Statistical Discovery of Transcriptomic Cancer Signatures using Multimodal Local Search

Emile Zakiev, Johann Dreo, Mara Santarelli, Benno Schwikowski

▶ **To cite this version:**

Emile Zakiev, Johann Dreo, Mara Santarelli, Benno Schwikowski. Statistical Discovery of Transcriptomic Cancer Signatures using Multimodal Local Search. PGMO Days, Programme Gaspard Monge pour l'Optimisation, la recherche opérationnelle et leurs interactions avec les sciences des données., Nov 2022, Palaiseau, France. hal-04110700

HAL Id: hal-04110700
https://hal.science/hal-04110700

Submitted on 30 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT PASTEUR

Statistical Discovery of Transcriptomic Cancer Signatures using Multimodal Local Search

• E. Zakiev/ J. Dreo • 29/11/2022

The problem

What causes cancer to be resistant?

While killing most of the cells, monodrug chemotherapy doesn't kill all, allowing remaining cells proliferate

The problem

What causes cancer to be resistant?

While killing most of the cells, monodrug chemotherapy doesn't kill all, allowing remaining cells proliferate

Why not mixing multiple drugs to be sure?

The problem

What causes cancer to be resistant?

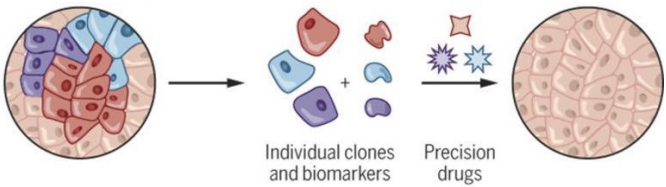
While killing most of the cells, monodrug chemotherapy doesn't kill all, allowing remaining cells proliferate

Why not mixing multiple drugs to be sure? Because of excessive toxicity

The problem

What causes cancer to be resistant?

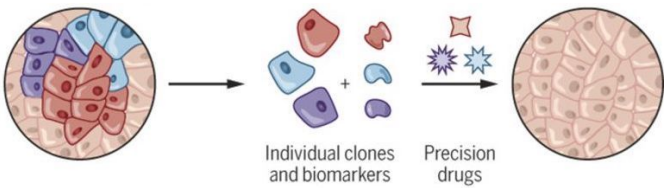
Need for precision drugs fine-tailored to each subpopulation of cells in each patient



The problem

What causes cancer to be resistant?

Need for precision drugs fine-tailored to each subpopulation of cells in each patient



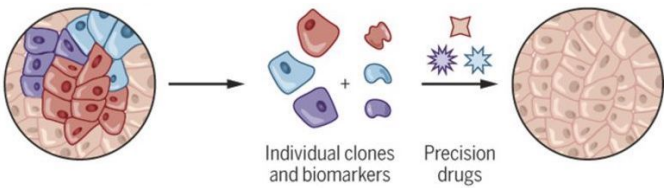
- the subpopulations of cells defined by their scRNAseq profile

INSTITUT PASTEUR

The problem

What causes cancer to be resistant?

Need for precision drugs fine-tailored to each subpopulation of cells in each patient

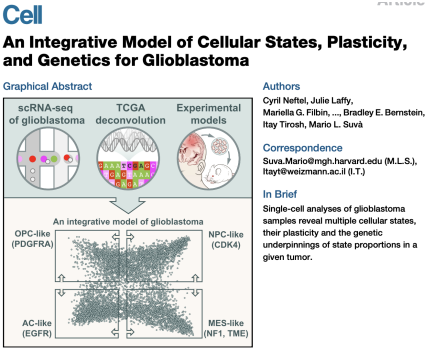


- the subpopulations of cells defined by their scRNAseq profile
  - by their characteristic *signatures* (ensembles of genes)

INSTITUT PASTEUR

Inspired by...

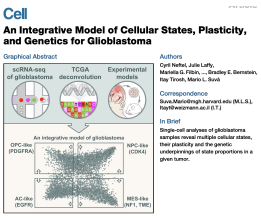
...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**



INSTITUT PASTEUR

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**


Cell

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

Graphical Abstract

scRNA-seq of glioblastoma | TCGA deconvolution | Experimental models

An integrative model of glioblastoma

OPC-like (PDGFRA) | NPC-like (CDK4) | AC-like (EGFR) | MES-like (NF1, TMD)

Authors

Cyril Neftel, Julie Laffy, Mariella G. Filbin, ..., Bradley E. Bernstein, Itay Tirosh, Mario L. Suvà

Correspondence

Suva.Mario@mgh.harvard.edu (M.L.S.), Itay@weizmann.ac.il (I.T.)

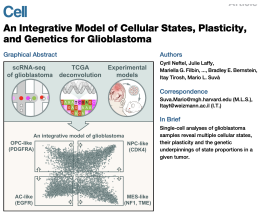
In Brief

Single-cell analyses of glioblastoma samples reveal multiple cellular states, their plasticity and the genetic underpinnings of state proportions in a given tumor.

Were able to show benefit of certain gene inhibitors in glioblastoma xenografts on mice

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**

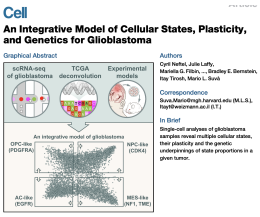


Were able to show benefit of certain gene inhibitors in glioblastoma xenografts on mice

- Hierarchical clustering of cells to form *partitions* of each sample into up-regulated and "the rest of the sample" parts

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**

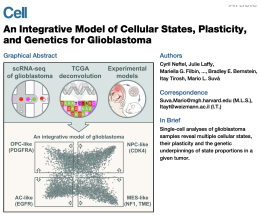


Were able to show benefit of certain gene inhibitors in glioblastoma xenografts on mice

- Hierarchical clustering of cells to form *partitions* of each sample into up-regulated and "the rest of the sample" parts
- Avoiding batch effects: integrated common genesets instead

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**


Cell

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

Graphical Abstract

scRNA-seq of glioblastoma | TCGA deconvolution | Experimental models

An integrative model of glioblastoma

OPC-like (PDGFRA) | NPC-like (CDK4)

AC-like (EGFR) | MES-like (NF1, TIMP)

Authors

Cyril Neftel, Julie Laffy, Marietta G. Filbin, ..., Bradley E. Bernstein, Itay Tirosh, Mario L. Suvà

Correspondence

Suva.Mario@mgh.harvard.edu (M.L.S.), Itay@weizmann.ac.il (I.T.)

In Brief

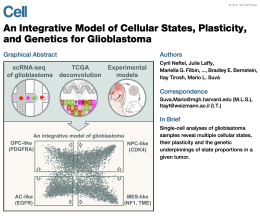
Single-cell analyses of glioblastoma samples reveal multiple cellular states, their plasticity and the genetic underpinnings of state proportions in a given tumor.

Problems:

- Global clustering (dominated by global similarities)

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**

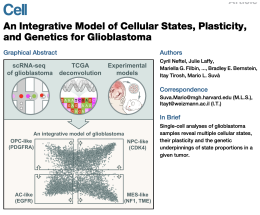


Problems:

- Global clustering (dominated by global similarities)
- Search in each sample separately, missing out on a synchronized unified search

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**

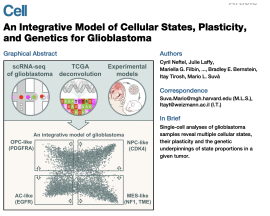


Problems:

- Global clustering (dominated by global similarities)
- Search in each sample separately, missing out on a synchronized unified search
- Dichotomous clustering: cell can belong to only 1 cluster (like many methods)

Inspired by...

...the glioblastoma paper by Neftel et al. (2019) where they searched for *signatures* using **global hierarchical clustering** of cells and genes in each sample **separately**



Problems:

- Global clustering (dominated by global similarities)
- Search in each sample separately, missing out on a synchronized unified search
- Dichotomous clustering: cell can belong to only 1 cluster (like many methods)
- Extremely ad-hoc, overabundance of dubious thresholds

Optimization search for the best contrast

Our method

- Is not a mere clustering of the cells

INSTITUT PASTEUR

Optimization search for the best contrast

Our method

- Is not a mere clustering of the cells
- Looks for signatures in all samples at the same time

Optimization search for the best contrast

Our method

- Is not a mere clustering of the cells
- Looks for signatures in all samples at the same time
- Each cell can support multiple programs

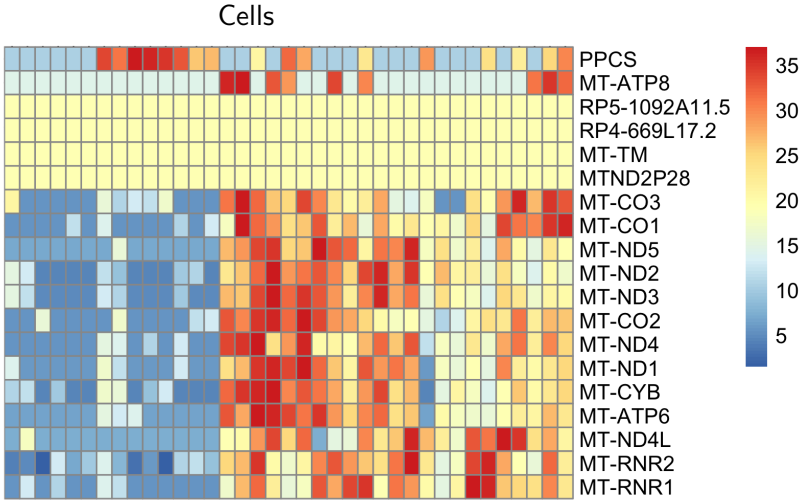
Optimization search for the best contrast

Our method

- Is not a mere clustering of the cells
- Looks for signatures in all samples at the same time
- Each cell can support multiple programs
- No ad-hoc thresholds, no parameters

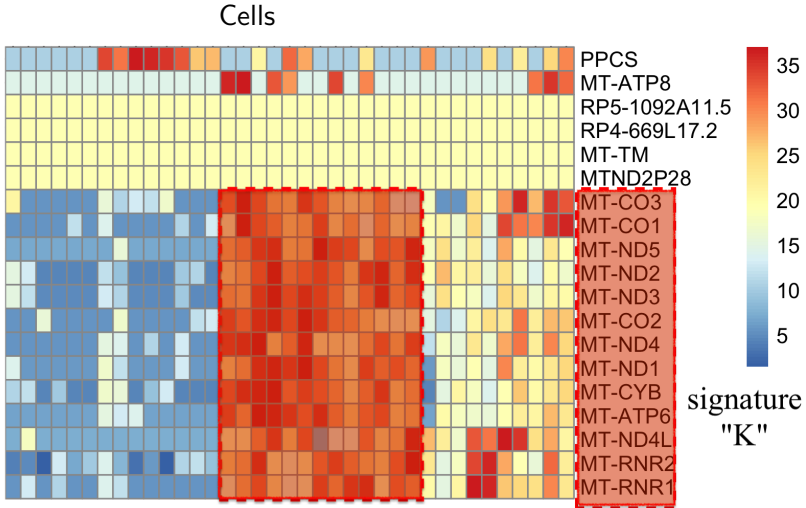
Optimization search for the best contrast

find such partition of sample's cells and genes that confers the highest contrast from the rest of the sample



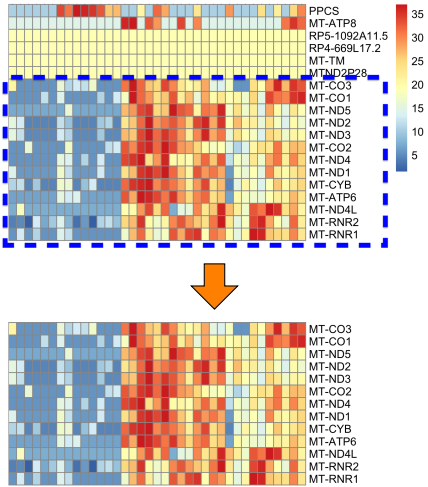
Optimization search for the best contrast

find such partition of sample's cells and genes that confers the highest contrast from the rest of the sample



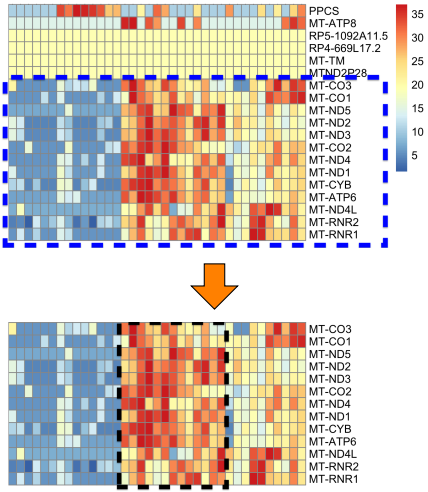
Optimization search for the best contrast

find such partition of sample's cells and genes that confers the highest contrast from the rest of the sample



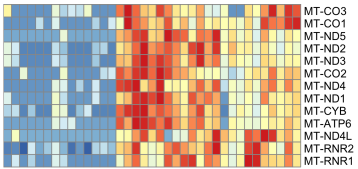
Optimization search for the best contrast

find such partition of sample's cells and genes that confers the highest contrast from the rest of the sample



INSTITUT PASTEUR

Friedman Statistic



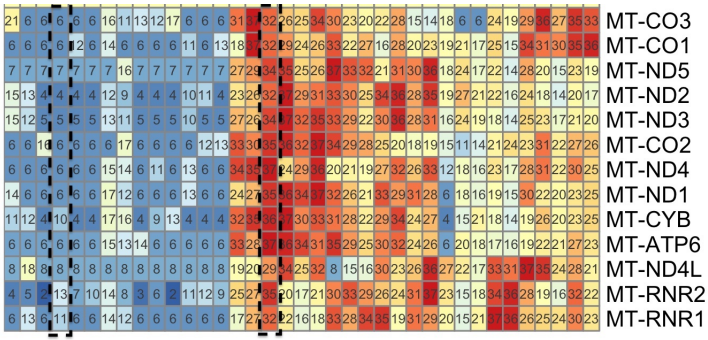
$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(R_{.j} - \frac{k+1}{2} \right)^2$$

where

$$R_j = \sum_{i=1}^{n} r_{ij} \quad \text{and} \quad R_{.j} = \frac{R_j}{n}.$$

INSTITUT PASTEUR

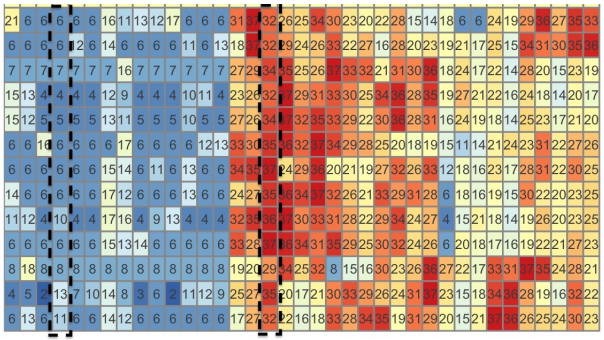
Friedman Statistic



$$R_j = \sum_{i=1}^{n} r_{ij} \quad \text{and} \quad R_{.j} = \frac{R_j}{n}.$$

INSTITUT PASTEUR

Friedman Statistic

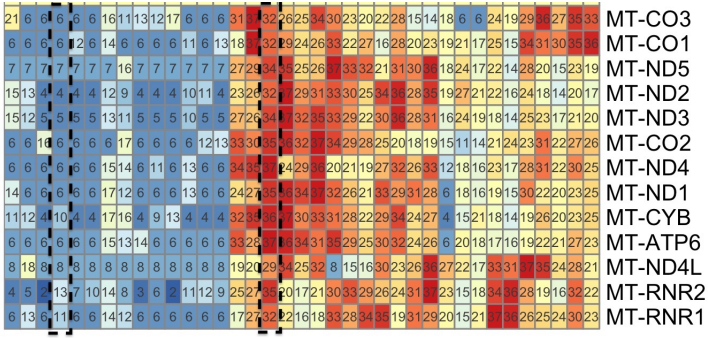


$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{k+1}{2} \right)^2$$

$$= 370.1$$

INSTITUT PASTEUR

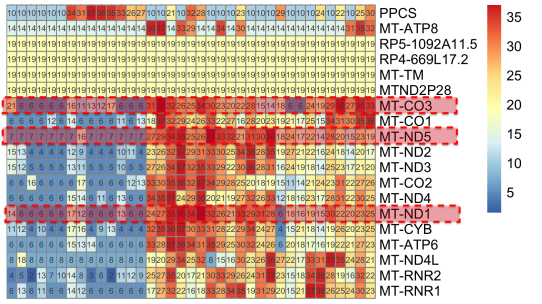
Friedman Statistic



$$-\log p(S, df) = -\log p(370.1, 36) = 129.7 \quad (1)$$

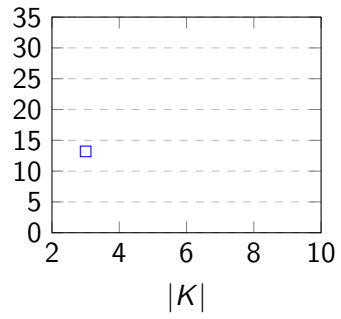
INSTITUT PASTEUR

Selecting Genes



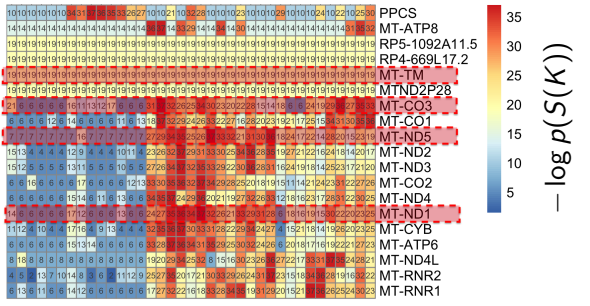
Currently selected genes highlighted

Friedman's logP vs # of genes



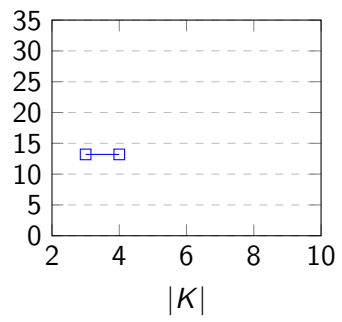
INSTITUT PASTEUR

Optimization Process



Currently selected genes highlighted

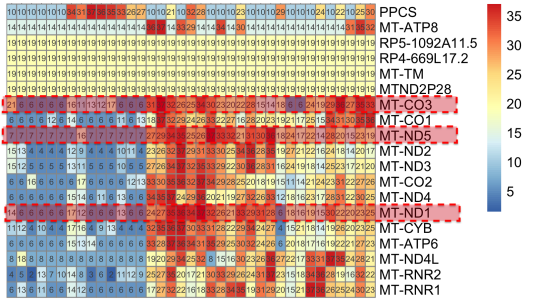
Friedman's logP vs # of genes



INSTITUT PASTEUR

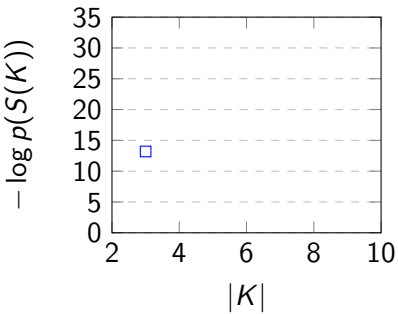
Selecting Genes

(rolling back)

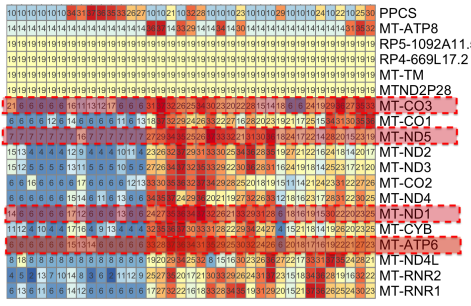


Currently selected genes highlighted

Friedman's logP vs # of genes

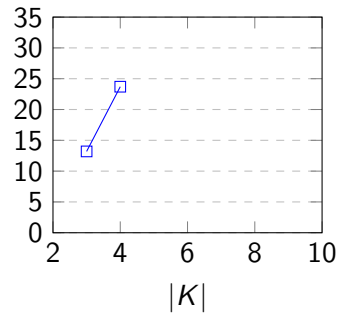


Selecting Genes

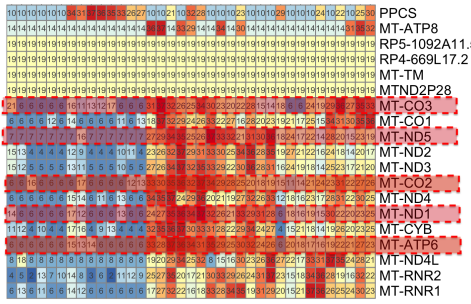


Currently selected genes highlighted

Friedman's logP vs # of genes

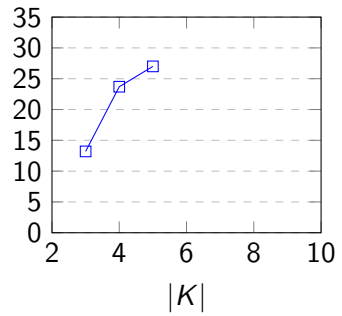


Selecting Genes



Currently selected genes highlighted

Friedman's logP vs # of genes

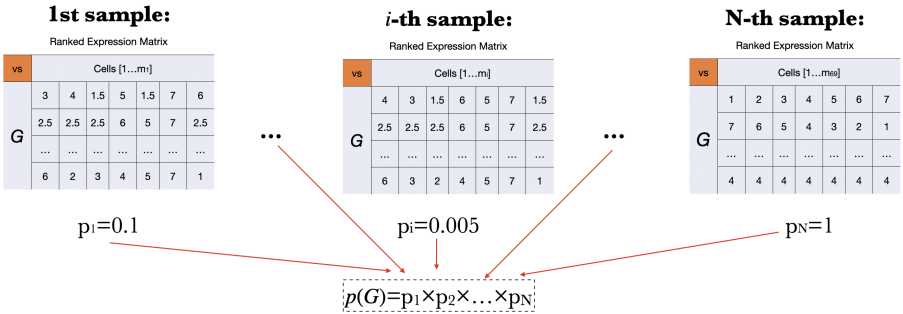


INSTITUT PASTEUR

Objective to Optimize

Find signatures $\hat{K}$ which maximize the objective function over N samples

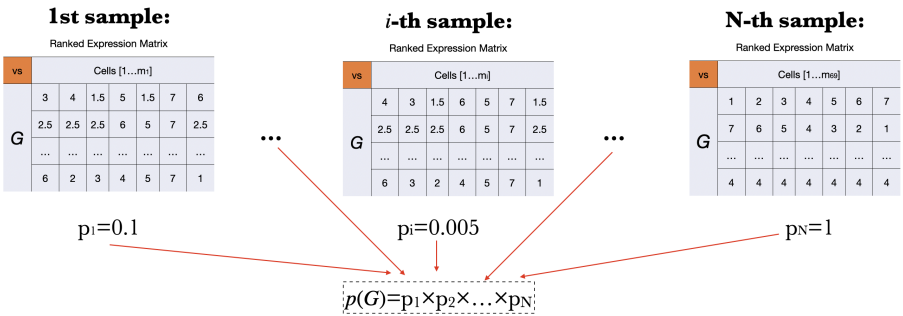
Easily expanded over multiple samples by multiplying individual p values



INSTITUT PASTEUR

Objective to Optimize

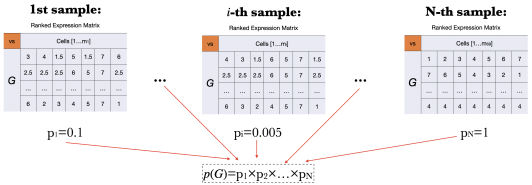
Find signatures $\hat{K}$ which maximize the objective function over N samples



INSTITUT PASTEUR

Objective to Optimize

Find signatures $\hat{K}$ which maximize the objective function over N samples



Optimize the objective function

$$g(K) = \sum_{i=1}^{N} (-\log p_i(K)) \tag{2}$$

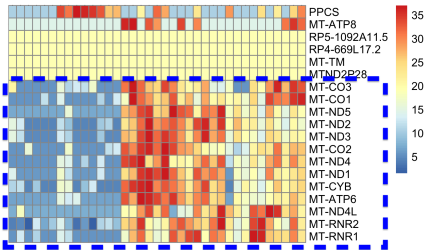
so that

$$\hat{K} = \arg\max_{K \in \mathfrak{G}} g(K) \tag{3}$$

INSTITUT PASTEUR

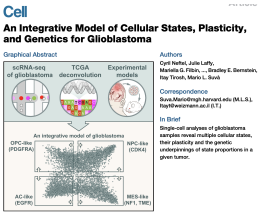
Objective to Optimize

Optimal signature $\hat{K}$



Results

Glioblastoma Smartseq2 data from Nefel et al.

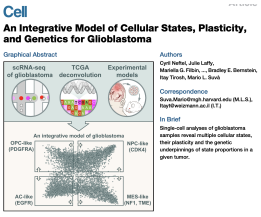


27 samples, 8k cells x 23 k genes. Known ground truth, known signatures

INSTITUT PASTEUR

Results

Glioblastoma Smartseq2 data from Nefel et al.

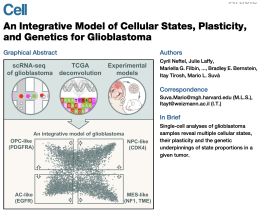


27 samples, 8k cells x 23 k genes. Known ground truth, known signatures

- Our method discovers similar signatures as Neftel et al, Seurat, HARMONY, LIGER and cNMF

Results

Glioblastoma Smartseq2 data from Nefel et al.



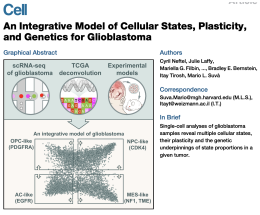
27 samples, 8k cells x 23 k genes. Known ground truth, known signatures

- Our method discovers similar signatures as Neftel et al, Seurat, HARMONY, LIGER and cNMF
  - enrichments from MSigDB check out as well

INSTITUT PASTEUR

Results

Glioblastoma Smartseq2 data from Nefel et al.

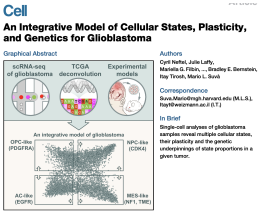


27 samples, 8k cells x 23 k genes. Known ground truth, known signatures

- Our method discovers similar signatures as Neftel et al, Seurat, HARMONY, LIGER and cNMF
  - enrichments from MSigDB check out as well
- On top of that it discovers small signatures not found by any of the methods above, and their enrichments are highly relevant to glioblastoma

Results

Glioblastoma Smartseq2 data from Nefel et al.


Cell

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

Graphical Abstract

scRNA-seq of glioblastoma | TCGA deconvolution | Experimental models

An integrative model of glioblastoma

OPC-like (PDGFRA) | NPC-like (CDK4) | AC-like (EGFR) | MES-like (NF1, TMIC)

Authors

Cyril Neftel, Julie Laffy, Marielle G. Filbin, ..., Bradley E. Bernstein, Itay Tirosh, Mario L. Suvà

Correspondence

Suva.Mario@mgh.harvard.edu (M.L.S.), Itay@weizmann.ac.il (I.T.)

In Brief

Single-cell analyses of glioblastoma samples reveal multiple cellular states, their plasticity and the genetic underpinnings of state proportions in a given tumor.

27 samples, 8k cells x 23 k genes. Known ground truth, known signatures

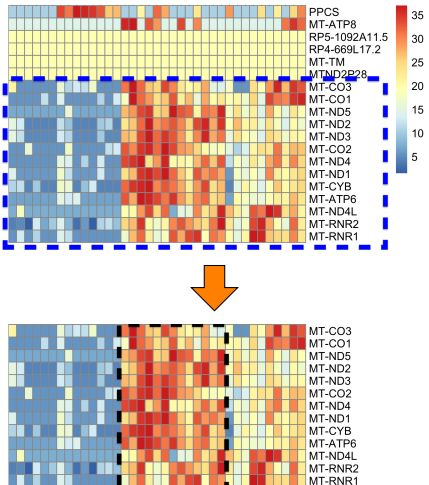
- Our method discovers similar signatures as Neftel et al, Seurat, HARMONY, LIGER and cNMF
  - enrichments from MSigDB check out as well
- On top of that it discovers small signatures not found by any of the methods above, and their enrichments are highly relevant to glioblastoma
- Our signatures are on average enriched for smaller pathways/genesets than the competitors' signatures

Solutions

	starting loss	No of iter	Final score	Geneset
1	start_loss was 91.7711	48 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
2	start_loss was 105.185	48 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
3	start_loss was 107.669	38 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
4	start_loss was 113.993	29 iterations locally	623.992	ATP1B2 BCAN CST3 DBI EDNRB GATM GPM6B HEPACA...
5	start_loss was 113.502	33 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
6	start_loss was 125.488	48 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
7	start_loss was 101.827	47 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
8	start_loss was 109.972	49 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
9	start_loss was 114.923	48 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...
10	start_loss was 122.534	47 iterations locally	969.918	AURKB BIRC5 BUB1B CDCA5 CDK1 CENPF FAM64A KIF...

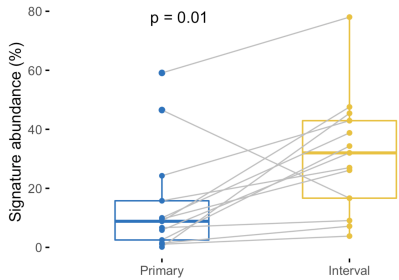
Addendum 1: Tricky part

how to define "abundance" of a signature in a given sample?



Addendum 1: Functional testing of the signatures

After obtaining all the signatures, test them for association with resistance.



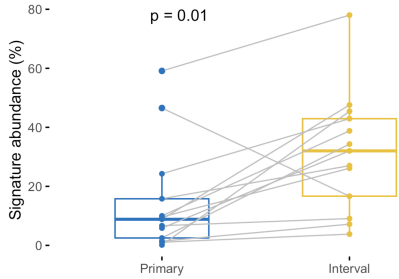
"a desirable signature"

Signature's **abundance** goes up in "after" treatment vs "before" treatment?
Potential chemoresistance signature

INSTITUT PASTEUR

Addendum 1: Functional testing of the signatures

After obtaining all the signatures, test them for association with resistance.



"a desirable signature"

Signature's **abundance** goes up in "after" treatment vs "before" treatment?
Potential chemoresistance signature

- Only 11 pairs :(

Addendum 1: Compound Objective Function

Sum over all samples of Friedman Statistic log-p values

$$g(K) = \sum_{i=1}^{n} (-\log p_i(K)) \tag{4}$$

where n is the number of samples

$$f(K) = \delta \cdot h(K) + g(K) \tag{5}$$

$$\hat{K} = \underset{K \in \mathfrak{G}}{\arg\max}\, f(K) \tag{6}$$

INSTITUT PASTEUR

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster
- Performance-oriented code, as few overheads as possible
  - Aggressive optimization -03 flag during compilation

INSTITUT PASTEUR

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster
- Performance-oriented code, as few overheads as possible
  - Aggressive optimization -03 flag during compilation
- Partial update of the Friedman Statistic (not a complete recalculation)

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster
- Performance-oriented code, as few overheads as possible
  - Aggressive optimization -03 flag during compilation
- Partial update of the Friedman Statistic (not a complete recalculation)
- I borrowed R's "under the hood" Cpp implementation of $\log p$ of χ^2 distribution
  - very fast - uses Chebyshev rational approximations for $Erfc(x)$ calc

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster
- Performance-oriented code, as few overheads as possible
  - Aggressive optimization -03 flag during compilation
- Partial update of the Friedman Statistic (not a complete recalculation)
- I borrowed R's "under the hood" Cpp implementation of $\log p$ of χ^2 distribution
  - very fast - uses Chebyshev rational approximations for $Erfc(x)$ calc

Resulting algo scales as O(n) where n is the number of cells

Addendum 2: A Brief Note on Performance

How we make the algo work faster

- embarrassingly parallel, ready-to-deploy on an HPC cluster
- Performance-oriented code, as few overheads as possible
  - Aggressive optimization -03 flag during compilation
- Partial update of the Friedman Statistic (not a complete recalculation)
- I borrowed R's "under the hood" Cpp implementation of log p of χ^2 distribution
  - very fast - uses Chebyshev rational approximations for $Erfc(x)$ calc

Resulting algo scales as O(n) where n is the number of cells
One full iteration on Pasteur's Maestro Cluster core:
49 microseconds for 27 samples of 8000 cells
169 microseconds for 63 samples of 25k cells

Addendum 3: Friedman Statistic Quick Update

Friedman statistic with n genes:

$$S'_n = \frac{12\sum_{j=1}^{m} R_j^2 - 3n^2m(m+1)^2}{nm(m+1) - [1/(m-1)]\sum_{i=1}^{n}\{(\sum_{j=1}^{g_i} t_{i,j}^3) - m\}} \tag{7}$$

Denote

$A_n = 12\sum_{j=1}^{m} R_j^2$, $B_n = 3n^2m(m+1)^2$, $C_n = nm(m+1)$,
$D_n = [1/(m-1)]\sum_{i=1}^{n}\{(\sum_{j=1}^{g_i} t_{i,j}^3) - m\}$,

so that

$$S'_n = \frac{A_n - B_n}{C_n - D_n} \tag{8}$$

INSTITUT PASTEUR

Addendum 3: Friedman Statistic Quick Update

Denote

$A_n = 12\sum_{j=1}^{m} R_j^2$, $B_n = 3n^2m(m+1)^2$, $C_n = nm(m+1)$,
$D_n = [1/(m-1)]\sum_{i=1}^{n}\{(\sum_{j=1}^{g_i} t_{i,j}^3) - m\}$,

so that When adding a gene: n $\rightarrow n+1$

$\mathrm{A}_{n+1} = 12\sum_{j=1}^{m}(R_{j,n} + r_j)^2 = A_n + 24\sum_{j=1}^{m} R_{j,n}r_j + 12\sum_{j=1}^{m} r_j^2$
$B_{n+1} = 3(n+1)^2m(m+1)^2 = B_n + 3m(m+1)^2(2n+1)$
$C_{n+1} = (n+1)m(m+1) = C_n + m(m+1)$
$D_{n+1} = D_n + [1/(m-1)]\{(\sum_{j=1}^{g_{n+1}} t_{n+1,j}^3) - m\}$

$$S'_{n+1} = \frac{(A_n + 24\sum_{j=1}^{m} R_{j,n}r_j + 12\sum_{j=1}^{m} r_j^2) - (B_n + 3m(m+1)^2(2n+1))}{(C_n + m(m+1)) - (D_n + [1/(m-1)]\{(\sum_{j=1}^{g_{n+1}} t_{n+1,j}^3) - m\})} \quad (9)$$

INSTITUT PASTEUR

Challenge: non-fixed size signatures

- in reality we use fixed size signatures, swapping genes in and out

Challenge: non-fixed size signatures

- in reality we use fixed size signatures, swapping genes in and out
- size-restriction-free variant ends up in absorbing all the genes in the dataset
  - even when penalizing the Friedman's S by the number of currently selected genes n :
    - $S' = S/n^{alpha}$ (where alpha is a real number [0.5...2])

INSTITUT PASTEUR