Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model

Anthony Gitter

Cancer Bioinformatics (BMI 826/CS 838)

May 5, 2015

All figures from Cho2013 unless noted otherwise

Class business

- Project presentations Thursday
- Guidelines on website
- Project report due May 11

- How to schedule presentation order?

Inspiration from CMapBatch

	Chris rank	1
	Jiayue rank	4
	Network stratification project rank	√4 (1)
	Anita rank	7
	Vee rank	6
	Survival prediction project rank	√42 (3)
	Taylor rank	3
	Haixiang rank	5
Outlier	~~Erkin rank~~	~~2~~
	Clustering pipeline project rank	√15 (2)

Subtyping in cancer

- Substantial differences across tumors even within one type of cancer
  - Molecular alterations
  - Survival outcomes
  - Response to therapy

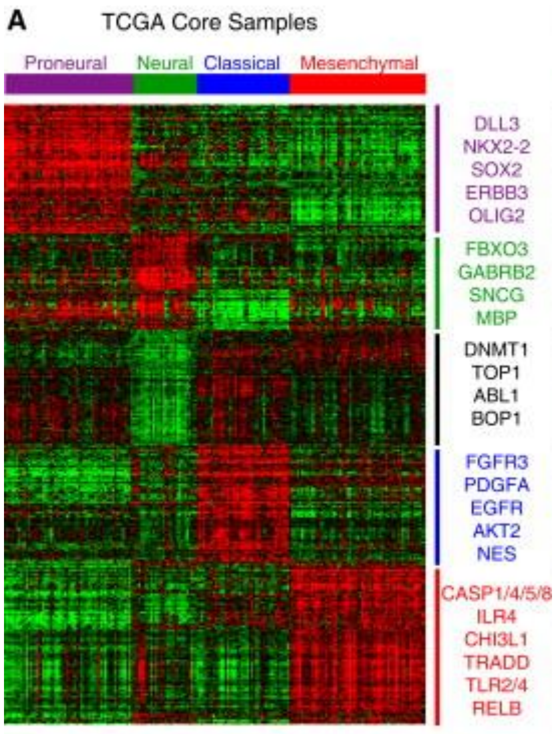
Traditional subtyping

- Learn gene expression signature to distinguish classes
  - AML vs ALL
  - PAM50 for breast cancer
  - Glioblastoma (GBM) Verhaak2010

GBM subtypes

- Learn class centroids with ClaNC (classification to nearest centroids)
    - t-test statistic to identify genes
    - 210 genes per class in GBM

- Neural subtype has been criticized



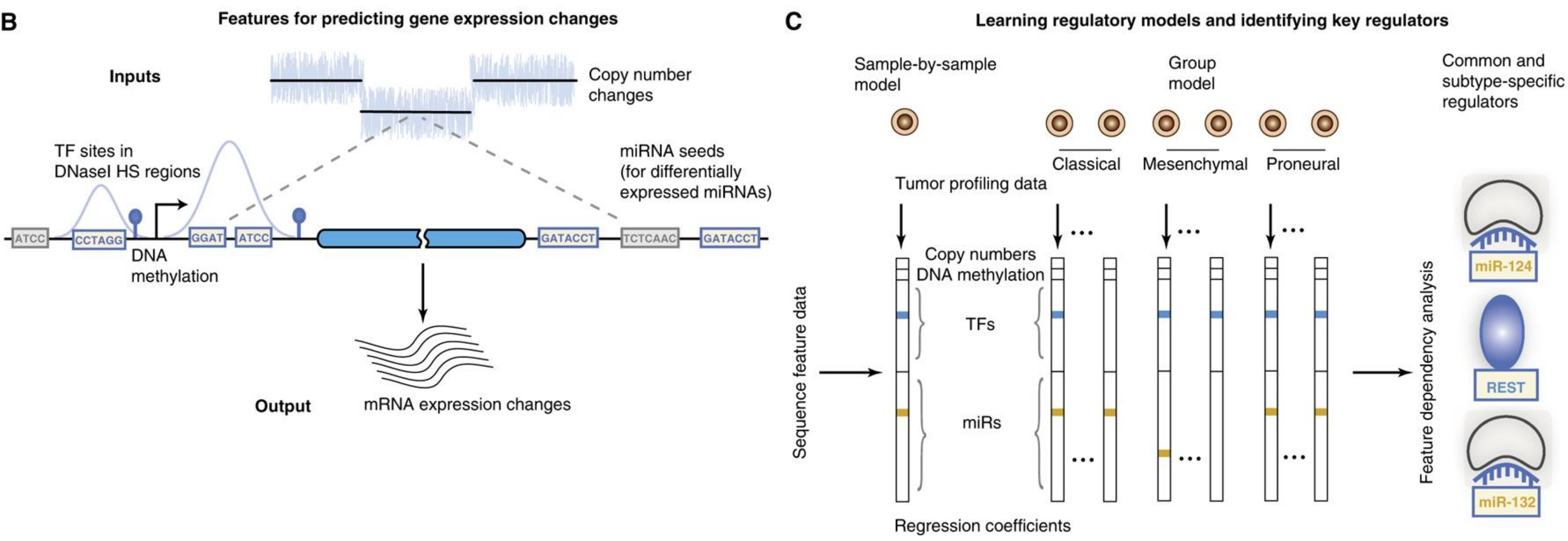
Verhaak2010

Many analyses depend on subtypes

- MutSig or other enrichment tests

Many analyses depend on subtypes

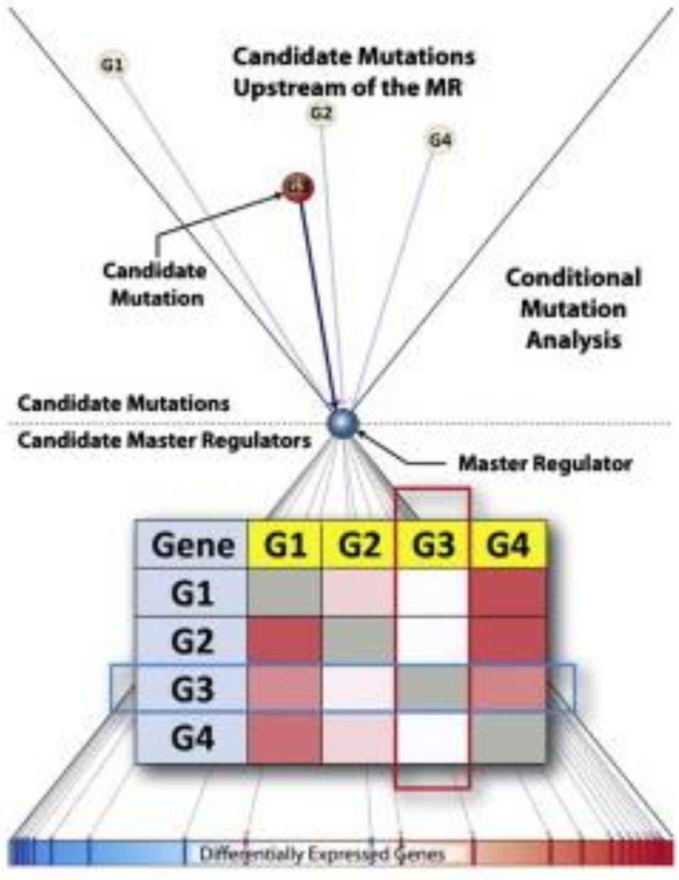
- Group lasso in regulator regression



[Setty2012]

Many analyses depend on subtypes

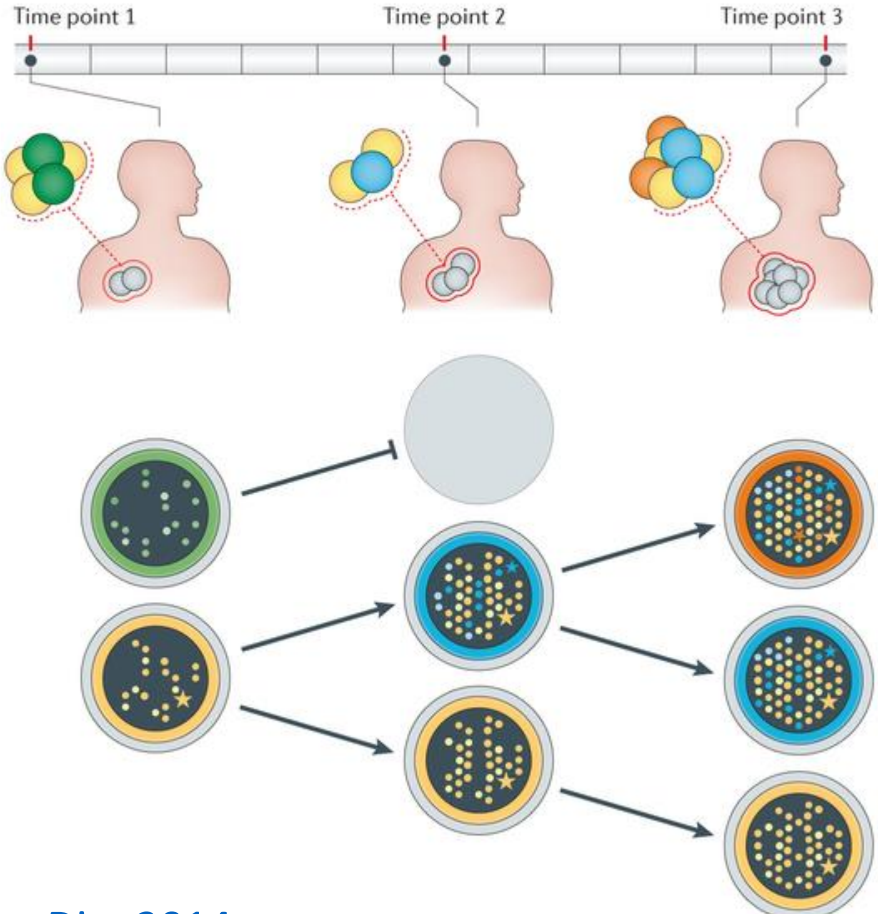
- DIGGIT functional CNV association test



[Chen2014]

Problem with subtype classifiers

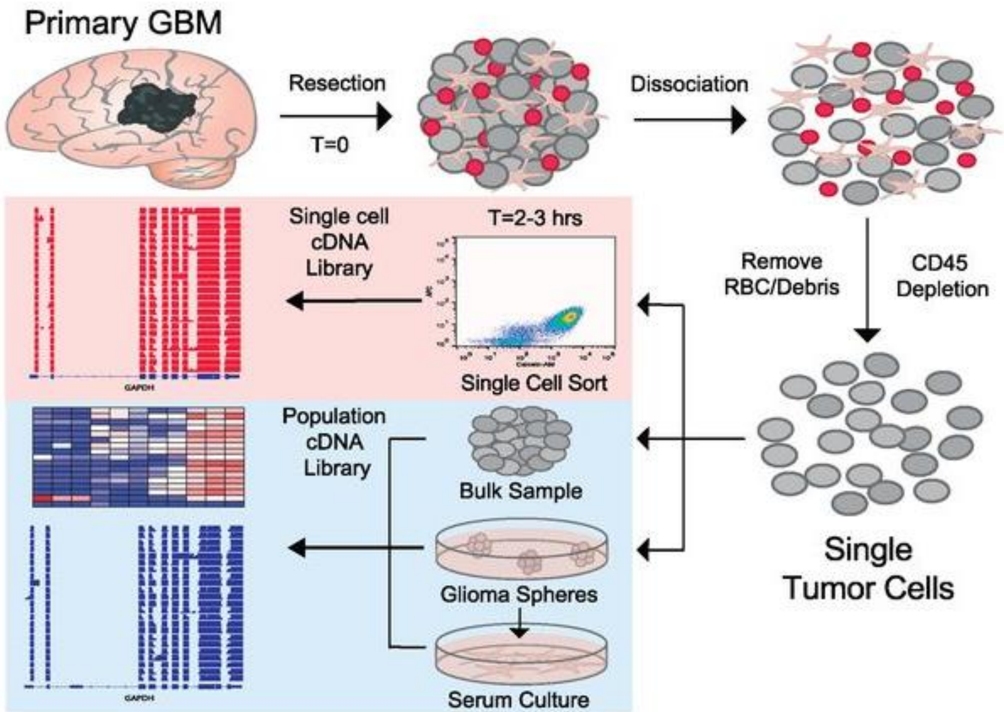
- Cancer and individual tumors are heterogeneous



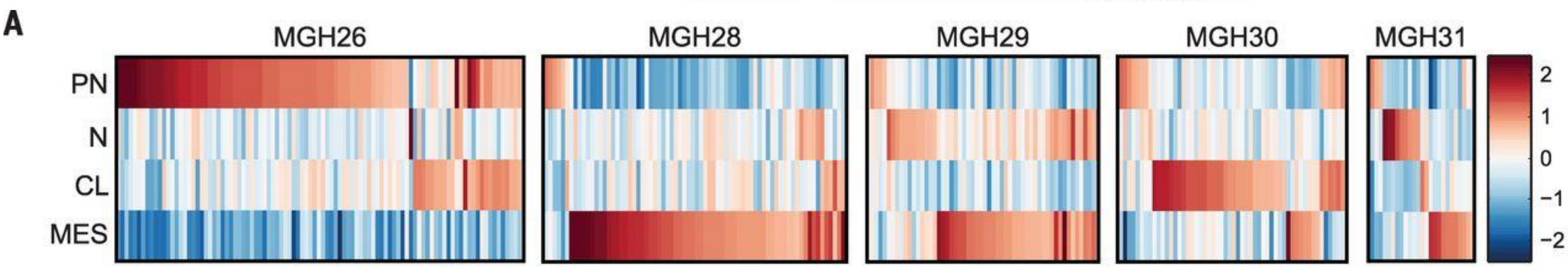
Ding2014

Heterogeneity in expression classification

- Single-cell RNA-seq shows a **single GBM tumor** is composed of cells from **multiple subtypes**



Patel2014



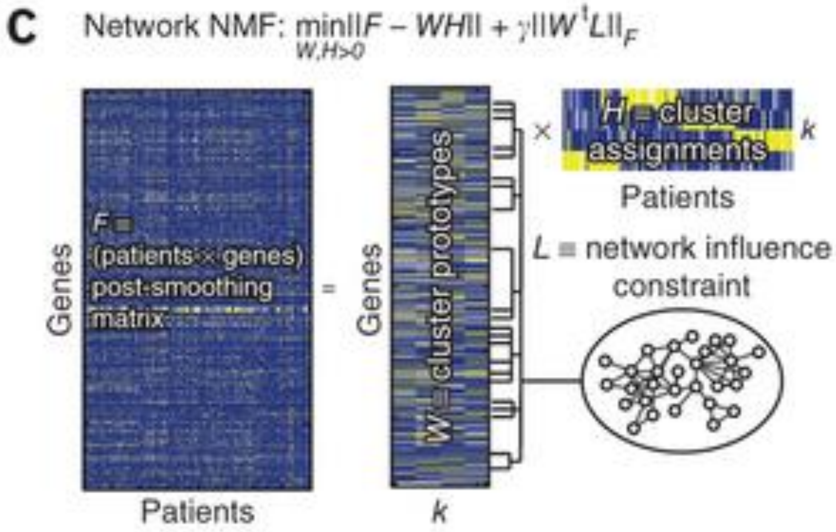
Prob_GBM: mixtures of subtypes

- Patients are mixtures of subtypes
- Subtypes are mixtures of genomic factors

- Sound familiar?

Relation to Non-negative Matrix Factorization

- Network-based stratification
- Similar concepts, different strategies



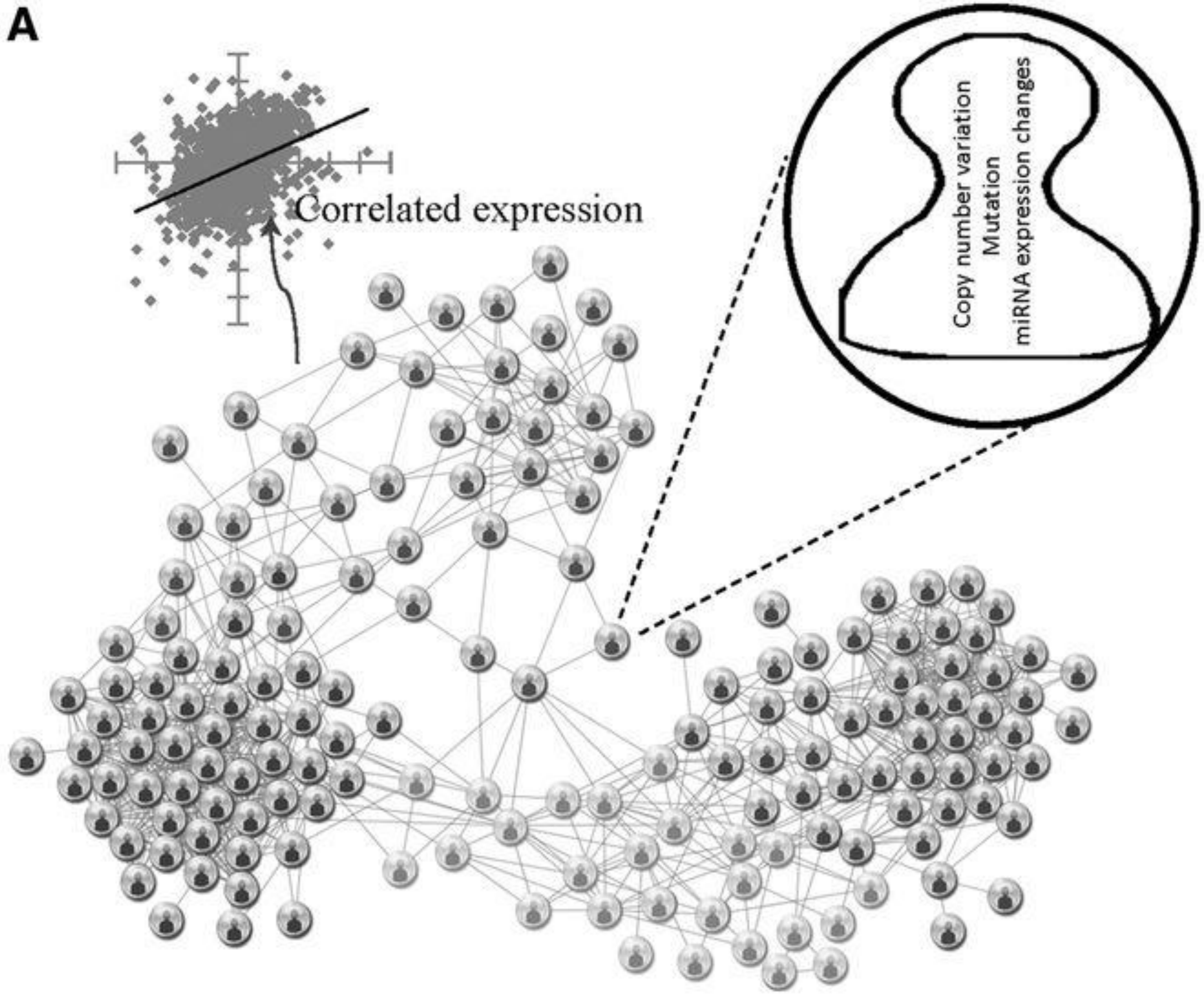
Hoffree2013

Prob_GBM model

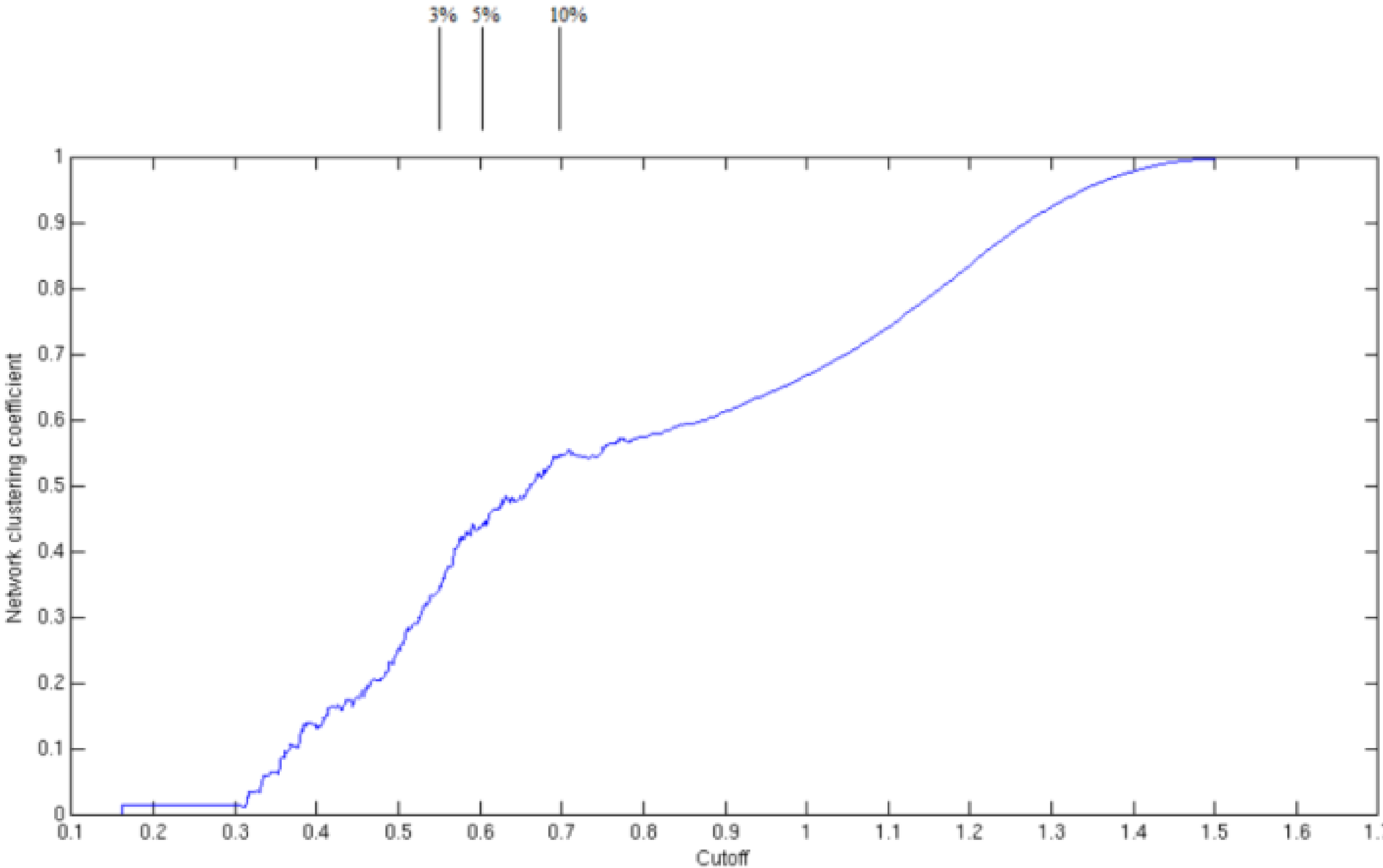
- Gene expression is a molecular level phenotype
  - Treated as effect of disease, not cause
- Patient-patient similarity based on expression

- Genomic factors cause disease
  - Mutations, CNV, miRNAs
- Expression similarities explained by genomic similarities

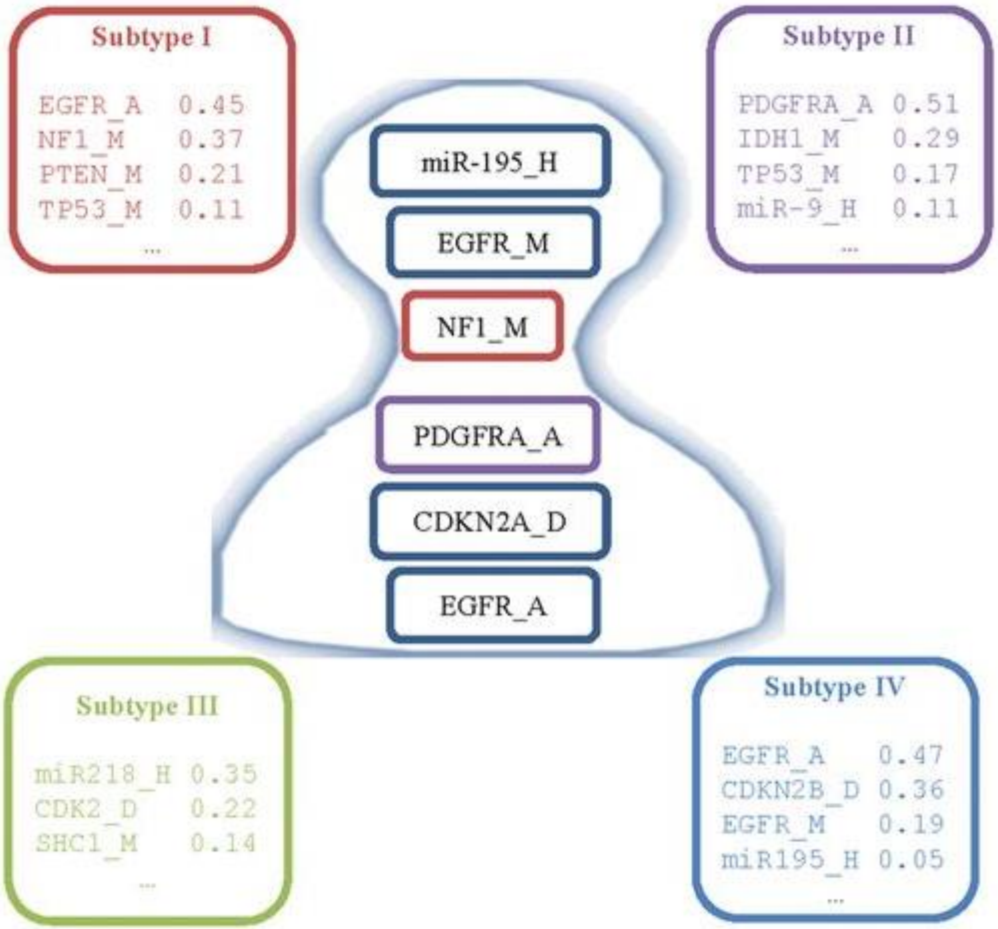
Build patient-patient similarity network



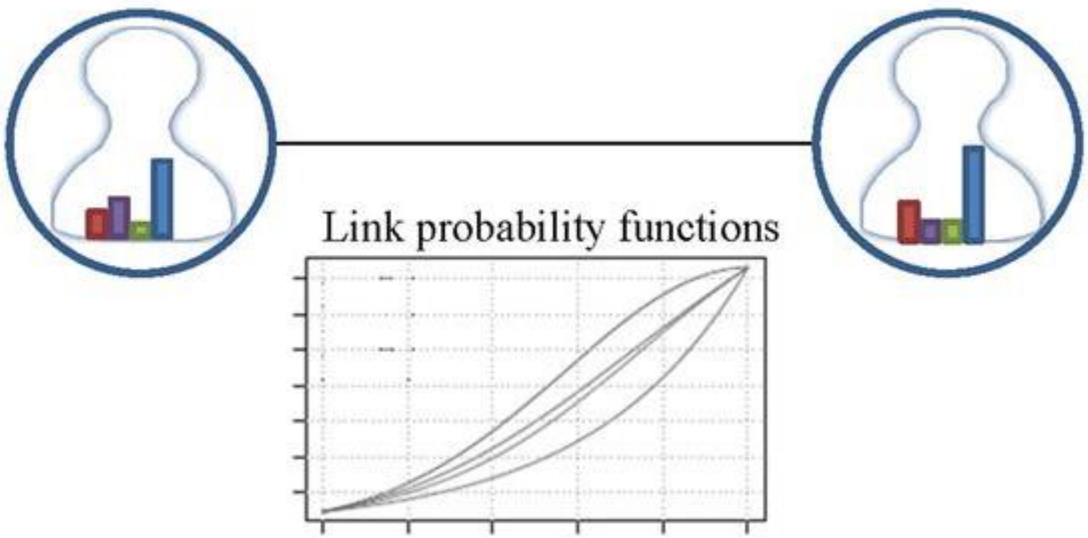
Choose co-expression threshold



Learn subtype distributions

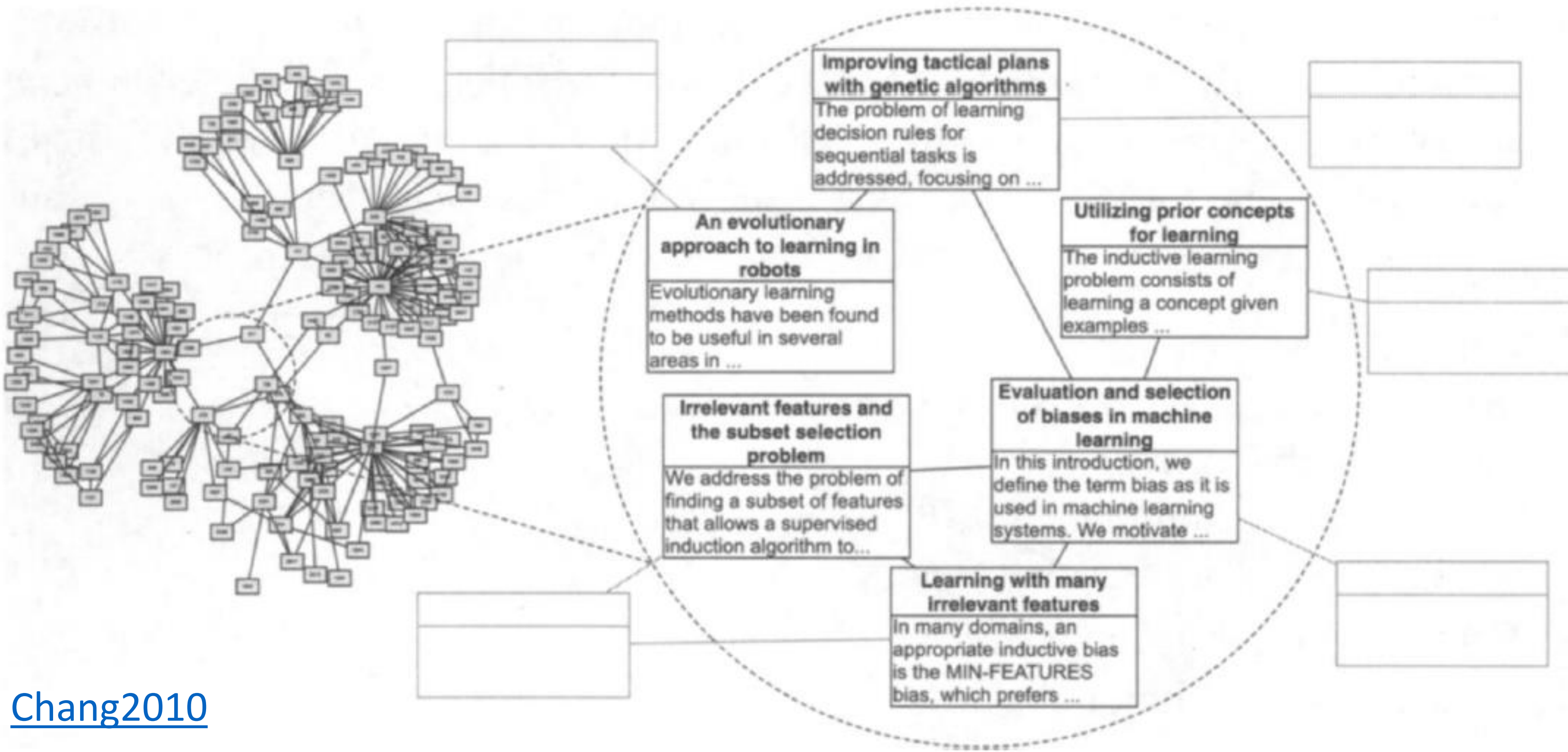


Likelihood of edge between similar patients from subtype assignments



Inspired by relational topic model

- Documents are bags of words
- Document-document citation network



[Chang2010]

Mapping to cancer domain

- Documents = patients
- Bag of words = bag of genomic alterations
- Document citation link = patient-patient co-expression above some threshold

Generative probabilistic model

1. For each patient d:
   (a) Draw subtype proportions $\theta_d | \alpha \sim \mathrm{Dir}(\alpha)$.
   (b) For each "gene" $w_{d,n}$:
      i. Draw assignment $z_{d,n} | \theta_d \sim \mathrm{Mult}(\theta_d)$.
      ii. Draw "gene" $w_{d,n} | z_{d,n}, \boldsymbol{\beta}_{1:K} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$.
2. For each pair of patients d, d':
   (a) Draw binary link indicator

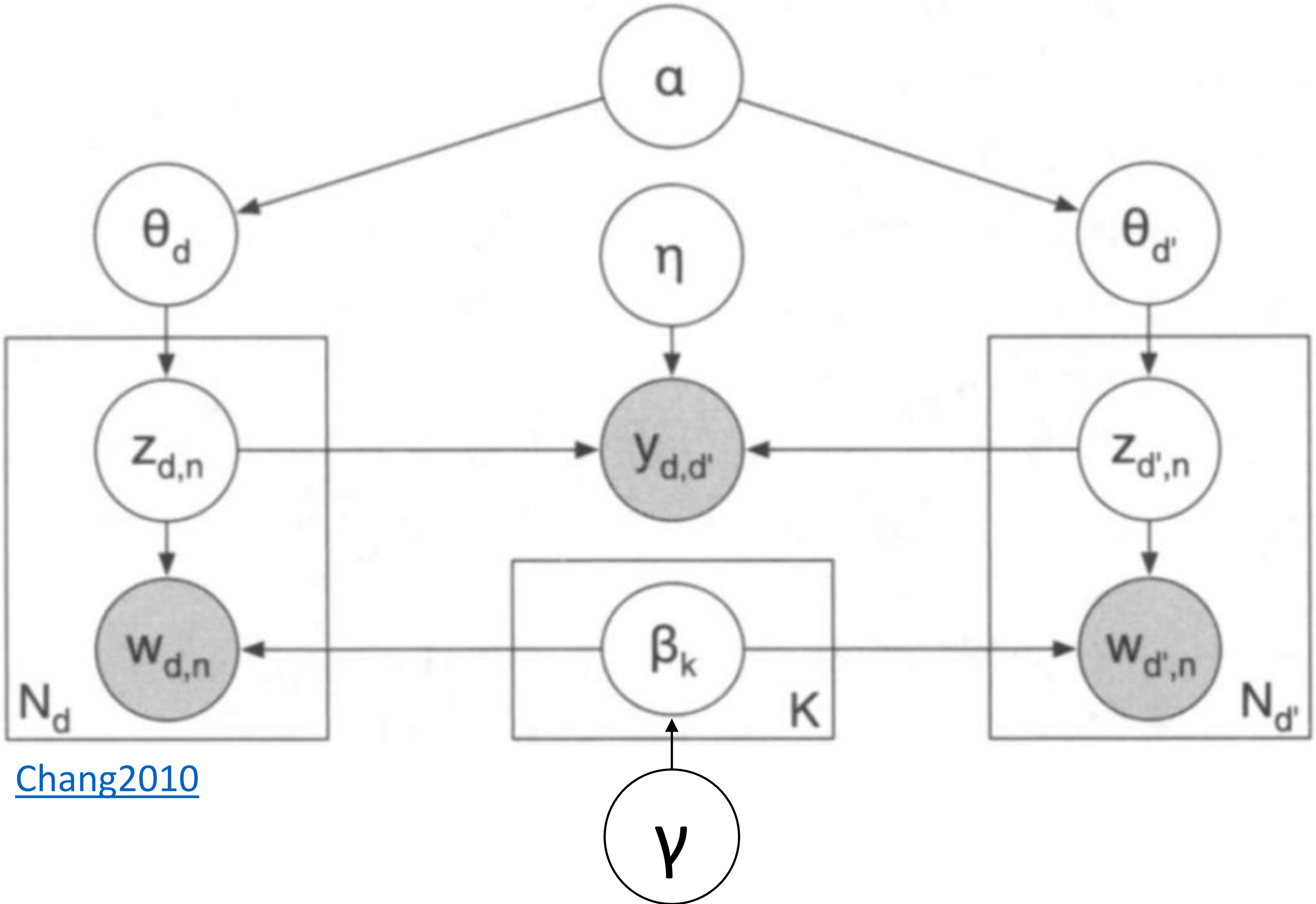
$$y_{d,d'} | \mathbf{z}_d, \mathbf{z}_{d'} \sim \psi(\cdot | \mathbf{z}_d, \mathbf{z}_{d'}, \boldsymbol{\eta}),$$

where $\mathbf{z}_d = \{z_{d,1}, z_{d,2}, \ldots, z_{d,n}\}$.

$d \rightarrow p$
$w \rightarrow g$

Chang2010

Generative probabilistic model



Chang2010

Prob_GBM distributions

- Joint distribution

$$p(\mathrm{B}, \Theta, Z, \mathrm{G}, \mathrm{L}) = \prod_k p(\beta_k) \prod_p p(\theta_p) \times \left(\prod_n p(z_{p,i}|\theta_p) p(g_{p,i}|\beta_{z_{p,i}}) \right) \prod_{p,p'} \psi(l_{p,p'}|\mathbf{z}_p, \mathbf{z}_{p'}).$$

- Posterior distribution of the latent variables

$$p(\mathrm{B}, \Theta, \mathrm{Z}|\mathrm{G}, \mathrm{L}) = \frac{p(B, \Theta, \mathrm{Z}, \mathrm{G}, \mathrm{L})}{p(\mathrm{G}, \mathrm{L})}.$$

Model estimation

- Cannot maximize posterior exactly
- Gibbs sampling generates samples from this distribution

- Two Gibbs sampling references:
  - 1 page summary
  - 231 slide tutorial

Latent variables of interest

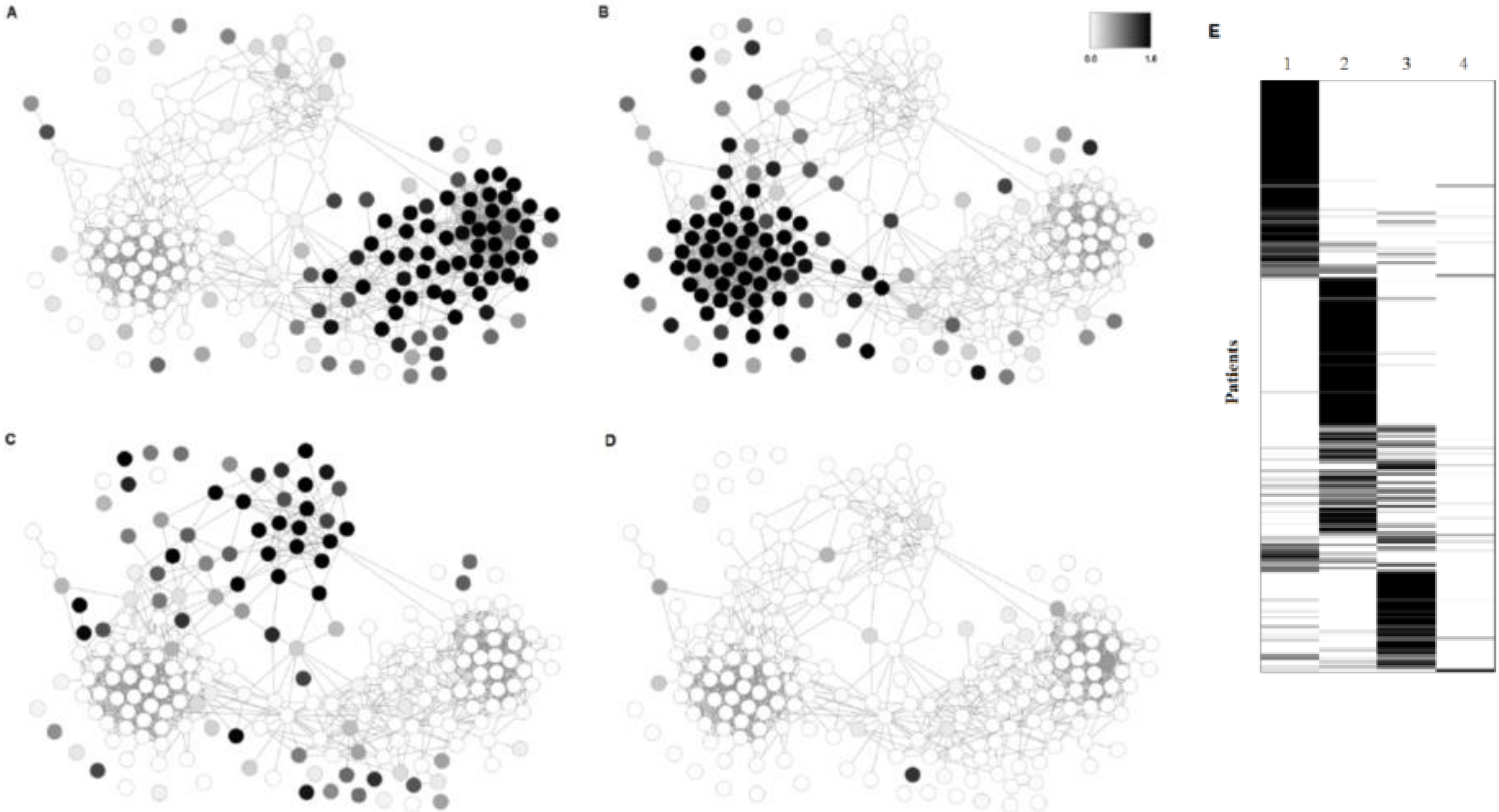
$$\hat{\theta}_k^p = \frac{c_k^p + \alpha}{\sum_{k=1}^K c_k^p + K\alpha}.$$

Subtype distributions per patient *p*

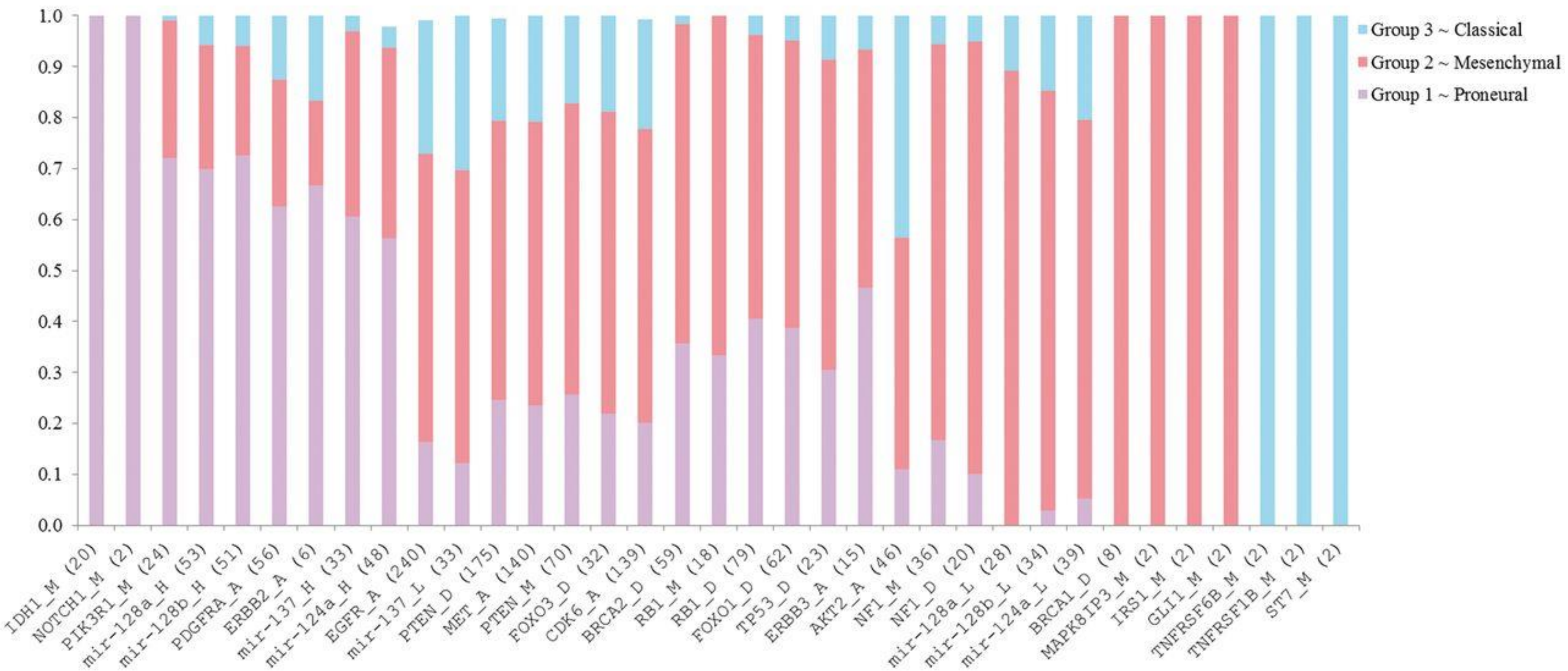
$$\hat{\beta}_k^n = \frac{c_k^n + \gamma}{\sum_{k=1}^K c_k^n + N\gamma}.$$

Distributions of genomic alteration *n* under subtype *k*

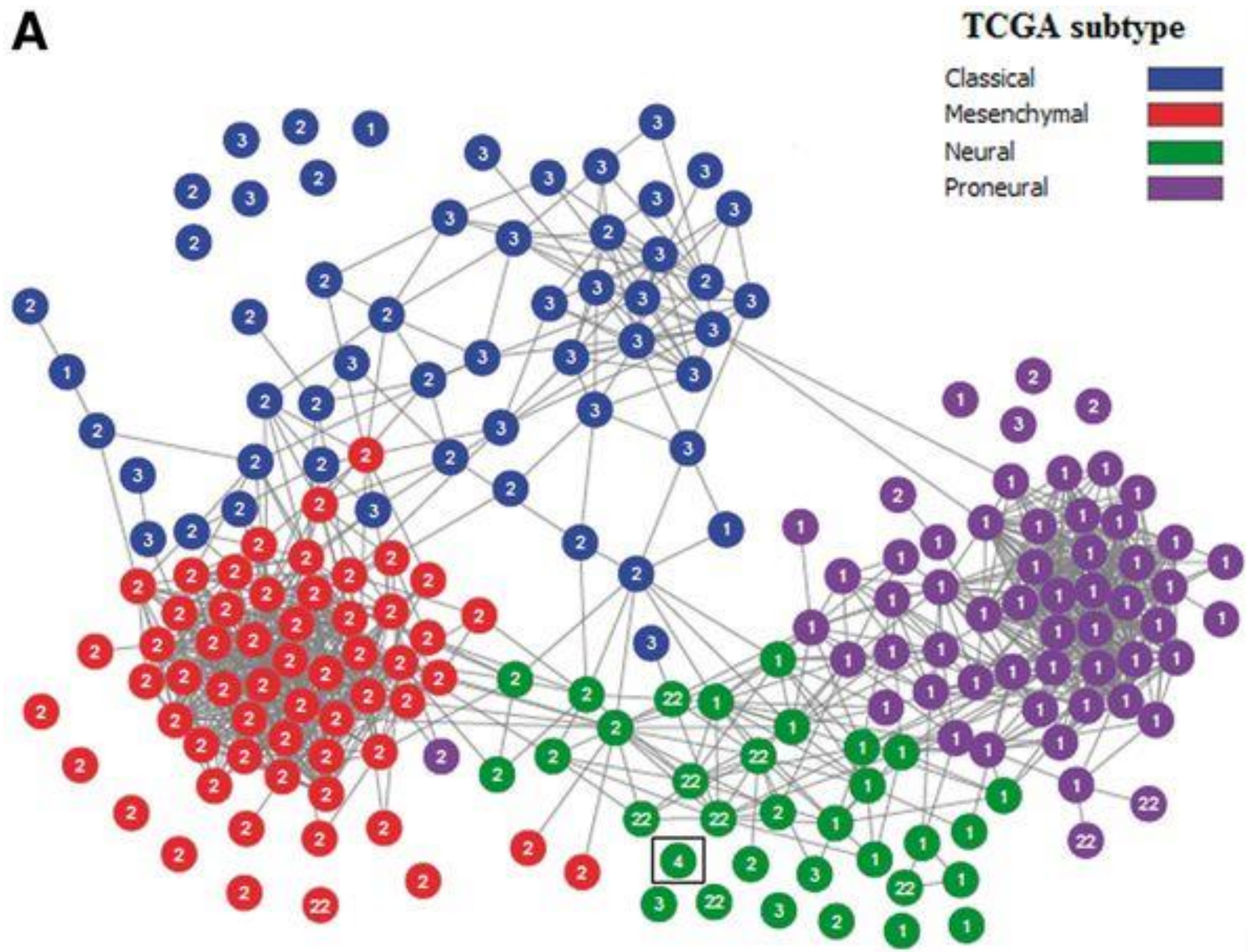
Visualizing patient distributions



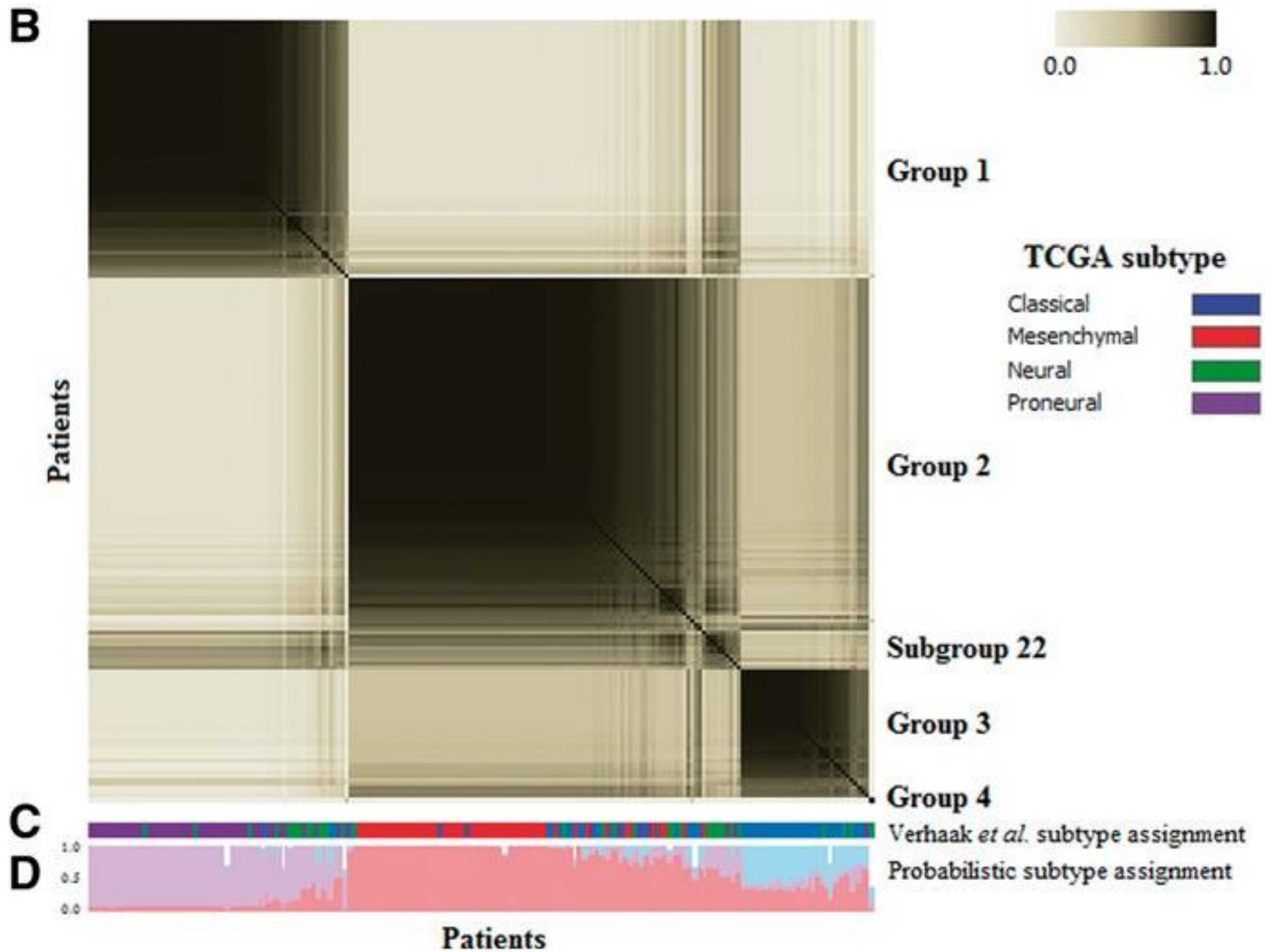
Visualizing genomic alteration distributions



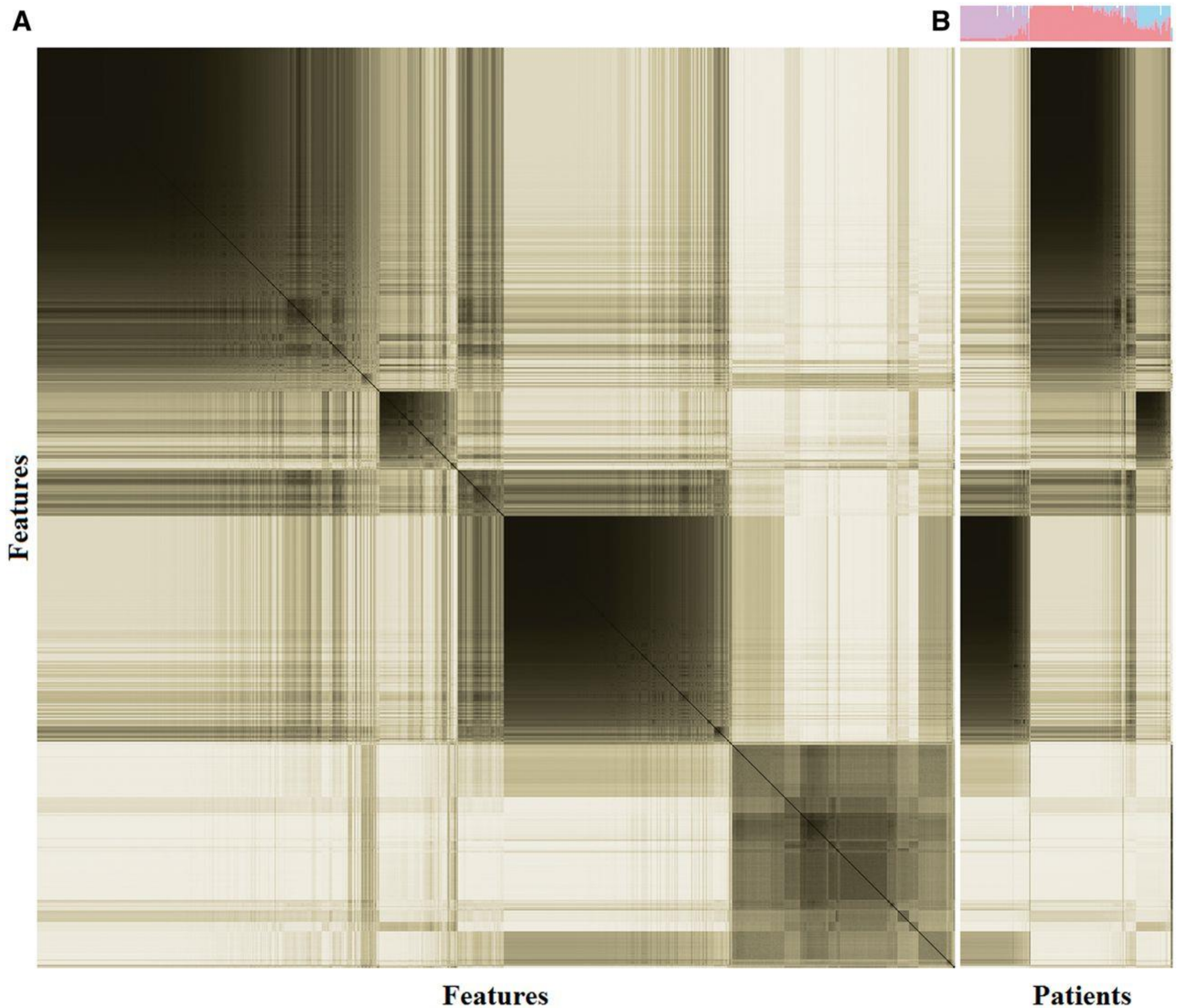
Assigning patients to subtypes



Neural is mixture of subtypes



Stability of subtype assignments



Ultimate patient-subtype, alteration-subtype associations