Signature Analyzer: scaling Bayesian NMF to millions of individuals with GPUs

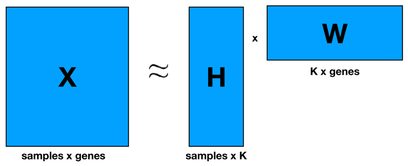
Shankara Anand[1], François Aguet*[1], Amaro Taylor-Weiner*[1,2], Justin Cha[1], Nicholas J. Haradhvala[1], Sager Gosai[1,2], Jaegil Kim[1], Kristin Ardlie[1], Eliezer M. Van Allen[1,3,4], and Gad Getz[#1,4,5]

BROAD INSTITUTE

[1] Broad Institute of Harvard and MIT, Cambridge, MA, USA; [2] Harvard University, Cambridge, MA, USA; [3] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA USA; [4] Harvard Medical School, Boston, MA; [5] MGH Cancer Center and Dept. of Pathology, Boston, MA

Introduction

- Current methodologies for analyzing genomic data, **notably non-negative matrix factorization,** were designed for tens to thousands of samples.
- Dataset sizes are reaching millions of samples or individual cells.
- CPU implementations use thousands of cores, costing money and time.
- GPUs reduce **run time** and **lower costs** by offering more compute power.

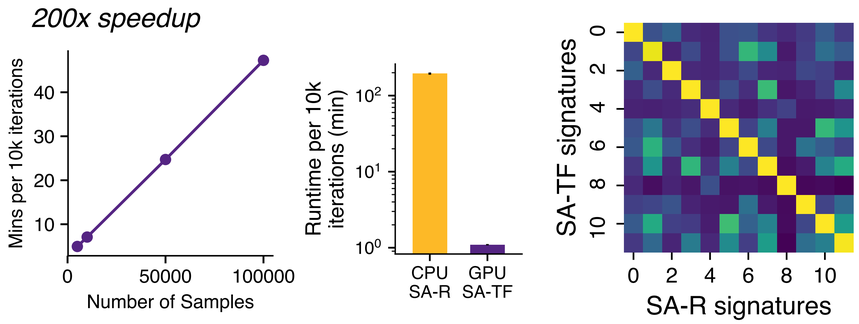


Non-negative Matrix Factorization

*We aim to implement a Bayesian variant of non-negative matrix factorization at **scale**.*

Transitioning to GPUs

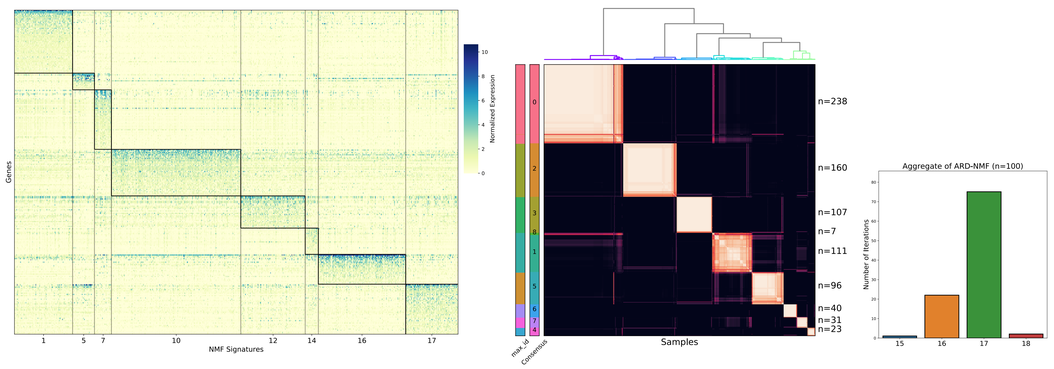
- Open source libraries, such as PyTorch and TensorFlow, make method optimization easy.
- Native execution on GPU and CPU enables cheap development and simple transition to GPU without specialized environments. (*see Links)*



(**Left**) *Running **S**ignature **A**nalyzer scales linearly as a function of number of samples.* (**Center**) *Average run time to compute 10,000 iterations of ARD NMF using **S**ignature **A**nalyzer in R (yellow) and **S**ignature **A**nalyzer-**GPU** (purple).* (**Right**) *Correlation heat map of signatures derived from both methods using the same input mutation counts matrix.*

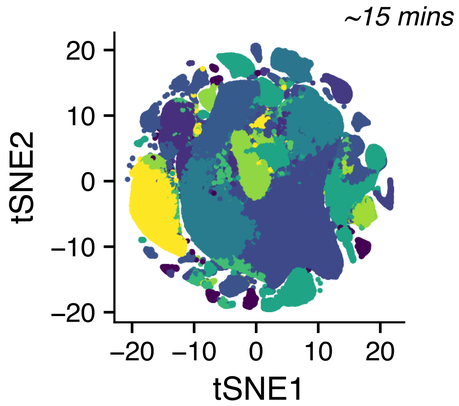
Clustering Large RNA-Seq Cohorts

~ 1000 Bulk RNA-Seq



(**Left**) *Best ARD-NMF solution for Bulk RNA-Seq decomposition of tumor dataset. We identify 8 de novo transcriptional programs.* (**Center**) *Consensus solution of 100 runs. At scale, we can generate consensus statistics for large transcriptional datasets.* (**Right**) *Distribution of latent dimensions over 100 runs (~ 40 mins).*

~ 1 mil single-cell RNA-Seq



(**Above**) *t-Distributed Stochastic Neighbor Embedding (t-SNE) of 1e6 embryonic mouse brain cells. Colors indicate transcriptional programs identified using SA-GPU.*

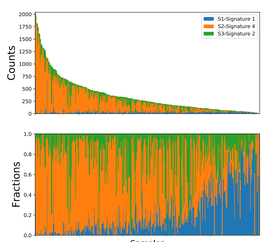
Mapping Mutational Signatures

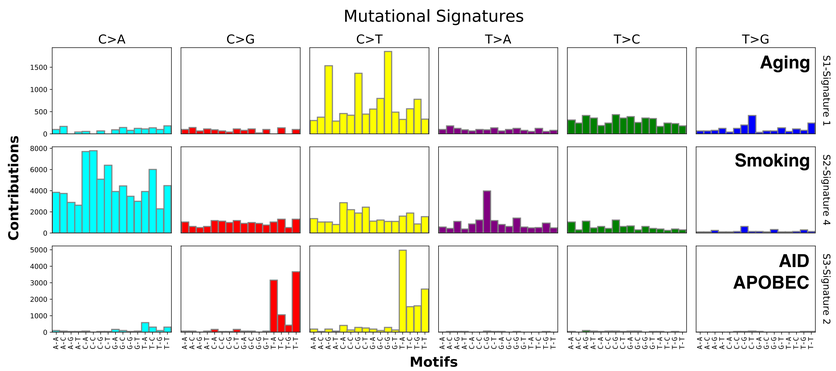
We support mapping & plotting API for:

- COSMIC2 Single-Base Substitutions
- COSMIC3 Single-Base Substitutions (WGS/WES)
- COSMIC3 Doublet-Base Substitutions
- COSMIC3 Indels

LUAD MC3 I ARD-NMF decomposition of mapped, 96-base context vector encoding for single-base substitutions. (**Right**) Patient-level signature contributions extracted. (**Far-Right**) Bar-plot of 96 single-base substitution contributions labeled by canonical phenotypes.

~ 500 patients





Links

Paper (*Genome Biology*)

Taylor-Weiner, A., Aguet, F., Haradhvala, N.J. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20,** 228 (2019) doi:10.1186/s13059-019-1836-7

Signature Analyzer (Github)

A GPU-enabled implementation of a Automatic Relevance Determination (ARD) non-negative matrix factorization; mapping for mutational signatures and use with expression matrices

Discussion

- Simply re-implementing current methods using open source GPU compatible libraries results in orders of magnitude increase in speed. (*see Links*)
- Scaling methods in this way made possible hypotheses involving more complex models, larger datasets, with more accurate empirical measurements.
- We demonstrate extraction of **transcriptional programs** across millions of single-cells and fast decomposition & mapping of **somatic mutational signatures**
- Aggregating across multiple runs of stochastic methods allows for robust statistics
- We anticipate a transition to GPU-based computing for computational genomics methods.

This work was supported by the Common Fund of the Office of the Director of the National Institutes of Health contract HHSN268201000029C awarded to The Broad Institute and by grant 2018-182729 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.