Cancer stratification and prognosis from mutations using gene networks

Marine Le Morvan
Advised by **Jean-Philippe Vert** & **Andrei Zinoyev**

CBIO - Mines Paristech, INSERM U900 - Curie institute, Paris, France

Novembre 26^{th}, 2019

Cancer genomics

- Genome sequencing
- Mutations in cancer
- Towards precision medicine

Cancer stratification and survival prediction from mutation data

A word on my current work

Outline

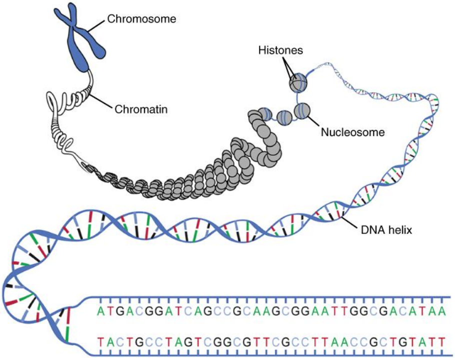
Cancer genomics

- Genome sequencing
- Mutations in cancer
- Towards precision medicine

Cancer stratification and survival prediction from mutation data

A word on my current work

The human genome



- Human DNA is contained into 22 pairs of chromosomes, plus X and Y.
- Each chromosome is a long molecule of DNA.
- A total of 3 billion nucleotides A, T, C, G (*Les Misérables* - V. Hugo - $\times$1000).
- It is estimated that the human genome contains 20000-25000 protein coding genes.

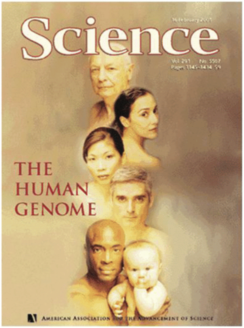
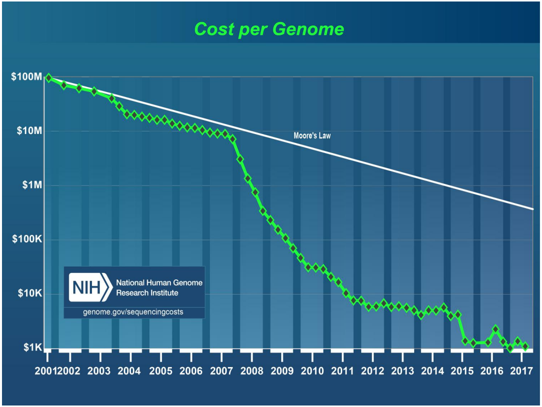


Figure: Covers of Science and Nature in 2001 announcing that the human genome has been sequenced (almost completely) for the first time.

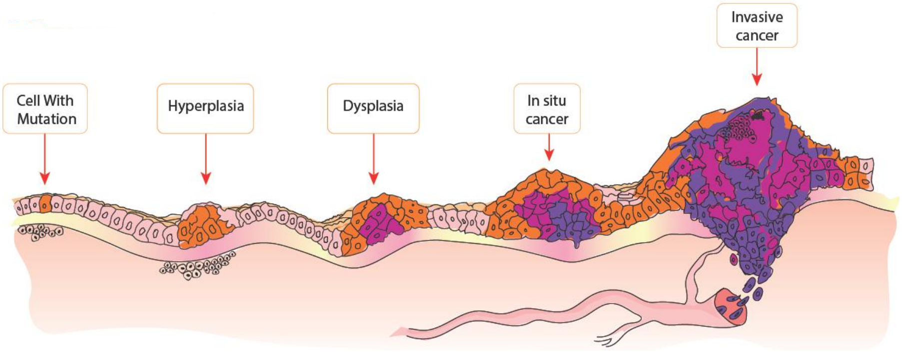
- The Human Genome Project cost 3 billion dollars over the period 1990-2003.
- It was the first *reference* genome for *Homo Sapiens*, assembled from the genomes of a few donors.

Sequencing cost for a whole genome



- Sequencing costs have plummeted since 2007 thanks to the advent of next-generation sequencing.
- In 2019, a whole genome can be sequenced in a day for around 1000$, and less for whole exome (i.e only genes).

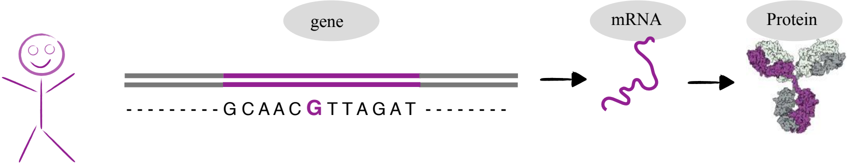
Cancer



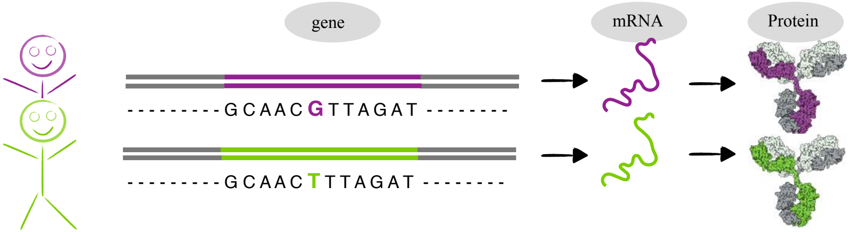
soure: https://strandls.com/what-is-cancer

- The mechanisms that lead to the onset cancer are not fully understood yet.
- The disease is driven by genetic alterations in cancer cells that induce uncontrolled cell proliferation.
- Disease alterations mean point mutations, insertions, deletions, copy number variations, methylation changes, ...

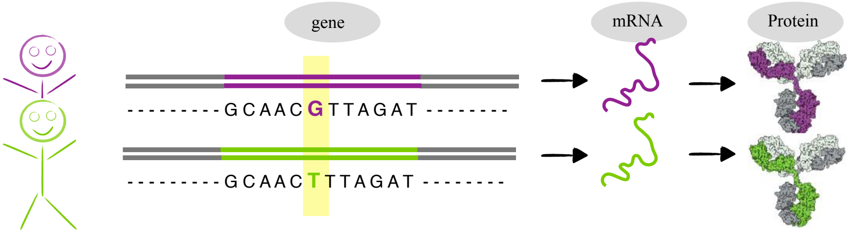
Nucleotide variations and mutations



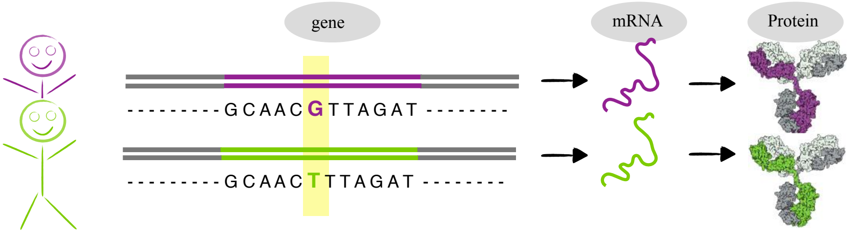
Nucleotide variations and mutations



Nucleotide variations and mutations



Nucleotide variations and mutations



- Single Nucleotide Polymorphism (SNP):
  - ✓ Variation compared to a reference genome in typically $>1\%$ of the population.
- Germline mutation:
  - ✓ Variation compared to a reference genome in typically $<1\%$ of the population.
- Somatic mutation:
  - ✓ Variation compared to one's germline cells. Appears during one's lifetime and is not present in all cells.
  - ✓ Somatic mutations play an important role in the onset of many cancers.

Driver mutations

- Somatic mutations naturally occur in a lifetime and accumulate with age.
- The number of mutations in protein coding genes widely varies across cancers, from a 10s to 1000s.
- Recent studies have estimated that cancer cells have on average between 1 and 10 driver mutations depending on cancer types.
- A central topic in cancer research: distinguish driver from passenger mutations (Proto-oncogenes and tumor-supressor genes).

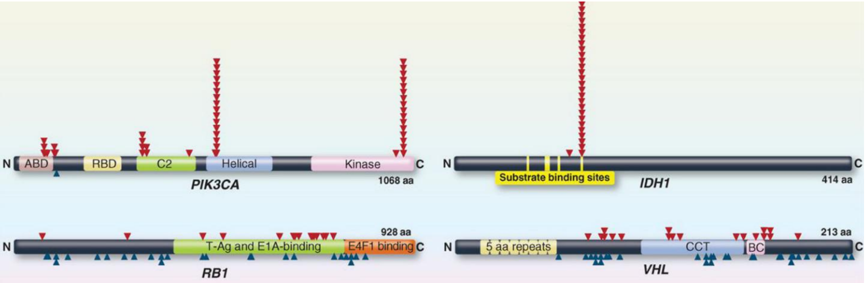
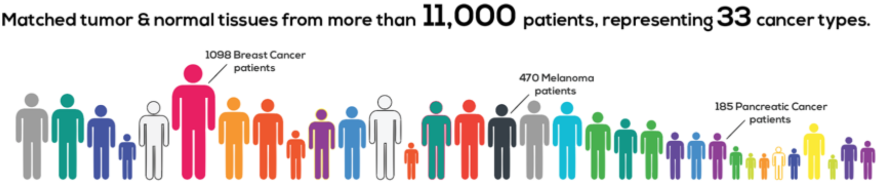


Figure: Vogelstein et al., 2013

The Cancer Genome Atlas (TCGA)



- Large scale tumor sequencing efforts from 2006 to 2018.
- Provide in particular somatic mutation data, as well as patients clinical records.

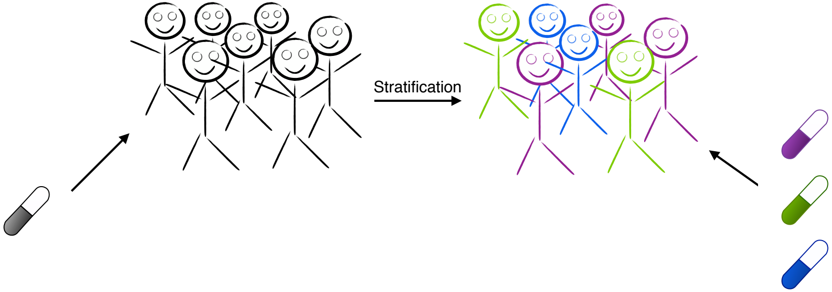
Precision medicine

Precision medicine:

- aims at integrating the genetic specificities of an individual with its conventional medical record to adapt treatment, or prevention strategies.

Examples of research questions:

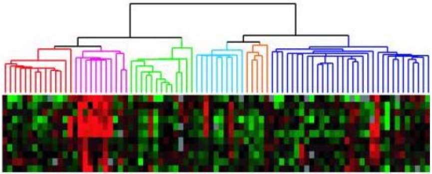
- Patient stratification
- Prediction of survival, relapse, metastasis, drug toxicity, drug resistance, ...



Patient stratification: one example

Cancers are usually classified by tissue of origin (breast, lung, ...). However, it slowly evolves towards a classification based on molecular descriptors.

- Within tumor type heterogeneity: example of breast cancer.



Basal-like: PARP inhibitors

ERBB2+: HER2-targeted drugs

Normal Breast-like

Luminal Subtype C

Luminal Subtype B: Hormonal manipulation, HER2-targeted drudge CDK inhibitors

Luminal Subtype A: Hormonal manipulation, CDK inhibitors

Figure: Breast cancer patient stratification obtained via unsupervised clustering of gene expression profiles. Figure adapted from Perou et al., with treatment information from Jeanne De Lartigue.

- Recent studies have also highlighted shared alterations across cancer types (drug repurposing).

Outline

Cancer genomics

- Genome sequencing
- Mutations in cancer
- Towards precision medicine

Cancer stratification and survival prediction from mutation data

A word on my current work

NetNorM

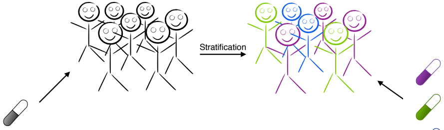
PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

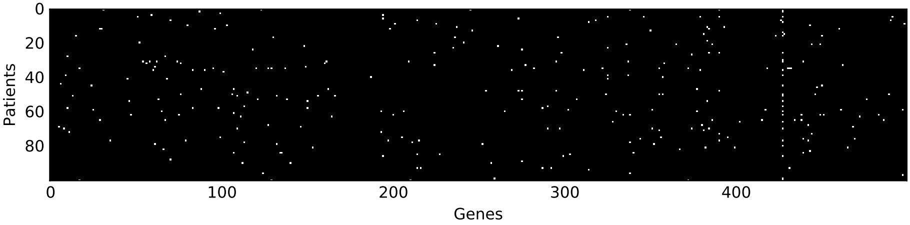
NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis

Marine Le Morvan[1,2,3], Andrei Zinovyev[1,2,3], Jean-Philippe Vert[1,2,3,4]*

- A new representation of somatic mutation profiles,
- based on gene networks,
- to improve patients stratification and survival prediction.



Somatic mutations



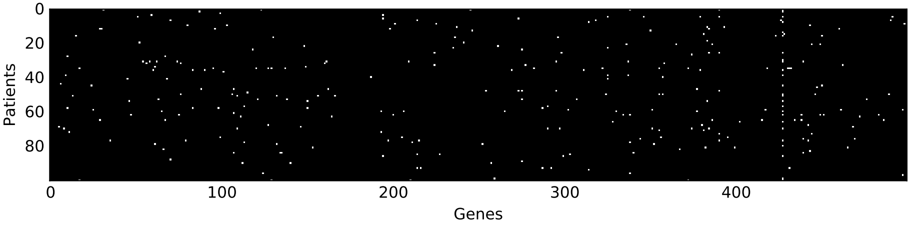
The raw data:

- Binary mutation profiles where a 1 stands for the presence of one (or more) mutation in a given gene for a given patient
- yield poor survival prediction performances,
- are not well suited for patient stratification.

Challenges:

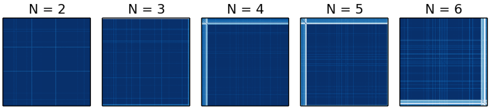
- High dimension (around $\approx 20,000$ genes).
- Low mutation frequency.
- Patients share few mutations in common.

Somatic mutations

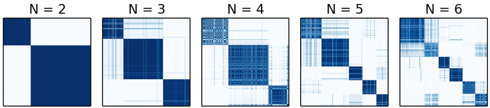


- Patient stratification:
  - ✓ Non-negative Matrix factorisation (NMF)
  - ✓ Consensus clustering

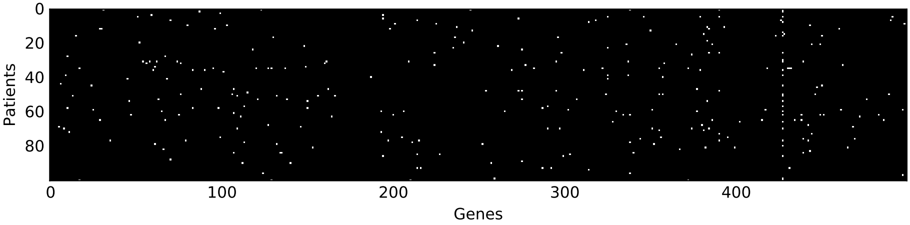
- Observed behaviour:



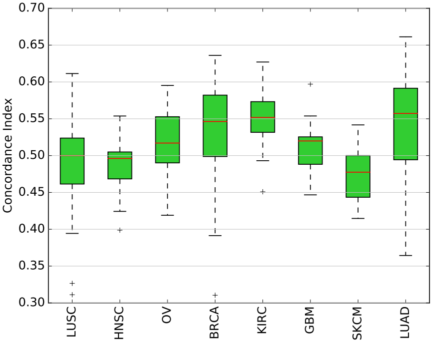
- Desired behaviour:



Somatic mutations

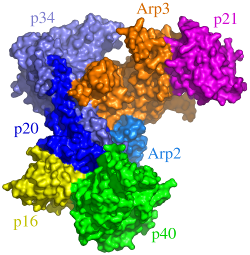


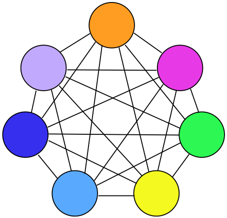
- Survival prediction:
  - ✓ Sparse survival SVM [VanBelle]
  - ✓ 4 × 5-fold cross-validation



Gene-gene interaction networks

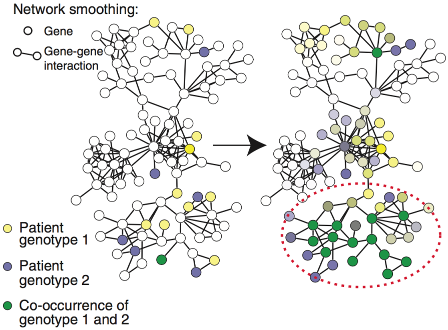
- An idea is to use protein-protein interaction networks to create an overlap between patients.
- Many types of interactions recorded:
  - ✓ Complexes and physical interactions
  - ✓ Biochemical reactions (phosphorylation, . . .)
  - ✓ Catalysis
  - ✓ Regulatory interactions
  - ✓ . . .





- Hypothesis: if two mutations in different genes are close on the gene network, they may cause similar downstream effects.

Previous work: Network-based stratification (NBS)



[Hofree2013Network]

- **Assumption**

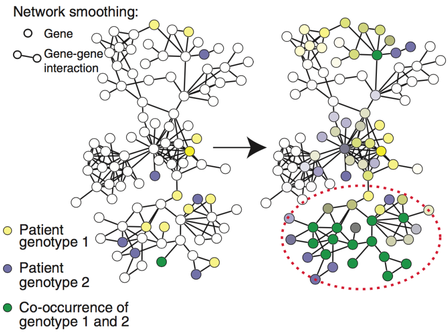
  Even if two tumors have no mutations in common, the same subnetworks may be affected.

- **Method**

  1. Network smoothing.
     Diffusion process.
     Each mutation profile (row of the mutation matrix) is smoothed independently.

  3. Non-Negative matrix factorisation (NMF).

Previous work: Network-based stratification (NBS)



[Hofree2013Network]

- **Assumption**

  Even if two tumors have no mutations in common, the same subnetworks may be affected.

- **Method**

  1. Network smoothing.
     Diffusion process.
     Each mutation profile (row of the mutation matrix) is smoothed independently.
  2. Quantile normalisation (QN)
     The i^{th} smallest value of all samples (patients) is set to the median of all i^{th} smallest values across samples.
  3. Non-Negative matrix factorisation (NMF).

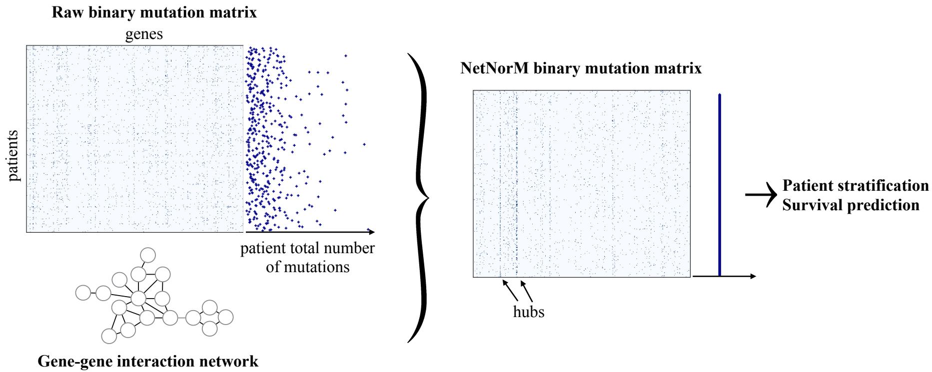
Overview of NetNorM - 1/2

- Quantile normalisation:
  - ✓ has no obvious biological motivation.
  - ✓ it modifies the smoothed mutation profiles so that the interpretation in terms of shared mutated subnetworks is not so straightforward after QN.
  - ✓ QN is crucial for NBS to work

- We propose NetNorM a new representation of mutation profiles:
  - ✓ inspired from the crucial role of QN in NBS,
  - ✓ and try to identify and predictive signals created.

- We compare the different representations of mutations (raw binary, NBS, NetNorM) for two tasks:
  - ✓ survival prediction,
  - ✓ patient stratification.

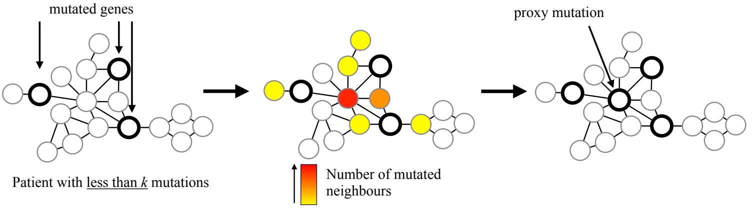
Overview of NetNorM - 1/2



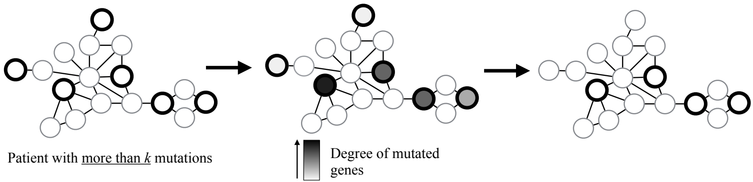
NetNorM replaces $\boldsymbol{x} \in \{0,1\}^p$ by a representation with more information shared between samples $\phi(\boldsymbol{x}) \in \mathcal{H}$ where $\mathcal{H} = \left\{ \boldsymbol{x} \in \{0,1\}^p : \sum_{i=1}^{p} \boldsymbol{x}_i = k \right\}$ and relies on a gene network to remove/add mutations. k is a parameter chosen by cross-validation.

Toy example with $k = 4$: (in reality, k is around of few 10s to a few 100s)

1. **Add** mutations to patients with fewer than k mutations.



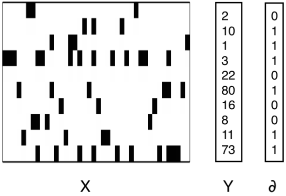
2. **Remove** mutations from patients with more than k mutations.



Large-scale efforts to collect exome somatic mutation profiles

Data used in this study:

- 3,378 samples with survival information (somatic mutations in exomes - silent mutations removed)
- from 8 cancer types
- downloaded from TCGA and cBioPortal.



✓ $\boldsymbol{X}$: mutation matrix

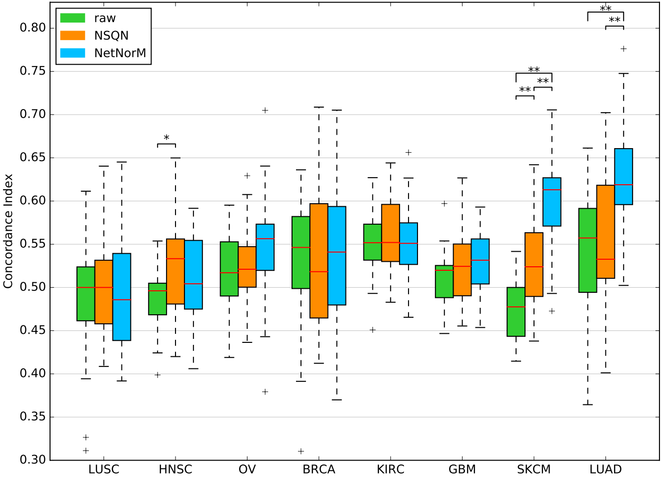
✓ $\boldsymbol{y}$: months of survival since diagnosis

✓ $\boldsymbol{\delta}$: censoring status (1: deceased, 0: alive)

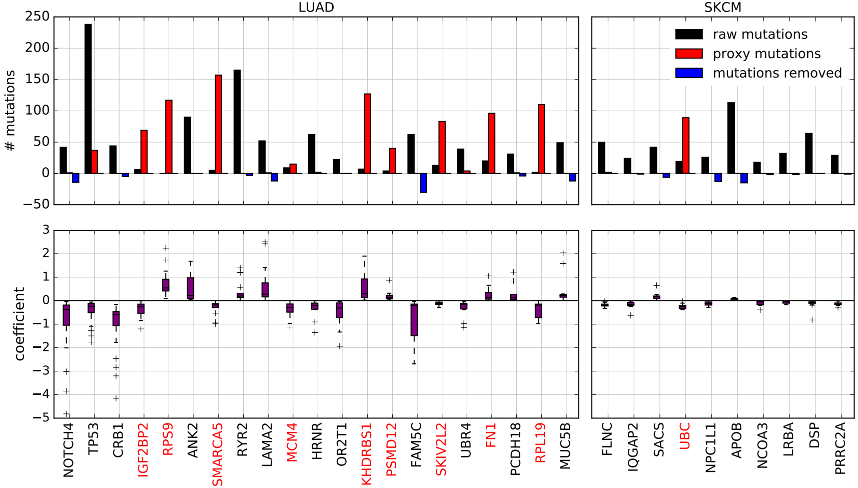
Cancer type	Patients	Genes
LUAD (Lung adenocarcinoma)	430	20 596
SKCM (Skin cutaneous melanoma)	307	17 461
GBM (Glioblastoma multiform)	265	14 748
BRCA (Breast invasive carcinoma)	945	16 806
KIRC (Kidney renal clear cell carcinoma)	411	10 608
HNSC (Head & Neck squam. cell carcinoma)	388	17 022
LUSC (Lung squamous cell carcinoma)	169	13 589
OV (Ovarian serous cystadenocarcinoma)	363	10 192

Comparison of survival prediction performances

- ✓ We assume $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$
- ✓ Sparse survival SVM [VanBelle]
- ✓ 4×5-fold cross-validation
- ✓ Gene network: Pathway Commons.

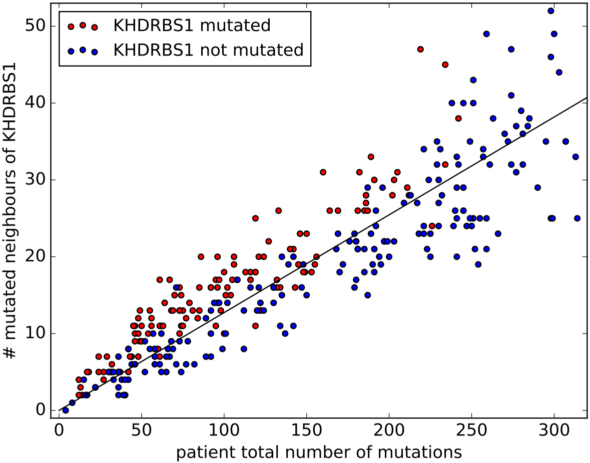


Genes frequently selected in survival prediction models



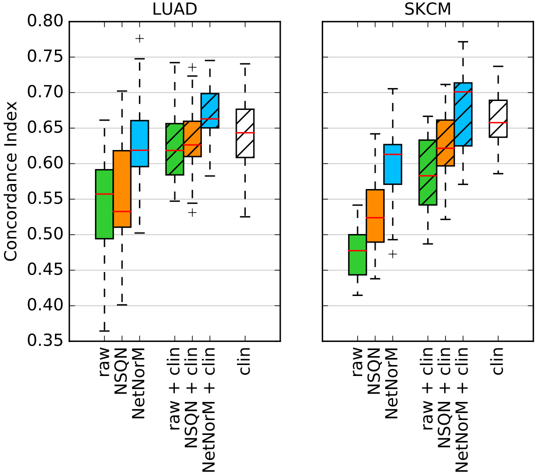
Genes selected at least 10 times out of 20 folds

Proxy mutations encode local and global mutational burden

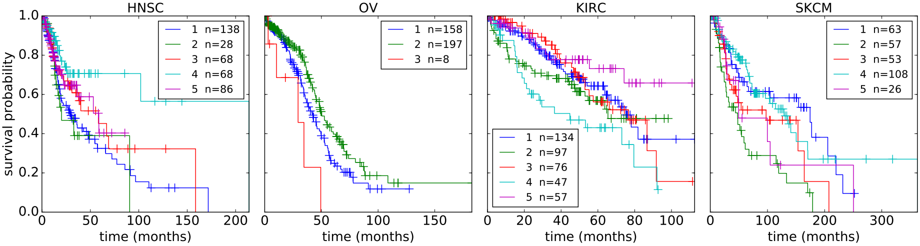


Mutations in KHDRBS1 are almost only proxy mutations.

- ✓ Models are learned on mutations and clinical data separately and subsequently averaged.
- ✓ Clinical data alone outperforms mutation data alone.
- ✓ There is information in mutation data, as captured by NetNorM, that allows to improve on clinical data alone.



- Unsupervised patient stratification:
  - ✓ With NMF + consensus clustering.
  - ✓ Number of clusters tested vary from 2 to 6.
  - ✓ The logrank test (case > 2 subgroups) tests whether or not there is at least one subgroup whose survival distribution is different from the others.



Conclusion

- Somatic mutation profiles are challenging because:
  - ✓ Low mutation frequency.
  - ✓ Few shared mutations among patients.
  - ✓ Large variability in the total number of mutations.
- Network smoothing/local averaging sometimes helps
  - ✓ but with current methods, looking at direct neighbours is good enough.
- Normalising for the total number of mutations is important
  - ✓ with NSQN or NetNorM.
  - ✓ NetNorM creates a signal related to local and global mutational burden.

Outline

Cancer genomics

- Genome sequencing
- Mutations in cancer
- Towards precision medicine

Cancer stratification and survival prediction from mutation data

A word on my current work

Supervised learning with missing values

Missing values are ubiquitous in various fields/experiments: electronic health records, polls, sensor data, . . .

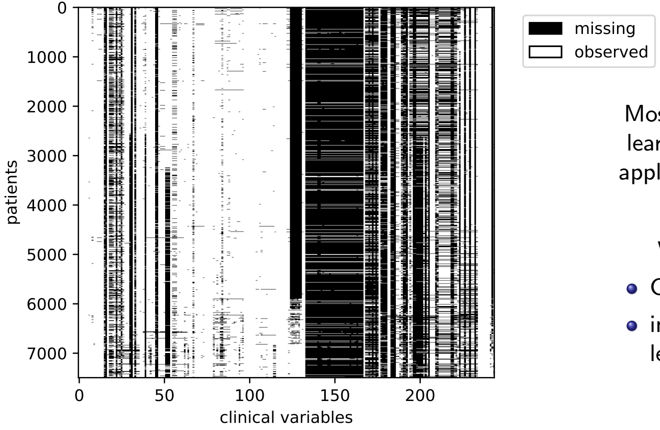


Figure: Traumabase clinical health records.

Most off-the-shelf machine learning models cannot be applied with missing values.

What can be done:

- Complete-case analysis?
- imputation prior to learning?

Linear regression with missing values

- Setting:
  - Linear regression model of the complete data X:

$$Y = \sum_{j=1}^{d} \beta_j X_j$$

  - We aim to find a predictor $\hat{f}_n$ which minimizes the least squares loss:

$$\hat{f}_n \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f(Z_i))^2$$

    where Z is the incomplete data.

- We show that the best possible regression function (Bayes predictor):
  - is not linear, and characterize its form.
  - can be computed by a linear regression model on an expanded feature set, or approximated with a single layer perception.

Conclusion

- We are always interested in applying the newly developed methodology to real datasets (for now Traumabase - electronic health records, probably paleoclimatology dataset).
- Don't hesitate to contact us if you are faced with missing data problems!

Thank you for your attention!