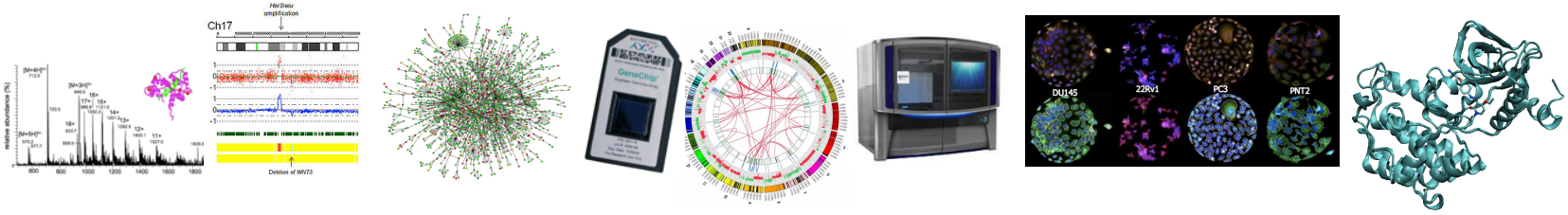
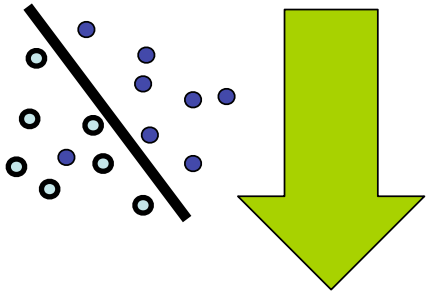
Patient stratification from somatic mutation profiles using gene networks

Jean-Philippe Vert

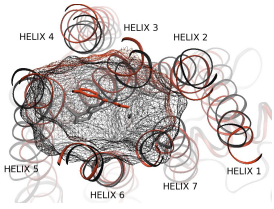
NCI - Institut Curie Symposium, April 4, 2018

Team's rationale

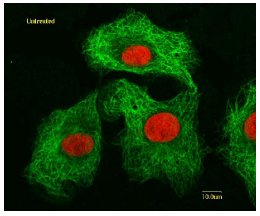




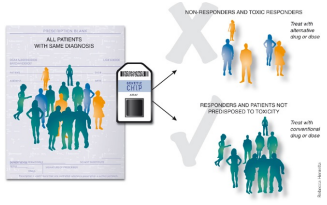
Machine learning
Learning with complex data
Regularization
Scalable algorithms



Molecules
(Epi)-Genomics
Systems biology
Drug design

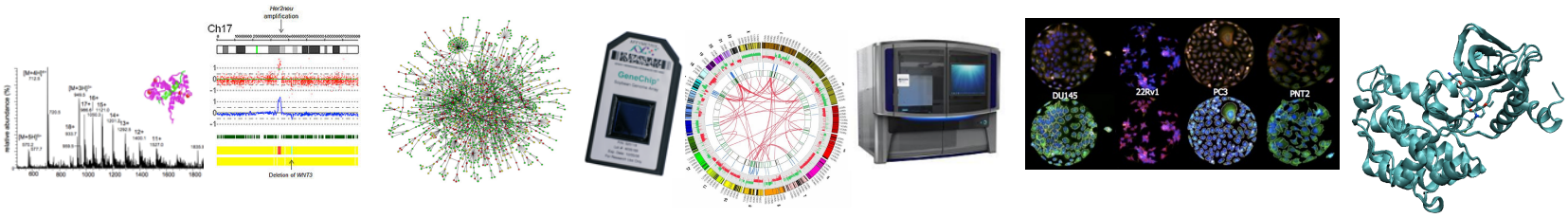


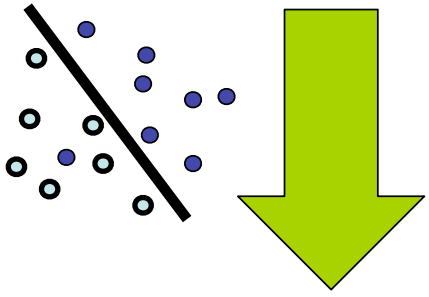
Cells
High-content screening
Single-cell genomics
Tumour heterogeneity



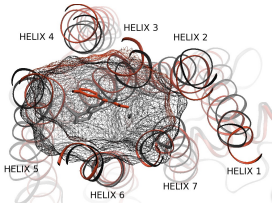
People
Precision medicine
GWAS
Patient monitoring

Team's rationale

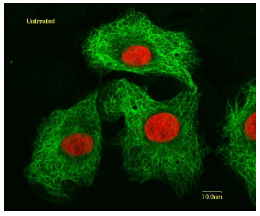




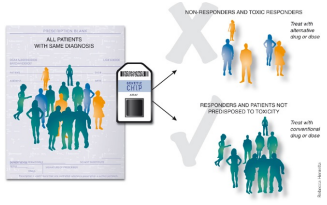
Machine learning
Learning with complex data
Regularization
Scalable algorithms



Molecules
(Epi)-Genomics
Systems biology
Drug design



Cells
High-content screening
Single-cell genomics
Tumour heterogeneity



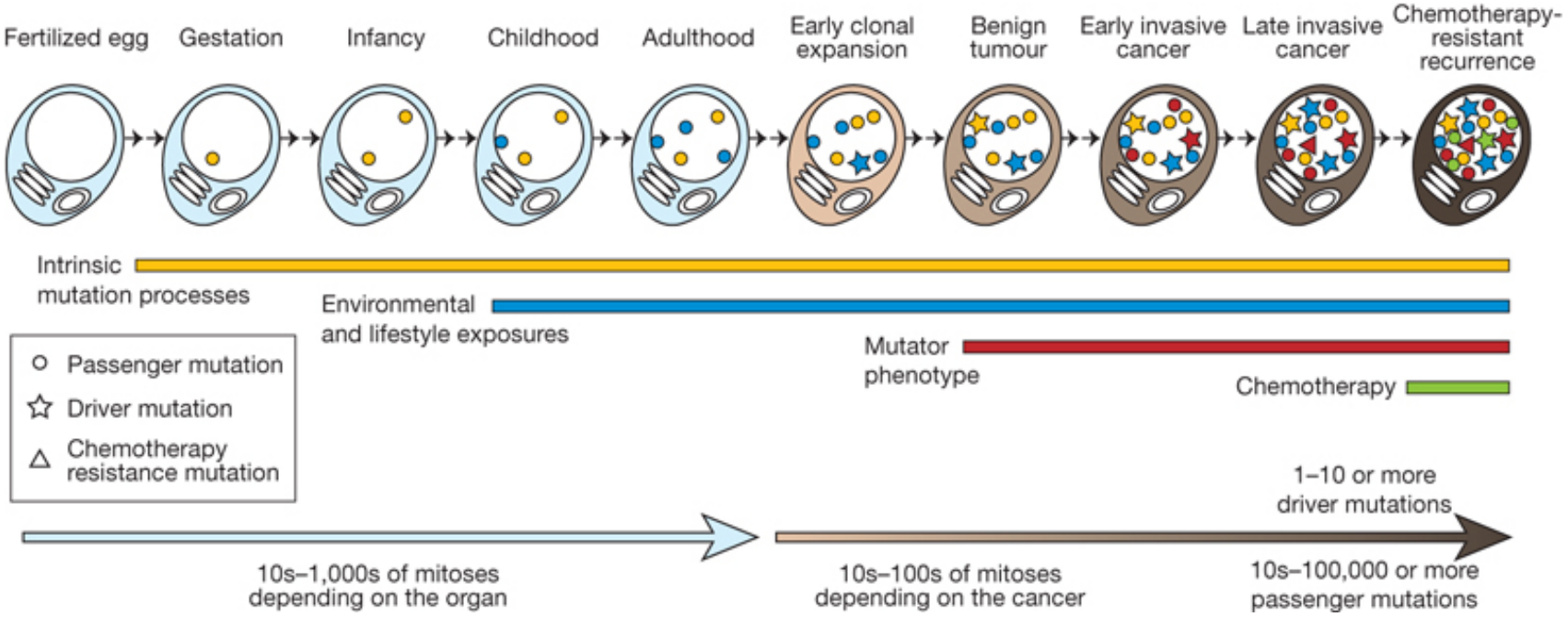
People
Precision medicine
GWAS
Patient monitoring

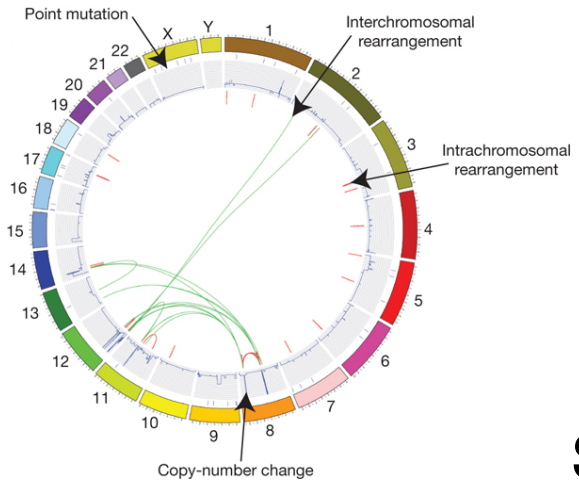
Joint work with

Marine Le Morvan

Andrei Zinovyev

Somatic mutations in cancer

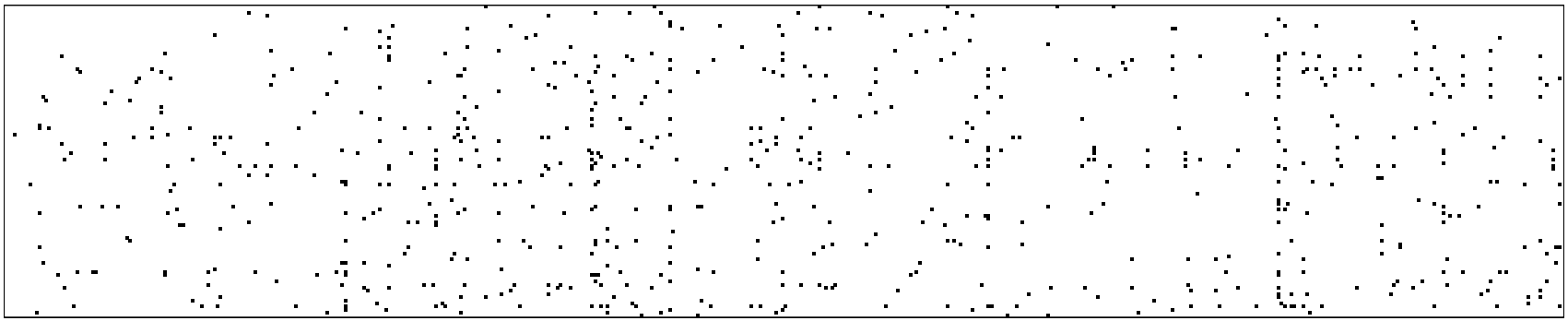




Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

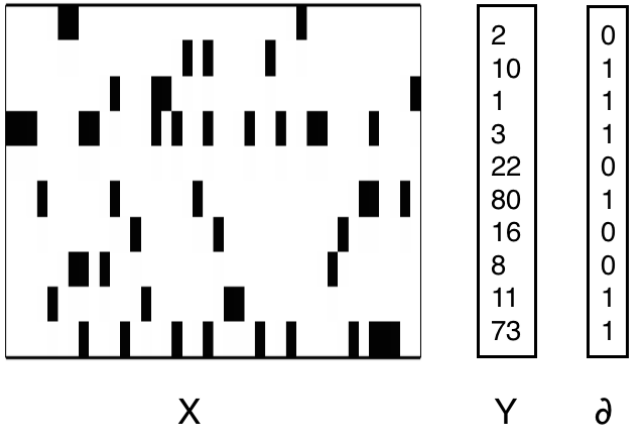
- 3,378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

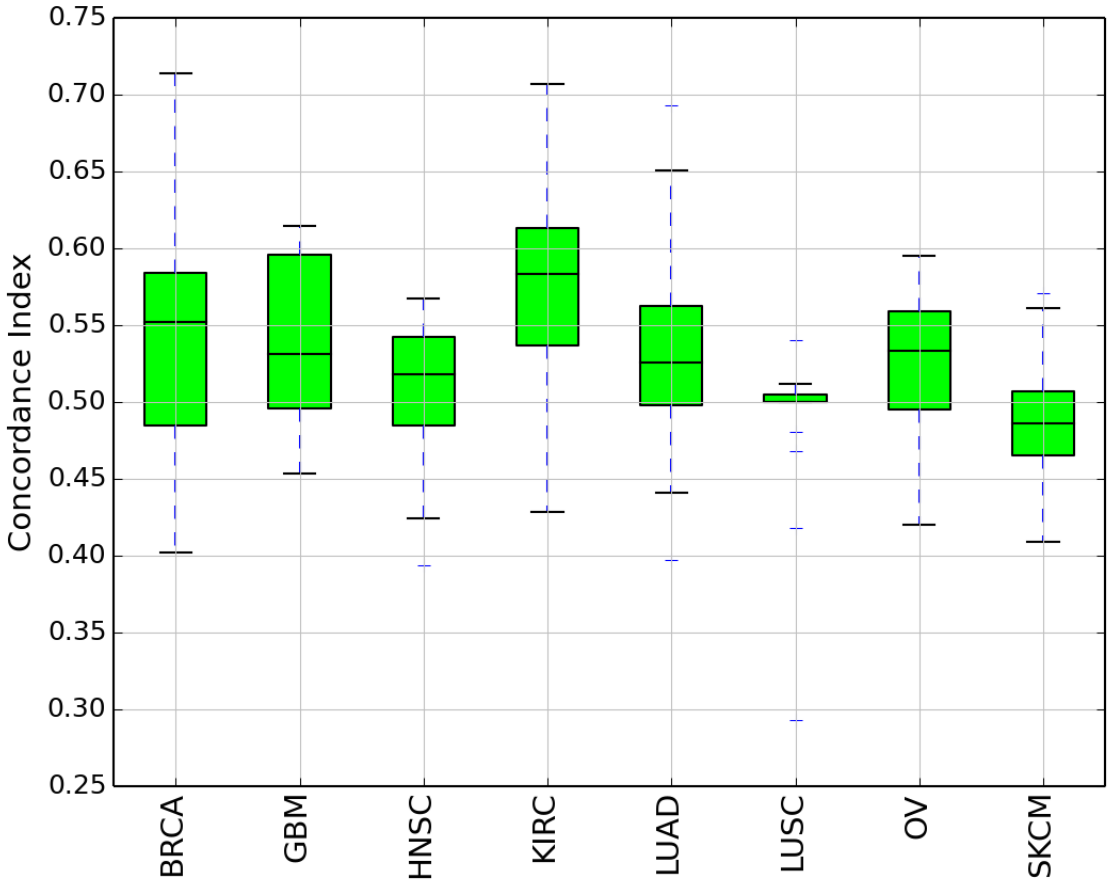


Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC (Lung squamous cell carcinoma)	169	13 590
OV (Ovarian serous cystadenocarcinoma)	363	10 195

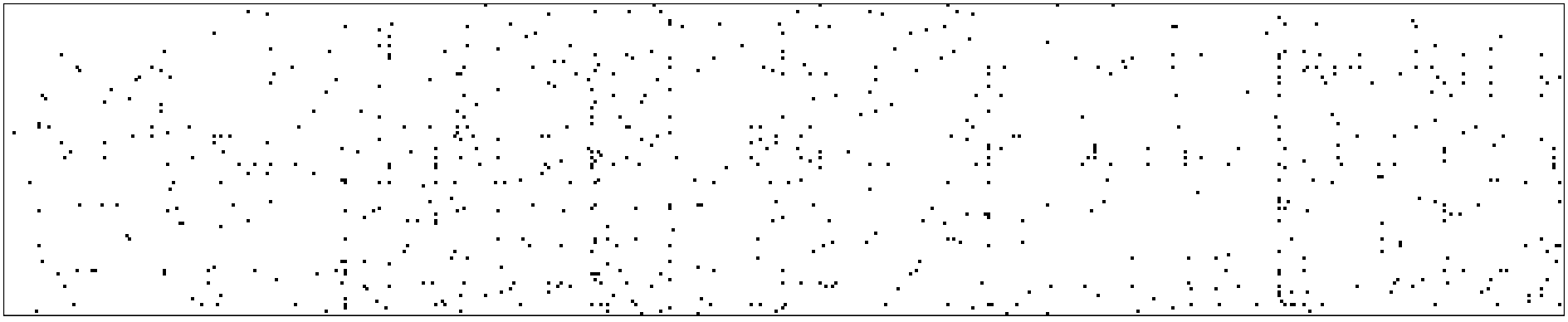
Survival prediction from raw mutation profiles

- Each patient is a binary vector: each gene is mutated (1) or not (0)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times





Approach: change representation?



Can we replace

$$x \in \{0,1\}^p \quad \text{with } p \text{ very large, very sparse}$$

by a representation with more information shared between samples

$$\Phi(x) \in \mathcal{H}$$

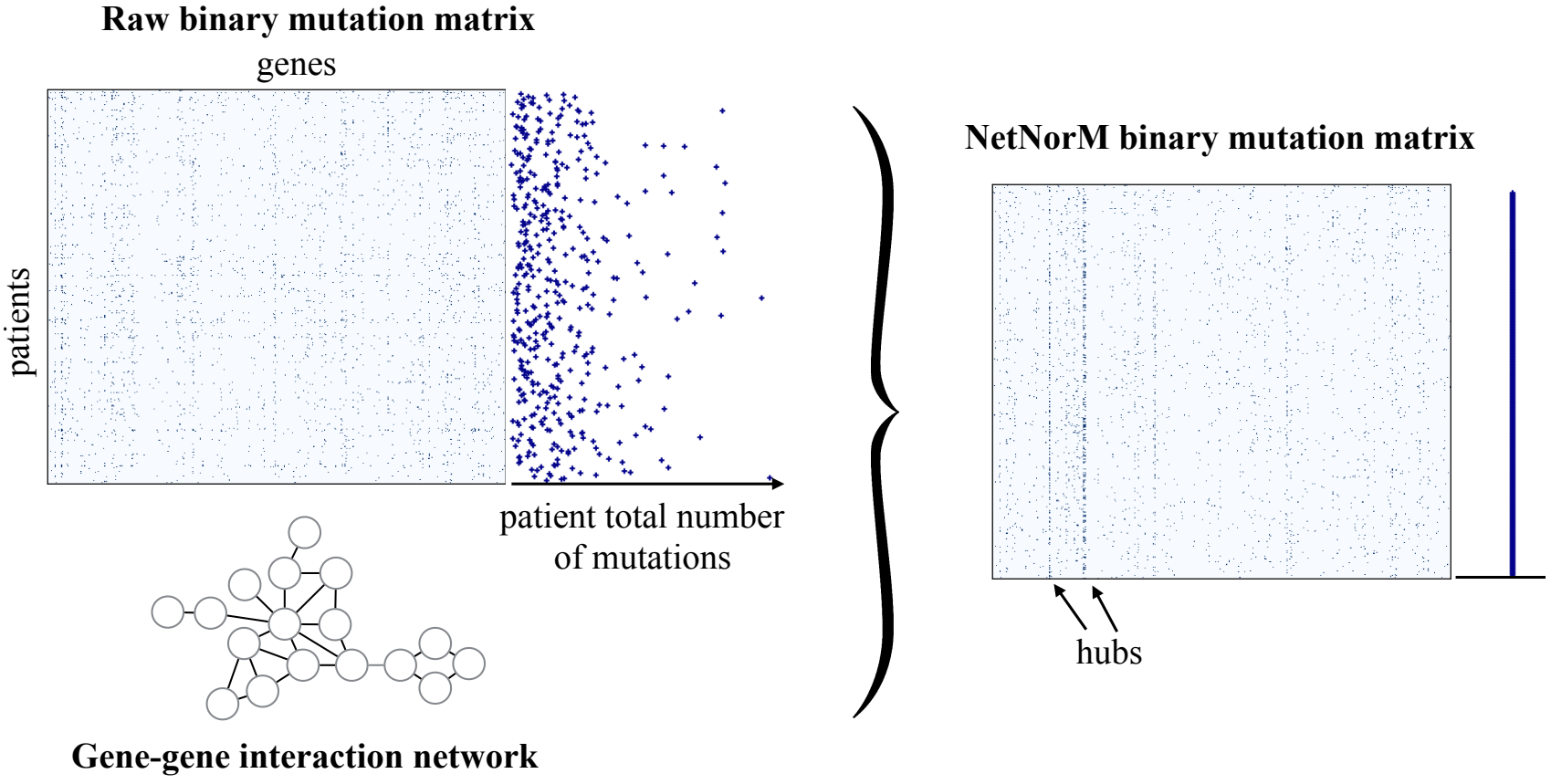
that would allow better supervised and unsupervised classification?

NetNorm Overview (Le Morvan et al., 2017)

Take

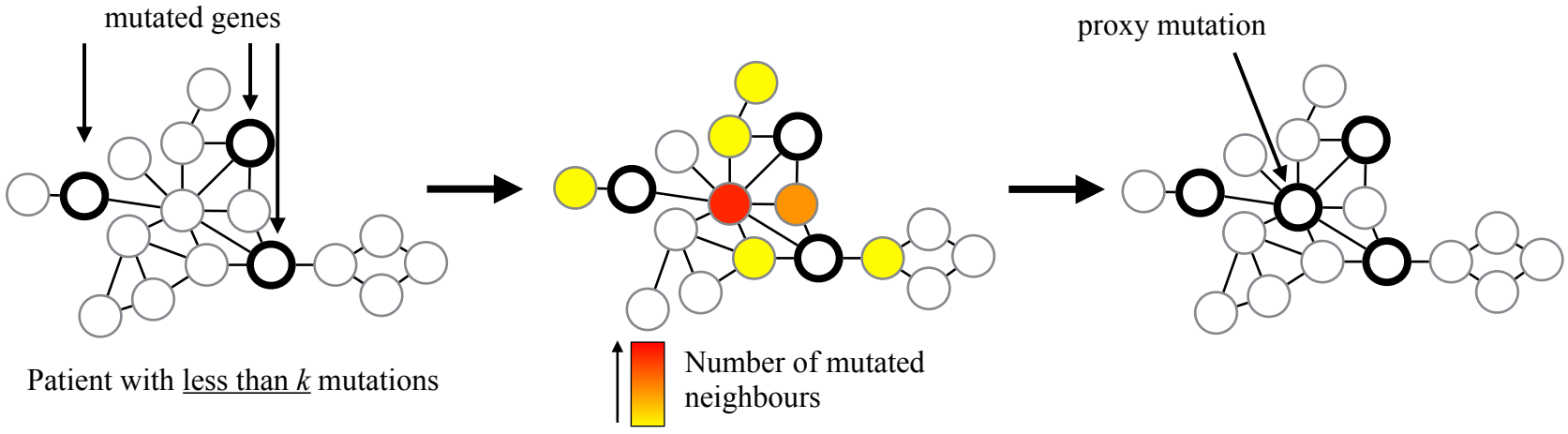
$$\mathcal{H} = \left\{ x \in \{0,1\}^p \, : \, \sum_{i=1}^{p} x_i = K \right\}$$

and use a gene network to transform x to $\phi(x) \in \mathcal{H}$ by adding/removing mutations

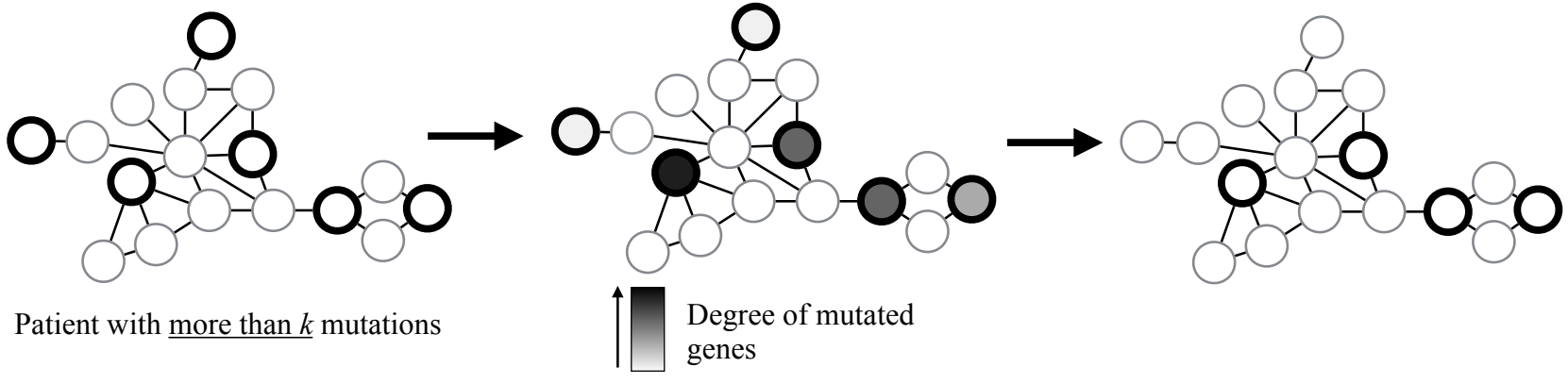


NetNorm detail (k=4)

1. Add mutations for patients with few (less than *K*) mutations



2. Remove mutations for patients for many (more than *K*) mutations



In practice, *K* is a free parameter optimized on the training set, typically a few 100's.

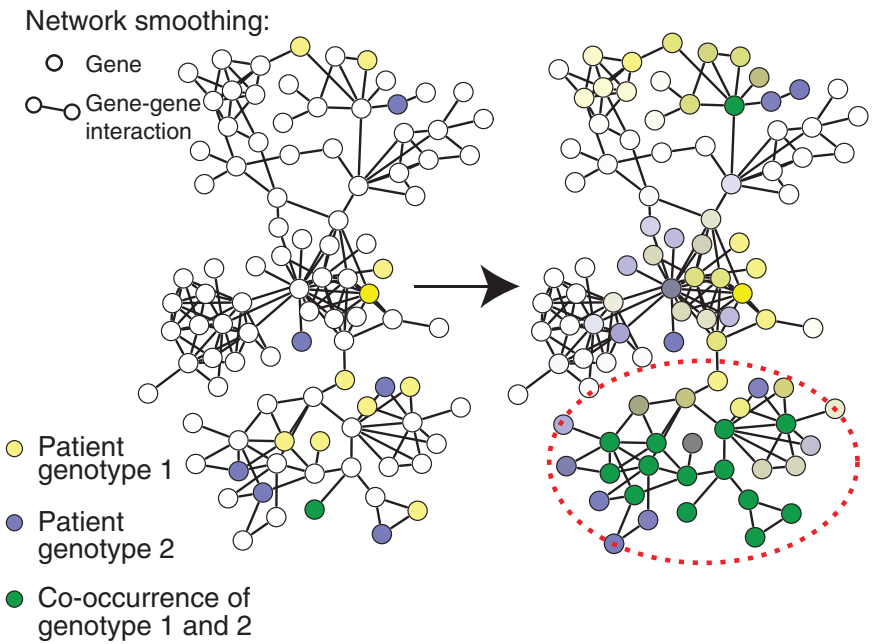
Related work (Hofree et al., 2013)

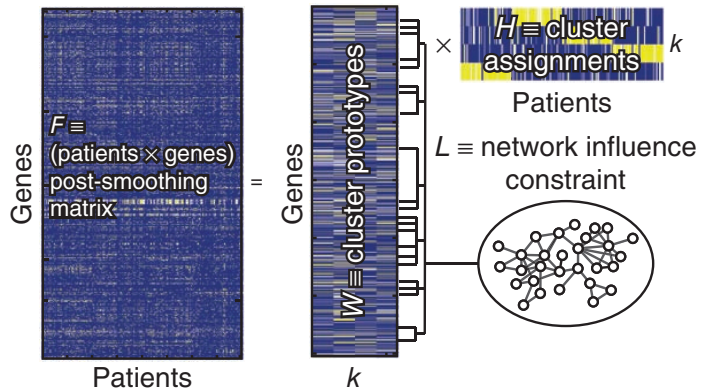
Network-based stratification of tumor mutations

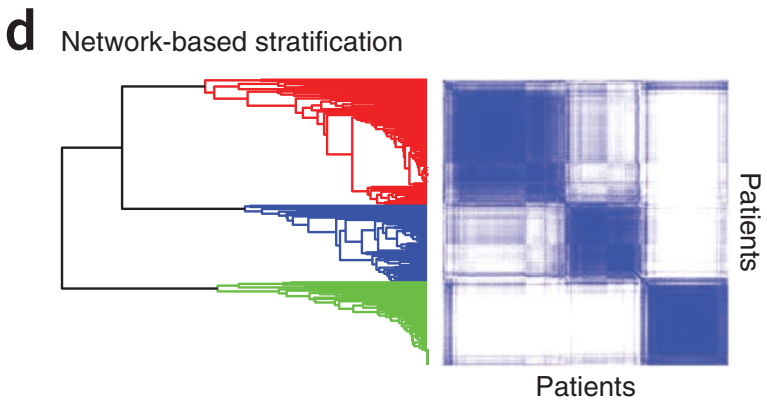
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. Correspondence should be addressed to T.I. (tideker@ucsd.edu).

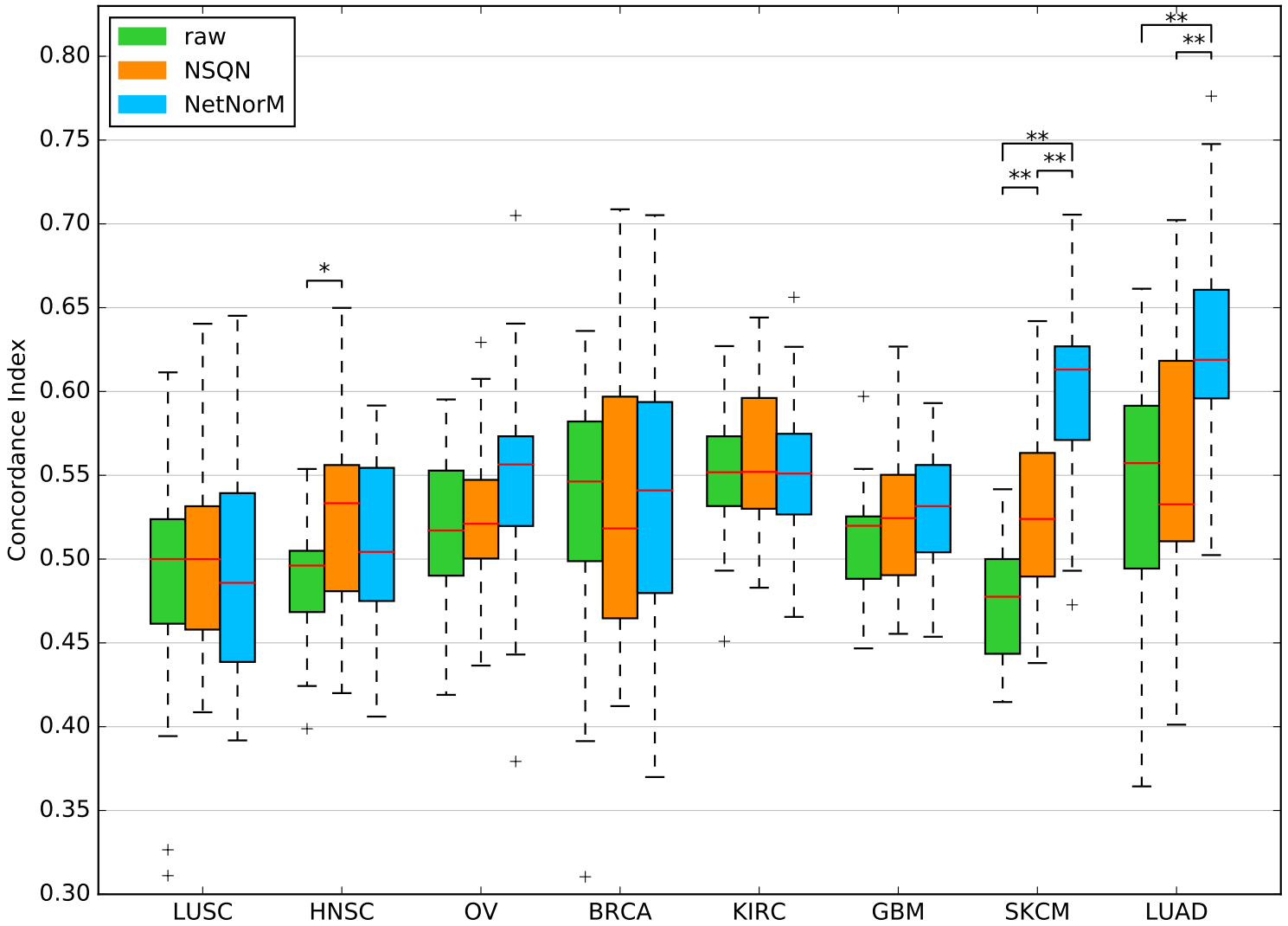
RECEIVED 14 FEBRUARY; ACCEPTED 12 AUGUST; PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2651







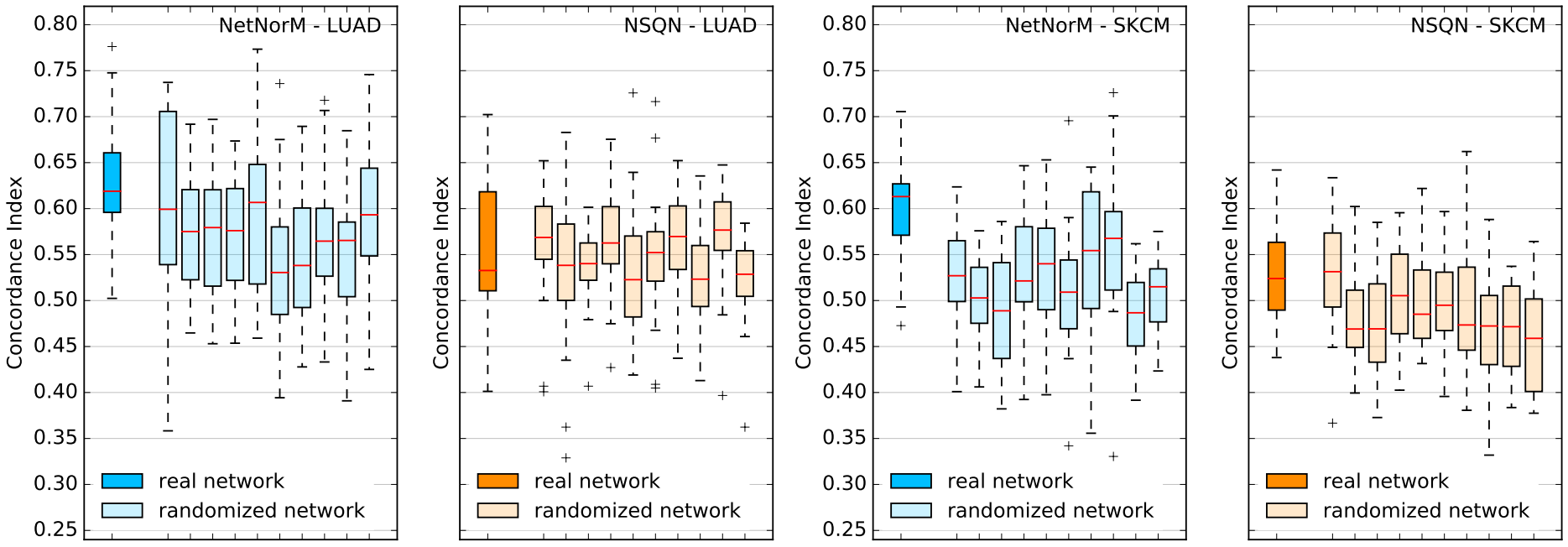
Results: survival prediction



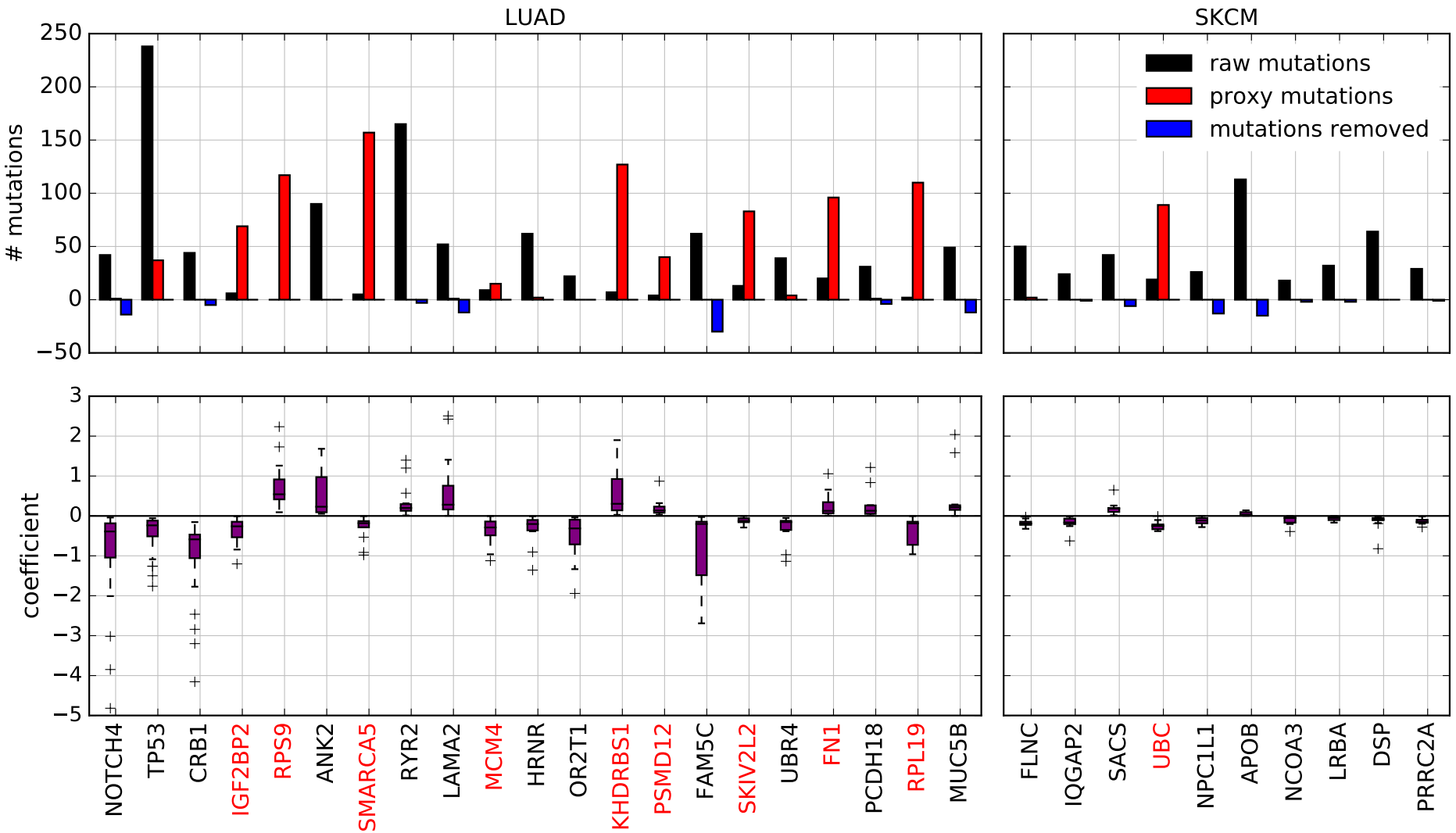
Use Pathway Commons as gene network.
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)

NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:

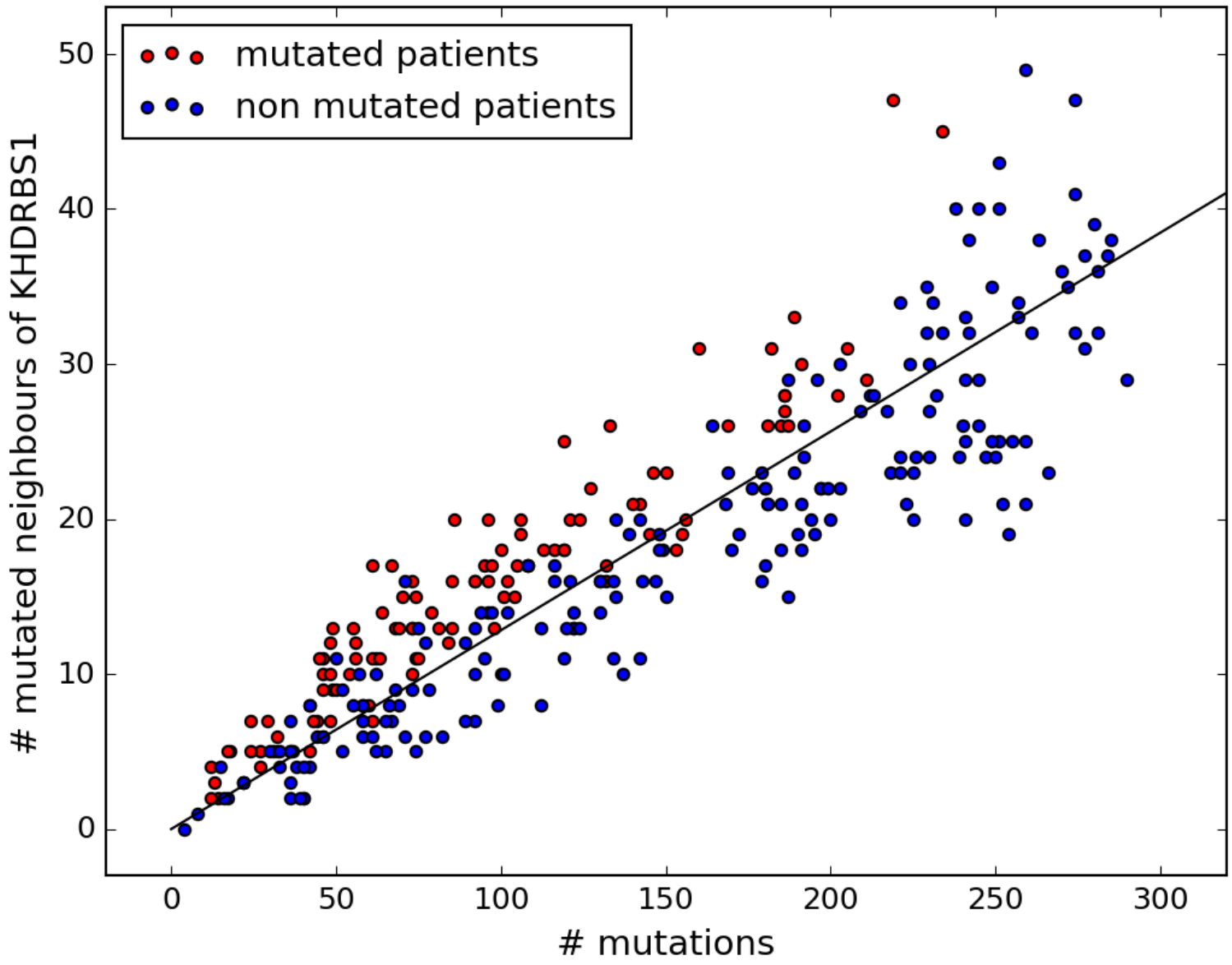


Selected genes represent "true" or "proxy" mutations

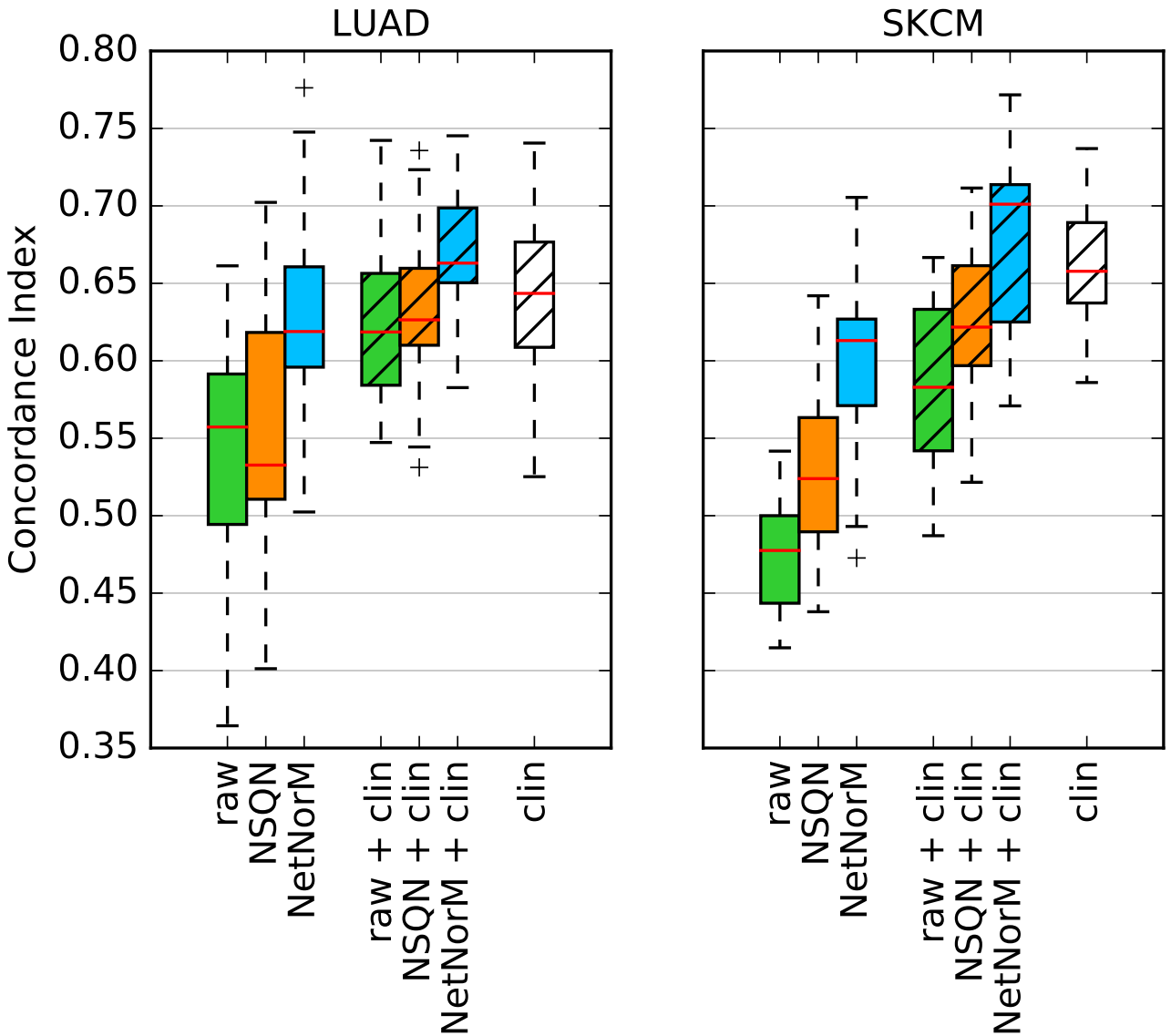


Genes selected in at least 50% of the cross-validated sparse SVM model

Proxy mutations encode both total number of mutations and local mutational burden

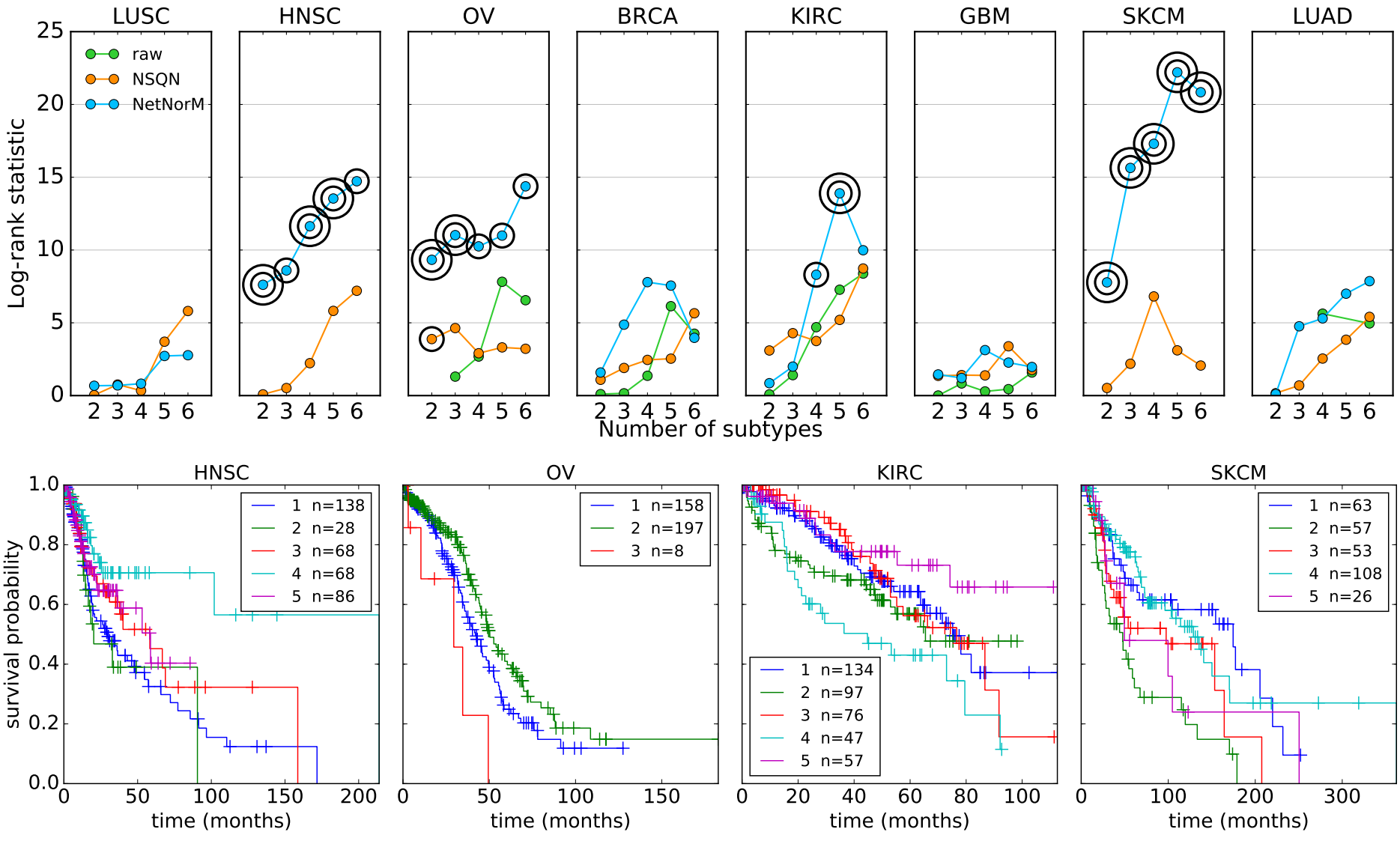


Adding good old clinical factors



Combination by averaging predictions

Performance on unsupervised patient stratification



Conclusion

- Somatic mutation profiles are challenging because
  - Little overlap between patients
  - Large variability in number of mutations
- Network smoothing / local averaging sometimes helps
  - but with current methods, looking at the direct neighbors is good enough
- Normalizing for total number of mutations is important
  - through QN or NetNorm, for example
  - this is not for biological reasons, but for mathematical reasons
  - Much room for improvement to find a good representation $\Phi(x)$
- Try it!
  - `https://github.com/marineLM/NetNorM`

Thanks

The Adolph C. and Mary Sprague Miller Institute for Basic Research in Science
University of California, Berkeley

References

M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nat Methods*, 10(11):1108–1115, Nov 2013. doi: 10.1038/nmeth.2651. URL http://dx.doi.org/10.1038/nmeth.2651.

Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR:W&CP*, pages 1935–1944, 2015. URL http://jmlr.org/proceedings/papers/v37/jiao15.html.

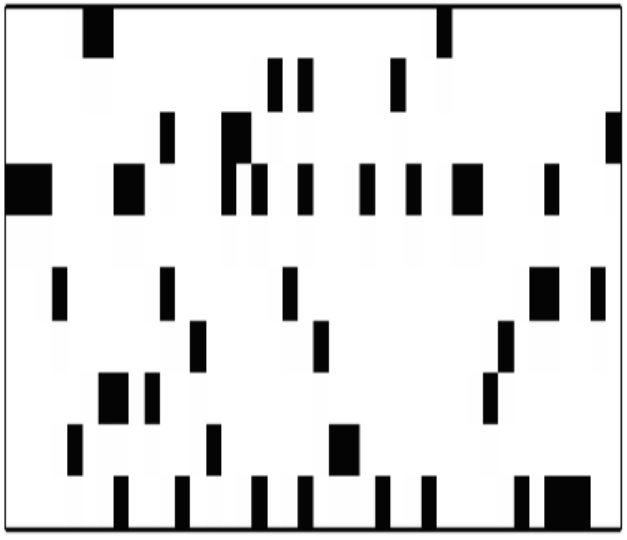
Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. doi: 10.1109/TPAMI.2017.2719680. URL http://dx.doi.org/10.1109/TPAMI.2017.2719680.

W. R. Knight. A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.*, 61(314):436–439, 1966. URL http://www.jstor.org/stable/2282833.

M. Le Morvan, A. Zinovyev, and J.-P. Vert. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comp. Bio.*, 13(6):e1005573, 2017. URL http://hal.archives-ouvertes.fr/hal-01341856.

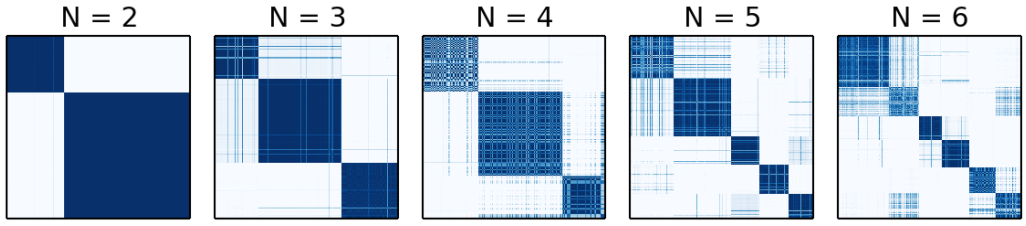
M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr 2009. doi: 10.1038/nature07943. URL http://dx.doi.org/10.1038/nature07943.

Patient stratification (unsupervised) from raw mutation profiles

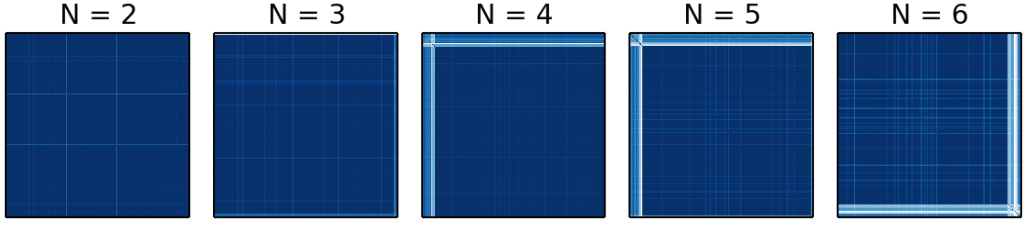


✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:



✓ Observed behaviour:

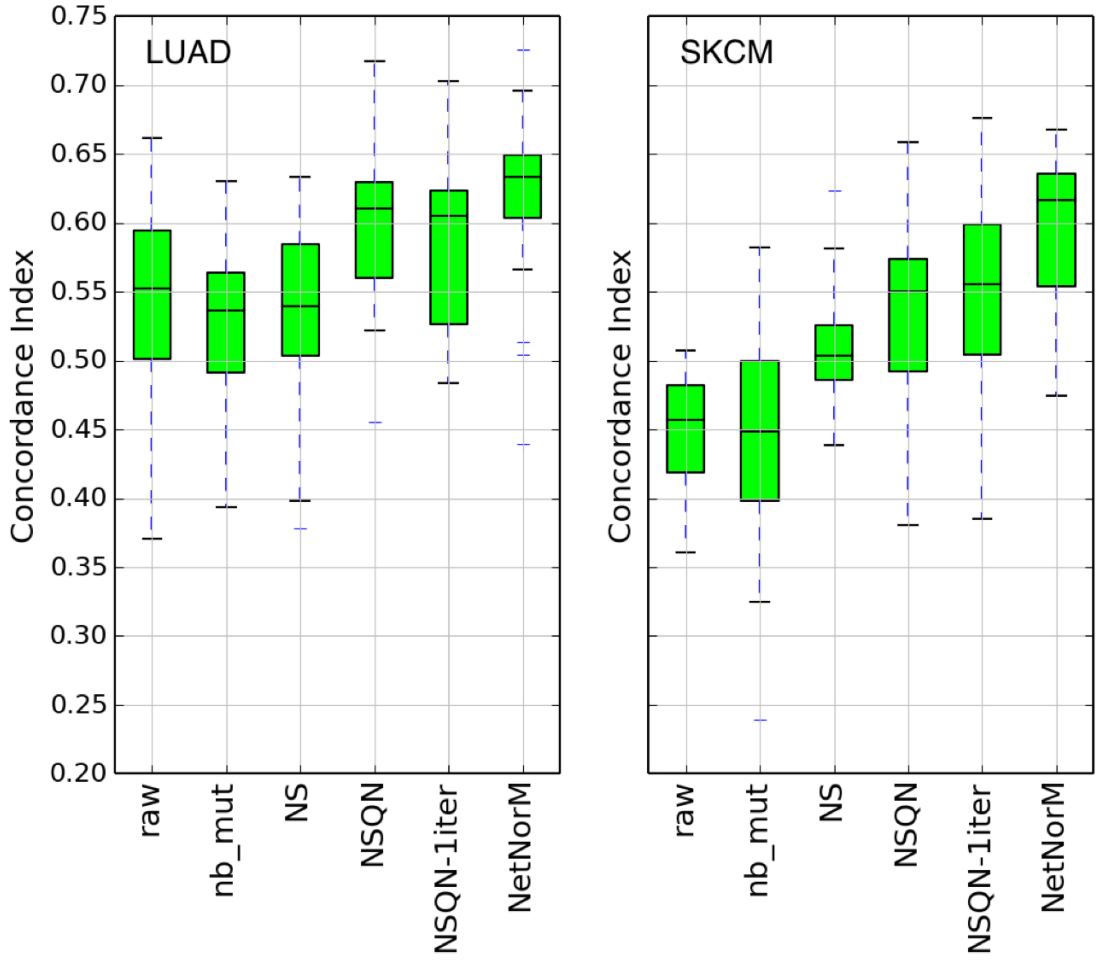


Patients share very few mutated genes!

QN matters...

Both NetNorm and NSQN transforms follow a 2-step a approach:

1. Smooth the raw data onto the gene network (NS)
2. Quantile normalize the smoothed profile (QN)



QN after network smoothing