

Paper Presentation:

Hofree, *et al.*

Network-based stratification of tumor mutations.

Nature Methods. (2013).

Addison Hu

CB&B 555

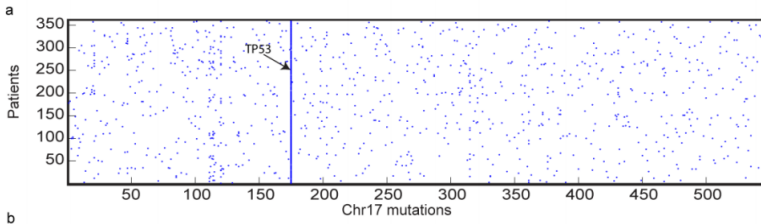
8 December 2016

Biological Motivation: Cancer Subtype Discovery

- Many forms of cancer consist of multiple subtypes with different causes and clinical outcomes.
- Identification of these subtypes are of key interest.
- In *tumor stratification*, a population of tumors is divided into meaningful subtypes based on similarity of molecular profiles.
- Past work in this regard has focused on mRNA expression data; shortcomings include RNA sample quality and lack of reproducibility between biological replicates.

Background: Somatic Mutation Profile

- High-throughput sequencing is used to identify mutations that have become enriched in the tumor cell population.
- Assumption: this set of mutations contain the causal drivers of tumor progression; therefore, similarities and differences in mutations across patients can guide tumor stratification.
- Problem: Somatic mutation profiles are highly sparse; typically fewer than 100 mutated bases are found in the entire exome.



Hypothesis: Use Gene Network Information

- Authors sought to integrate somatic tumor genomes with gene interaction networks.
- Assumption: Although two tumors may not have any mutations in common, they may share the networks affected by these mutations.
- Idea: Use network knowledge to cluster somatic mutation profiles into tumor subtypes that both provide biological insight and are tied to clinical outcomes (e.g., patient survival time, emergence of drug resistance).

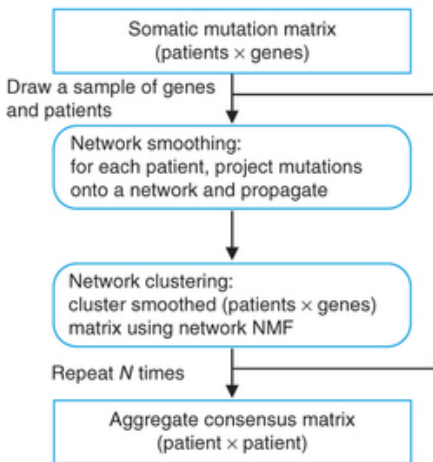
Data

- Data obtained from the Cancer Genome Atlas
- Techniques demonstrated with ovarian, uterine, and lung cancer cohorts
- Somatic mutations for each patient are represented as a series of binary states $\{0, 1\}$ on genes; a 1 denotes that a mutation has occurred for a gene relative to the germ line.
- Each patient's mutation profile is projected onto a human gene interaction network obtained from public databases (e.g., STRING¹, Pathway Commons²).

¹Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* (2011).

²Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* (2011).

Network-based Stratification: Overview



Network smoothing

After mapping the patient mutation profile onto a molecular network, the state of each gene is in $\{0, 1\}$.

Network propagation smooths the mutation signal across the network by simulating a random walk on the network:

$$F_{t+1} = \alpha F_t A + (1 - \alpha) F_0$$

where F_0 is a patient-by-gene matrix, A is a row-normalized adjacency matrix of the gene interaction network, and α is a tuning parameter that controls the distance that a mutation signal is allowed to diffuse through the network.

Because each row of F corresponds to a patient, we are essentially running a random walk for each patient.

Network smoothing:

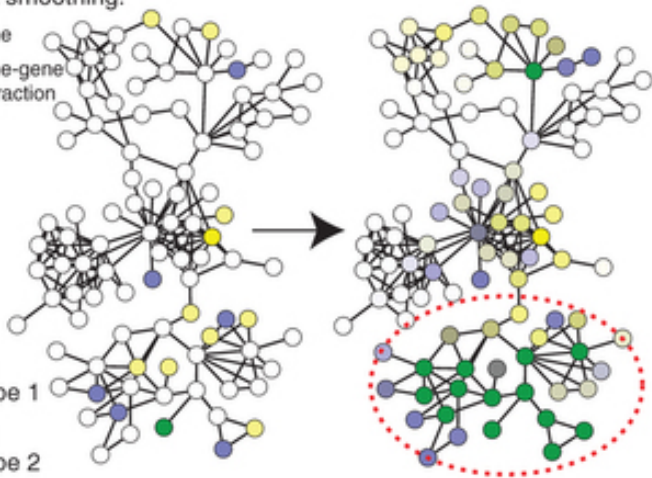
○ Gene

○—○ Gene-gene
interaction

● Patient
genotype 1

● Patient
genotype 2

● Co-occurrence of
genotype 1 and 2



Network-Regularized NMF

Non-negative matrix factorization is an unsupervised technique in which a data matrix $F \in \mathbf{R}_+^{n \times d}$ is decomposed into matrices

$$W \in \mathbf{R}_+^{n \times k}, H \in \mathbf{R}_+^{k \times d}$$

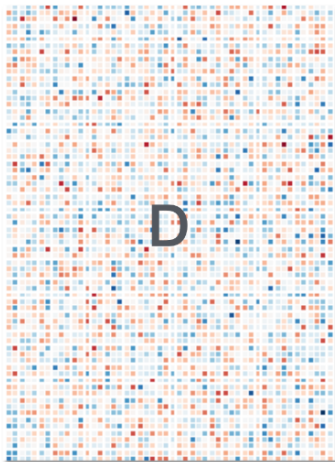
$$\underset{W, H}{\text{minimize}} \quad \|F - WH\|_F^2 + \text{tr}(W^T K W)$$

$$\text{subject to} \quad W \in \mathbf{R}_+^{n \times k}, H \in \mathbf{R}_+^{k \times d}$$

Here, a regularization term derived from the adjacency matrix is used to constrain the basis vectors in W to respect the local network neighborhoods.³

The authors chose $k = 12$ subtypes.

³Deng, *et al.* Non-negative Matrix Factorization on Manifold. in *8th IEEE Int. Conf. Data Mining* (IEEE, 2008).



$[n \times v]$

\approx



$[n \times k]$



$[k \times v]$

Matrix Factorization

Consensus Clustering

- Authors performed network-regularized NMF 1,000 times on subsamples⁴ of the dataset, then turned the resultant clustering outcomes into a co-clustering matrix of patients.
- The co-clustering matrix records the frequency that each patient pair shared membership in the same subtype over across all clustering iterations.
- Consensus clusters were then derived from this matrix through techniques such as average linkage hierarchical clustering.

⁴For each subsample, take 80% of the patients and 80% of the mutated genes at random without replacement

RESULTS & VALIDATION

Performance on Tumor Mutation Data

- The authors used NBS to stratify patients profiled by TCGA (Cancer Genome Atlas) full-exome sequencing for uterine, ovarian, and lung cancers.
- In each of the three cancers, NBS resulted in robust subtype structure; standard consensus clustering (not based on network structure) was unable to stratify the patient cohort.
- Results held for all three gene networks used.⁵

⁵STRING, HumanNet, PathwayCommons.

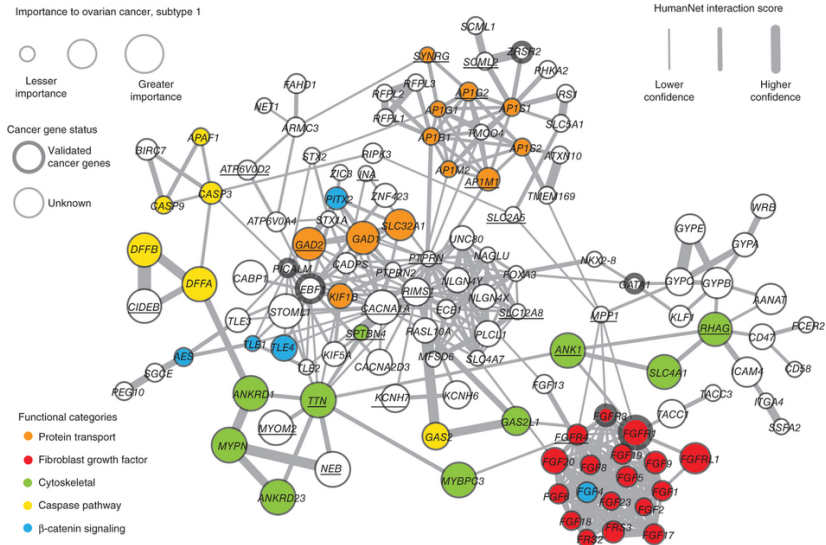
NBS-Identified Subtype & Observed Clinical Data

- **Uterine cancer:** identified subtypes closely associated with the recorded subtype on a histological basis (microscopic analysis)
- **Ovarian & Lung cancers:** identified subtypes were significant predictors of patient survival time, independently of clinical covariates (e.g., tumor stage, age, mutation rate, etc.)
- Permuting the mapping between mutated genes and the network yielded poorer performance of network-based stratification.
- Predictions based on NBS-derived subtypes were competitive with those derived from other TCGA data types, e.g., copy-number variation, methylation, mRNA expression, microRNA expression, protein profiles.

Network Region Identification by Subtype

The authors also sought to identify regions of the gene network most responsible for discriminating the somatic mutation profiles of tumors of different subtypes, focusing on ovarian cancer.

- For each subtype, the authors identified the genes for which the network-smoothed mutation state differed significantly for patients of that subtype versus the others.
- This partitioning of genes was projected onto the HumanNet network.



Conclusion

- Incorporating network knowledge improves tumor subtype stratification techniques.
- Performance of technique was measured against subtype identifiability and clinical outcome predictability for uterine, ovarian, and lung cancers.
- Identifiability depended on network structure; permuting the mapping between mutations and the network resulted in poorer performance.