Quiver Laplacians, Feature Selection, & Chromatin Accessibility

Otto Sumray
AlToGeLis 24, Dresden

based on arxiv:2404.06993

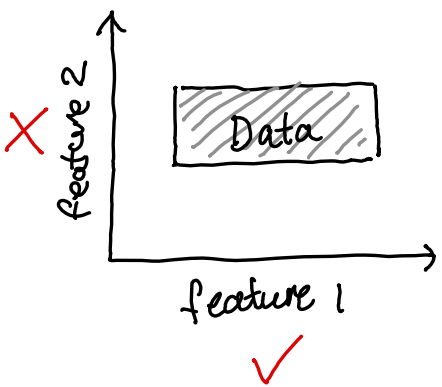
Feature selection

→ reduce dimensionality

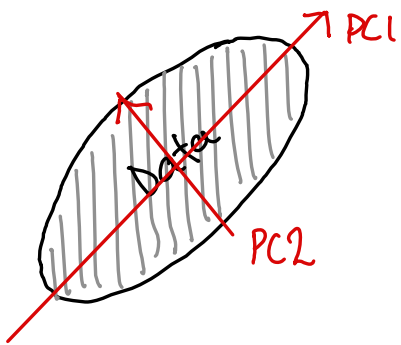
→ find features that best explain the data

e.g.

Most varying features



Principal component Analysis



Chromatin Accessibility

If genes are like recipes...

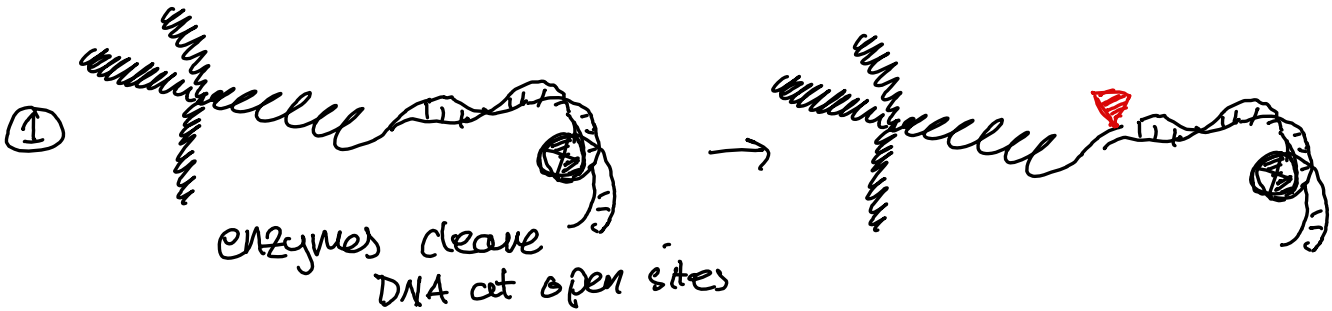
CAGCATG... = | 1 cup Glu
1 cup His
⋮ |

... Chromatin is like the recipe book

The chromatin is "open" or "accessible"
in certain regions to allow protein complexes to regulate gene expression

- enhancers
- promoters
- e.g. transcription factors

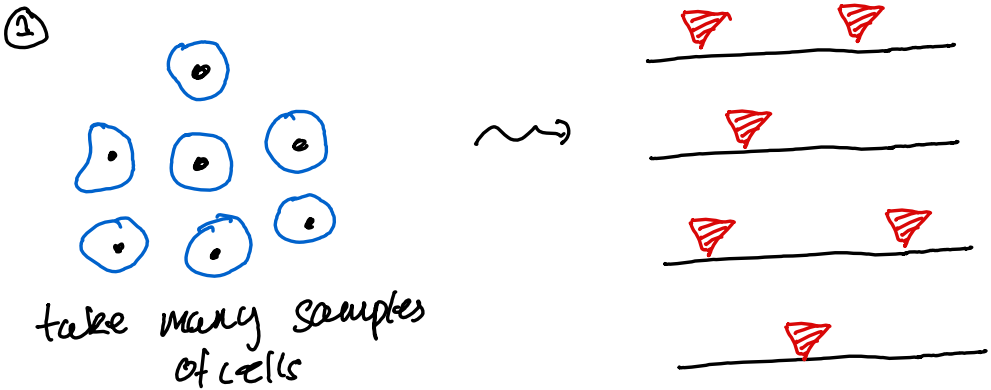
Chromatin accessibility can be measured (ATAC-seq)



endpoints of fragments give accessible sites

problem is... there are 3 billion genomic positions!
and noisy, sparse etc...

Feature selection: Peak calling



Now 100,000s of features

Often we're interested in subsets of the data

↓ e.g. cell types

e.g. in blood:

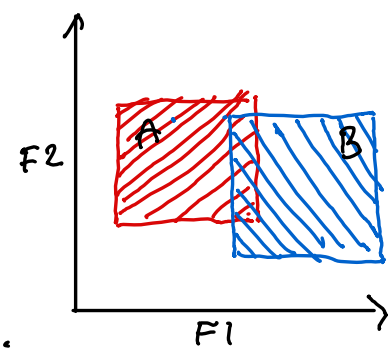
- T cells
- B cells
- NK cells
- Monocyte cells.
- ⋮

X finite set of points

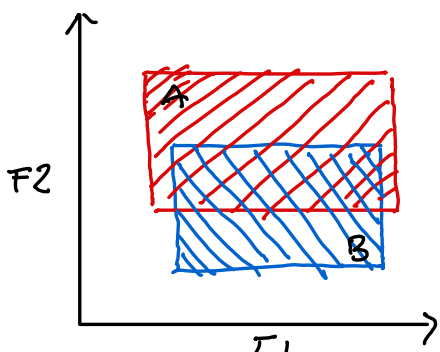
$\mathcal{U} = \{U_i : i \in I\}$

cover of X : $\bigcup_{i \in I} U_i = X$

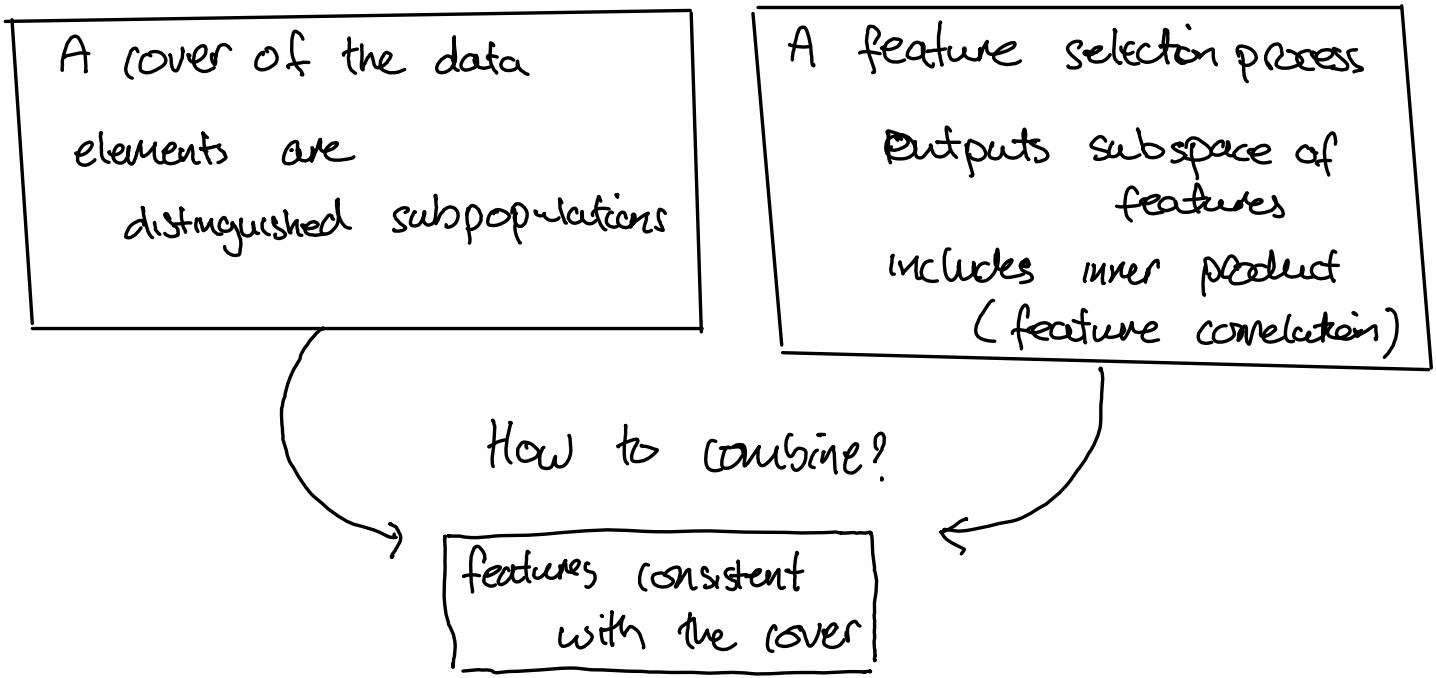
But features relevant for X may not be relevant for each U_i (& vice versa)



F1 higher variance for $A \cup B$ but not A or B

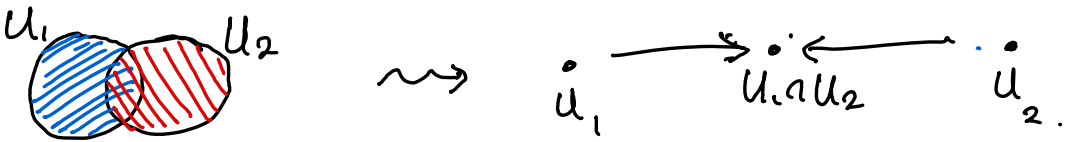


F1 higher variance for A, B but not $A \cup B$

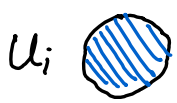

A cover of the data
elements are
distinguished subpopulations
A feature selection process
outputs subspace of
features
includes inner product
(feature correlation)
How to combine?
features consistent
with the cover

A cover defines a directed graph

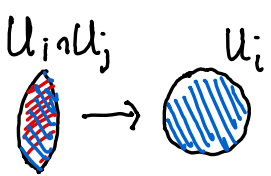
vertices	$\sigma \subset \mathcal{U} = \{U_i : i \in I\}$	AND	$\bigcap \sigma \neq \emptyset$	
edges	$\tau \to \sigma$	iff	$\tau \subset \sigma$	(hence $\bigcap \tau \supseteq \bigcap \sigma$)

E.g. 

Feature selection attaches a f.d. vector space to each node



⇝ Space of features relevant for U_i



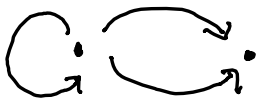
⇝ Orthogonal projection
features$(U_i \cap U_j)$ ⟶ features(U_i)

Q a finite quiver

finite sets

V vertices

E edges



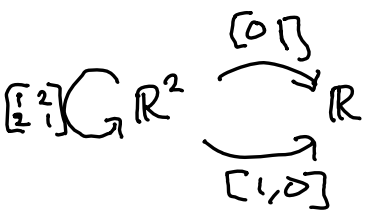
Source & target maps

$s, t : E \longrightarrow V$

A. Fdtlilb - representation of Q.

$A_v \in$ Fdtlilb $\quad \forall\, v \in V$

$A_e : A_{s(e)} \longrightarrow A_{t(v)} \quad \forall\, e \in E$



space of global sections of $(\mathcal{G}, \mathbb{A}_\bullet)$

$$\Gamma(\mathcal{G}, \mathbb{A}_\bullet) = \left\{ (x_v)_{v \in V} \in \prod_{v \in V} \mathbb{A}_v \;:\; \forall e \in E \quad \mathbb{A}_e(x_{s(e)}) = x_{t(e)} \right\}$$

e.g. for $\mathbb{R} \xrightarrow{1} \mathbb{R} \xleftarrow{-1} \mathbb{R}$

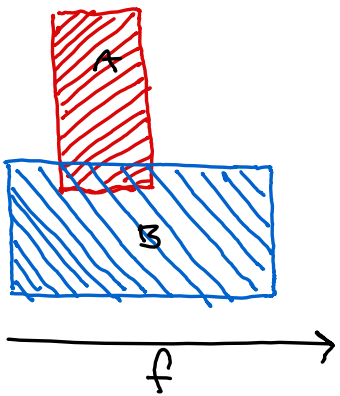
$(1, 1, 1)$ not a section

$(1, 1, -1)$ is a section

For a cover and feature selector

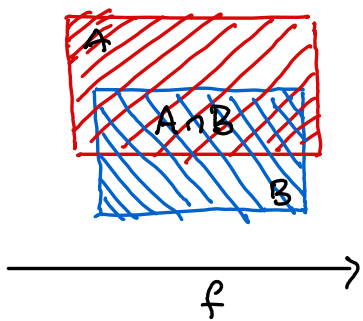
(globally) compatible features

are global sections of the corresponding quiver rep.



Not a globally compatible feature

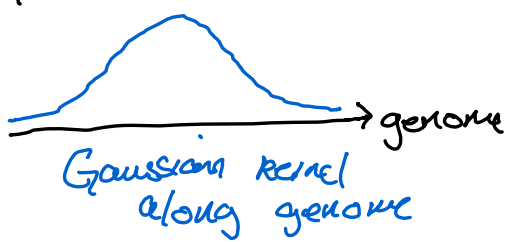
relevant for B and A∩B but not for A!



Is a globally compatible feature

Recap so far

Data	chromatin accessibility profiles for thousands of single cells
Cover	different cell types
Feature selection	significantly accessible sites
Quiver representation	Quiver from $\subseteq$ relations of cover (nerve of.) valued with space of selected features inner product
Global sections	globally compatible features



How to compute sections?

One way: "boundary map"

$(\mathcal{Q}, \mathcal{A}_\bullet)$

$$B_{\mathcal{A}} : \prod_{v \in V} \mathcal{A}_v \longrightarrow \prod_{e \in E} \mathcal{A}_e$$

block form

$$(B_{\mathcal{A}})_{e,v} = \begin{cases} \mathcal{A}_e - \mathrm{Id}_v & \text{if } s(e) = t(e) = v \\ \mathcal{A}_e & \text{or if } s(e) = v \\ -\mathrm{Id}_v & \text{or if } t(e) = v \\ 0 & \text{o/w.} \end{cases}$$

$$\boxed{\ker B_{\mathcal{A}} = \Gamma(\mathcal{Q}; \mathcal{A}_\bullet)}$$

each (block row)

$$[0\,0\,0 \cdots 0\ \mathcal{A}_e\ 0\,0\,0 \cdots 0\ -\mathrm{Id}\ 0\,0\,0]$$

(other ways to compute, see Seigel, Harrington, Nanda 2021)

Data is noisy! want approximate sections

$$x \in \prod_{v \in V} \mathbb{A}_v, \quad \|x\| = 1.$$

$$\| B_{\mathbb{A}} x \|_2^2 \leq \varepsilon \iff x^* B_{\mathbb{A}}^* B_{\mathbb{A}} x \leq \varepsilon$$

Define the <u>Laplacian</u> of $(\mathcal{Q}, \mathbb{A}_\bullet)$

as $L_{\mathbb{A}} := B_{\mathbb{A}}^* B_{\mathbb{A}}$.

$$\boxed{\ker L_{\mathbb{A}} = \Gamma(\mathcal{Q}, \mathbb{A}_\bullet)}$$

$L_{\mathbb{A}}$ is PSD: $0 \leq \lambda_1(L_{\mathbb{A}}) \leq \cdots \leq \lambda_n(L_{\mathbb{A}})$

$n = \sum_{v \in V} \dim \mathbb{A}_v$

if $L_{\mathbb{A}} x = \lambda_i x \Rightarrow x^* L_{\mathbb{A}} x = \lambda_i$

$\Rightarrow$ eigenvectors for small eigenvalues are approx. sections

Stability fix Q. $A_\bullet$, $\hat{A}_\bullet$ Q-reps

$$\tau = (\tau_v : A_v \longrightarrow \hat{A}_v)_{v \in V}$$

Thm OS, Harrington, Nanda 24 $1 \le k \le n - \mathrm{Null}(\tau)$

$$\lambda_k(\hat{A}_\bullet) \le \kappa(\tau)^2 \lambda_{k+\mathrm{Null}(\tau)}(A_\bullet) + \partial(\tau)\|\tau^+\| \left[2\kappa(\tau) \lambda_{k+\mathrm{Null}(\tau)}(A)^{1/2} + \partial(\tau)\|\tau^+\| \right]$$

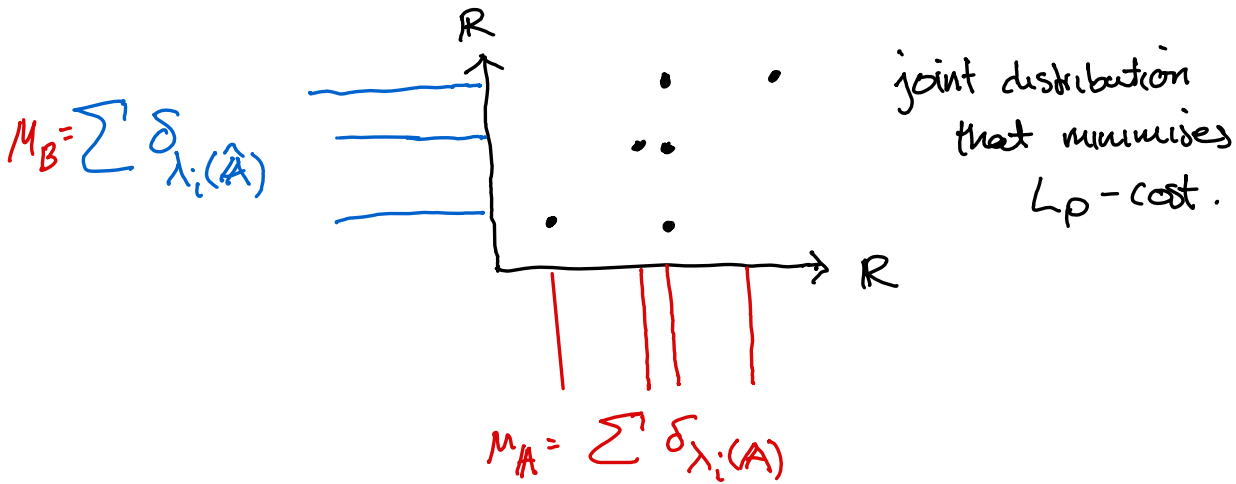
→ generalized condition number $\kappa(\tau) = \|\tau\| \cdot \|\tau^+\|$

→ defect. $\partial(\tau) = \left[\sum_{e \in E} \|A_e \tau_{s(e)} - \tau_{t(e)} A_e\|^2 \right]^{1/2}$.

"how close to a morphism" of quiver reps

To compare $\lambda_\bullet(A_\bullet)$ and $\lambda_\bullet(\hat{A}_\bullet)$

need Wasserstein metric W_p



joint distribution that minimises L_p-cost.

Corr. $\mathcal{Q}$ quiver from cover

$A_\bullet, \hat{A}_\bullet$ reps from two different feature selectors

$$W_1(\mu_{A}, \mu_{\hat{A}}) \leq \frac{\#\text{edges } \mathcal{Q}}{m} \left[f(\varepsilon) \cdot c + g(\varepsilon) \cdot (m - c) \right]$$

for f, g trigonometric polynomials

$$\varepsilon = \max_v \operatorname{dist}_{\text{Grass}}(A_v, \hat{A}_v)$$

$$m = \max \text{ total rank of } A, \hat{A}$$

$$c = \sum_v |\dim A_v - \dim \hat{A}_v|$$

Eigenvalue bounds for changing the quiver Q

Thms OS, HATH, UN 24.

Does not preserve $\Gamma(Q, A_\bullet)$

① Removing edges

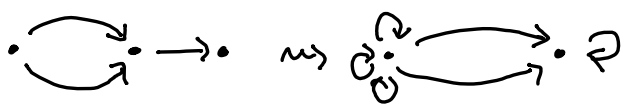
② Removing vertices (and assoc. edges)

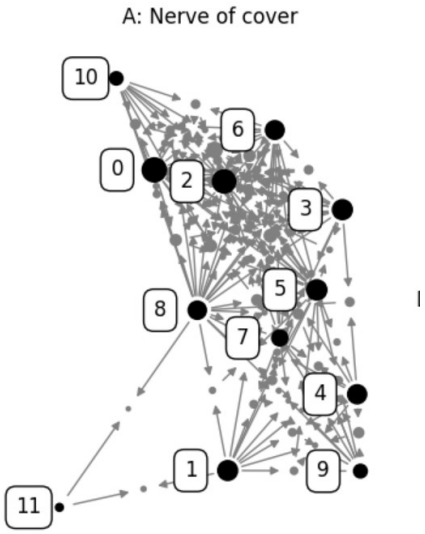
Does preserve $\Gamma(A_\bullet)$

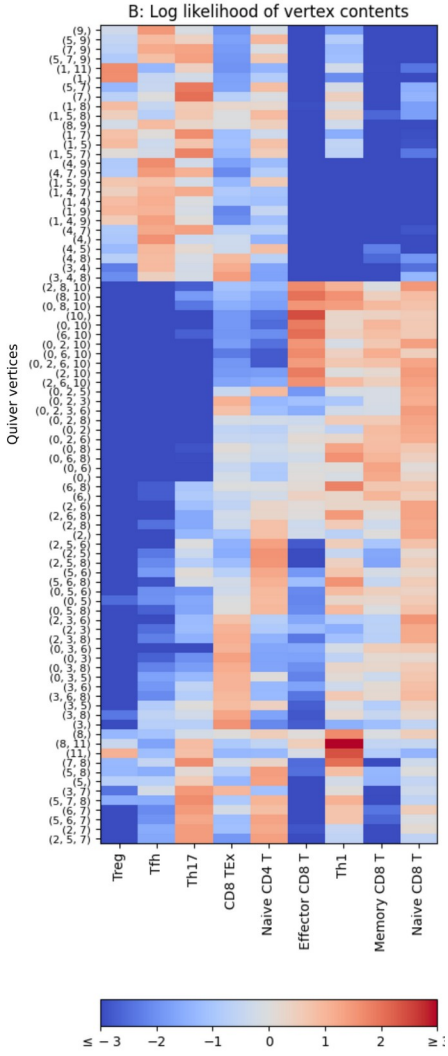
[for certain $A_\bullet$]

③ Edge "homotopy"

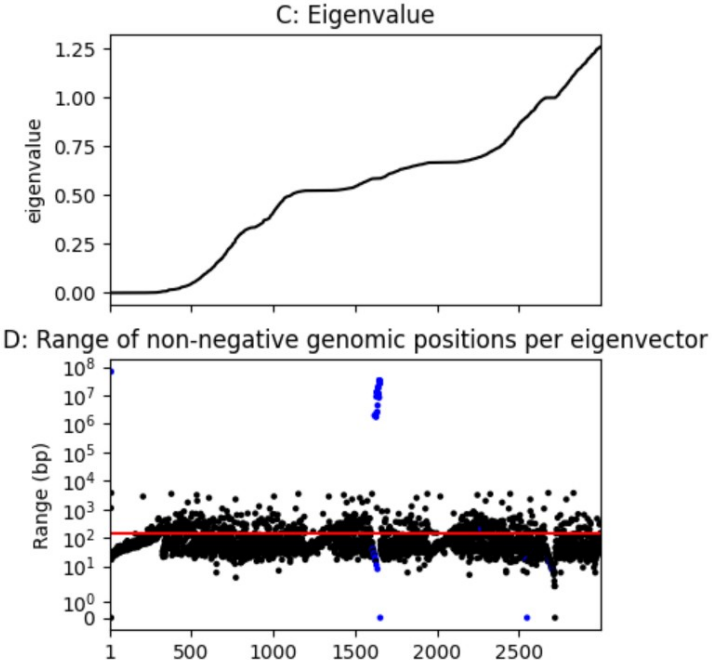
④ Reducing vertices (Kron reduction / Schur complement)

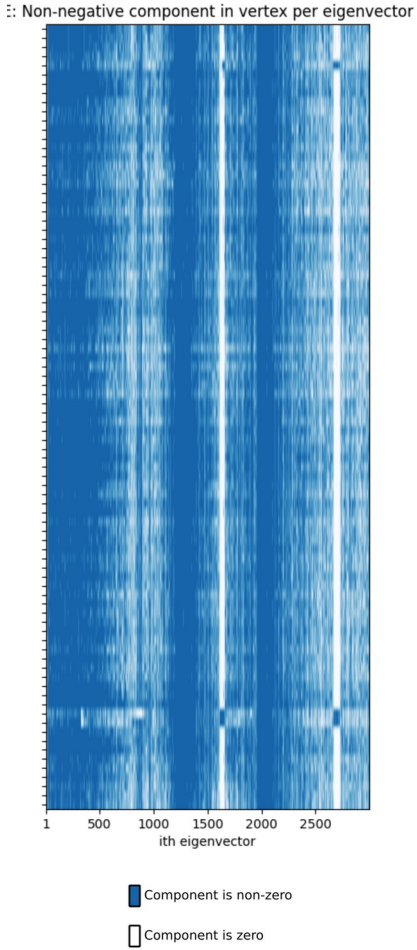


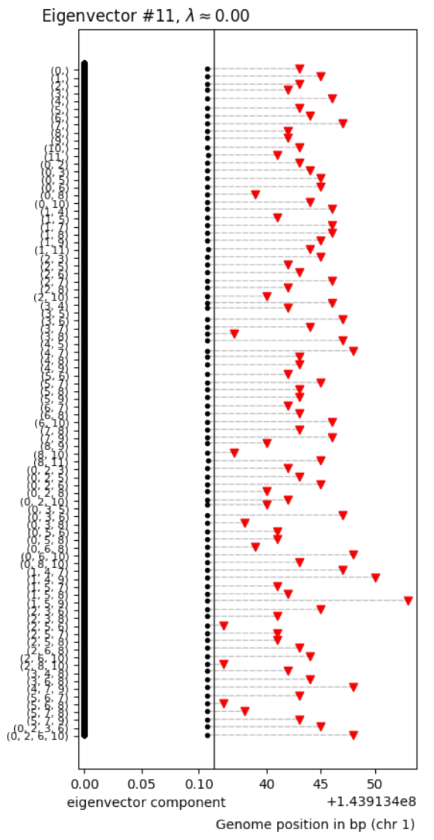


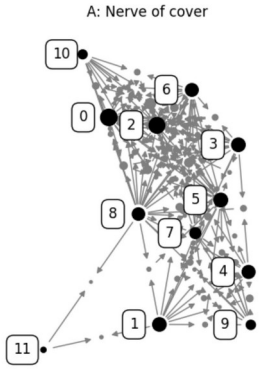


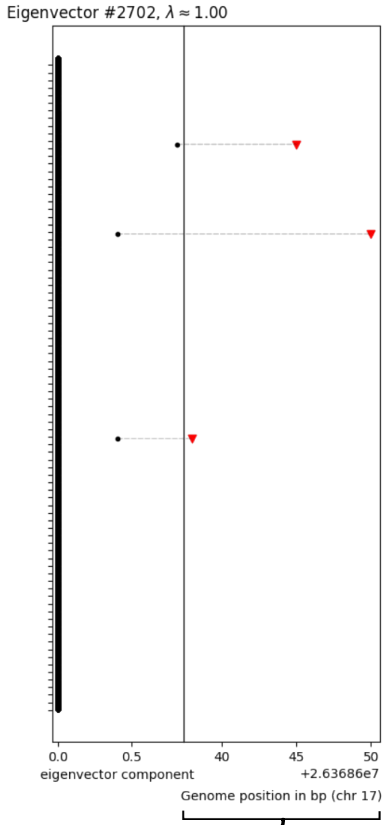
✓ 30,000
T cells infiltrating BCC tumour

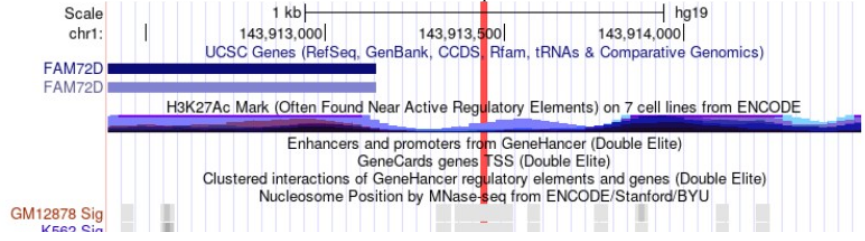

C: Eigenvalue
eigenvalue
1.25
1.00
0.75
0.50
0.25
0.00
D: Range of non-negative genomic positions per eigenvector
Range (bp)
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
1
500
1000
1500
2000
2500

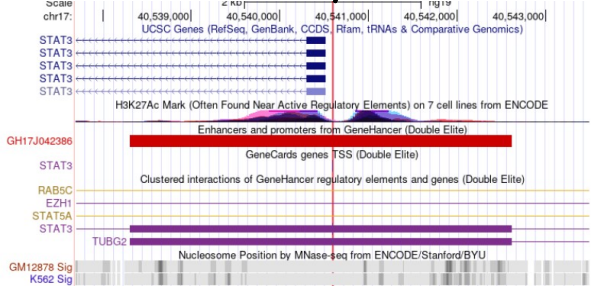

: Non-negative component in vertex per eigenvector
1
500
1000
1500
2000
2500
ith eigenvector
Component is non-zero
Component is zero


Eigenvector #11, λ ≈ 0.00
0.00
0.05
0.10
40
45
50
eigenvector component
+1.439134e8
Genome position in bp (chr 1)


A: Nerve of cover
10
6
0
2
3
8
5
7
4
1
9
11


Eigenvector #2702, λ ≈ 1.00
0.0
0.5
40
45
50
eigenvector component
+2.63686e7
Genome position in bp (chr 17)


Scale
1 kb
hg19
chr1:
143,913,000
143,913,500
143,914,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
FAM72D
FAM72D
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Enhancers and promoters from GeneHancer (Double Elite)
GeneCards genes TSS (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
GM12878 Sig
K562 Sig


chr17:
40,539,000
40,540,000
40,541,000
40,542,000
40,543,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
STAT3
STAT3
STAT3
STAT3
STAT3
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
GH17J042386
Enhancers and promoters from GeneHancer (Double Elite)
GeneCards genes TSS (Double Elite)
STAT3
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
RAB5C
EZH1
STAT5A
STAT3
TUBG2
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
GM12878 Sig
K562 Sig