↗	Schuster-Böckler and Lehner (2012)
H3K4me1	Histone ChIP-seq	Marks poised and active enhancers; enriched at regulatory elements.	↘	Hodgkinson et al. (2012)
H3K4me3	Histone ChIP-seq	Marks active promoters; associated with transcription initiation.	↘	Hodgkinson et al. (2012)
H3K9me3	Histone ChIP-seq	Marker of constitutive heterochromatin; gene silencing/repression.	↗	Schuster-Böckler and Lehner (2012)
H3K27me3	Histone ChIP-seq	Repressive mark, associated with Polycomb-mediated gene silencing.	↗	Schuster-Böckler and Lehner (2012)
H3K36me3	Histone ChIP-seq	Marker of transcriptional elongation within gene bodies.	↘	Li et al. (2013)
CTCF	TF ChIP-seq	Insulator protein, key architectural TF regulating chromatin looping and gene expression.	↗	Katainen et al. (2015)
Replication timing	Repli-seq	Timing of DNA replication during S-phase of a cell; reflects chromatin state and genome organization.	↗ (later) ↘ (early)	Supek and Lehner (2015)
Methylation	WGBS	Level of DNA methylation; regulates gene expression and silencing.	↗ at CpGs	Bird (1980)
GC content	—	Proportion of G and C nucleotides a region; influences DNA stability and nucleosome positioning.	↘	Makova and Hardison (2015)

Estimated signatures and coefficients



- Sig5 (SBS3) HRD/BRCA: Covariates have varying effects on mutation rate.
- Sig6 (SBS1) Aging, CpG sites: **Accurate recovery** aided by methylation covariate.
- Replication timing: Late timing ⇒ high rate for Sigs 4, 6, 7 (APOBEC, CpG, ROS). Early timing ⇒ increased rate for Sig9, opposite of previous finding.
- H3K9me3, H3K27me3 (Heterochromatin, gene suppression): Increased rates, especially Sig7 (SBS13) ROS. H3K36me3: Decreased rates, especially Sigs 5, 6, 7.

Supervised analysis with known signatures: Illustration of varying exposures



Summary

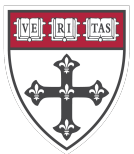
- Poisson process factorization (PPF) provides a framework for **joint modeling** of mutational signatures, position-specific exposures, and genomic covariates.
- We compute the MAP estimate using a computationally tractable **majorization-minimization algorithm** and use **Gibbs sampling** for posterior inference.
- The model provides insight into patient-specific information for precision treatment as well as general patterns of cancer biology.
- We envision that the PPF model will be **useful in many other applications** beyond cancer genomics.

Poisson process factorization for mutational signature analysis with genomic covariates

Jeffrey W. Miller

Department of Biostatistics, Harvard University

Nonparametric Bayesian Inference - Computational Issues
Jan 12, 2026 || ICERM || Brown University



HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Dana-Farber
Cancer Institute

Giovanni Parmigiani

Alessandro Zito

References I

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, et al. (2013). Signatures of mutational processes in human cancer. *Nature 500*(7463), 415–421.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I* (Second ed.). Probability and its Applications (New York). New York: Springer-Verlag.

Otlu, B., M. Díaz-Gay, I. Vermes, E. N. Bergstrom, M. Zhivagui, M. Barnes, and L. B. Alexandrov (2023, 2025/06/21). Topography of mutational signatures in human cancer. *Cell Reports 42*(8).

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 689–710.

Zito, A. and J. W. Miller (2024). Compressive Bayesian non-negative matrix factorization for mutational signatures analysis.

Maximum a posteriori (MAP) estimation

- **Signatures**: For k, i, update each probability r_{ik} in signature k as

$$r_{ik} \leftarrow \frac{r'_{ik}}{\sum_i r'_{ik}} \qquad r'_{ik} \leftarrow r_{ik}\left(\frac{\alpha_{ik}-1}{r_{ik}} + \sum_j \sum_{n=1}^{N_{ij}} \frac{\theta_{kj} e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)}}{\sum_{s=1}^K r_{is}\theta_{sj} e^{\boldsymbol{\beta}_s^\top \mathbf{x}(t_n)}}\right).$$

- **Baseline exposures**: For j, k, update each θ_{kj} as

$$\theta_{kj} \leftarrow \theta_{kj}\left(\frac{a-1}{\theta_{kj}} + \frac{\mu_k}{\int_0^T \frac{1}{2} c_j(t)(a+\mu_k e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)})\mathrm{d}t} \sum_i \sum_{n=1}^{N_{ij}} \frac{r_{ik} e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)}}{\sum_{s=1}^K r_{is}\theta_{sj} e^{\boldsymbol{\beta}_s^\top \mathbf{x}(t_n)}}\right).$$

- **Regression coefficients**: For $k = 1, \ldots, K$, update $\boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_k - \mathbf{H}_k^{-1}\mathbf{g}_k$, where

$$\mathbf{H}_k = -\sum_j \theta_{kj} \int_0^T \frac{1}{2} c_j(t) e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)} \mathbf{x}(t)\mathbf{x}(t)^\top \,\mathrm{d}t - \frac{1}{\sigma_k^2} I_p,$$

$$\mathbf{g}_k = -\sum_j \theta_{kj} \int_0^T \frac{1}{2} c_j(t) e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)} \mathbf{x}(t)\,\mathrm{d}t + \sum_{i,j} \sum_{n=1}^{N_{ij}} \frac{r_{ik}\theta_{kj} e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)}}{\sum_{s=1}^K r_{is}\theta_{sj} e^{\boldsymbol{\beta}_s^\top \mathbf{x}(t_n)}} \mathbf{x}(t_n) - \frac{1}{\sigma_k^2}\boldsymbol{\beta}_k.$$

- **Relevance weights and variances**: For each k, update

$$\mu_k \leftarrow \frac{a\sum_j \theta_{kj} \int_0^T \frac{1}{2} c_j(t)\,\mathrm{d}t + \varepsilon a J}{2aJ+1}, \qquad \sigma_k^2 \leftarrow \frac{\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k/2 + d_0}{p/2 + c_0 + 1}.$$

Gibbs sampler for posterior inference

- **Latent attributions**: For each i, j, and $n = 1, \ldots, N_{ij}$, sample

$$(W_{ij}(t_n) \mid r, \theta, \beta) \sim \text{Mult}\Big(1; p_{ij1}(t_n), \ldots, p_{ijK}(t_n)\Big), \qquad p_{ijk}(t_n) = \frac{r_{ik}\theta_{kj}e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)}}{\sum_{s=1}^K r_{is}\theta_{sj}e^{\boldsymbol{\beta}_s^\top \mathbf{x}(t_n)}}.$$

- **Signatures**: Let $M_{ik} = \sum_j \sum_{n=1}^{N_{ij}} W_{ijk}(t_n)$ be the number of mutations of type i assigned to signature k, and sample

$$(r_k \mid W, \beta, \theta) \sim \text{Dir}\Big(\alpha_{1k} + M_{1k}, \ldots, \alpha_{Ik} + M_{Ik}\Big).$$

- **Baseline exposures**: Let $S_{kj} = \sum_i \sum_{n=1}^{N_{ij}} W_{ijk}(t_n)$ be the number of mutations in patient j assigned to signature k, and sample

$$(\theta_{kj} \mid W, \beta, r) \sim \text{Ga}\left(a + S_{kj},\ \int_0^T \tfrac{1}{2} c_j(t)\Big(a/\mu_k + e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)}\Big)\mathrm{d}t\right).$$

- **Regression coefficients**: Perform elliptical slice sampling on

$$\pi(\boldsymbol{\beta}_k|\theta, W, \sigma^2) \propto \exp\left(-\sum_j \theta_{kj} \int_0^T \tfrac{1}{2}\, c_j(t)\, e^{\boldsymbol{\beta}_k^\top \mathbf{x}(t)}\mathrm{d}t + \sum_{ijn} W_{ijk}(t_n)\boldsymbol{\beta}_k^\top \mathbf{x}(t_n)\right)\mathrm{N}(\boldsymbol{\beta}_k; 0, \sigma_k^2 I_p).$$

- **Relevance weights and variances**: Sample

$$(\mu_k|\theta) \sim \text{InvGa}\left(a_0 + aJ,\ b_0 + a\sum_j \theta_{kj} \int_0^T \tfrac{1}{2} c_j(t)\mathrm{d}t\right), \quad (\sigma_k^2|\beta) \sim \text{InvGa}\Big(c_0 + \frac{p}{2},\ d_0 + \frac{\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k}{2}\Big).$$