Network-based Stratification of Tumor Mutations

Dongfang Wang

December 28, 2013

Content

Content

Introduction

Cancer

1. complex : driven by a combination of genes
2. **heterogeneous:** gene combinations vary greatly between patients

Introduction

Cancer

1. complex : driven by a combination of genes
2. **heterogeneous:** gene combinations vary greatly between patients

TCGA:The Cancer Genomes Atlas

profile thousands of tumos at multiple layers of genome-scale information

1. +20 Cancer cohorts with 50-800 individuals
2. samples of different measurement types:
   - mRNA expression,Protein expression
   - Copy number variations,Single nucleoside polymorphisms,Methylation,miRNA
   - Patient genomes (somatic mutations)

Introduction

Cancer

1. complex : driven by a combination of genes
2. **heterogeneous:** gene combinations vary greatly between patients

TCGA:The Cancer Genomes Atlas

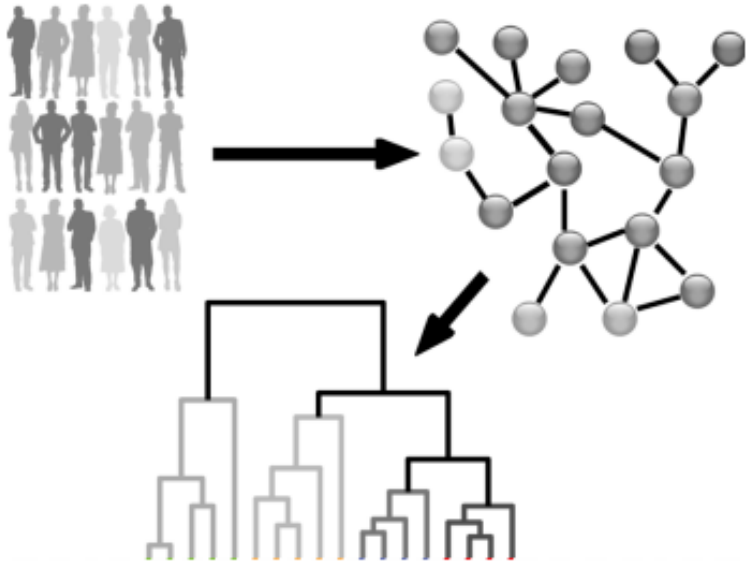
profile thousands of tumos at multiple layers of genome-scale information

1. +20 Cancer cohorts with 50-800 individuals
2. samples of different measurement types:
   - mRNA expression,Protein expression
   - Copy number variations,Single nucleoside polymorphisms,Methylation,miRNA
   - Patient genomes (somatic mutations)

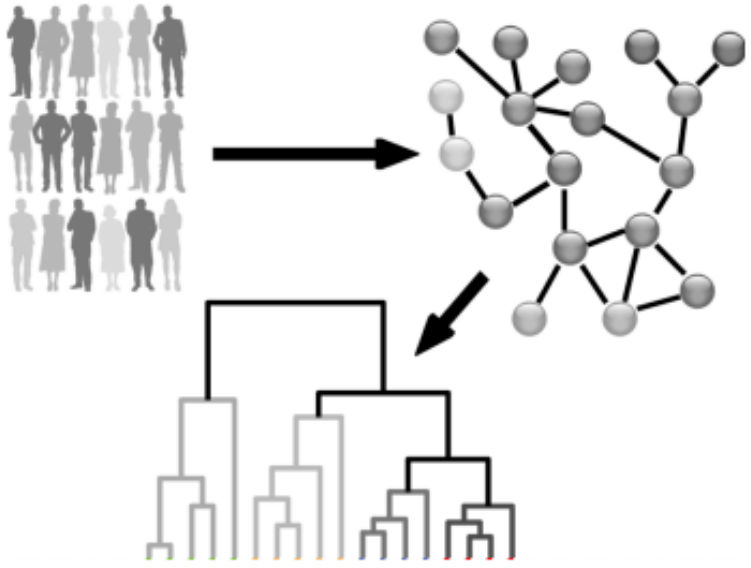
Informatic methods: integrate and interpret genome-scale information

1. molecular process driving tumor progression
2. need in the clinic (genome-scale tumor profiles are inable to derive clinicllay relevant data)

Stratification: driving cancer into subtypes



Stratification: driving cancer into subtypes



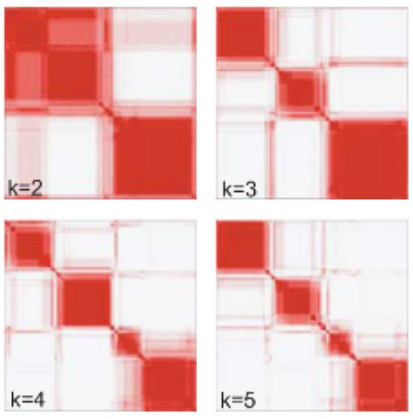
Why stratify?

subtype: clinical and biological meaning

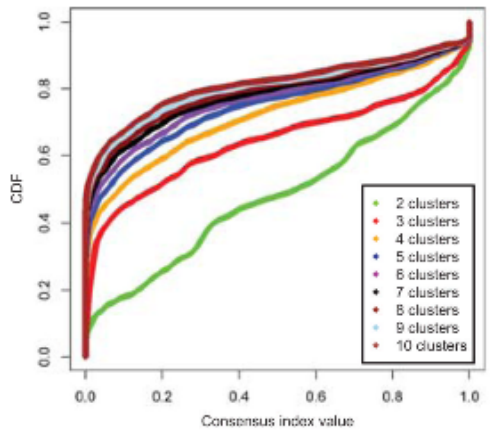
1. Better patient prognostics
2. Better undersatnding of tumor biology
3. subtype specific drug targets
4. Patient tailored treatment

Efforts to Stratify Using Gene Expression

A



B

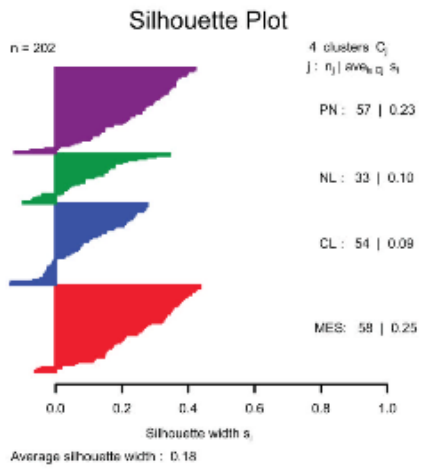


C

SigClust: All Pairwise

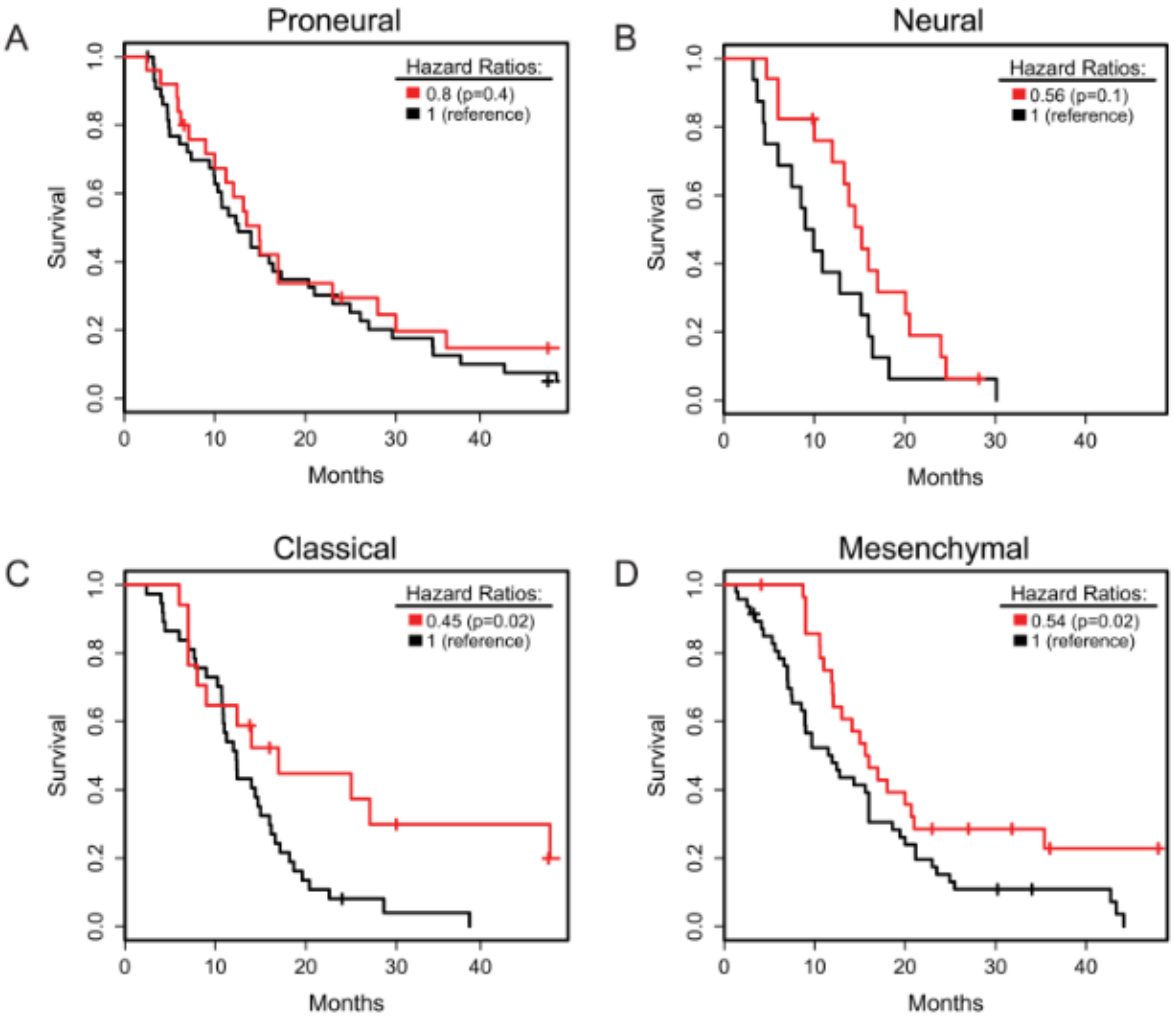
Cluster	PN	NL	CL	MES
PN				
NL	3.90E-03			
CL	1.21E-10	3.70E-04		
MES	5.93E-17	2.69E-07	5.17E-06	

D



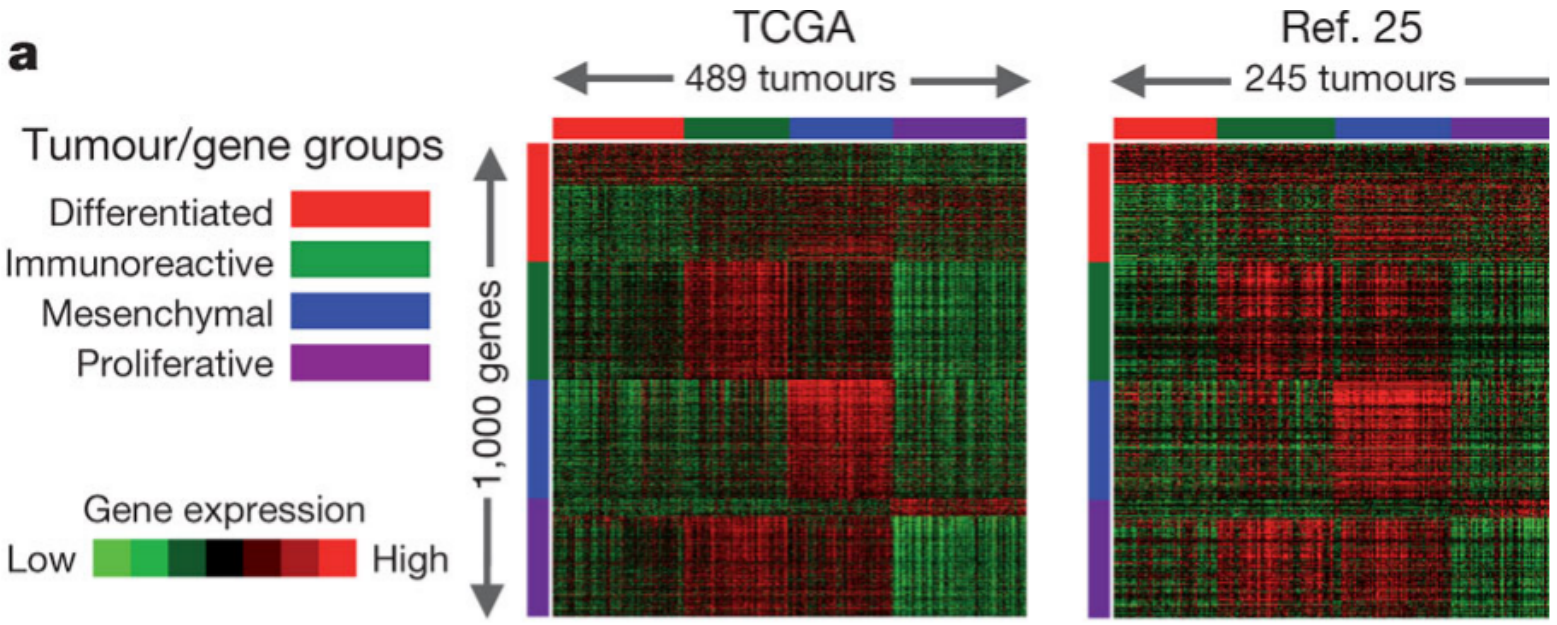
factor analysis
(202 samples&1740 genes)
consensus clustering

Verhaak, R.G.et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17, 98-110 (2010)



greatest benefit in Classical and no benefit in Proneural

Problems



No association to a clinical phenotype was reported (for these subgroups).
T.C.G.A.R.N. (TCGA), Integrated genomic analyses of ovarian carcinoma. Nature 474, 609-15 (2011).

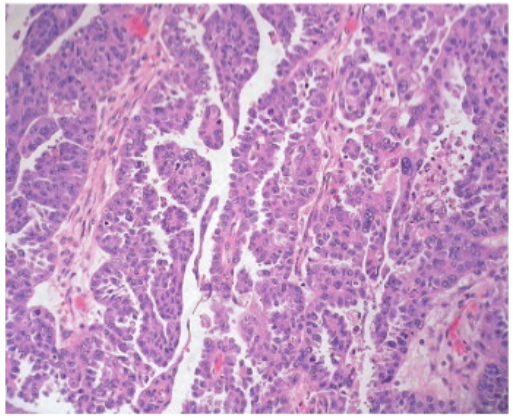
Problems

1. RNA sample quality
2. lack of reproducibility
3. ample oppurtunities of overfitting of data

Somatic Mutations

Somatic mutations:

- high-throughout sequencing
- compare the genome or exome of a patient's tumor to that of the germ line to identify mutations that have become enriched in the tumor cell population
- cantain causal drivers of tumor progression



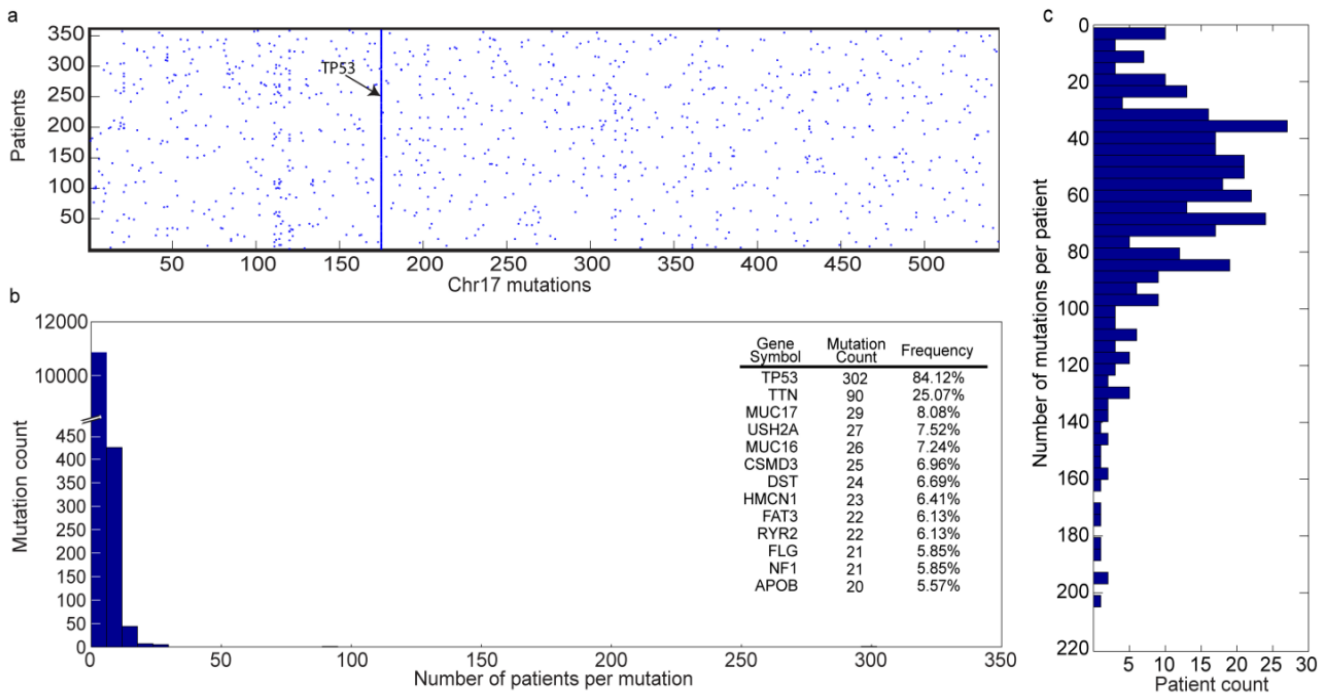
The Cancer Genome Atlas

Somatic mutations in high grade serous ovarian cancer

(359 matched patient/tumor exome sequenced with Illumina GAIIx;11,231 somatic mutations)

Challenges

1. extremely sparse
2. heterogeneous, clinically identical patients share no more than a single mutation



Gene Symbol	Mutation Count	Frequency
TP53	302	84.12%
TTN	90	25.07%
MUC17	29	8.08%
USH2A	27	7.52%
MUC16	26	7.24%
CSMD3	25	6.96%
DST	24	6.69%
HMCN1	23	6.41%
FAT3	22	6.13%
RYR2	22	6.13%
FLG	21	5.85%
NF1	21	5.85%
APOB	20	5.57%

Hypothesis

1. Cancer is a disease not of individual mutatios, nor of genes, but of combinations of genes acting in molecular networks corresponding to hallmark process
2. although two tumors may not have any mutations in common, they may **share the networks affected by these mutations**

NATURE METHODS | ARTICLE OPEN

Network-based stratification of tumor mutations

Matan Hofree, John P Shen, Hannah Carter, Andrew Gross & Trey Ideker

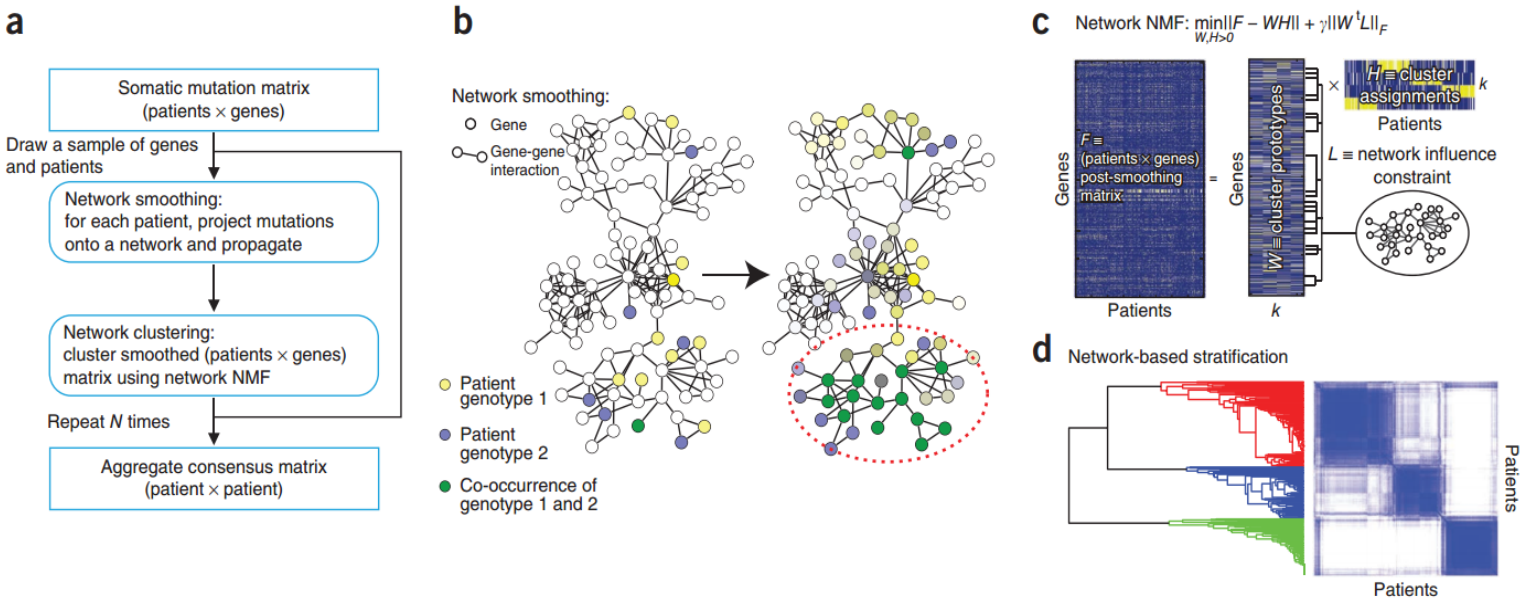
Affiliations | Contributions | Corresponding author

Nature Methods **10**, 1108–1115 (2013) | doi:10.1038/nmeth.2651
Received 14 February 2013 | Accepted 12 August 2013 | Published online 15 September 2013

Content

Overview of NBS Methods

1. Somatic mutation matrix
2. Network smoothing
3. Netwrok cluster
4. Consensus clusering



Data Source

Patient Mutation Profile

- TCGA data portal
- ovarian cancer, uterine endometrial cancinoma, lung adenocarcinoma somatic mutation data
- binary matrix; (0,1) states on genes

Molecular Network Data

Gene interaction networks:

- STRING v.9
- HumanNet v.1
- PathwayCommons

Network Smoothing

1. Mapping a patient mutation profile onto a molecular network
2. "Network propagation": smooth the mutation signal across the network

Network Smoothing

1. Mapping a patient mutation profile onto a molecular network
2. "Network propagation": smooth the mutation signal across the network

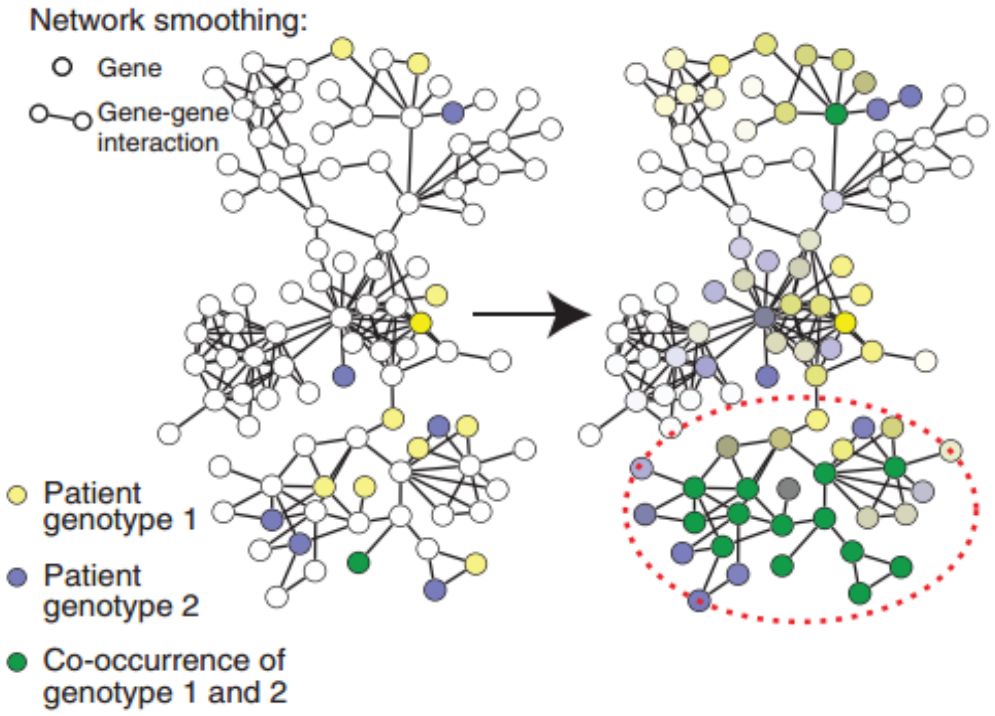
Random walk with restarts on a network

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

- F_0 : a patient-by-gene matrix
- A: degree-normalized adjacency matrix of the gene interaction network
- α: tuning parameter governing the distance that a mutation signal is allowed to diffuse through the network during propagation

1. α : only a minor effect on the results of the NBS over a sizeable range (0.5-0.8)
2. rows are **quantile normalized** to ensure that the smoothed mutation profile for each patient follows the same distribution

An Intuition for Network Smoothing

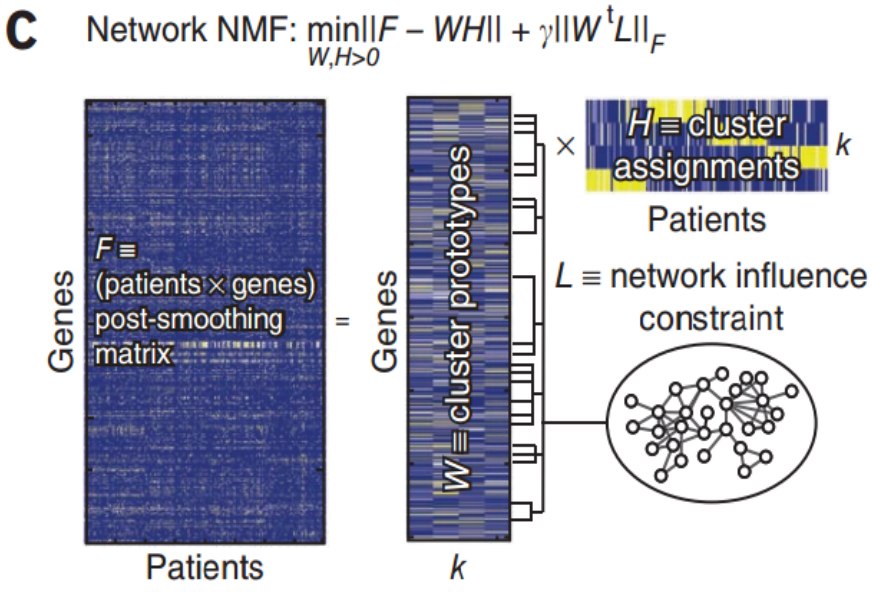


Genotype A: [...,0,0.3,1,0.4,0.4,1,0.2,0,0,0,0.3,0,0,0,0.4,0,1,0.1,0,...]
Genotype B: [...,0,0.2,1,0.2,0.4,0.2,0,0,0.4,1,0,0,0,0,0.2,0.4,0.5,0,...]

Network Clustering

Network-regularized NMF

$$\min_{W,H>0} \quad ||F - WH||^2 + \gamma trace(W^T KW)$$

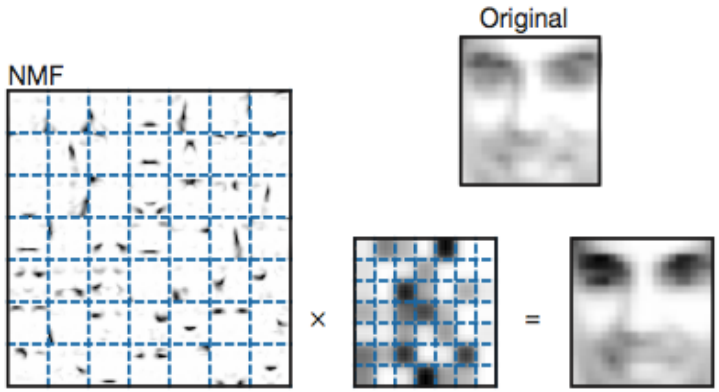


1. NMF: learn the parts of the object
2. Network-regularized: respect the structure of an underlying gene interaction network

Non-negative matrix factorization

$$\mathbf{V} \approx \mathbf{WH}$$

1. $\mathbf{V}$ is an $n \times m$ matrix, and $\mathbf{M} \in R^{n \times r}$, $\mathbf{H} \in R^{r \times m}$.



2. all elements of the three matrices are **non-negative**
3. $\mathbf{W}$: every column is a **basis image**
   $\mathbf{H}$: **encoding**; coefficients of the linear combination of basis images

$$\vec{V}_i = \mathbf{W}\vec{H}_i$$

Network-regularized

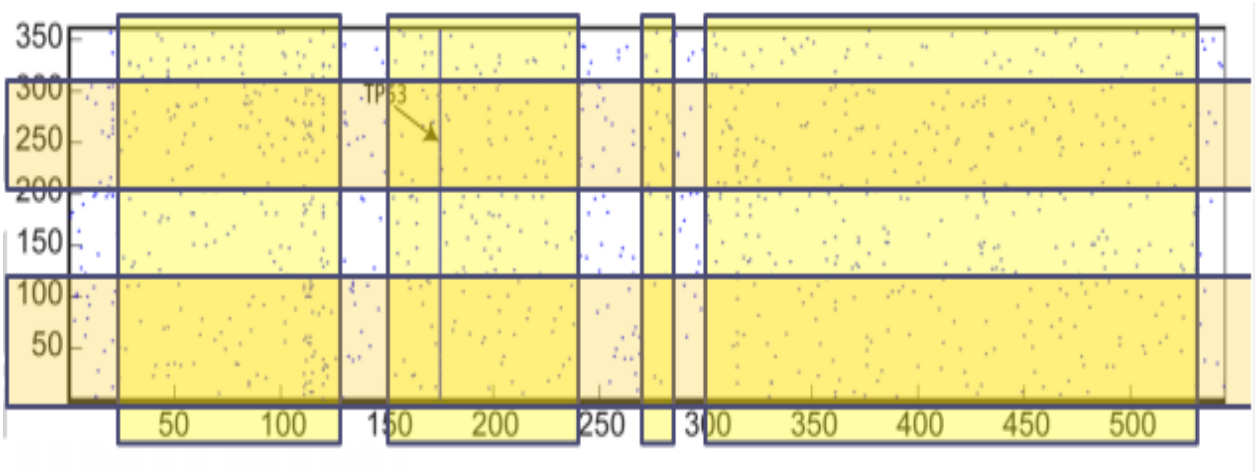
$$\min_{W,H>0} \quad ||F - WH||^2 + \gamma trace(W^T KW)$$

$$trace(W^T KW) = \sum_{ij} k_{ij} (w_i)^T (w_j)$$

constrain the basis vectors in W to respect local network neighbourhoods

Consensus Clustering

1. Draw a bootstrap sample of genes form $patients \times genes$ matrix



2. re-sampling 1000 times (80% of the patients and 80% of the mutates genes at random without replacement)
3. A similarity matrix of patients (how many times of a pair of patients appears in the same subtype)
4. stratify the patients

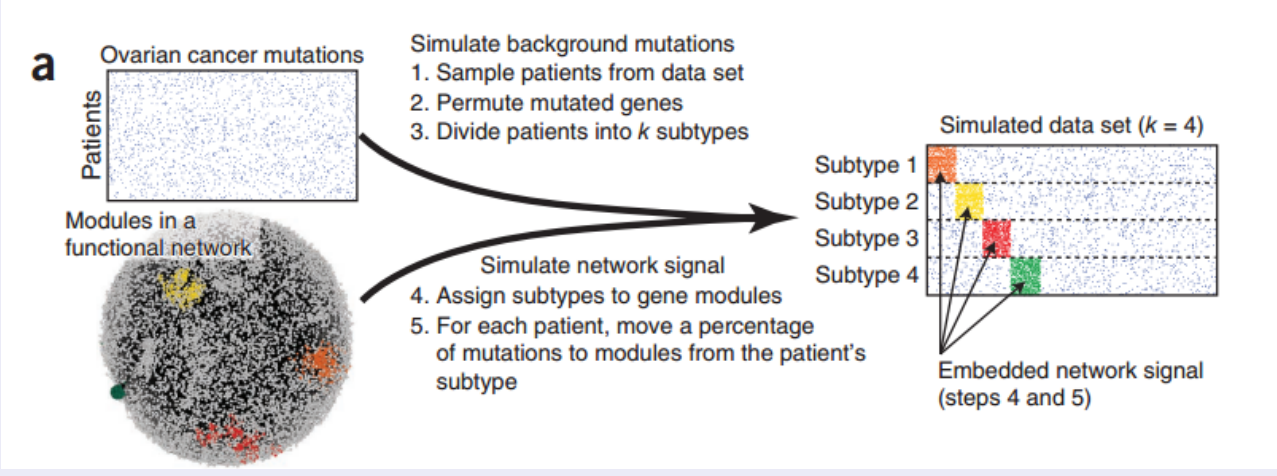
Content

Simulations

determine the ability of NBS to recover substypes from somatic mutation profiles

How to generate 'ground truth' ?

1. generate background mutation matrix (no subtype signal)

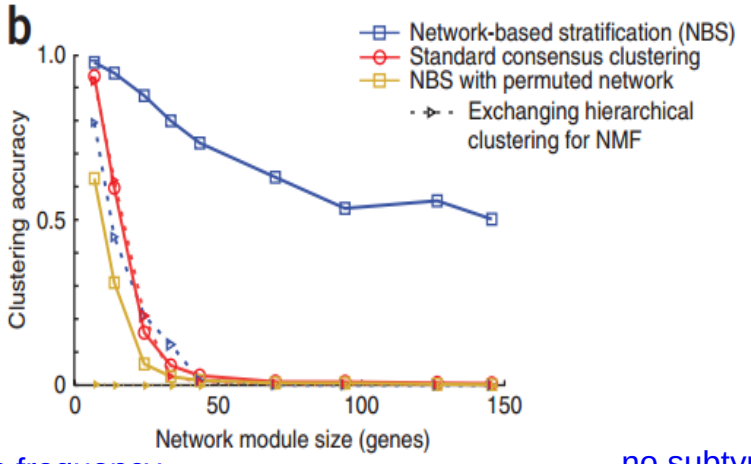


2. add network-based signal
   - establiah a set of network communities
   - assign a small number of netwrok modules to each subtype $\rightarrow$ 'driver' subnetwork
3. reassign a fraction of the patient's mutations to genes covered by the driver modules for that patient's subtype

A reasonable model of a pathway-based genetic disease

1. driven by genetic circuits corresponding to a molecular network
2. altered by mutations at multiple genes
3. many additional mutations that are noncausal 'passengers'

Simulation Results

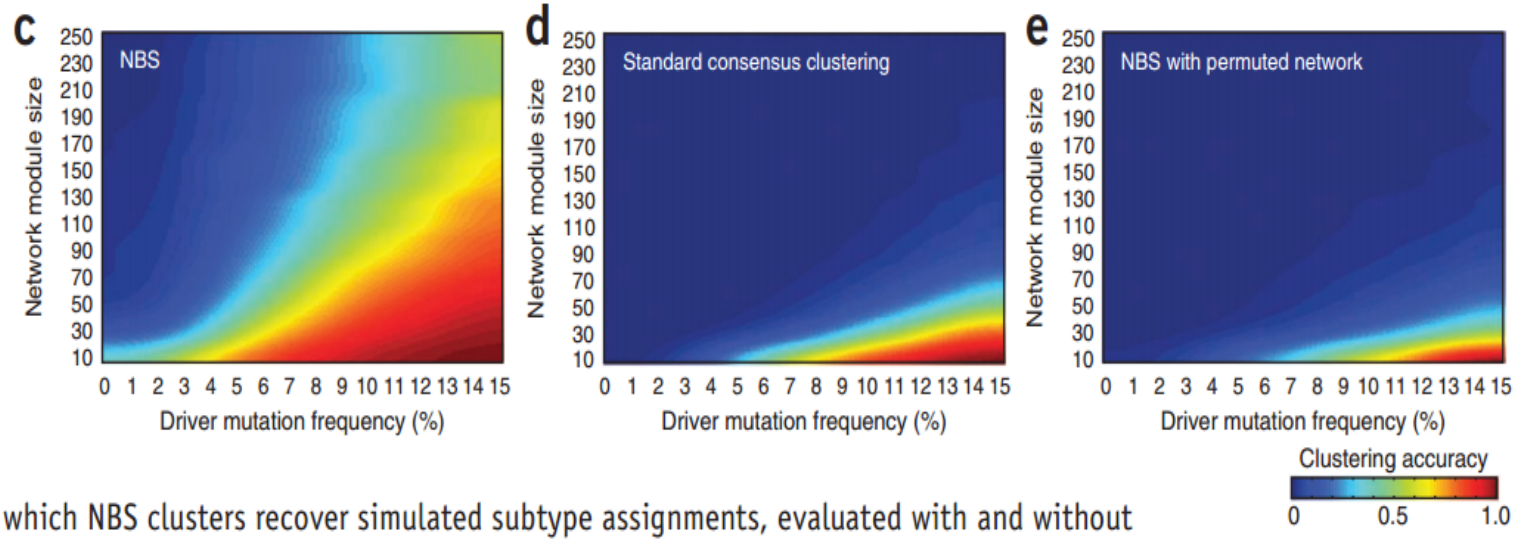


network-based

Non-negative Matrix Factorization

large module size & low mutation frequency

no subtypes (permuted network)



ı which NBS clusters recover simulated subtype assignments, evaluated with and without

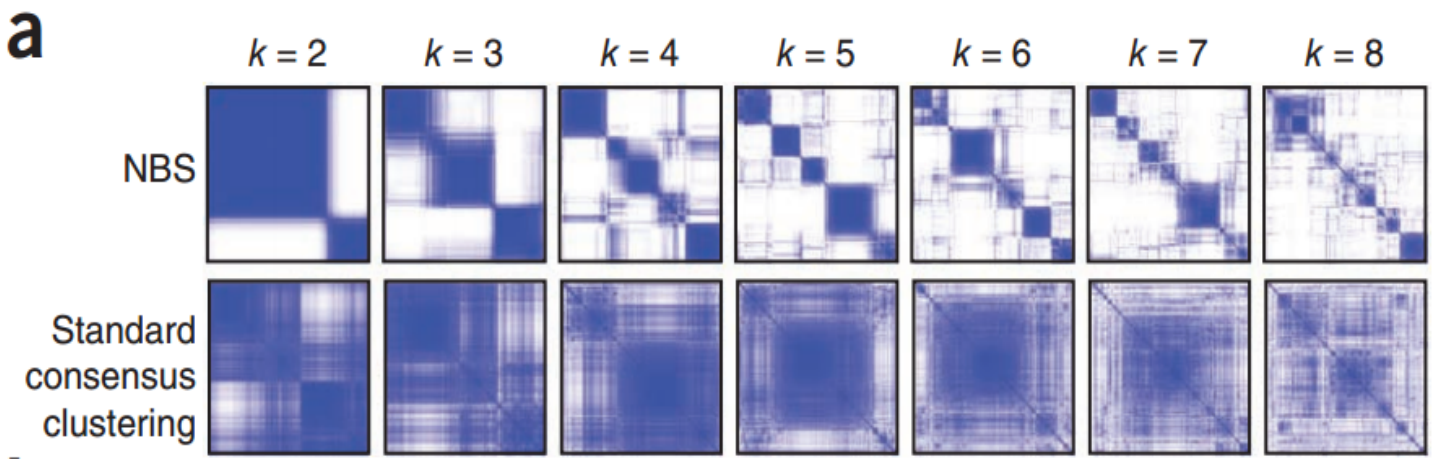
Performance Analysis

Overview

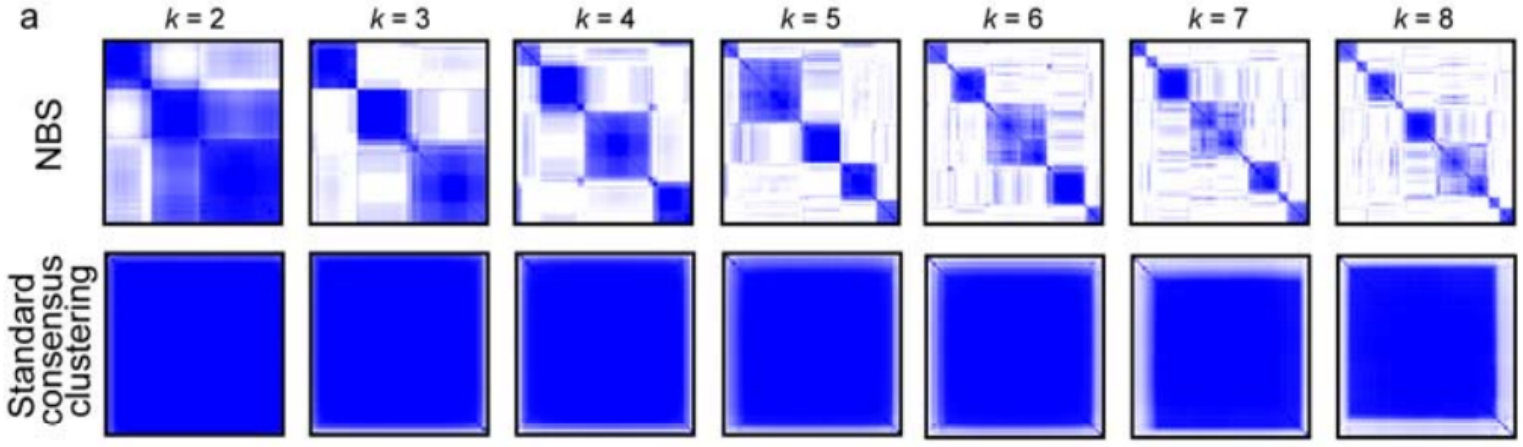
1. result of clustering
2. clinical and biological meaning
3. prediction

Diverse Cancer types

uterine ; ovarian; lung cancer (co-clustering matrix)

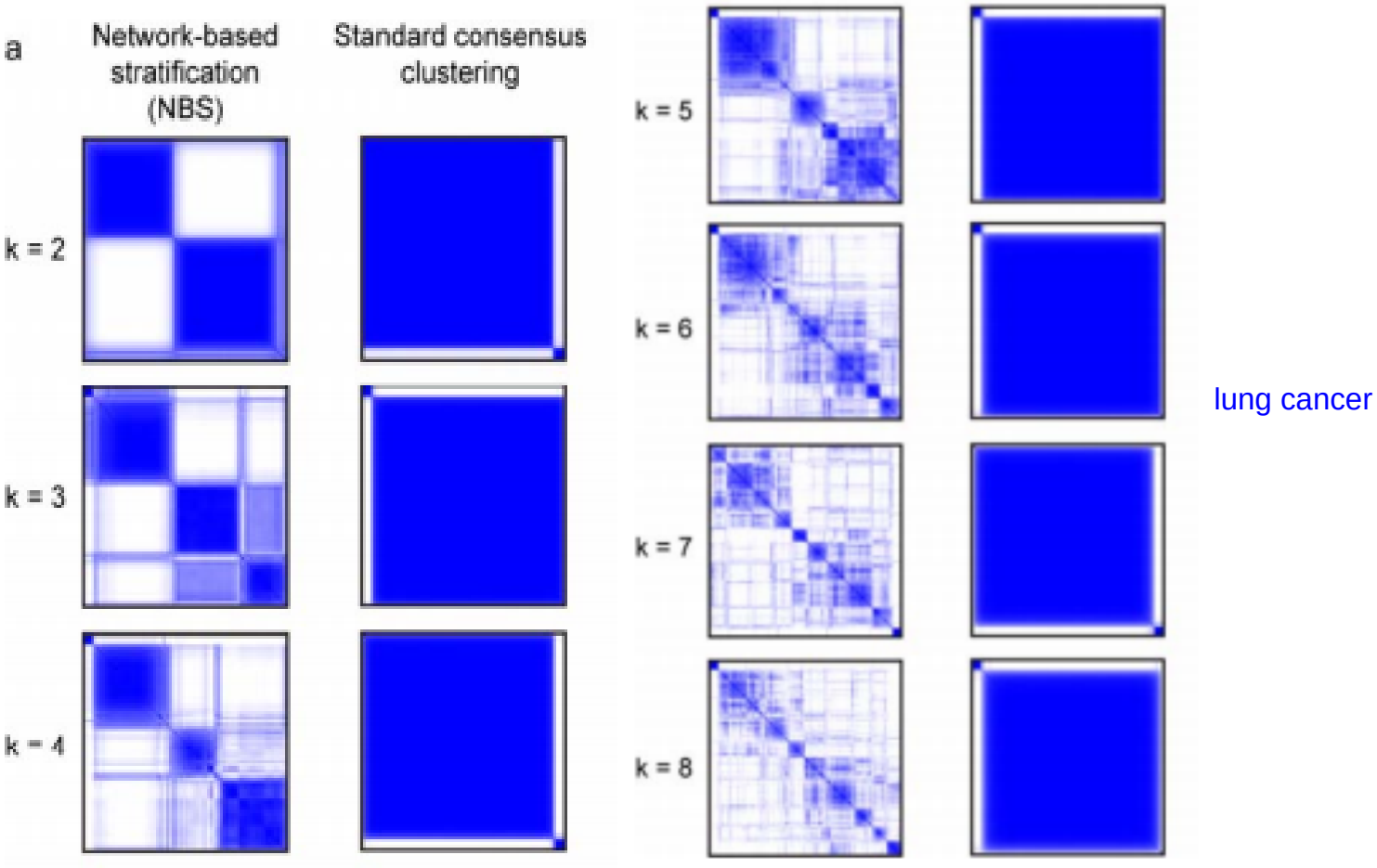


uterine



ovarian

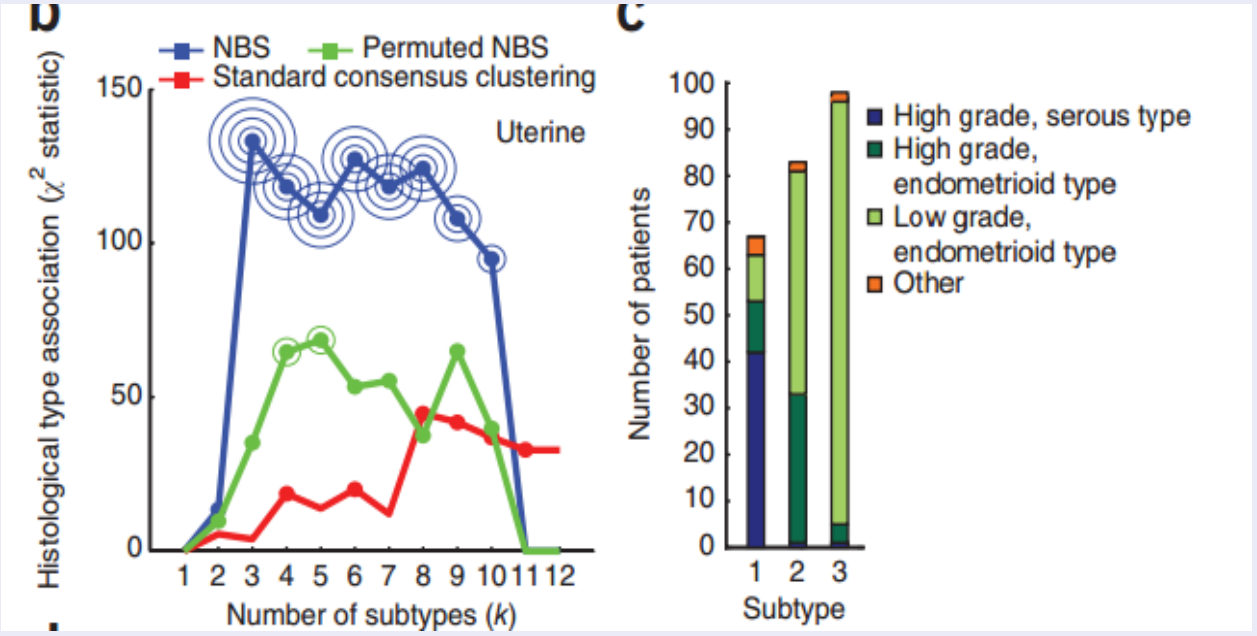
Diverse Cancer types



lung cancer

Biological Importance of Idetified Subtypes

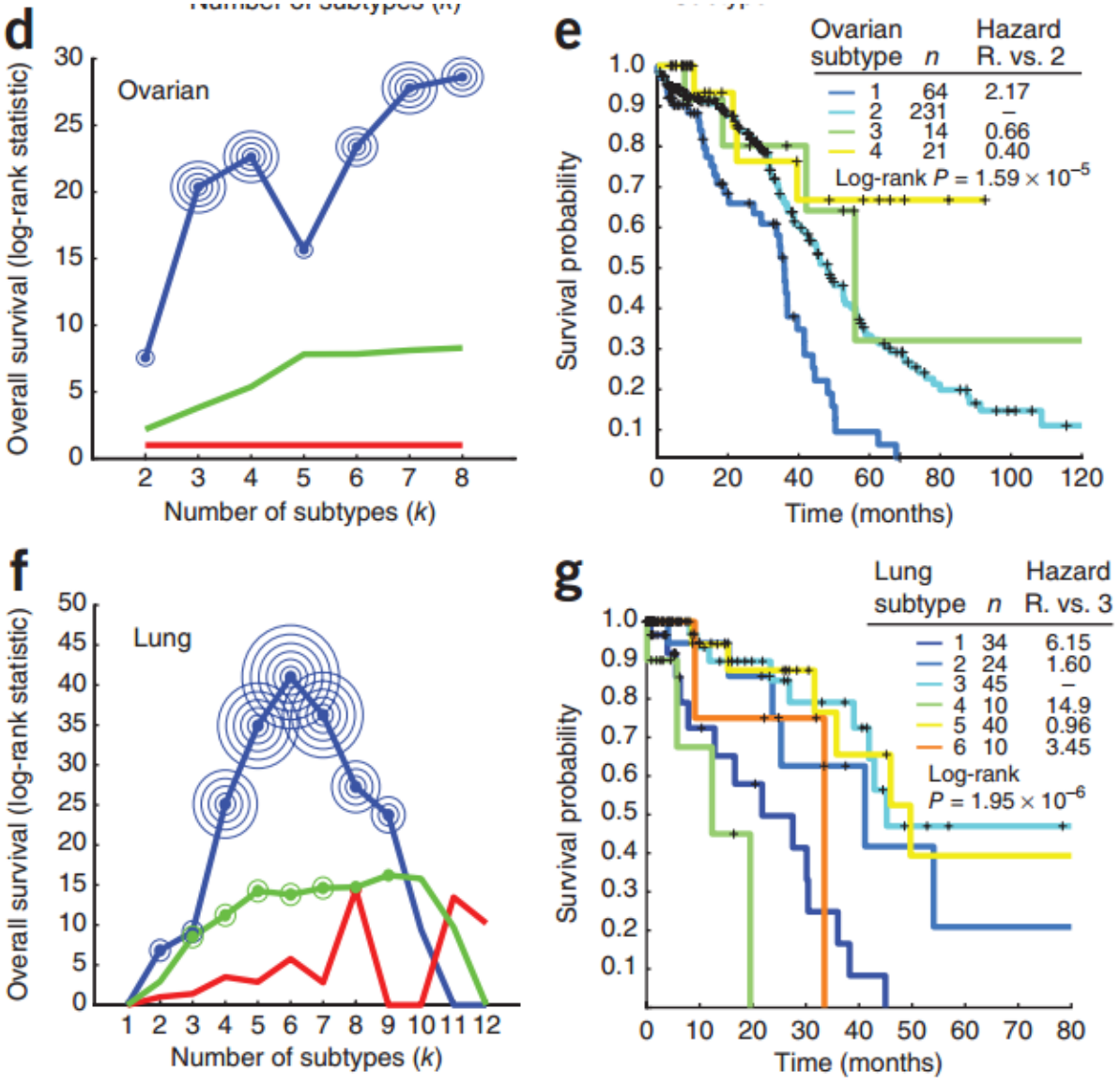
uterine: histological basis



p-value of significance (concentric circles surrounding a data point)

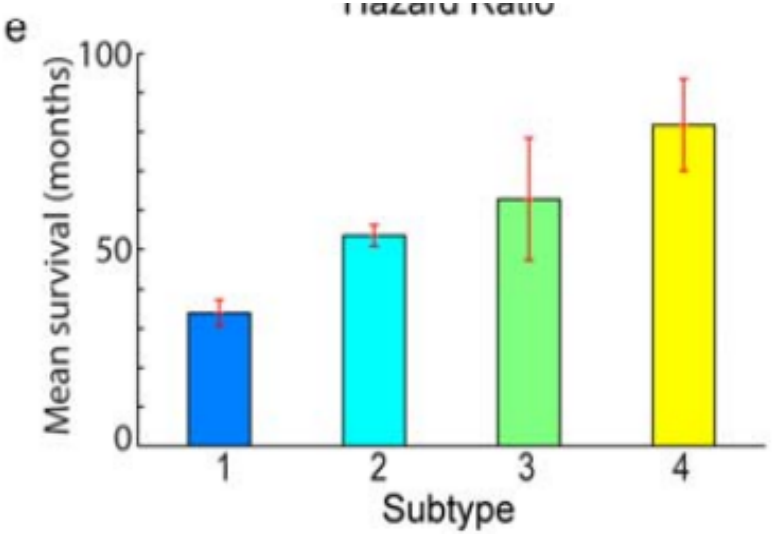
chi-square statistic: determine if a distribution of observed frequencies differs from the theoretical expected frequencies

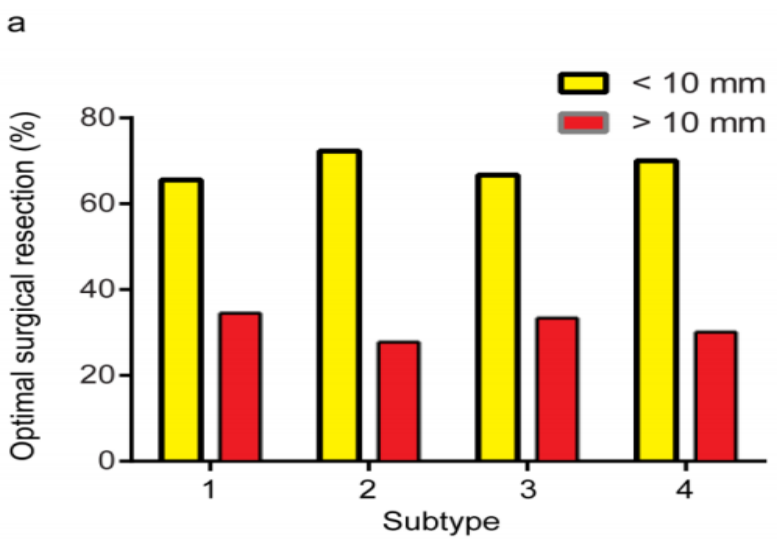
Survival Analysis

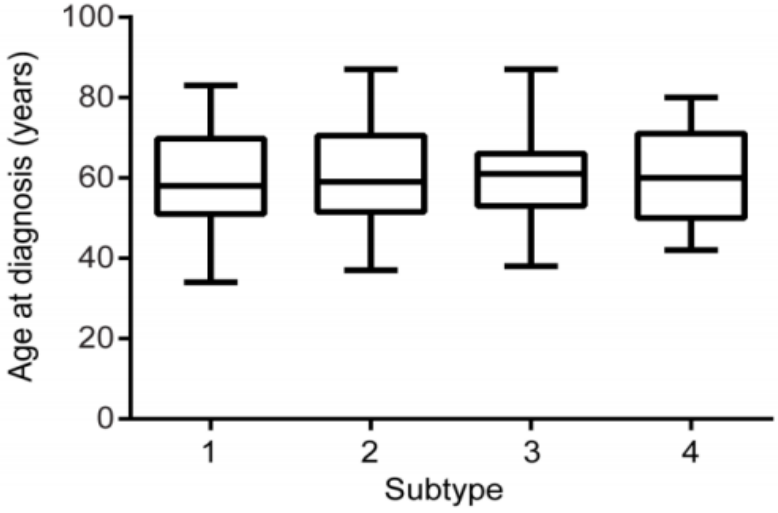


log-rank statistic: When are two KM curves statistically equivalent?

Independent of other clinical marker







Derived from other data type

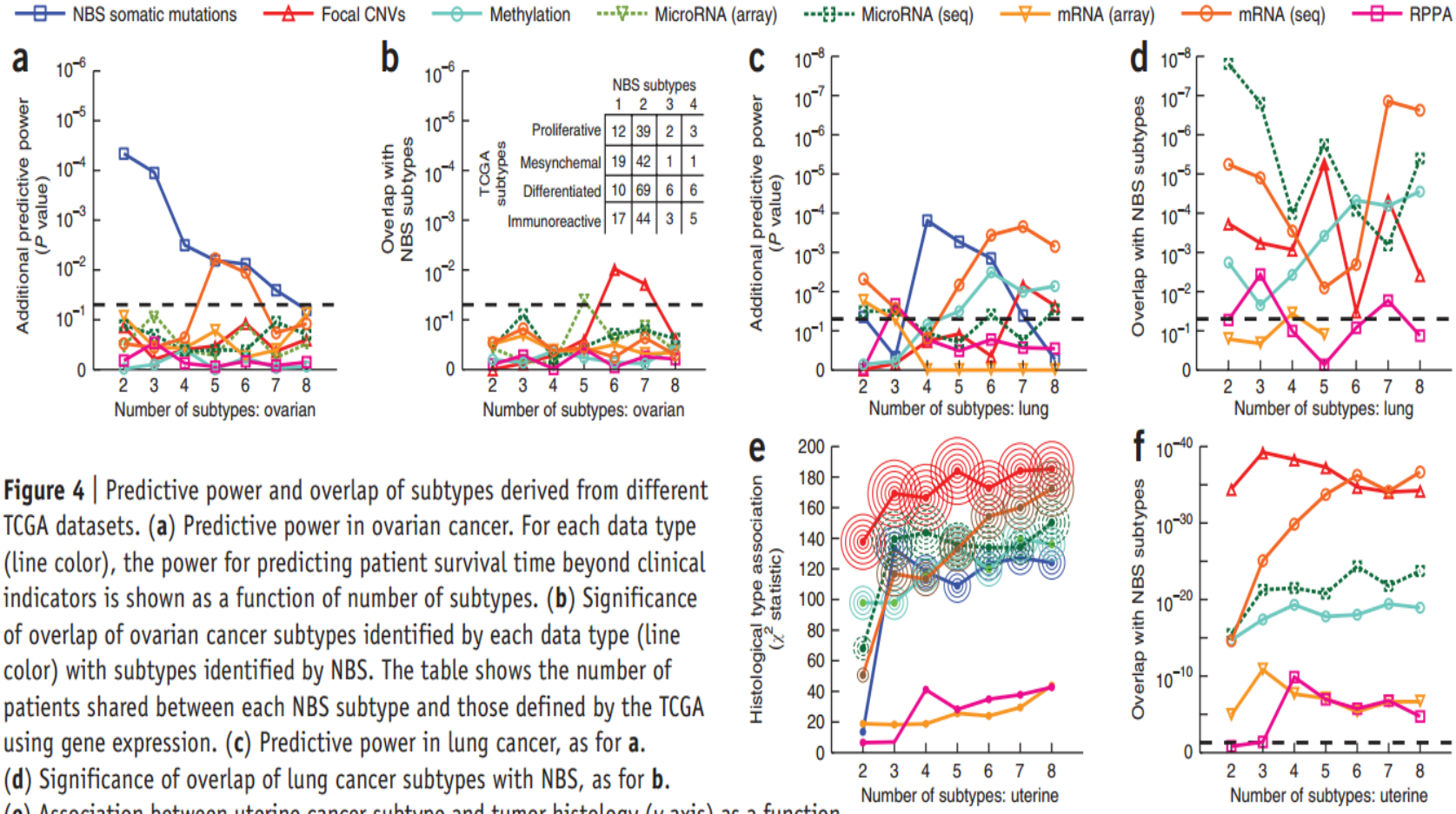
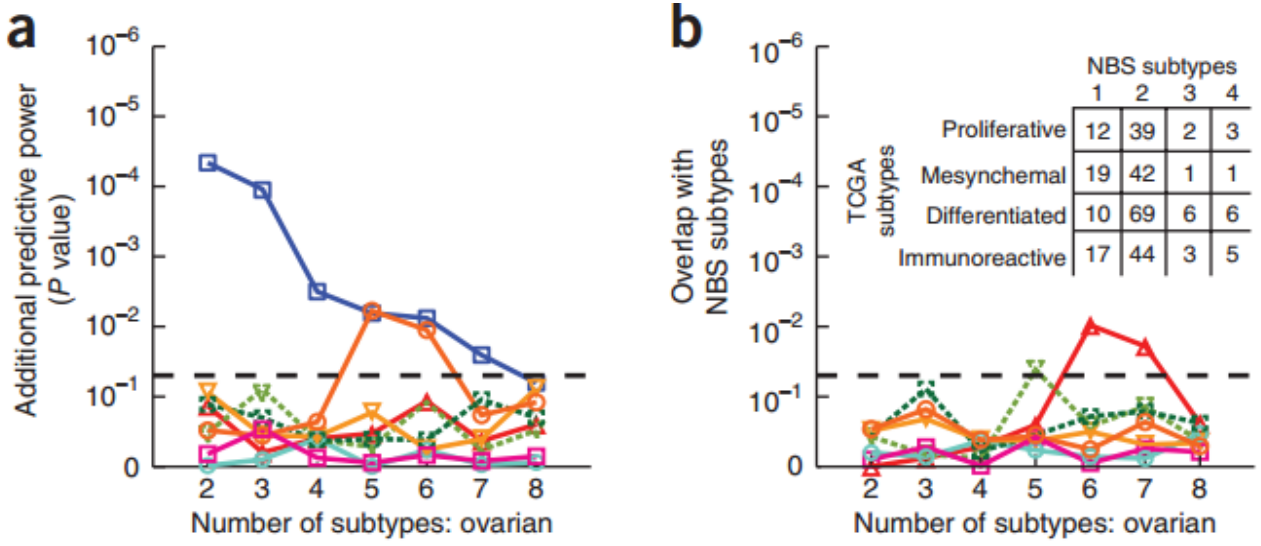


Figure 4 | Predictive power and overlap of subtypes derived from different TCGA datasets. (**a**) Predictive power in ovarian cancer. For each data type (line color), the power for predicting patient survival time beyond clinical indicators is shown as a function of number of subtypes. (**b**) Significance of overlap of ovarian cancer subtypes identified by each data type (line color) with subtypes identified by NBS. The table shows the number of patients shared between each NBS subtype and those defined by the TCGA using gene expression. (**c**) Predictive power in lung cancer, as for **a**. (**d**) Significance of overlap of lung cancer subtypes with NBS, as for **b**. (**e**) Association between uterine cancer subtype and tumor histology (*y* axis) as a function of the number of subtypes. *P* value of significance is indicated by concentric circles as in **Figure 3**. Colors are as in other panels, symbols have been omitted for clarity. (**f**) Significance of overlap of uterine cancer subtypes with NBS, as for **b**. Dashed horizontal lines indicate the $P = 0.05$ threshold of significance.

Compare with Other Data Type

CNV, methylation,mRNA expression,miRNA expression and protein profiles

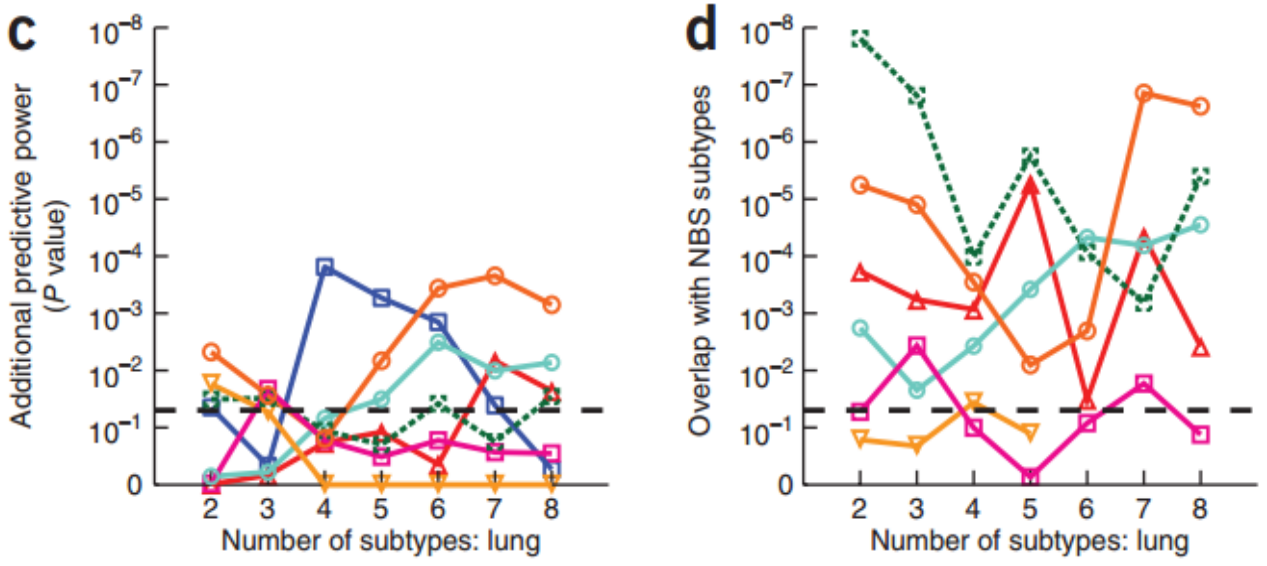
1. **ovarian cancer**: all other data types had inferior ability to predict survival beyond what would be predicted from clinical covariate and led to different subtype assignments



Compare with Other Data Type

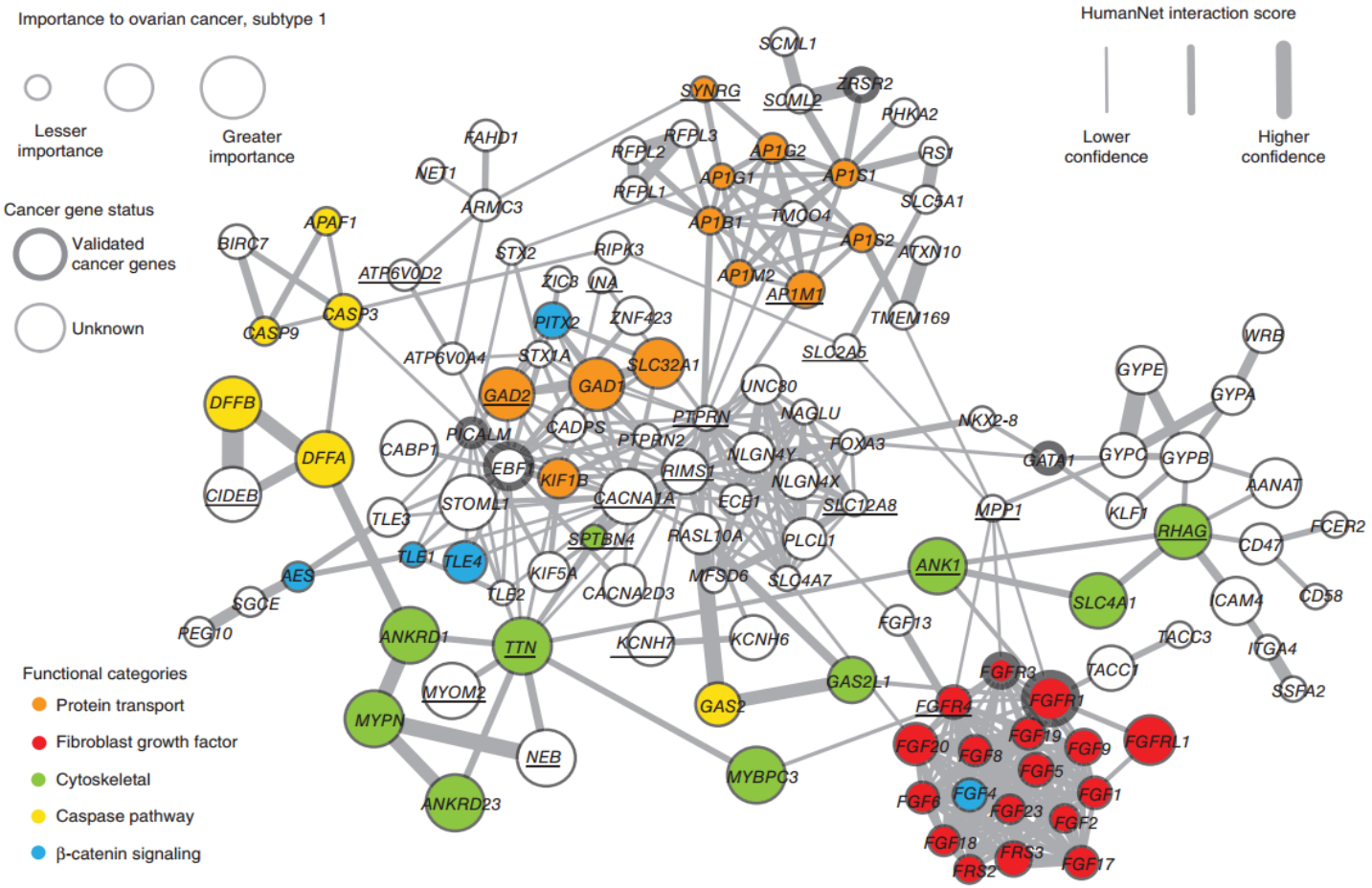
CNV, methylation,mRNA expression,miRNA expression and protein profiles

1. **lung cancer**:both NBS subtypes and those based-on RNA-seq had good predictive power and had some overlaps



Distinct Network Modules

Identify the regions of the network that are most responsible for discrimating the somatic mutation profiles of tumors of different subtypes



Wilcoxon rank-based test

Distinct Network Modules

Ovarian Subtype 1

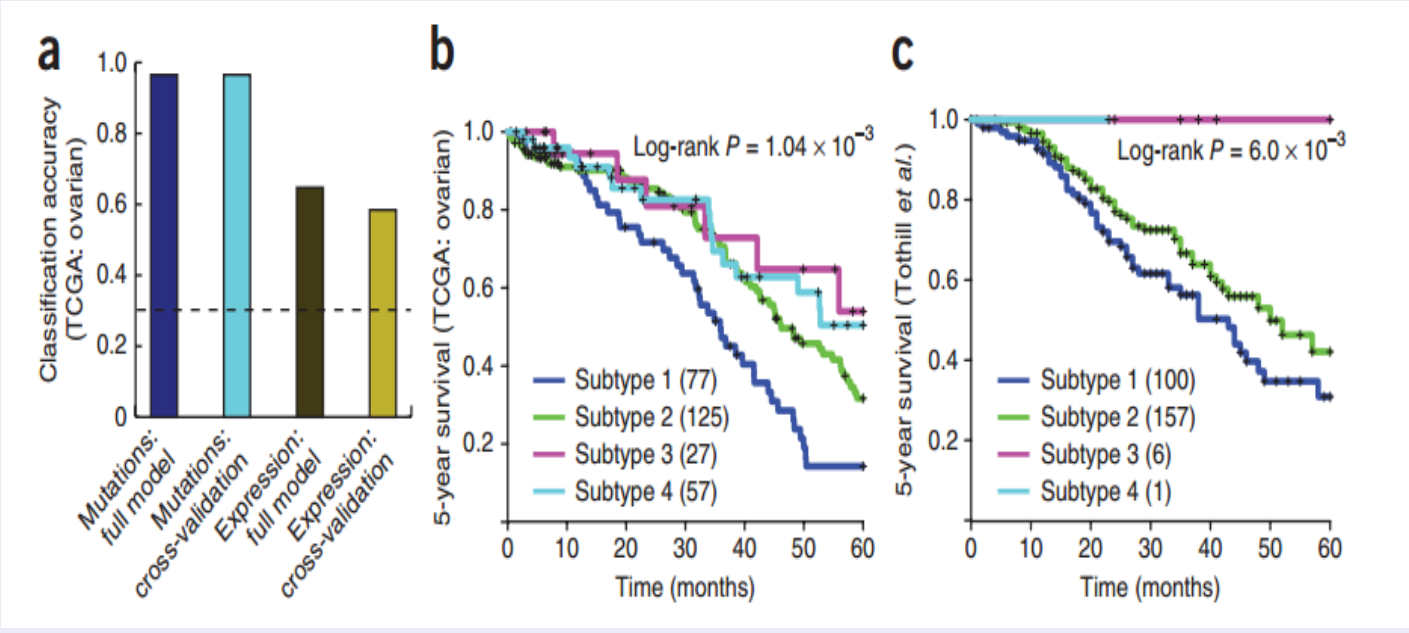
1. worst overall survival time and shortest platinum-free interval
2. contain over 20 genes in the FGF signaling pathway which has been previously implicated as a driver of tumor progression and associated with resistance to platinum

NBS not only can atratify patients into clinically informative subtypes but may help identify the molecular network regions commonly mutated in each subtype

Predictive Signatures

- Applicable to new patients not in the TCGA ?
- assign a patient to one of the existing NBS subtypes

Nearest shrunken centroid approach



Nearest shrunken centroid approach

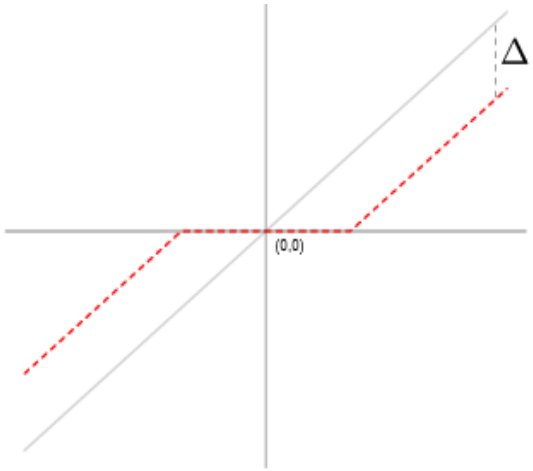
- Compute a standardized centroid for each class. This is the average gene expression for each gene in each class divided by the within-class standard deviation for that gene.

$$genes: \ i = 1,2,\cdots,p \quad samples: \ j = 1,2,\cdots,n \quad class: \ k = 1,2,\cdots,C$$

$$\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k} \qquad \bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$$

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{s_i} \qquad s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2$$

- shrinkage each d_{ik} towards 0 by soft-thresholding



$$d'_{ik} = sign(d_{ik})(|d_{ik}| - \Delta)_+$$

- shrunken centroid

$$\bar{x}'_{ik} = \bar{x}_i + s_i d'_{ik}$$

- Nearest centroid classification takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample.

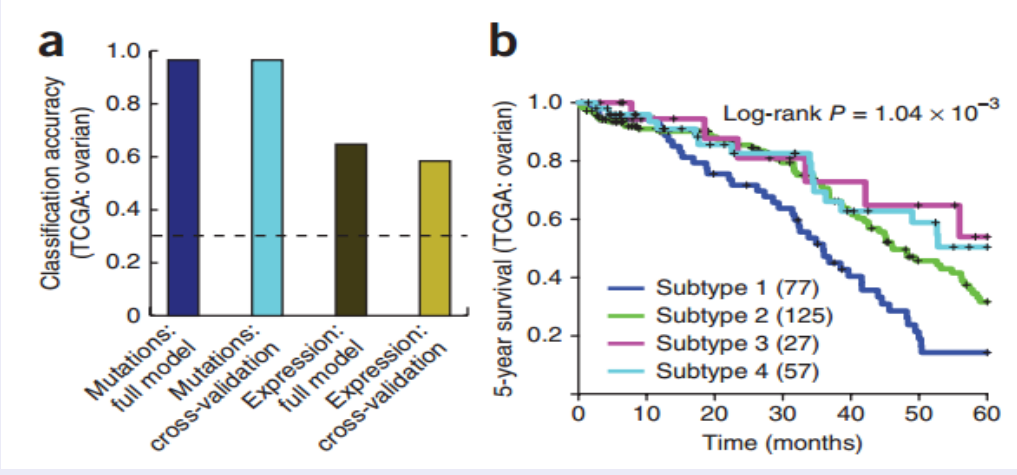
Why shrinkage ?

1. it can make the classifier more accurate by reducing the effect of noisy genes
2. it does automatic gene selection. In particular, if a gene is shrunk to zero for all classes, then it is eliminated from the prediction rule.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA 99, 65676572 (2002).

How about use mRNA expression data to assign a new patient?

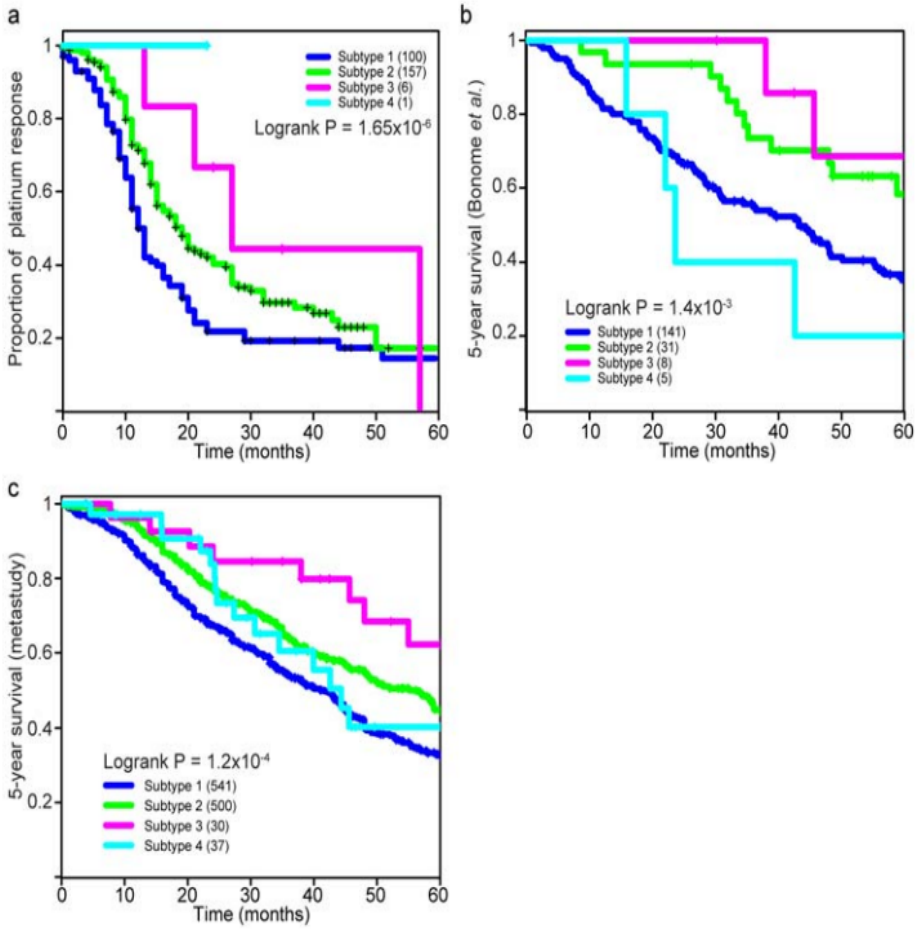
- mRNA expression data are presently much more widely available than are full genome or exome sequences



- classifier defined by NBS
- shrunken centroids: mRNA expression
- patients assigned by mRNA expression

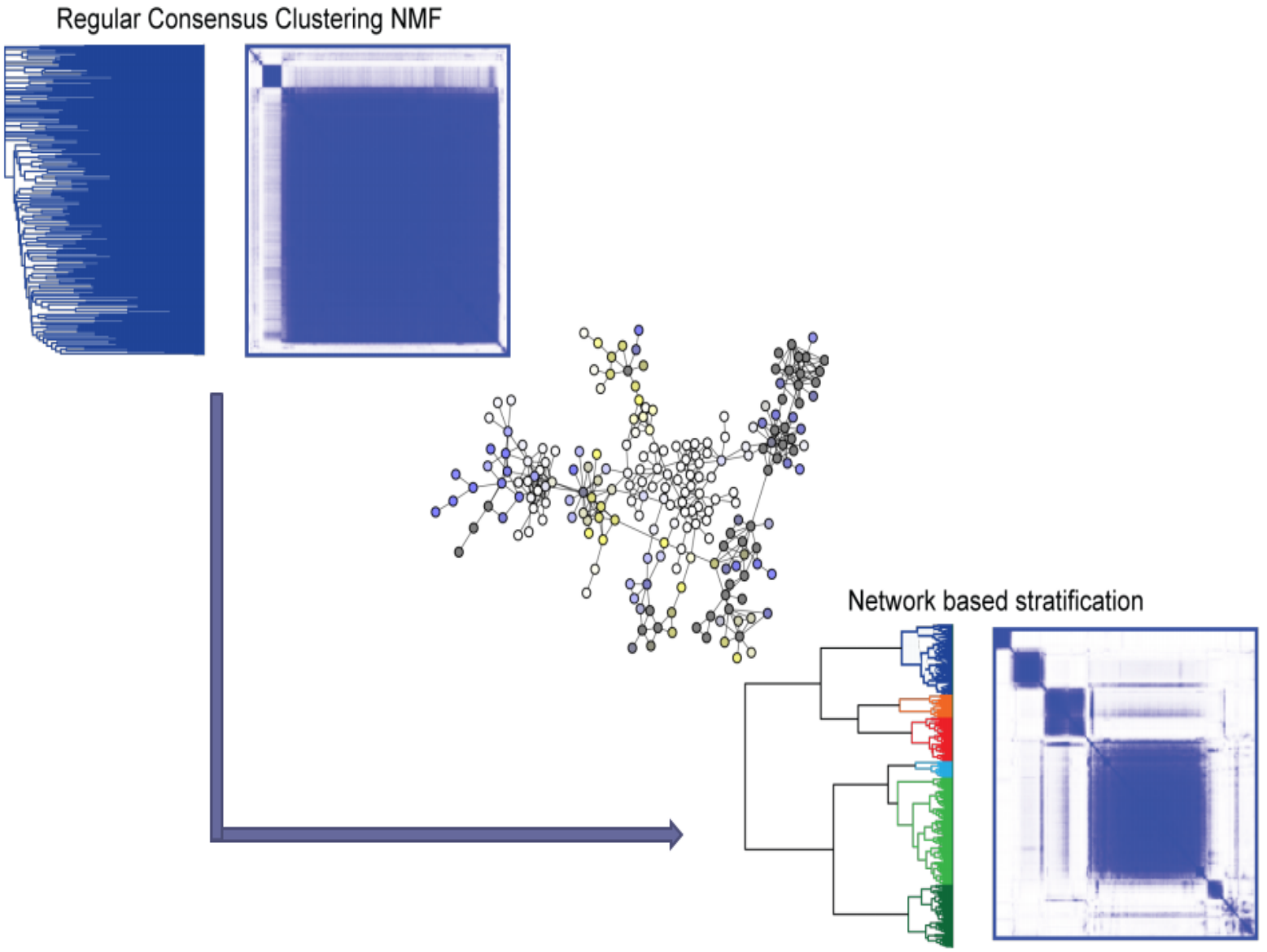
Gene Expression as a Surrogate Biomarker

TCGA and two independent studies (include gene expression profiles but lack somatic mutation profiles)



Content

Summary


Regular Consensus Clustering NMF
Network based stratification

Conclusion

1. Network-based stratification recovers biologically relevant subtypes
2. Somatic mutation subtypes are different from those recovered from other molecular profiles.
3. These subtypes can be recapitulated using gene expression.
4. Each subtype seems to have specific effected subnetworks.

Discussion

Somatic Mutation

1. differential measurements between tumor and normal tissues, whereas expression profiles are absolute measurements
   "base-line" : germ line ; leave only tumor-specific
2. Causal genetic events underlying tumor progression

Method

1. Use network as prior knowledge
2. NMF clustering

A comprehensive research

Simulation analysis;clinically and biologically meaningful;predictive signature

Discussion

How to perform effective pattern recognition from massive biological data?

1. **Generative model:**
   mostly used;
   data reduction;
   low dimension
   Assume latent variables

2. **Why discriminative model?**
   sample space;
   separability;
   unsupervised;
   feature selection;
   class label learning