Paper Presentation:

Hofree, *et al.*

Network-based stratification of tumor mutations.

Nature Methods. (2013).

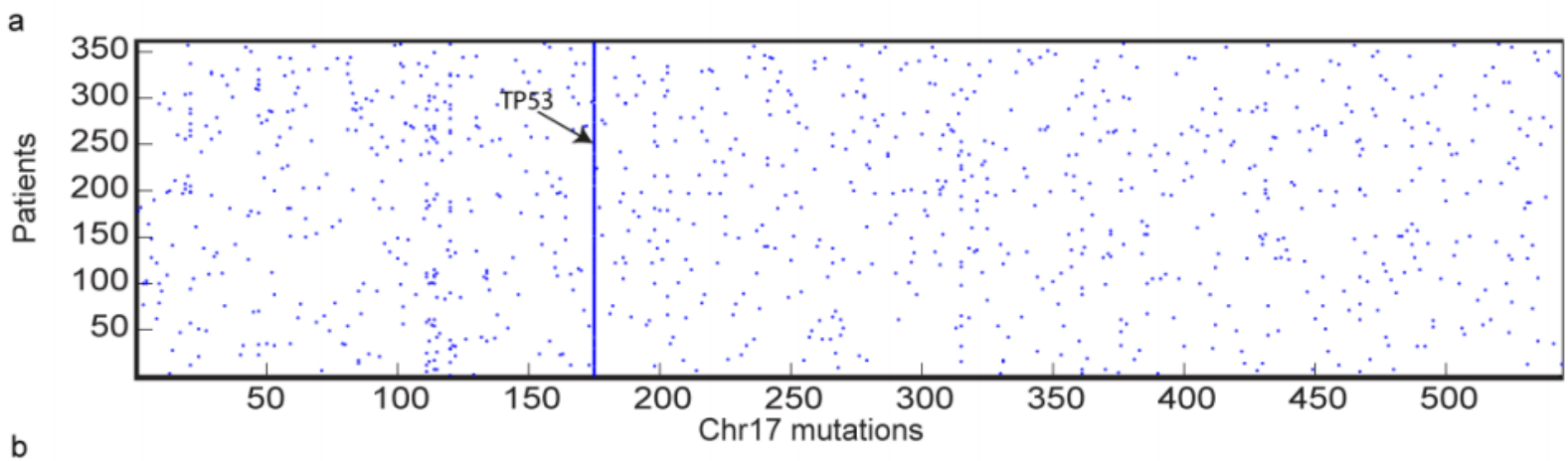
Addison Hu
CB&B 555
8 December 2016

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

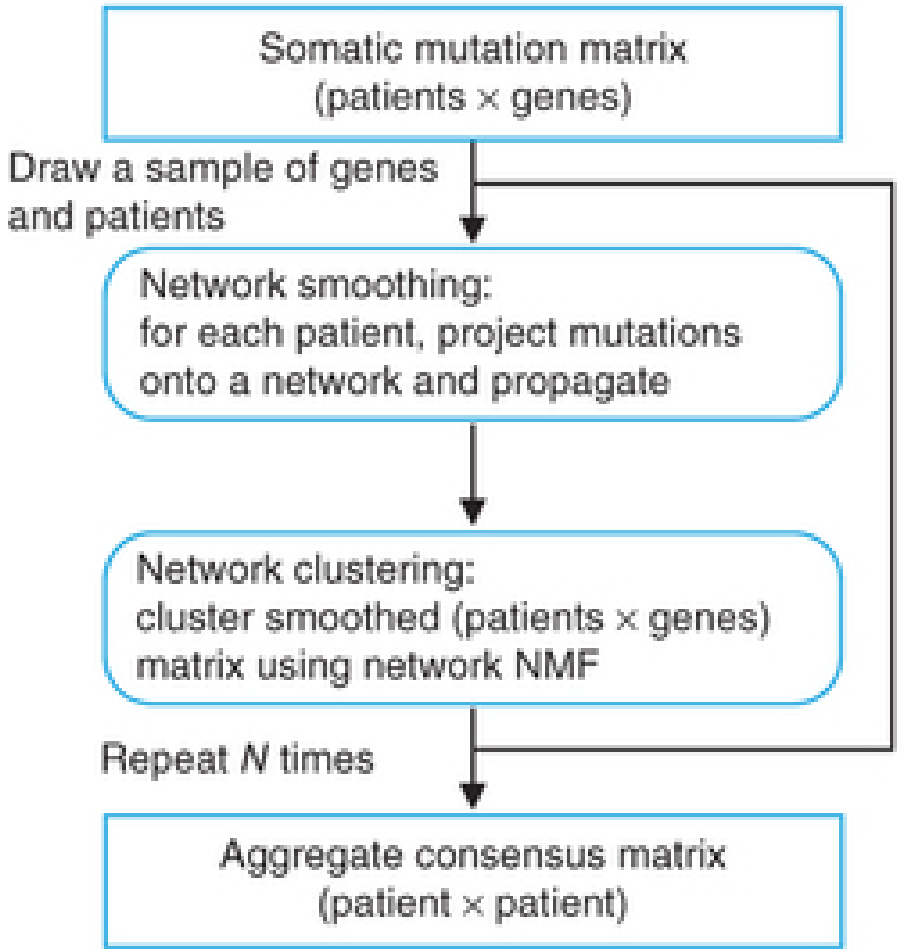
- Data obtained from the Cancer Genome Atlas
- Techniques demonstrated with ovarian, uterine, and lung cancer cohorts
- Somatic mutations for each patient are represented as a series of binary states $\{0, 1\}$ on genes; a 1 denotes that a mutation has occurred for a gene relative to the germ line.
- Each patient's mutation profile is projected onto a human gene interaction network obtained from public databases (e.g., STRING[1], Pathway Commons[2]).

---

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

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

Network-based Stratification: Overview



Network smoothing

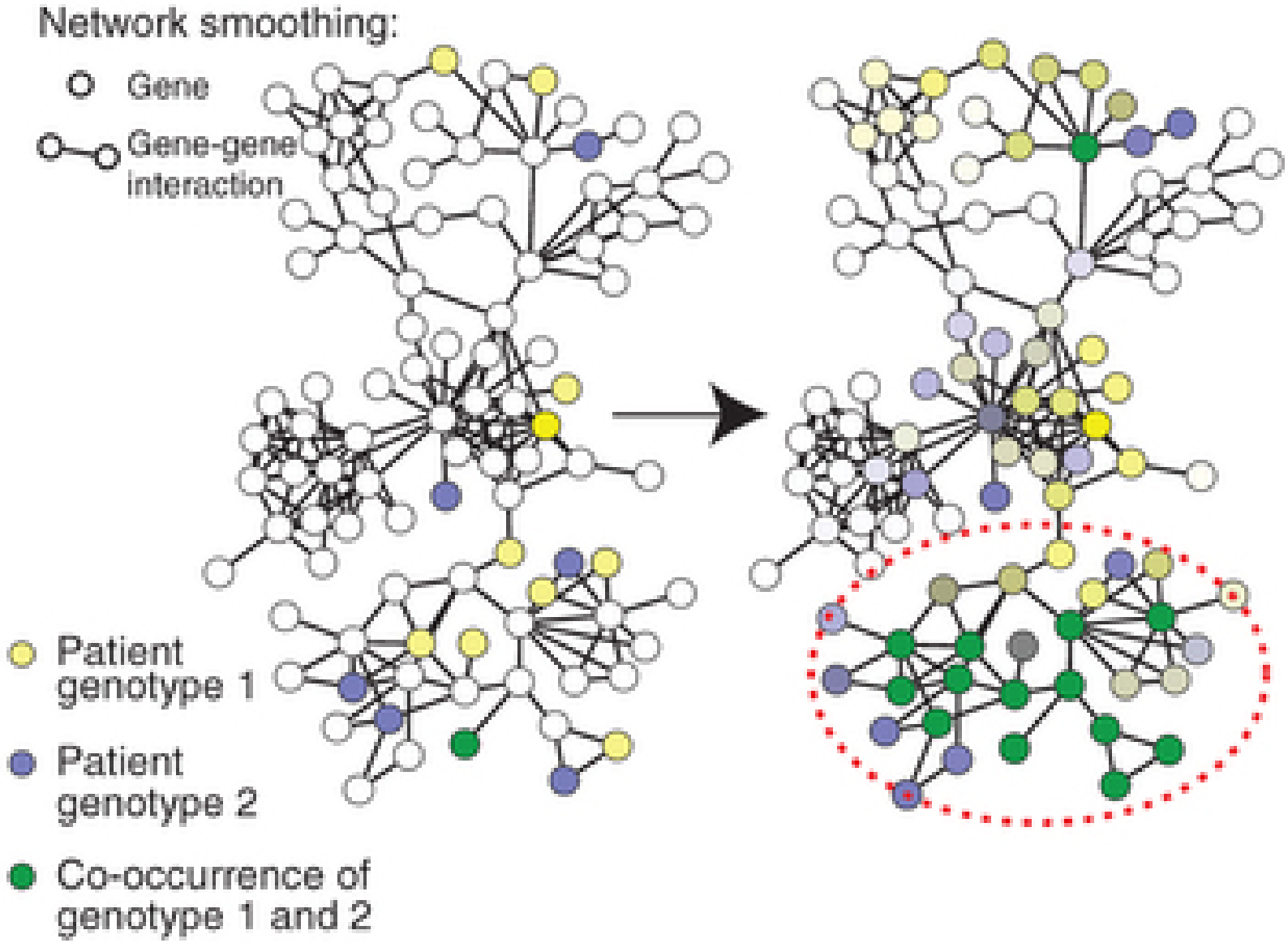
After mapping the patient mutation profile onto a molecular network, the state of each gene is in $\{0, 1\}$.

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

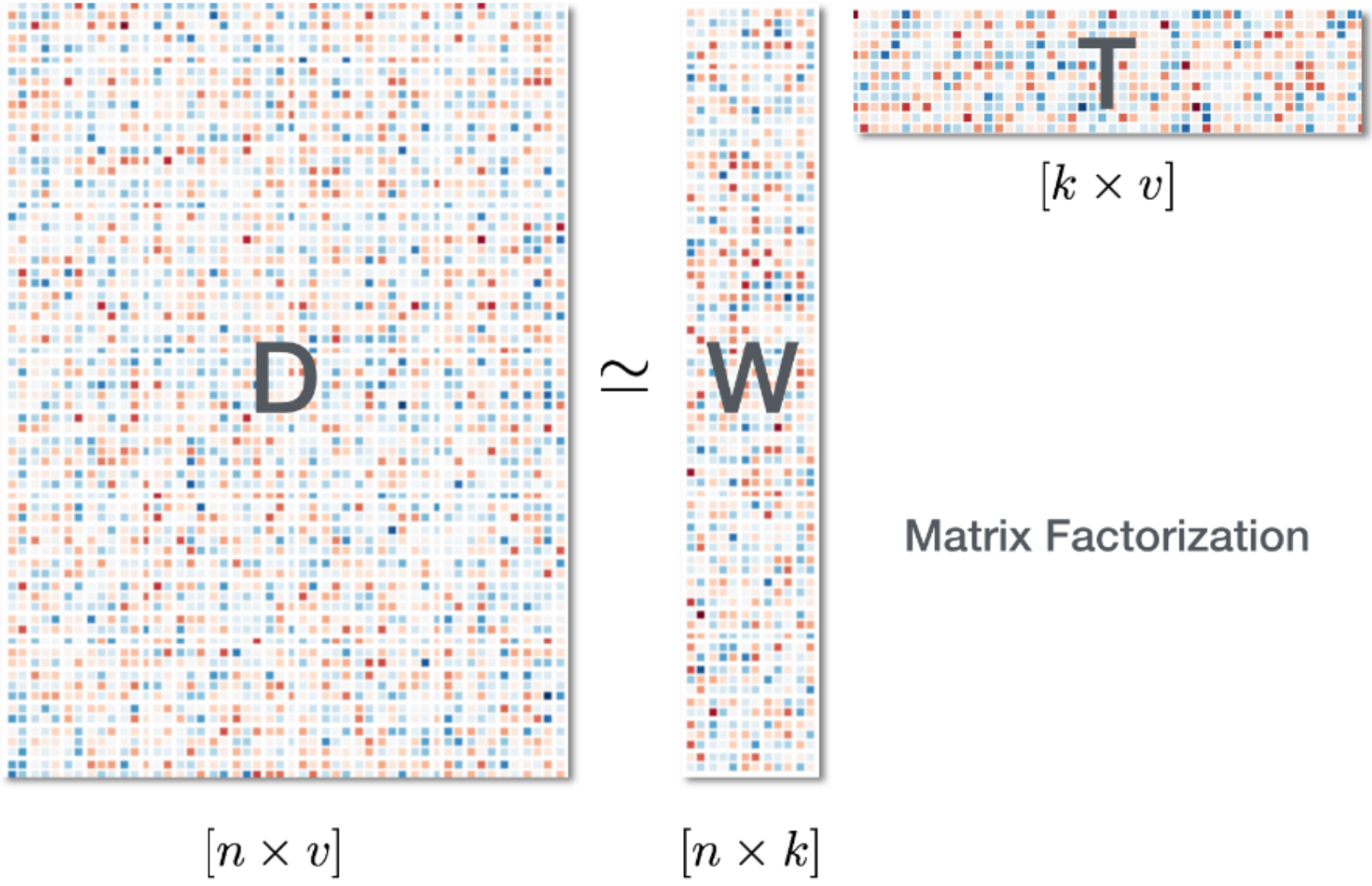
Non-negative matrix factorization is an unsupervised technique in which a data matrix $F \in \mathbf{R}_+^{n \times d}$ is decomposed into matrices $W \in \mathbf{R}_+^{n \times k}, H \in \mathbf{R}_+^{k \times d}$

$$\begin{array}{ll} \underset{W,H}{\text{minimize}} & \|F - WH\|_F^2 + \operatorname{tr}(W^T K W) \\ \text{subject to} & W \in \mathbf{R}_+^{n \times k}, H \in \mathbf{R}_+^{k \times d} \end{array}$$

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.

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


D
≃
W
T
$[k \times v]$
$[n \times v]$
$[n \times k]$
Matrix Factorization

Consensus Clustering

- Authors performed network-regularized NMF 1,000 times on subsamples[4] 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.

[4]For 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]

[5] STRING, HumanNet, PathwayCommons.

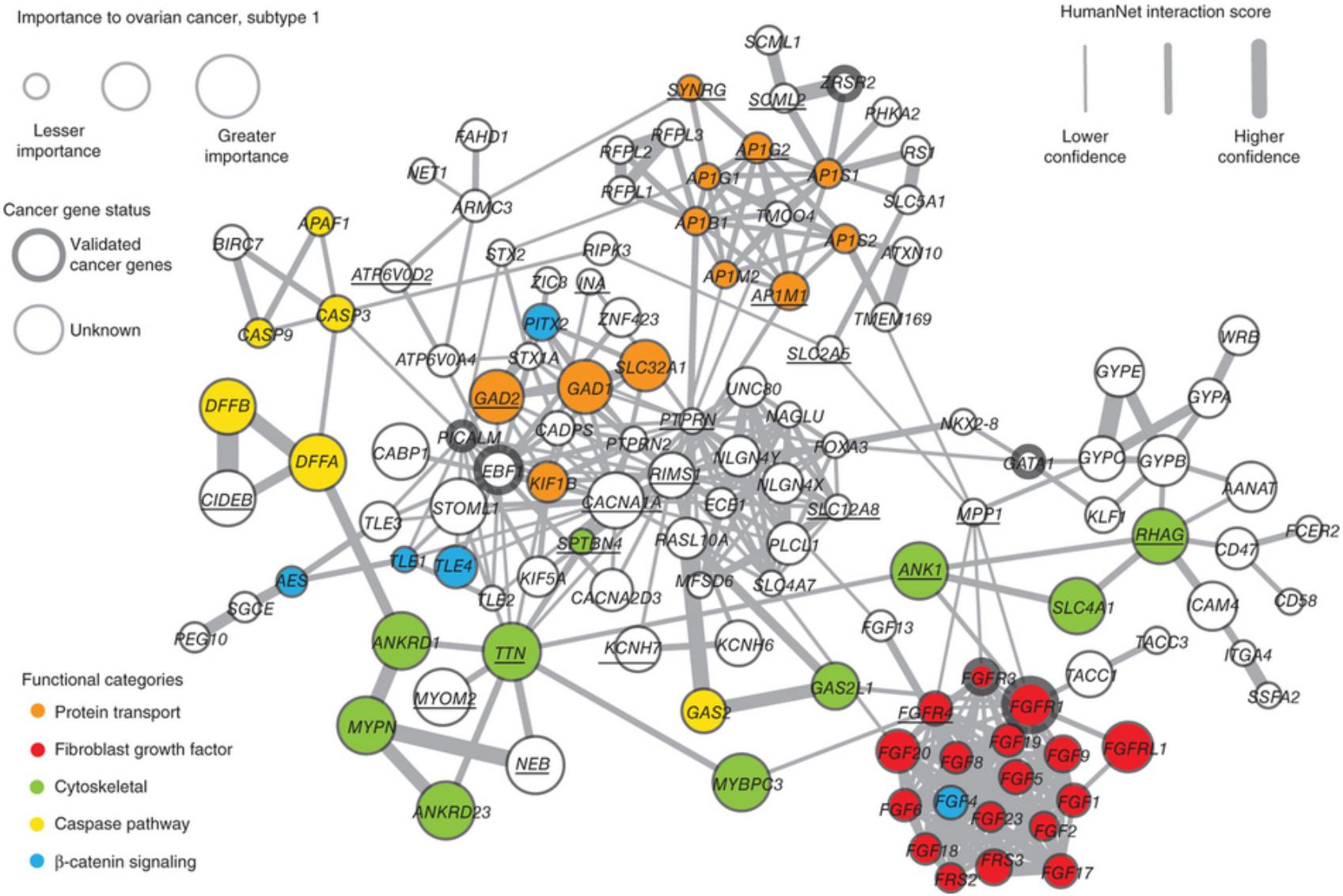
NBS-Identified Subtype & Observed Clinical Data

- **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.


Importance to ovarian cancer, subtype 1
Lesser importance
Greater importance
Cancer gene status
Validated cancer genes
Unknown
HumanNet interaction score
Lower confidence
Higher confidence
Functional categories
Protein transport
Fibroblast growth factor
Cytoskeletal
Caspase pathway
β-catenin signaling
SCML1
SYNRG
SCML2
ZRSR2
PHKA2
FAHD1
RFPL3
RFPL2
AP1G2
RSI
NET1
AP1G1
AP1S1
RFPL1
ARMC3
SLC3A1
AP1B1
TMOO4
APAF1
AP1S2
ATXN10
BIRC7
STX2
RIPK3
ATR6V0D2
ZIC3
INA
AP1M2
AP1M1
CASP3
PITX2
ZNF423
TMEM169
CASP9
SLC2A5
ATP6V0A4
STX1A
SLC32A1
GAD1
GAD2
UNC80
DFFB
NAGLU
PTPRN
PICALM
CADPS
PTPRN2
DFFA
CABP1
FOXA3
EBF1
KIF1B
RIMS1
NLGN4Y
CIDEB
NLGN4X
STOML1
CACNA1A
ECB1
SLC12A8
TLE3
SPTBN4
RASL10A
PLCL1
TLE1
TLE4
KIF5A
MFSD6
SLO4A7
AES
TLE2
CACNA2D3
SGCE
PEG10
ANKRD1
TTN
KCNH7
KCNH6
MYOM2
GAS2
GAS2L1
MYPN
NEB
MYBPC3
ANKRD23
NKX2-8
GATA1
MPP1
KLF1
ANK1
SLC4A1
GYPE
GYPA
WRB
GYPC
GYPB
AANAT
RHAG
CD47
FCER2
CAM4
CD58
TACC3
ITGA4
TACC1
SSFA2
FGF13
FGFR3
FGFR4
FGFR1
FGFRL1
FGF20
FGF8
FGF19
FGF9
FGF5
FGF6
FGF4
FGF23
FGF1
FGF2
FGF18
FRS3
FRS2
FGF17

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.