Cancer stratification from mutation profiles

Jean-Philippe Vert

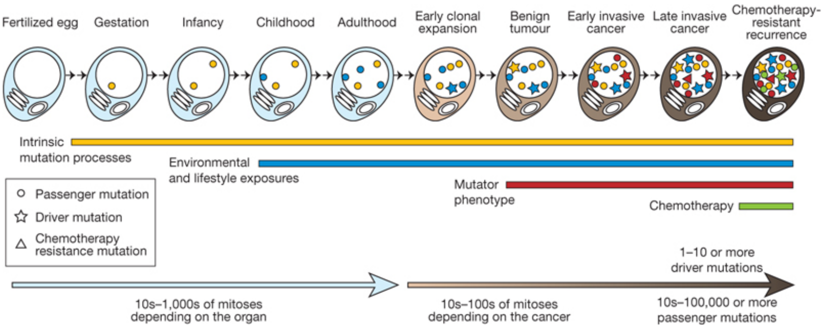
Banff, March 28, 2017

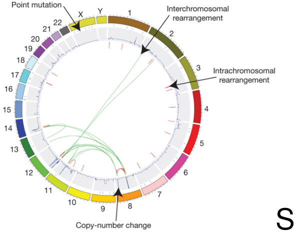
Joint work with

Marine Le Morvan

Andrei Zinovyev

Somatic mutations in cancer

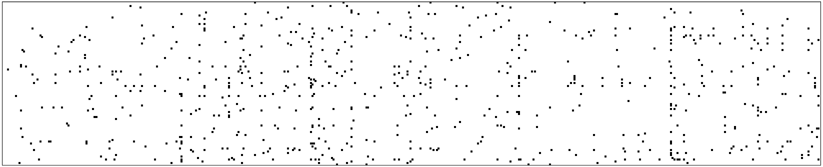

Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations



Stratton et al. (2009)

Large-scale efforts to collect somatic mutations

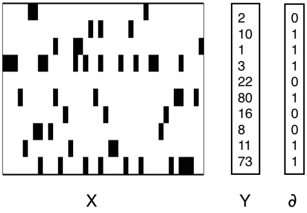
- 3, 378 samples with survival information from 8 cancer types
- downloaded from the TCGA / cBioPortal portals.

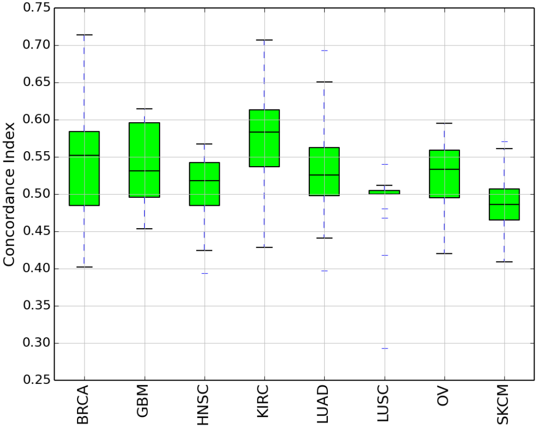


Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 463
GBM (Glioblastoma multiforme)	265	14 750
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 609
HNSC (Head and Neck squamous cell carcinoma)	388	17 022
LUSC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

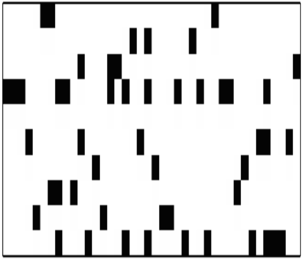
Survival prediction from raw mutation profiles

- Each patient is a binary vector: each gene is mutated (1) or not (2)
- Silent mutations are removed
- Survival model estimated with sparse survival SVM
- Results on 5-fold cross-validation repeated 4 times



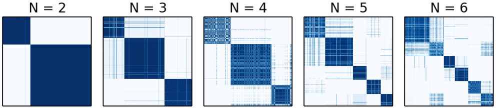


Patient stratification (unsupervised) from raw mutation profiles

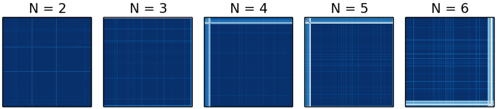


✓ Non-Negative matrix factorisation (NMF)

✓ Desired behaviour:

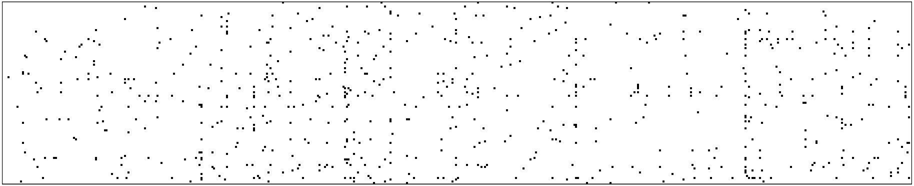


✓ Observed behaviour:



Patients share very few mutated genes!

Challenge



Can we replace

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

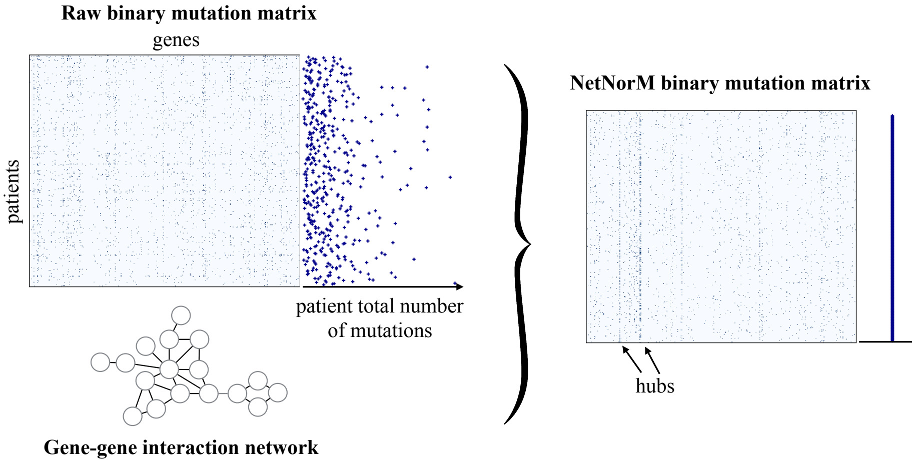
by a representation with more information shared between samples

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

that would allow better supervised and unsupervised classification?

NetNorm Overview (Le Morvan et al., 2016)

- Modify the binary vector $x \in \{0,1\}^p$ of each patient by adding or removing mutations, using a gene network as prior knowledge
- After Netnorm, all patients $\Phi(x) \in \{0,1\}^p$ have the same number of (pseudo-)mutations



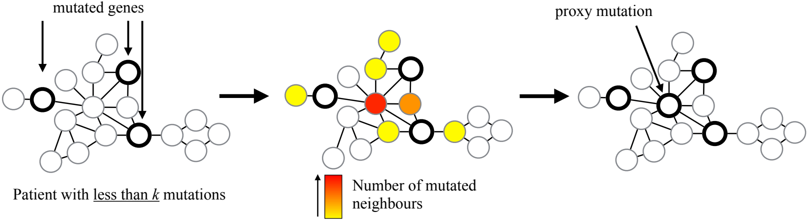
Raw binary mutation matrix

NetNorM binary mutation matrix

Gene-gene interaction network

NetNorm detail (k=4)

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



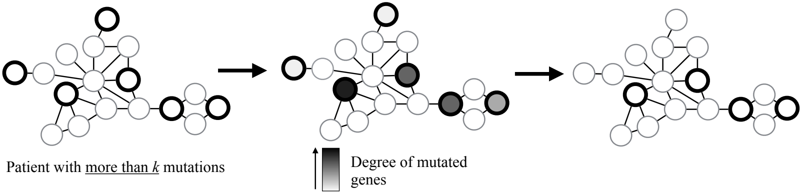
mutated genes

proxy mutation

Patient with <u>less than *k*</u> mutations

Number of mutated neighbours

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

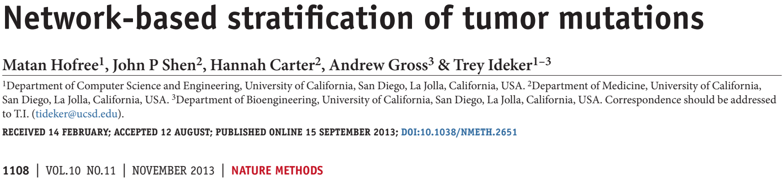


Patient with <u>more than *k*</u> mutations

Degree of mutated genes

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

Related work (Hofree et al., 2013)



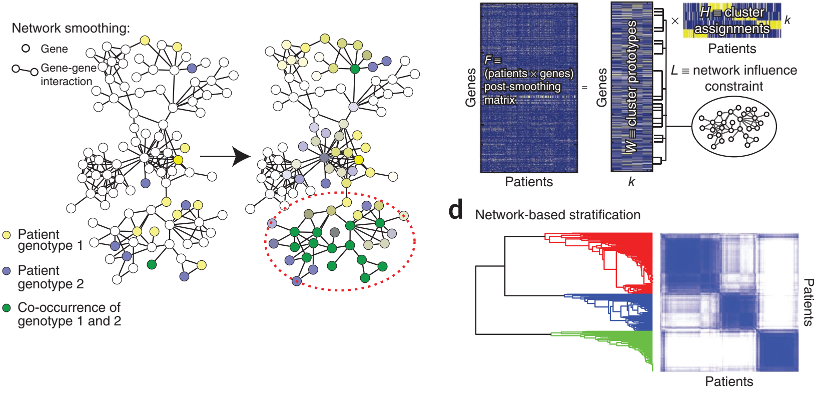
Network-based stratification of tumor mutations

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

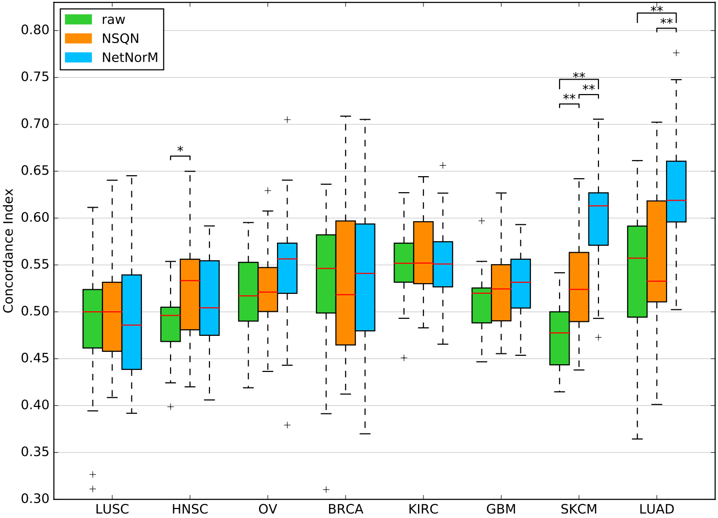
[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Department of Medicine, University of California, San Diego, La Jolla, California, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. 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

1108 | VOL.10 NO.11 | NOVEMBER 2013 | NATURE METHODS



Performance on survival prediction

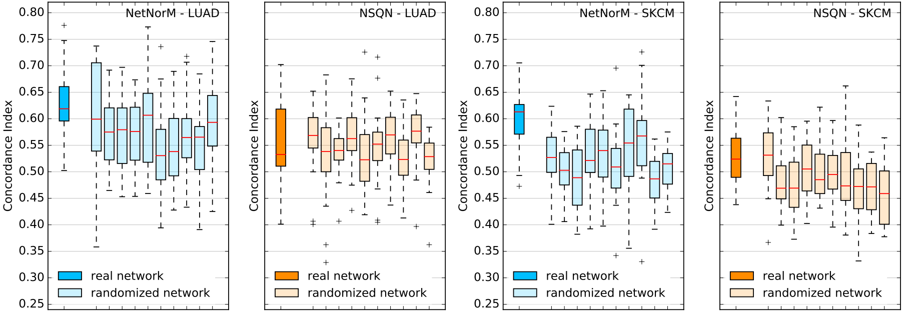


Use Pathway Commons as gene network.

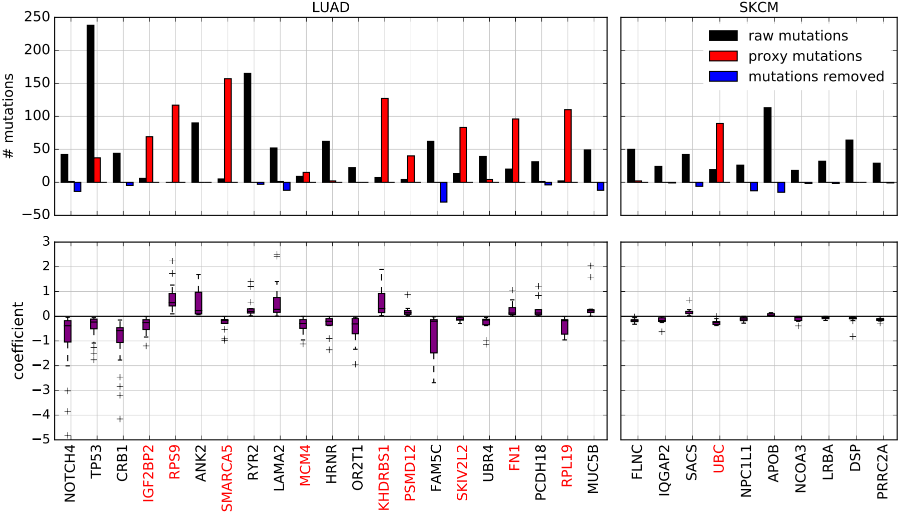
NSQN = Network Smoothing / Quantile Normalization (Hofree et al., 2013)

NetNorM and NSQN benefit from biological information in the gene network

Comparison with 10 randomly permuted networks:

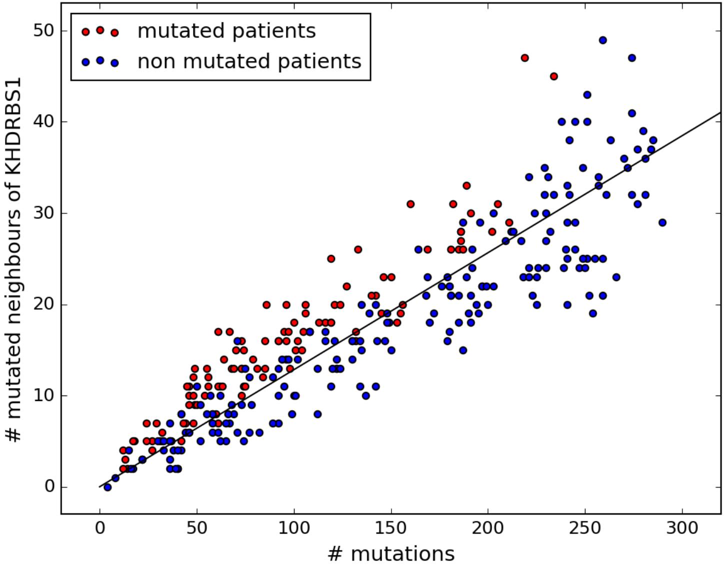


Selected genes represent "true" or "proxy" mutations

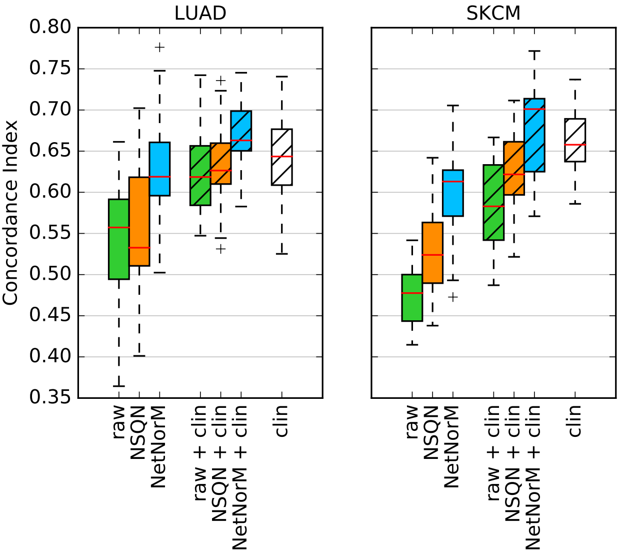


Genes selected in at least 50% of the cross-validated sparse SVM model

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

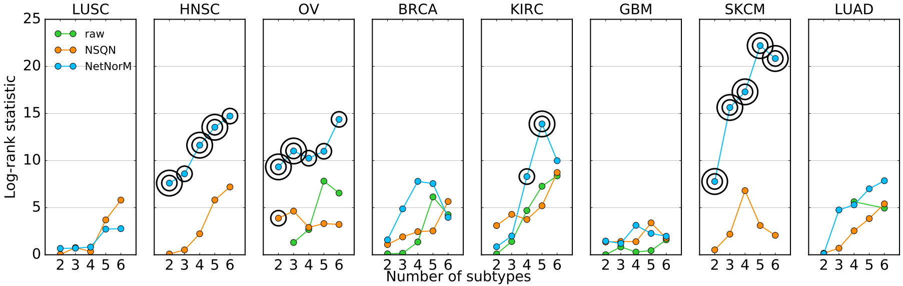


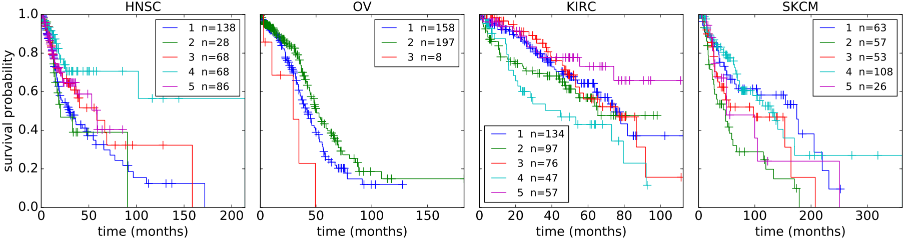
Adding good old clinical factors



Combination by averaging predictions

Performance on unsupervised patient stratification





Summary

- 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 $\Phi(x)$
- References
  - `https://hal.archives-ouvertes.fr/hal-01341856`
  - `https://github.com/marineLM/NetNorM`

Thanks

erc

SEVENTH FRAMEWORK PROGRAMME

INVESTISSEMENTS D'AVENIR

AGENCE NATIONALE DE LA RECHERCHE ANR

BIOMÉRIEUX

MINES ParisTech

institutCurie
Together, let's beat cancer.

Inserm
Institut national
de la santé et de la recherche médicale

PSL
RESEARCH UNIVERSITY

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

SIMONS INSTITUTE
for the Theory of Computing

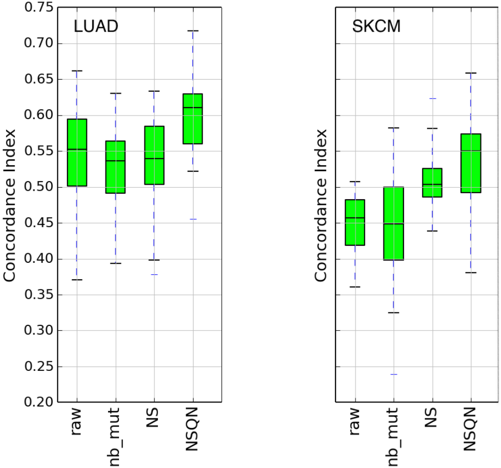
References

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.

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. Technical Report 01341856, HAL, 2016. URL http://hal.archives-ouvertes.fr/hal-01341856.

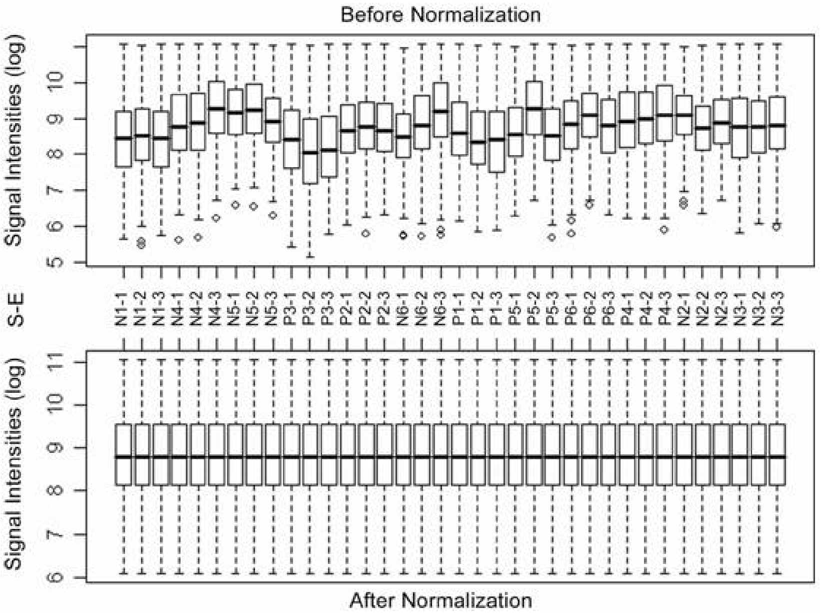
M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239): 719–724, Apr 2009. doi: 10.1038/nature07943. URL http://dx.doi.org/10.1038/nature07943.

NBS representation helps to predict survival

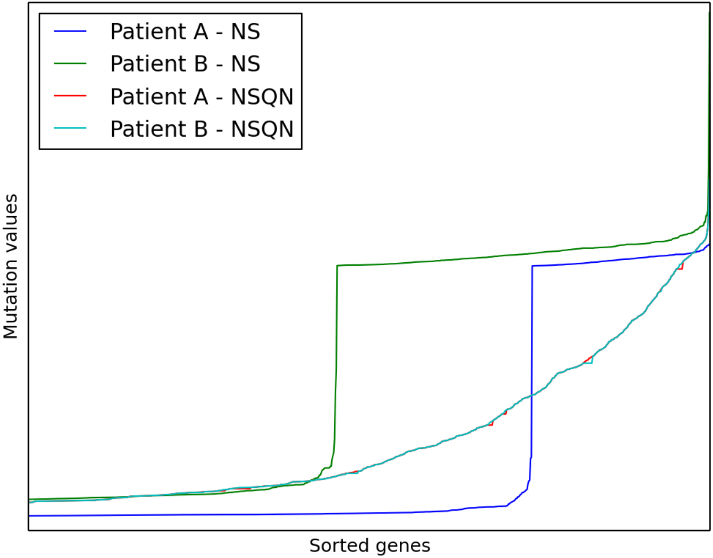


- NS = Network Smoothing
- QN = Quantile normalization
- NBS = NS+QN

What is QN?


Before Normalization
Signal Intensities (log)
5
6
7
8
9
10
S-E
N1-1
N1-2
N1-3
N4-1
N4-2
N4-3
N5-1
N5-2
N5-3
P3-1
P3-2
P3-3
P2-1
P2-2
P2-3
N6-1
N6-2
N6-3
P1-1
P1-2
P1-3
P5-1
P5-2
P5-3
P6-1
P6-2
P6-3
P4-1
P4-2
P4-3
N2-1
N2-2
N2-3
N3-1
N3-2
N3-3
Signal Intensities (log)
6
7
8
9
10
11
After Normalization

QN after network smoothing


Patient A - NS
Patient B - NS
Patient A - NSQN
Patient B - NSQN
Mutation values
Sorted genes