Presenter: Lee Sael

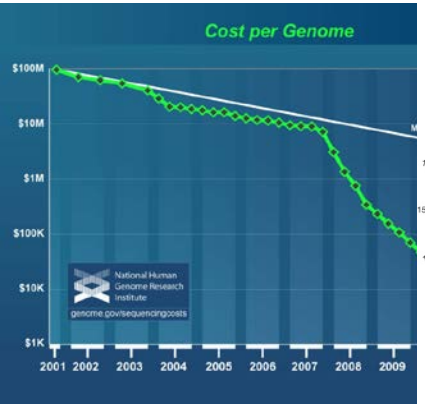
Collaborative work with POSTECH DM Lab. (Hwanjo Yu & Sungchul Kim)

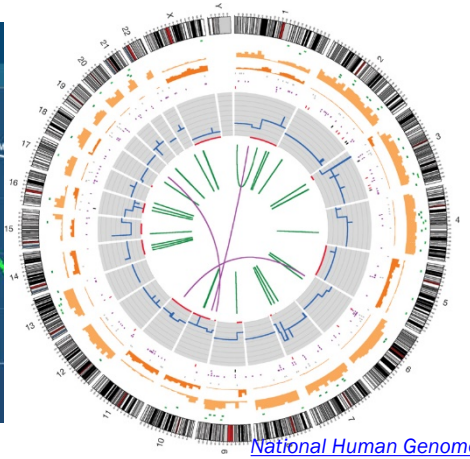
ORTHOGONAL NMF-BASED TOP-K PATIENT MUTATION PROFILE SEARCHING

Ref. Publication: Kim, S., Sael, L., & Yu, H. (2015). A mutation profile for top- k patient search exploiting gene-on tology and orthogonal non-negative matrix factorization. *Bioinformatics*, btv409.

FAST SOMATIC MUTATION PROFILE SEARCH – THE MOTIVATION

- Sequencing will become a common practice in medicine [1-3]
- Characterizing cancer patients with somatic mutations is a natural process for cancer studies because cancer is the result of accumulation of genetic alterations.
- Similarity search on mutation profiles can solve various translational bioinformatics tasks, including prognostics and treatment efficacy predictions for better clinical decision [4].



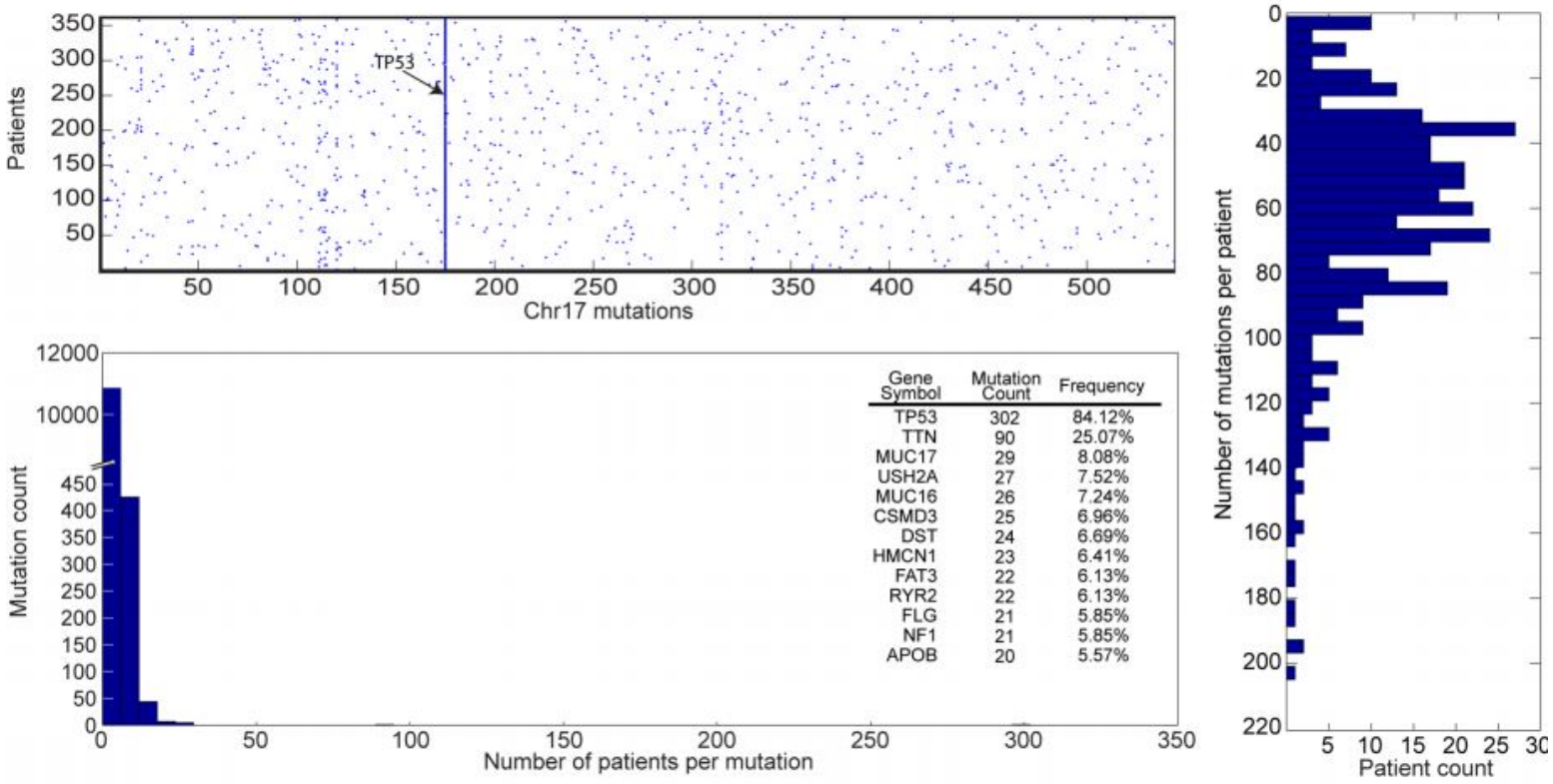


National Human Genome Research Institute (*NHGRI*)

ED Pleasance *et al. Nature* **000**, 1-6 (2009)
doi:10.1038/nature08658

CHALLENGE: SPARSITY AND HETEROGENEITY OF MUTATION DATA

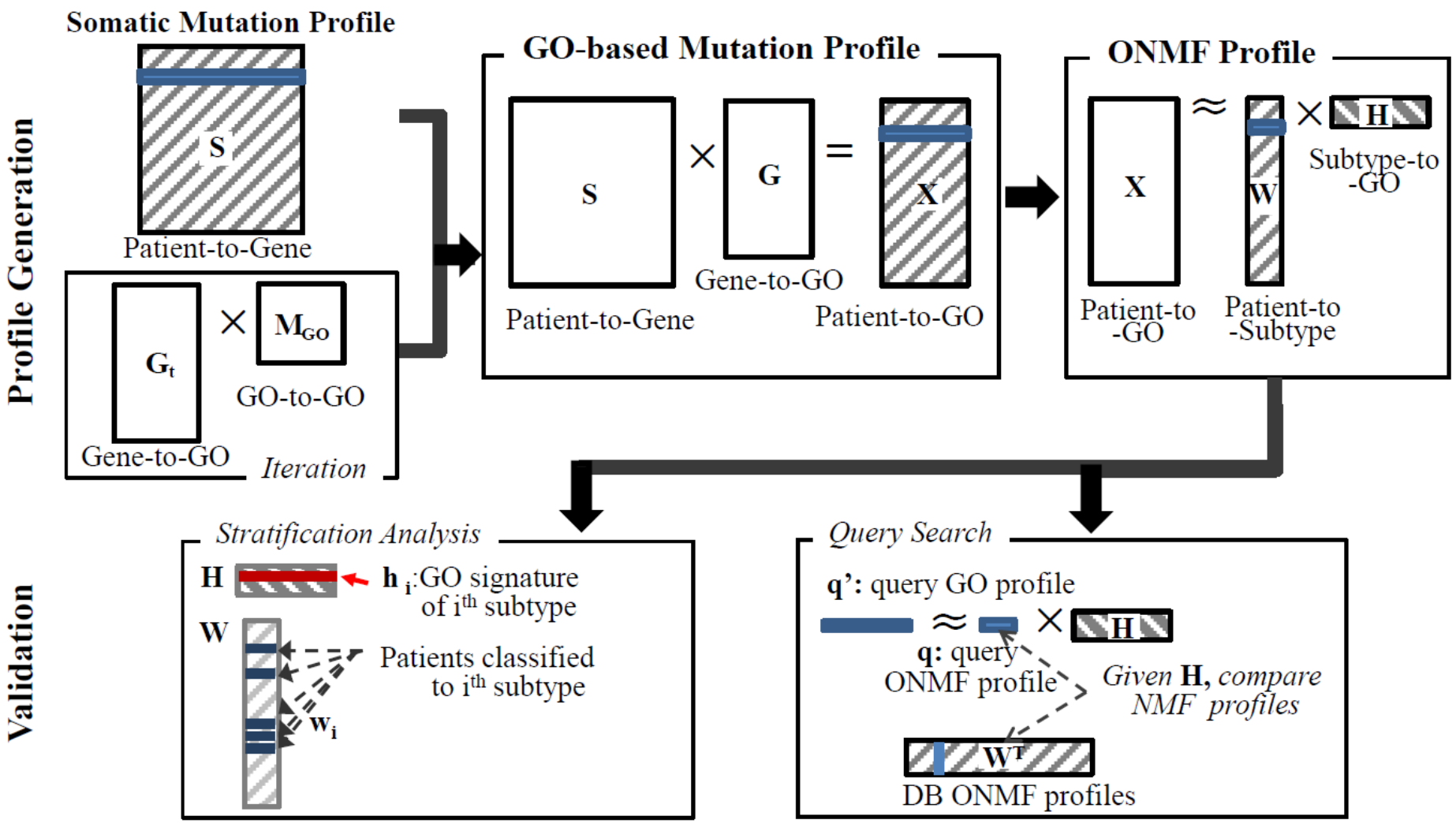
- Somatic mutation data are sparse in character, and for complex diseases, including cancer, mutations are genetically heterogeneous [5-6].



GO AND ONMF-BASED SOMATIC MUTATION PROFILE

- Goal
  - To provide a simple but effective mutation profile
- Method:
  - Exploit Gene-Ontology (GO) and orthogonal non-negative matrix factorization (ONMF)
- Target data
  - Somatic mutation data (from TCGA)
    - 5 different cancer types
- Characteristics of proposed profile
  - Compact representation of somatic mutation for cancer patients
  - Enable real-time search
  - Tolerant to heterogeneity
  - Directness in function interpretation
  - High predictive power for clinical features

OVERVIEW OF THE PROFILE GENERATION AND VALIDATION METHODS



SOMATIC MUTATION PROFILE, S

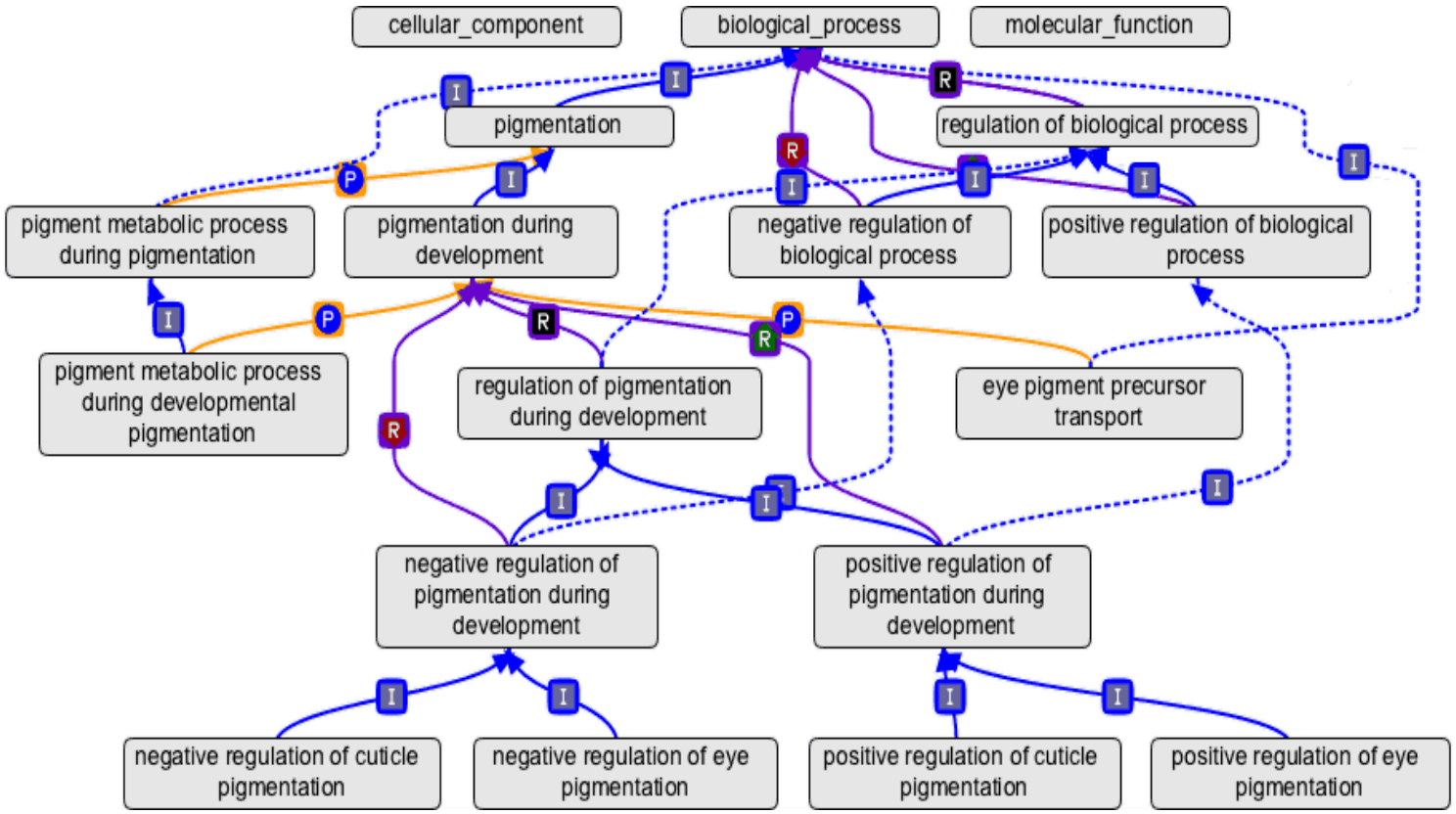
- For each patient, somatic mutations are represented as a profile of binary mutated states on genes.
- Types of mutation considered:
  - A single-nucleotide base change,
  - the insertion
  - deletion of bases

Patient	Gene	Variant type	Varient Class	Chorm.	Start/End Pos.	Ref_Allele
2352	NEK11	INS	Shift_Ins	19	58862932	-
2002	EGFR	DEL	Shift_Del	10	52575855	G
2002	TP53	SNP	Missense	10	52575855	A
2352	EGFR	SNP	Missense	3	9229467	T
A062	A2M	SNP	Silent			
…	…	…	…			

Patient	TP53	NEK11	EGFR	…
2352	0	1	0	
2002	1	0	1	
…	1	1	0	

GENE ONTOLOGY (GO)

- Terms in the Gene ontology (GO) are hierarchical representation of controlled vocabulary of gene and gene products [7-8].
-
- Biological terms in the same level may have different granularity in the GO hierarchy [9].
- We only use Biological Processes (BP) terms



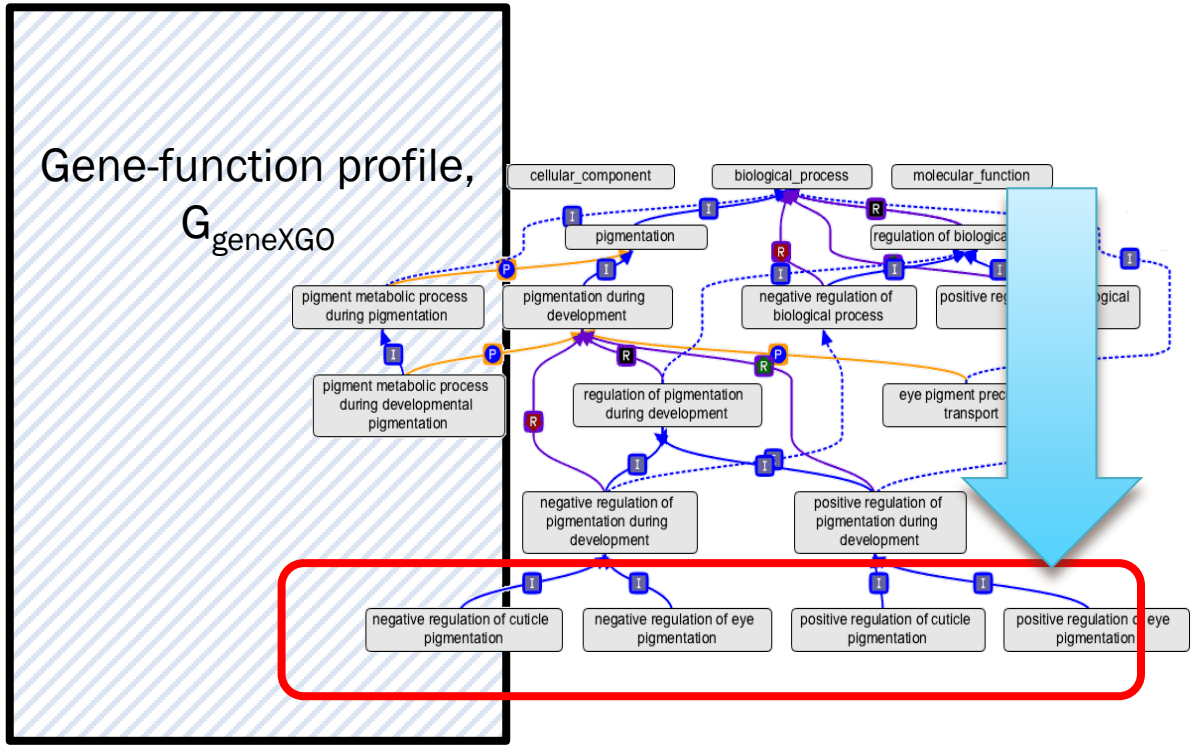
Adapted from a figure in Gene Ontology Consortium (geneontology.org)

GENE-FUNCTION PROFILE, $G_{GENEXGO}$

- Each gene is a binary vector of GO terms
  - 1 if annotated with the term,
  - 0 otherwise.
- Reducing correlation between GO terms by using only the most specific terms
  - Scores of non-leaf nodes are propagated to their descendant nodes until G_t converges

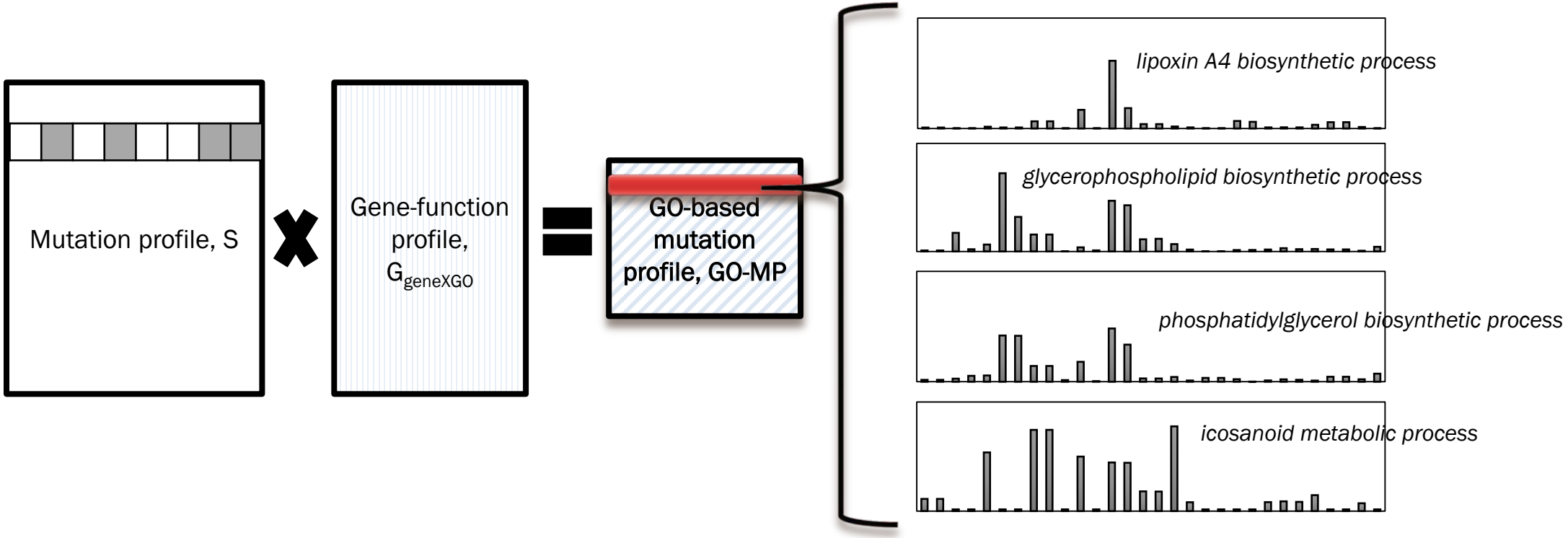
$$G_{t+1} = G_t \times M_{GO}$$

where G_t is the gene-function profile at the *t*-th iteration and M_{GO} is an adjacency matrix



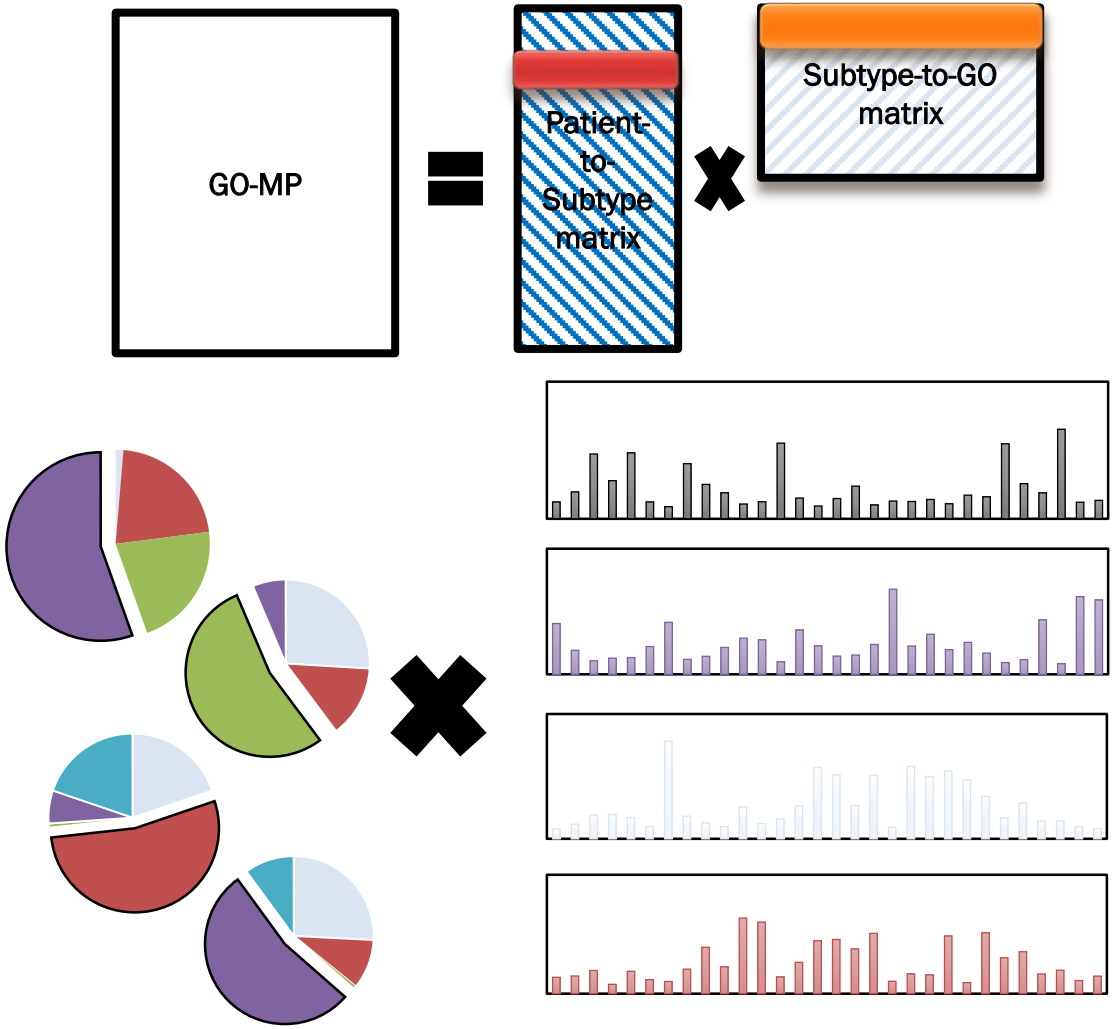
GO-BASED MUTATION PROFILE, GO-MP

- For each patient, GO-based somatic mutation profile is represented by a weighted sum of gene scores on each GO term.
  - Multiply Mutation Profile matrix S with Gene-GO Profile matrix.
  - $S \times G_{genexGO}$



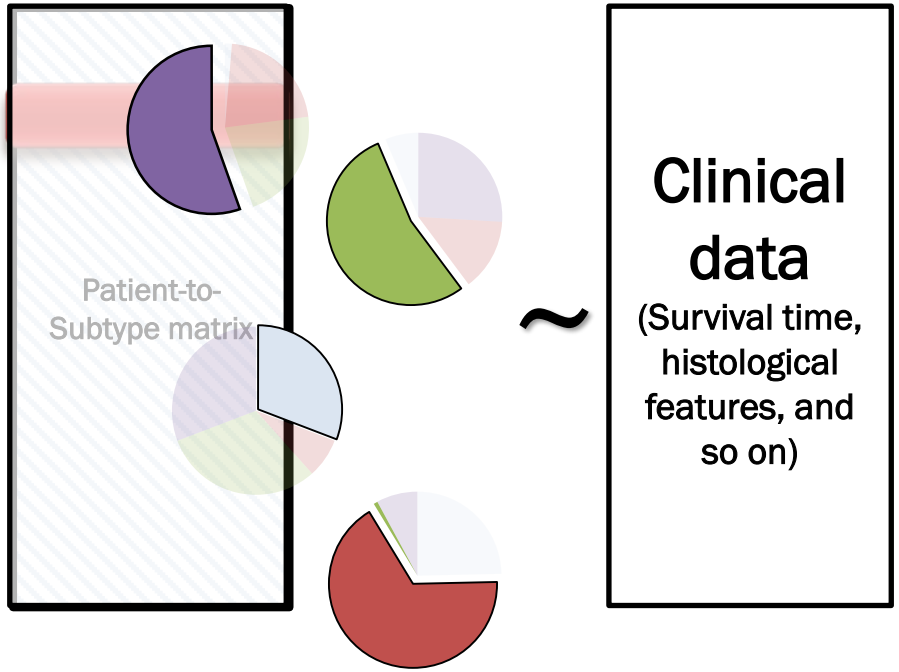
ONMF MUTATION PROFILE, ONMF-MP

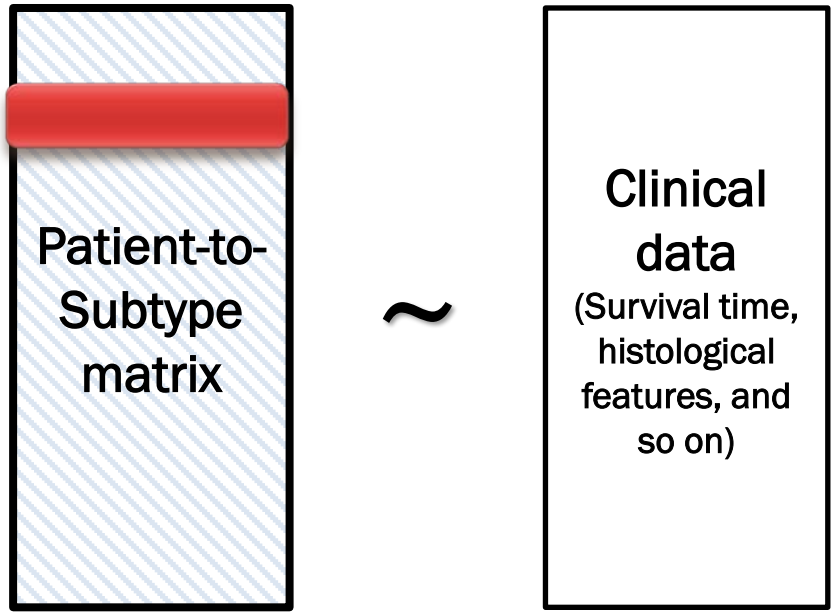
- Orthogonal Non-negative Matrix Factorization (ONMF)
  - $X \cong W \times H \qquad s.t. H\,H^T = I$
  - Generally, orthogonal constraints on NMF enhance the clustering quality
    - Similar basis vectors are avoided.
- ONMF mutation profile
  - The GO-MPs are further made compact by taking the encoding matrix W of ONMF on X as profile vectors.



PERFORMANCE VALIDATION

- Cancer stratification
  - Associations between the cancer subtypes and clinical features.
- Top-*k* search
  - Similarity of clinical profiles to determine whether the search results are correct.





EXPERIMENTAL RESULT

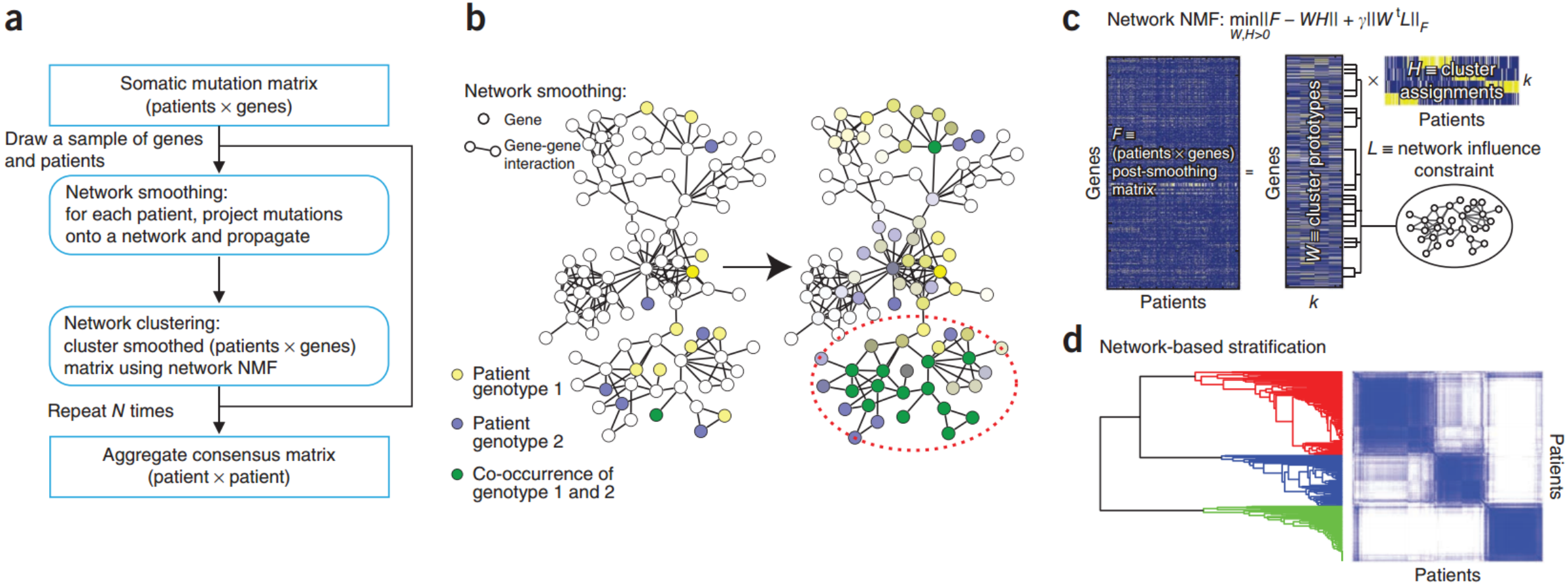
- Data set
  - Somatic mutation data of five tumor types downloaded from TCGA portal; UCEC, BRCA, OV, LUAD, GBM data

	UCEC	BRCA	OV	LUAD	GBM
# patients	247	772	441	516	291
# genes	9341	13078	12431	18067	9341

- Competitors
  - Cancer stratification - Network-Based Stratification (NBS). GOS (NMF on GO-MP), ORGOS (ONMF on GO-MP)
  - Top-*k* search – Somatic mutation profile, GO-MP, ONMF-MP

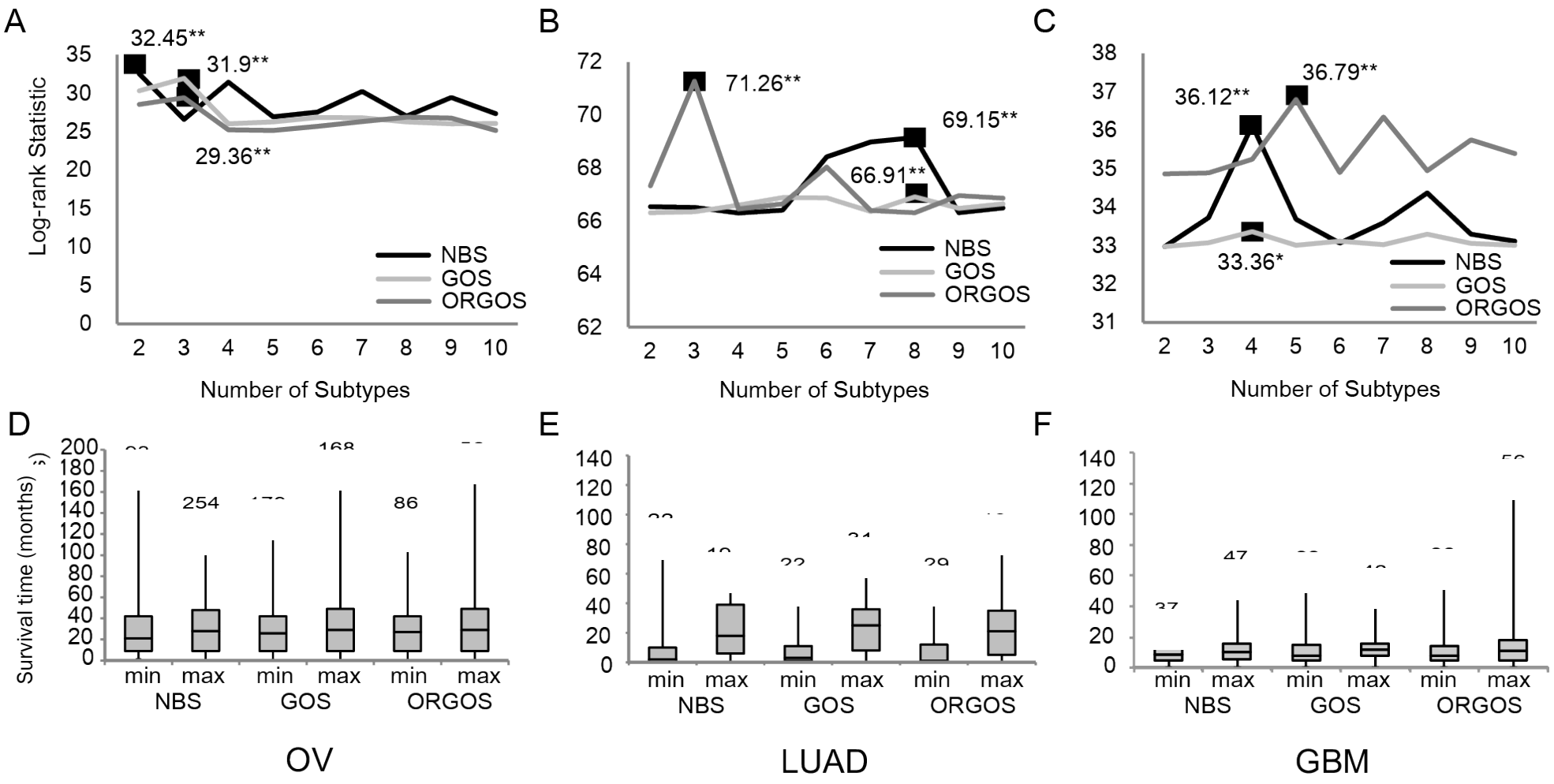
COMPARED METHOD NETWORK-BASED STRATIFICATION (NBS)

- A method to integrate somatic tumor genomes with gene networks

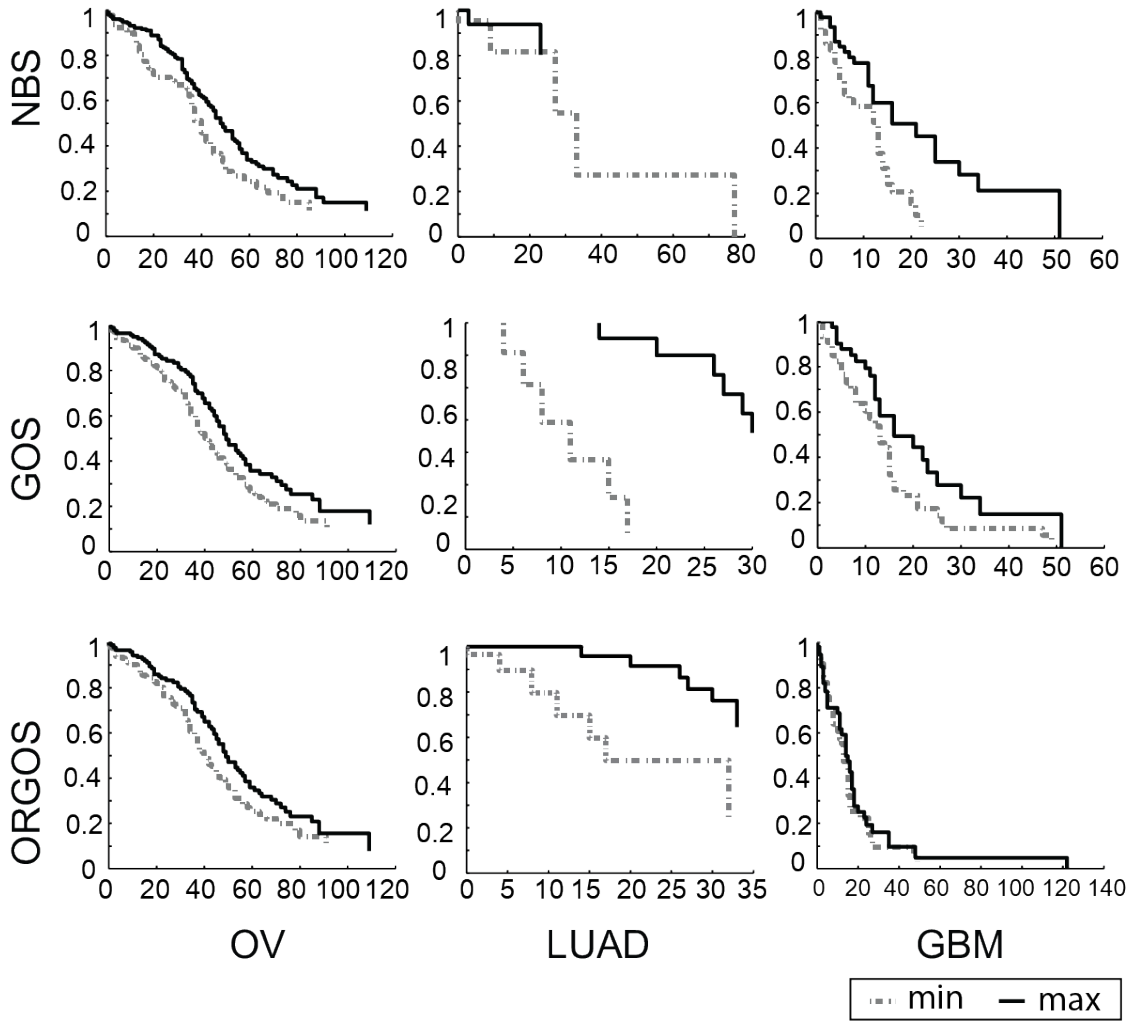


Matan Hofree, John P Shen, Hannah Carter, Andrew Gross & Trey Ideker, Network-based stratification of tumor mutations. (Nature 2013).

ASSOCIATION WITH PATIENT SURVIVAL

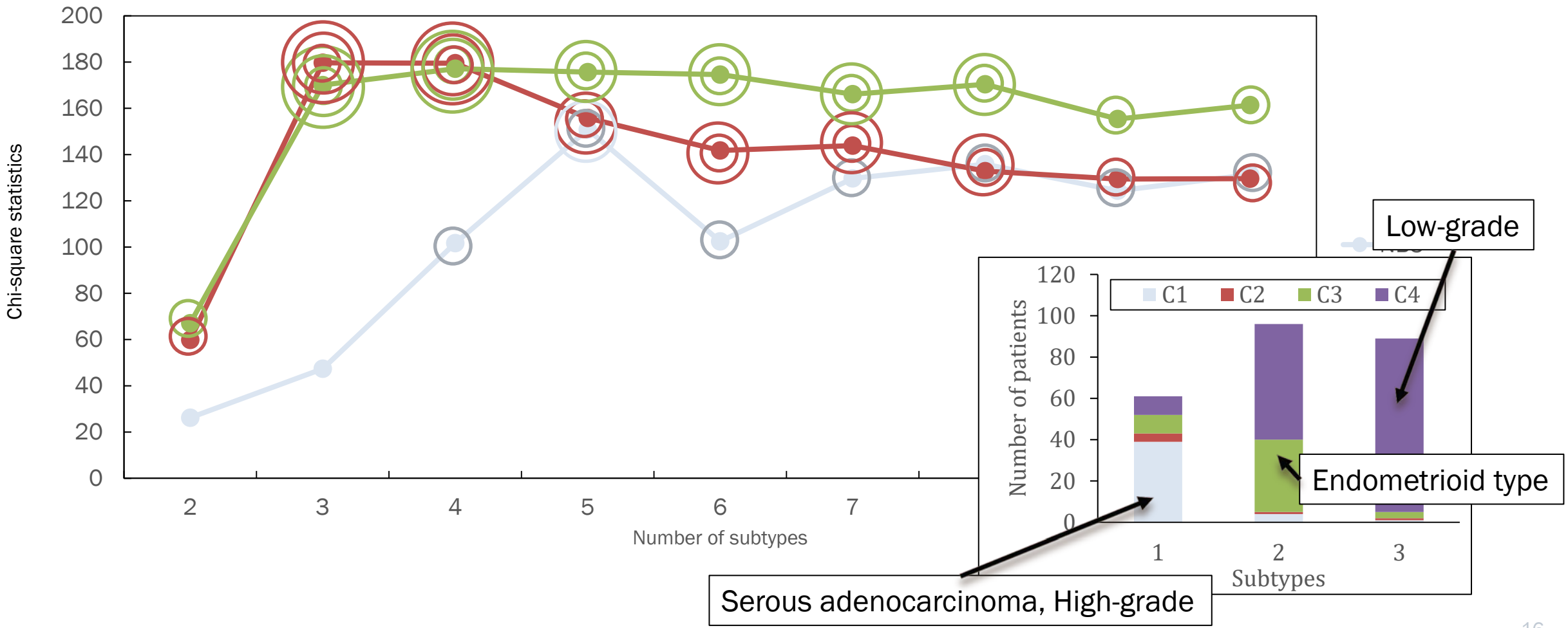


ASSOCIATION WITH PATIENT SURVIVAL

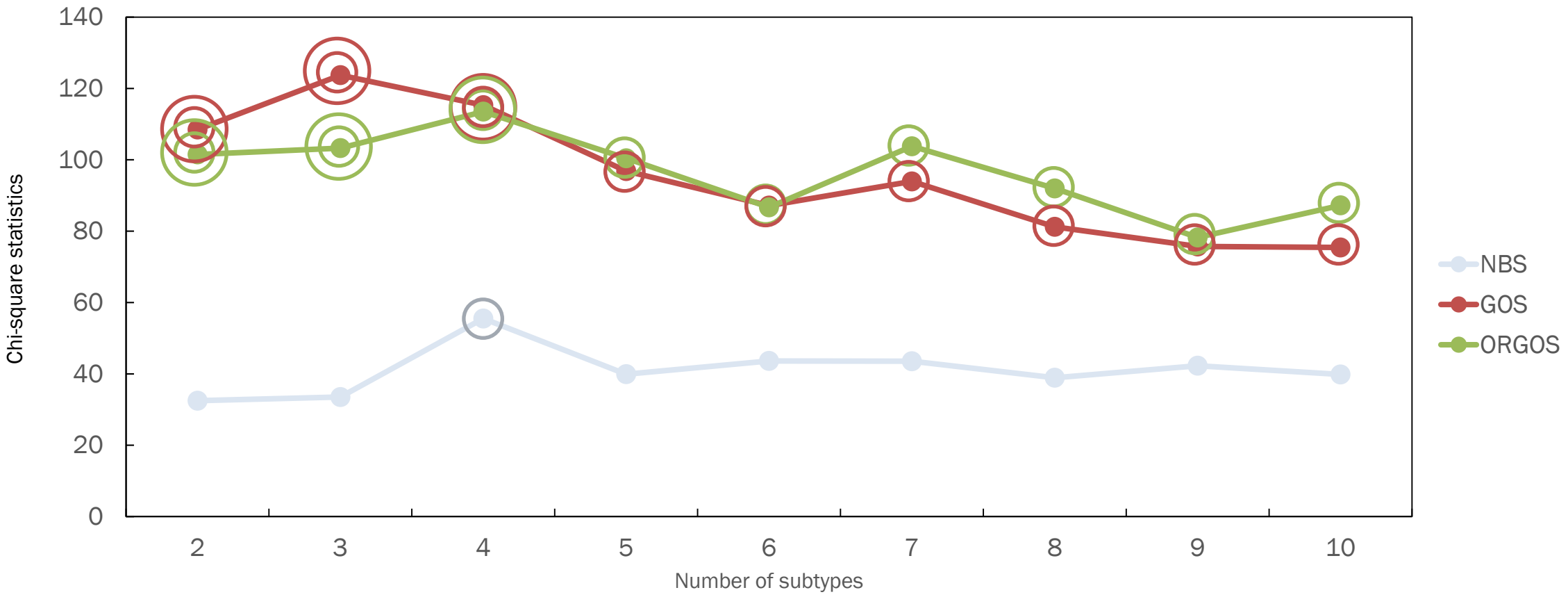


- In OV, three survival curves show similar pattern for the all three approaches.
- In LUAD, NBS produced inaccurate survival curves in which the min subtype shows longer survival pattern than the max subtype.
- In GBM data, NBS was successful at grouping the min survival while ORGOS was better at grouping the max survival.

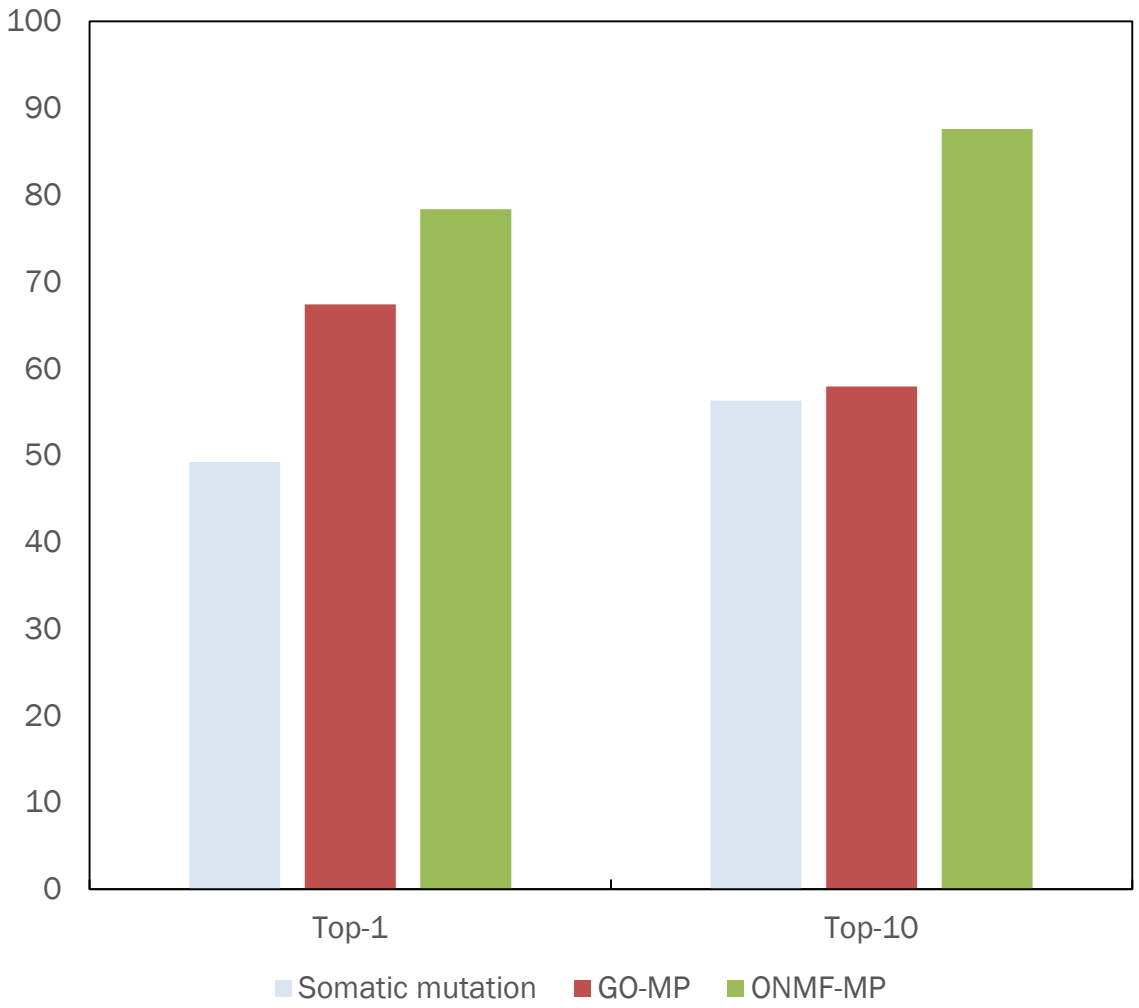
CHI-SQUARE STATISTICS OF SUBTYPES WITH HISTOLOGICAL BASIS FEATURE ON UCEC DATA



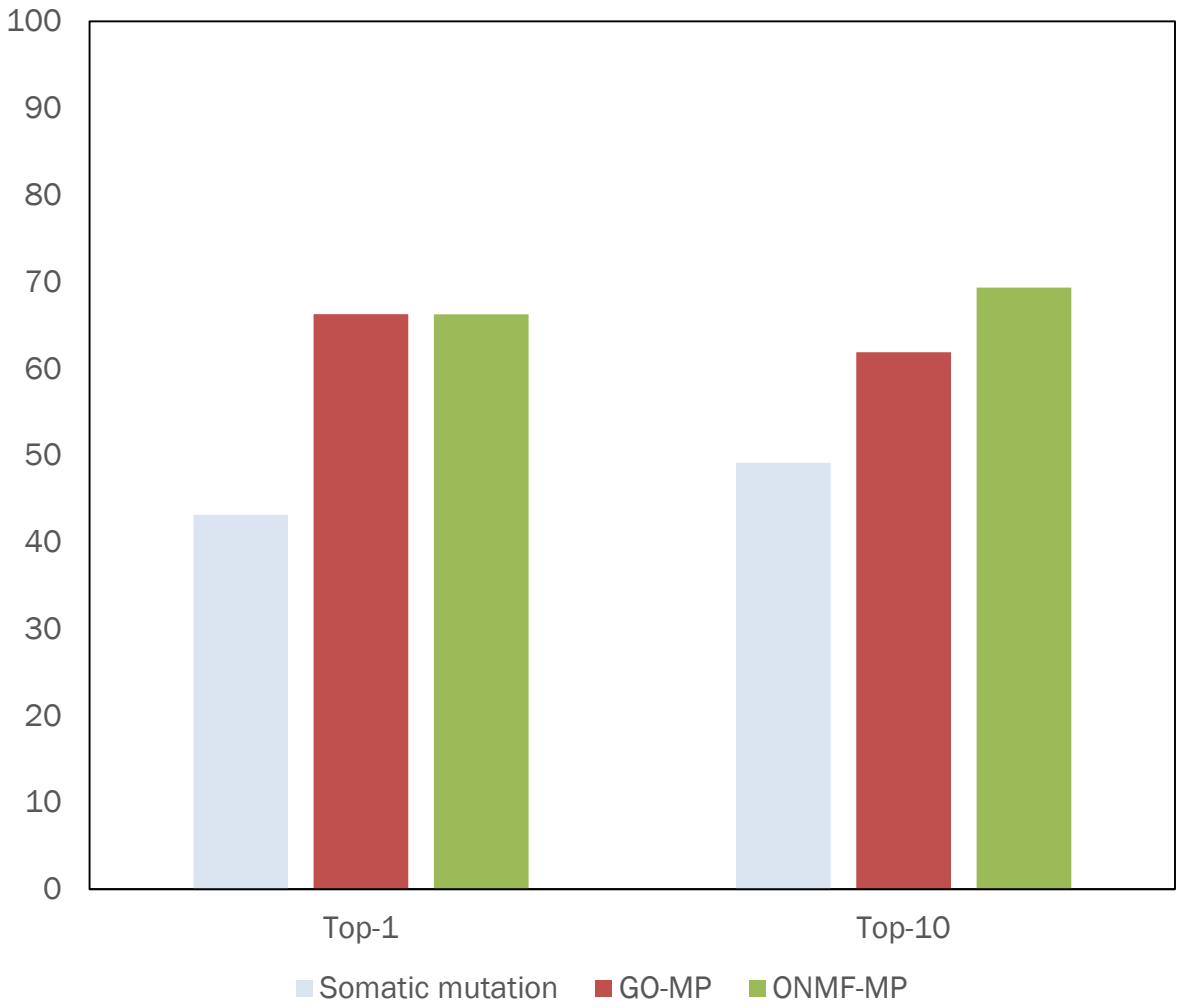
CHI-SQUARE STATISTICS OF SUBTYPES WITH ESTROGEN RECEPTOR STATUS ON BRCA DATA



TOP-*K* SEARCH ON SINGLE FEATURE

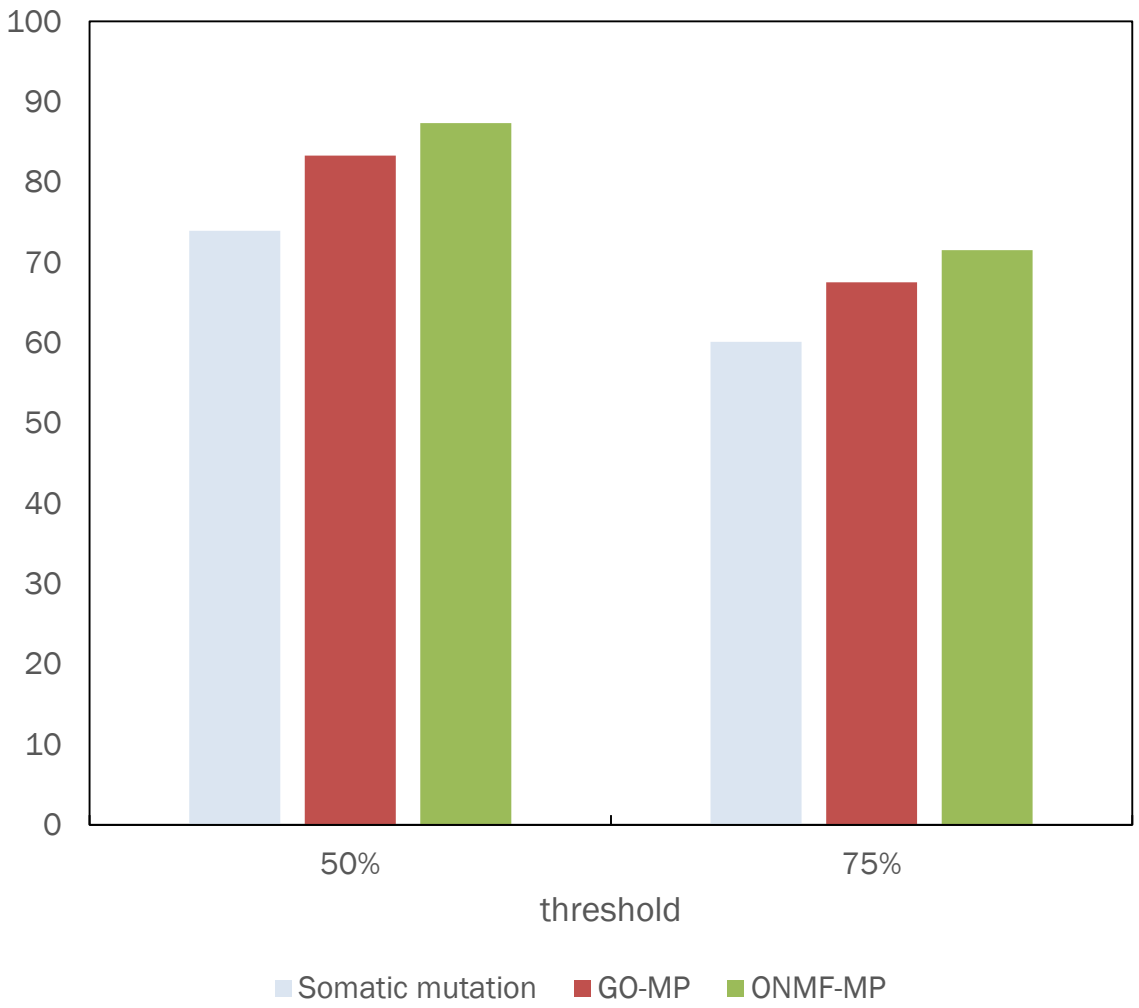


UCEC data; histological type

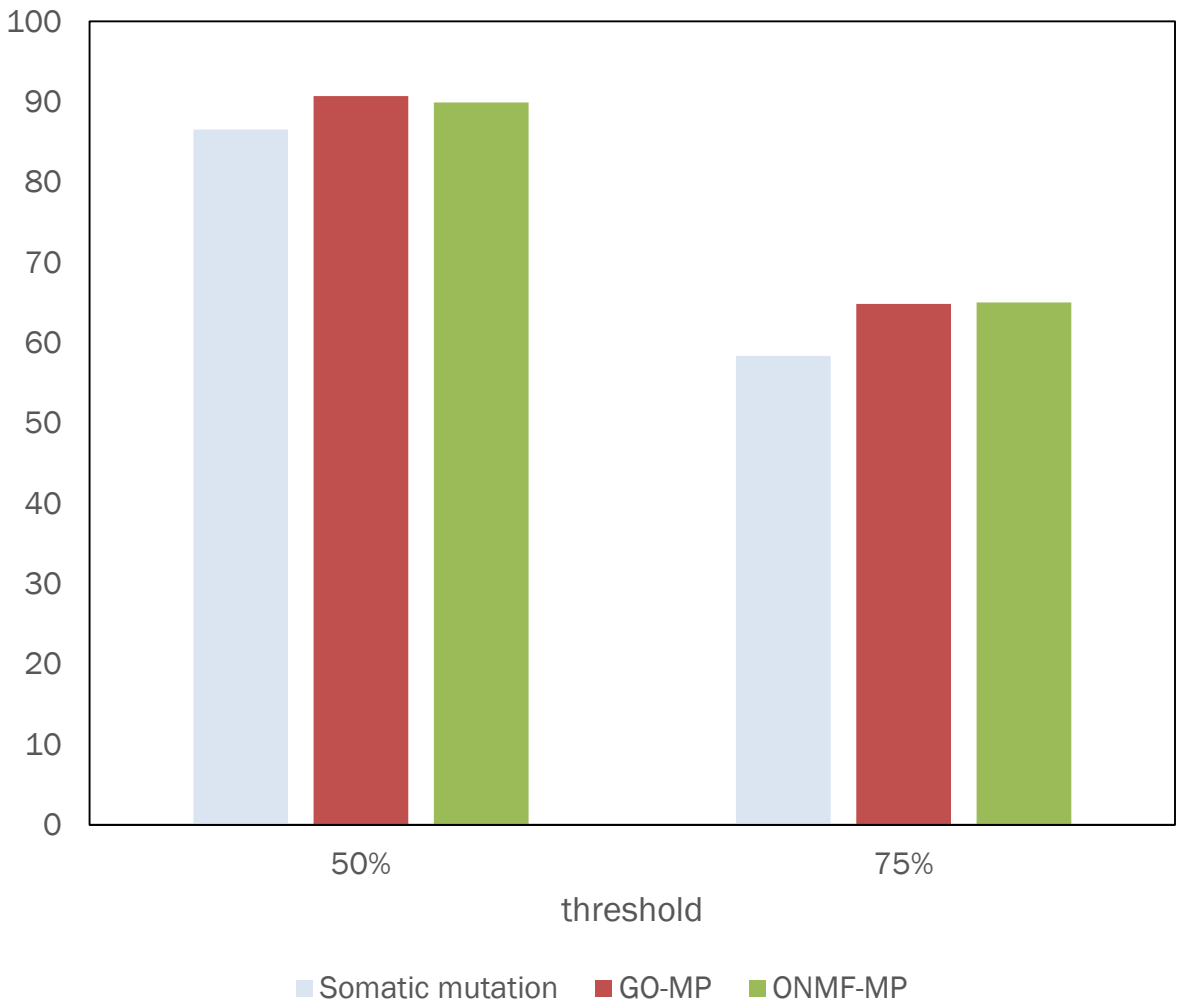


BRCA data; estrogen receptor status

TOP-10 SEARCH ON MULTIPLE FEATURES

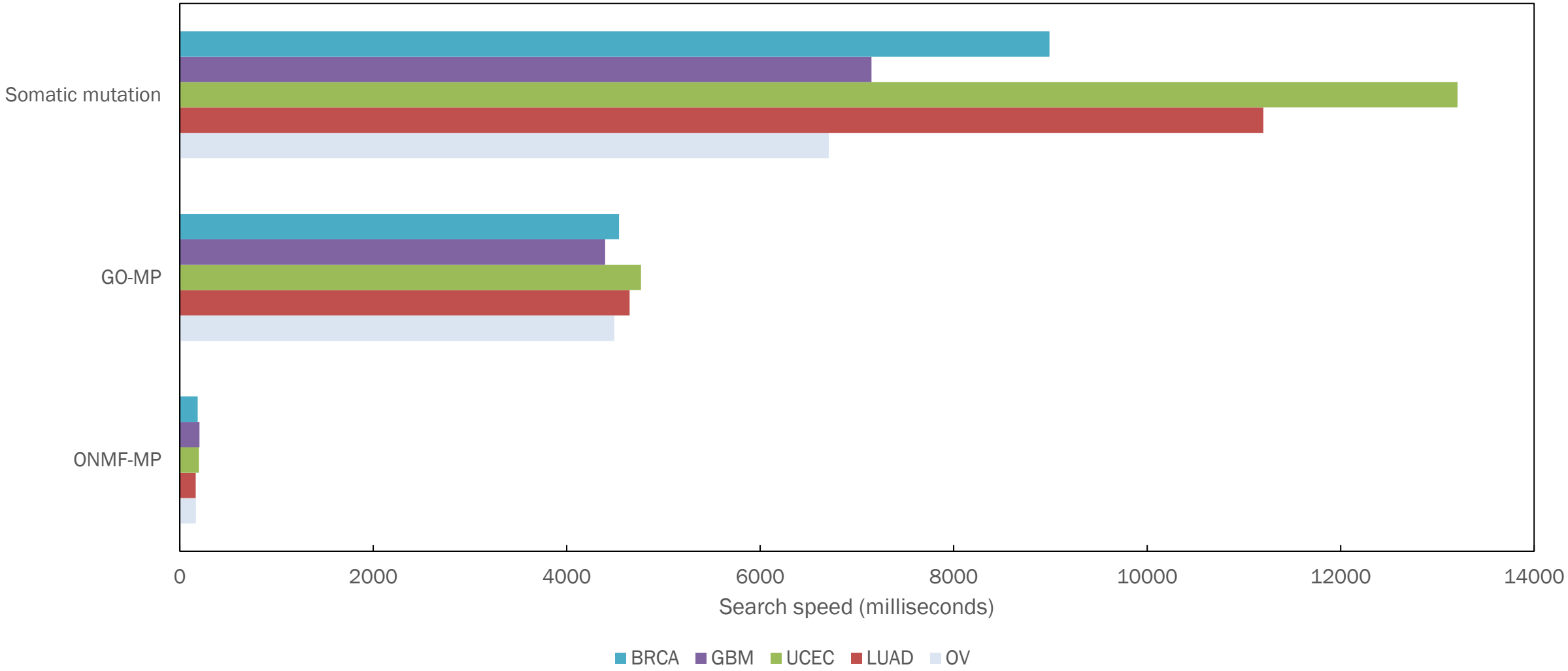


UCEC data

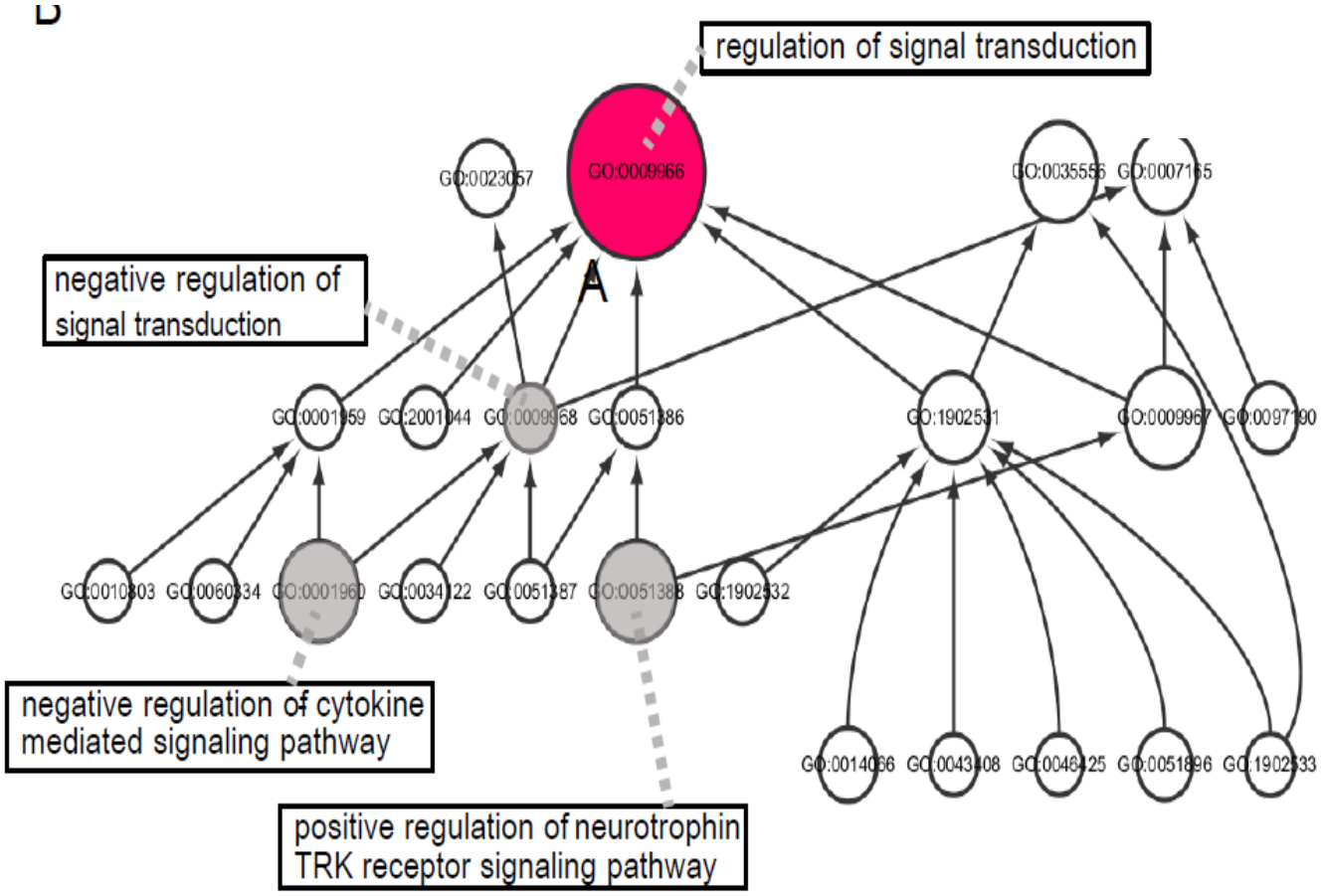


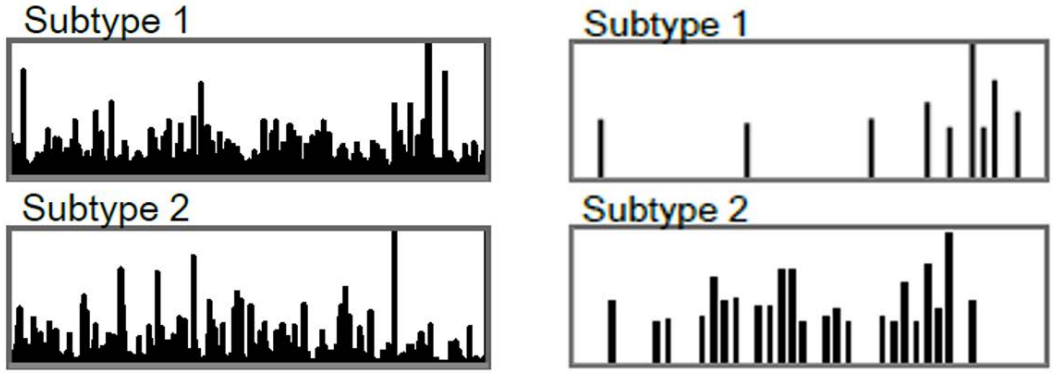
BRCA data

AVERAGE TOP-*K* SEARCH SPEED



PROPAGATION OF GO TERM SCORES



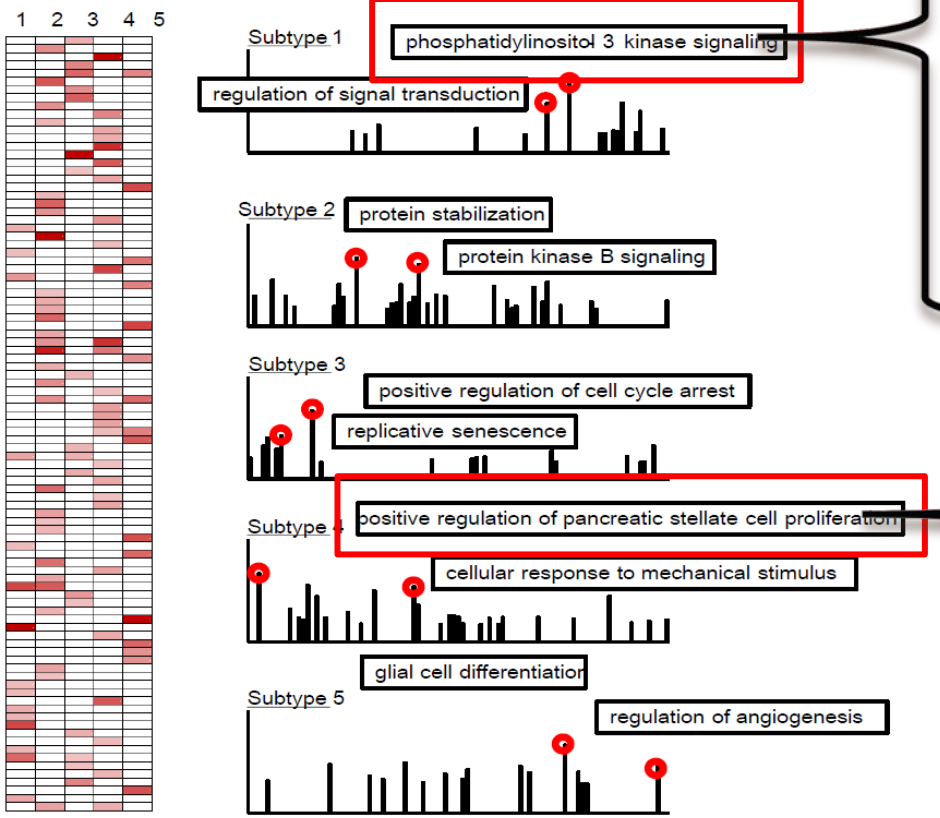


Algorithm 1: Identifying significant GO terms

Data: Initial score vector $\mathbf{w}_0$, GO terms $\mathbf{x}$
Result: A set of significant GO terms, $\mathbf{x}^*$

```
w* = w0
repeat
    foreach node i that is updated at the previous step do
        P = x[i].Parents(); % An index set of i-th GO term' ancestors
        scr = w*[i]/|P|
        if |P| == 1 then
            continue;
        end
        foreach p ∈ P do
            if ϵ < x[p].r then
                w*[p]+ = scr;
            else
                w*[p] = scr;
            end
        end
    end
until w* does not change;
return x* = GetSignificantGOterms(w*)
```

ANALYSIS OF SUBTYPES ON GO TERMS



"PI3K cascade is an important pathway that is involved in proliferation, invasion and migration in cancer [10-12].

"PI3K pathway influence GBM patients survival [13].

"Glioblastoma cancer and pancreatic cancer share network patterns that contain most of the candidate causative mutations [14].

"Pancreatic stellate cells are responsible for creating a tumor facilitatory environment that stimulates local tumor growth and distant metastasis [15].

CONCLUSION

- We suggest
  - Mutation profiles exploiting Gene Ontology and orthogonal NMF to obtain compact representation of mutation data and allow an efficient similar patient search.
- According to the results,
  - ONMF-MP allows us to efficiently search top-k patients that are clinically similar.
  - The tumor subtypes identified by using ONMF-MP are more closely associated with the clinical features than NBS.
    - The association of the subtypes with clinical feature in UCEC and BRCA data
    - The association of the subtypes with survival time in OV, LUAD, and GBM data

REFERENCES

1. The International Cancer Genome Consortium. International network of cancer genome projects. Nature 464, 993–996 (2010).
2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615 (2011).
3. Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. Science, 331(6024), 1553–1558.
4. Stuart, D. and Sellers, W. R. (2009). Linking somatic genetic alterations in cancer to therapeutics. Curr. Opin. Cell Biol., 21(2), 304–310.
5. Greenman, C. et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature, 446(7132), 153–8.
6. Mardis, E. R. (2012). Genome sequencing and cancer. Current Opinion in Genetics & Development, 22, 245–250.
7. Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol., 4(5), P3.
8. Khatri, P., Bhavsar, P., Bawa, G., and Draghici, S. (2004). Onto-Tools: an ensemble of web-accessible, ontology-based tools for the unctional design and interpretation of high-throughput gene expression experiments. Nucleic Acids Res., 32(Web Server issue), W449–456.
9. Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19(10), 1275–1283.
10. C. Jimenez, R. A. Portela, M. Mellado, J. M. Rodriguez-Frade, J. Collard, A. Serrano, C. Martinez-A, J. Avila, and A. C. Carrera. Role of the PI3K regulatory subunit in the control of actin organization and cell migration. J. Cell Biol., 151(2):249--262, Oct 2000.
11. Y. Samuels, O. Schmidt-Kittler, J.M. Cummins, L. Delong, I. Cheong, C. Rago, D.L. Huso, C. Lengauer, K.W. Kinzler, and B. Vogelsteinand V.E. Velculescu. Mutant pik3ca promotes cell growth and invasion of human cancer cells. Cancer Cell, 7:561--573, 2005
12. Z. Z. Zeng, Y. Jia, N. J. Hahn, S. M. Markwart, K. F. Rockwood, and D. L. Livant. Role of focal adhesion kinase and phosphatidylinositol 3'-kinase in in- tegrin bronectin receptor-mediated, matrix metalloproteinase-1-dependent invasion by metastatic prostate cancer cells. Cancer Res., 66(16):8091--8099, Aug 2006.
13. Y. Ruano, M. Mollejo, F. I. Camacho, A. Rodriguez de Lope, C. Fiano, T. Ribalta, P. Martinez, J. L. Hernandez-Moneo, and B. Melendez. Identication of survival-related genes of the phosphatidylinositol 3'-kinase signaling pathway in glioblastoma multiforme. Cancer, 112(7):1575--1584, Apr 2008.
14. G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. Genome Biol., 11(5):R53, 2010.
15. Z. Z. Zeng, Y. Jia, N. J. Hahn, S. M. Markwart, K. F. Rockwood, and D. L. Livant. Role of focal adhesion kinase and phosphatidylinositol 3'-kinase in in-tegrin bronectin receptor-mediated, matrix metalloproteinase-1-dependent invasion by metastatic prostate cancer cells. Cancer Res., 66(16):8091--8099, Aug 2006.