Freie Universität Berlin

Identification of Mutational Signatures Active in Individual Tumors

Sandra Krüger[1] Rosario M. Piro[1,2,3]

[1]Freie Universität Berlin
[2]Charité Universitätsmedizin Berlin
[3]German Cancer Consortium (DKTK)

NETTAB 2017 Workshop, Palermo, October 16-18, 2017

Outline

Background

- Mutational Processes And Mutational Signatures
- Construction Of Mutational Signatures

Mutational Signatures In Individual Tumors

- Alexandrov Signatures
- Shiraishi Signatures: `decompTumor2Sig`

Evaluation

- Test Procedure
- Results

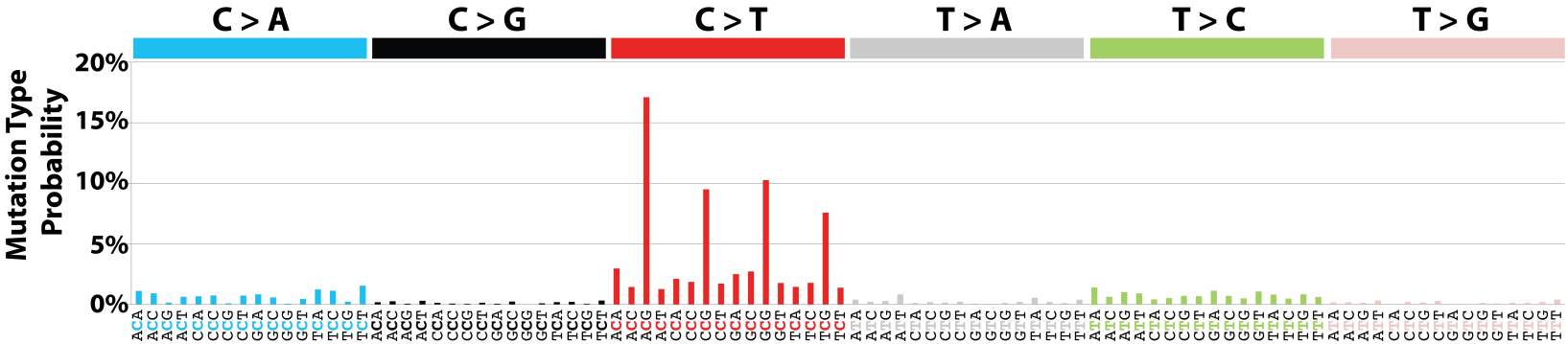
Mutational Processes And Somatic Mutations

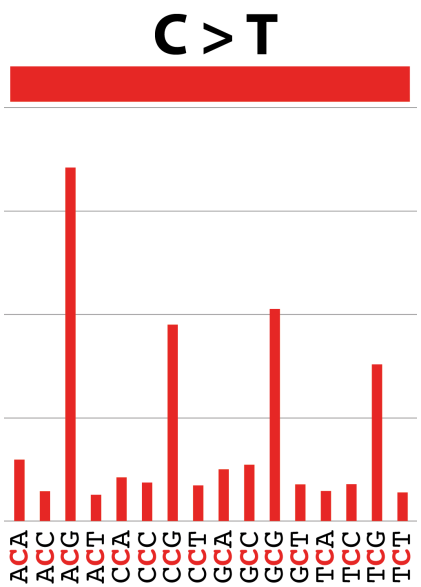
- Somatic mutations of individual tumors are caused by different mutational processes
- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type
- The sequence context of mutated bases is important!
- Examples
  - Lung cancers of tobacco smokers have a highly increased number of cytosine>adenine (C>A) transversions
  - Spontaneous deamination of 5-methylcytosine (age-related) causes cytosine>thymine (C>T) transitions in the context of CpGs
  - See, for example, Alexandrov and Stratton, Curr Opin Genet Dev 24:52–60, 2014
- Mutational processes can be represented by means of “mutational signatures”
  - Reflect the frequencies of base changes within their sequence context

“Alexandrov” Signatures

First published concept/notion of mutational signatures:

“Full” model (Alexandrov et al, Nature 500:415–421, 2013)



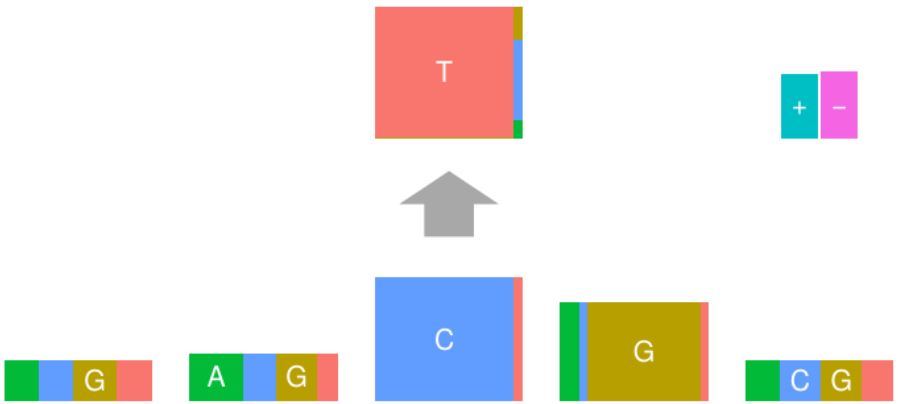


- Full dependency between mutated and adjacent bases
- For 3-base sequence contexts: $6 \times 4 \times 4 = 96$ parameters
- Can be described by a vector of 96 probabilities (i.e., sum is 1)

“Shiraishi” Signatures

Alternative concept/notion of mutational signatures:

“Independent” model (Shiraishi et al, PLoS Genet 11:e1005657, 2015)



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

Transcription strand

plus strand	minus strand
0.493	0.507

- Mutated base and adjacent bases as independent features
- For 5-base sequence contexts + transcriptional direction:
  $6 + 4 \times 4 + 2 = 24$ parameters
- Can be described by a table

How Are Mutational Signatures Derived?

(The following describes Alexandrov signatures; same for Shiraishi)

- The somatic mutations of a tumor are caused by multiple mutational processed. → We observe an overlap of multiple mutational signatures!

- Basic idea: the 96 mutation frequencies observed in tumor genome $\vec{g}$ can be described as the weighted sum of N signature vectors $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} w_k \vec{s}_k \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

  - **G** is the $96 \times G$-matrix of observed mutation frequencies in the tumors;
  - **S** is the $96 \times N$-matrix containing all signatures (one per column); and
  - **W** is the $N \times G$-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

How Are Mutational Signatures Derived?

- For a *set of G tumor genomes* we have:

$$\mathbf{G} = \mathbf{S} \times \mathbf{W}$$

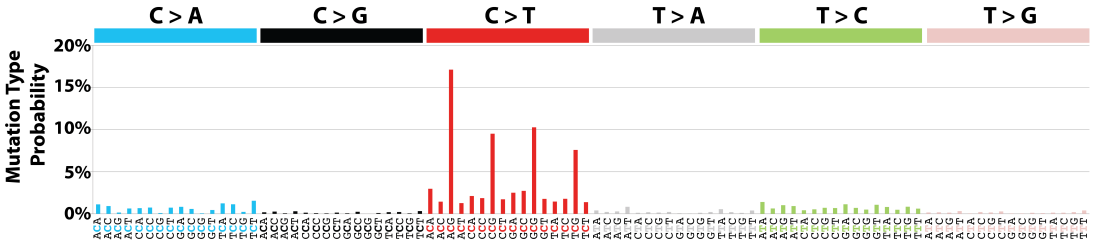
  - **G** is the 96 × *G*-matrix of observed mutation frequencies in the tumors;
  - **S** is the 96 × *N*-matrix containing all signatures (one per column); and
  - **W** is the *N* × *G*-matrix of weights (also called “exposures” or “contributions”) of the signatures in the single tumors

- Derive **S** and **W** at the same time!
  - Non-negative matrix factorization (Alexandrov et al, Cell Reports 3:246–259, 2013)
  - Principal component analysis (Gehring et al, Bioinformatics 31:3673–3675, 2015)

- Requires a large set of tumors!

- What if you what to determine which mutational processes contributed to the mutation load of a *single tumor*???
  (E.g., in a clinical setting)

Alexandrov Signatures In A Single Tumor



- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

$$\vec{g} = \sum_{k=1}^{N} (w_k \vec{s}_k) + \vec{\epsilon} \qquad \text{with} \quad \sum_{k=1}^{N} w_k = 1, w_k \geq 0$$

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)
- Goal: compute weights w_k and minimize the error terms (ϵ_k)!
  → weights measure the signatures' contributions to the mutational load of the tumor!
- Tool: `deconstructSigs` (Rosenthal et al, Genome Biol 17:31, 2016)
  - R package; constructs a solution by iteratively adding single mutational signatures to minimize the sum-squared error between $\vec{g}$ and $\sum_{k=1}^{N} (w_k \vec{s}_k)$

R Package decompTumor2Sig

```
### pmsignature (Shiraishi et al.) ###
library(pmsignature)
numsig = 4

# determine signatures from DB of tumor genomes
G <- readMPFile("myMutDB.MPF.txt.gz", numBases = 5, trDir = TRUE)
Param <- getPMSignature(G, K=numsig)

### decompTumor2Sig ###
library(decompTumor2Sig)

# read mutations of the individual tumor (function by pmsignature)
T <- readMPFile("queryTumor.MPF.txt.gz", numBases = 5, trDir = TRUE)

# convert mutation data to a matrix (same form as signatures)
genomes <- getGenomesFromMutationFeatureData(T, normalize=TRUE)

# predict exposures for tumor
signatures <- getSignatureListFromEstimatedParameters(Param)
exposures <- decomposeTumorGenomes(genomes, signatures)

exposures
#$PD4072a
#[1] 0.41464965 0.52231228 0.06303806 0.00000000
```

Quadratic Programming Approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following

Problem

$$\begin{aligned} \text{minimize} \quad & (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w}) \\ \text{subject to} \quad & \textstyle\sum_{s=1}^{k} w_s = 1, w_s \geq 0 \end{aligned}$$

Since $\vec{g}^T\vec{g}$ is constant and $(\mathbf{S}\vec{w})^T\vec{g} = \vec{g}^T\mathbf{S}\vec{w}$, we can simplify the problem:

$$\text{minimize} \quad -\vec{g}^T\mathbf{S}\vec{w} + \tfrac{1}{2}\vec{w}^T\mathbf{S}^T\mathbf{S}\vec{w} \qquad \text{subject to} \quad \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

- Classical quadratic programming problem!
- Can be easily solved using the R package `quadprog`

Estimation Of Accuracy

Dataset A:

- 21 breast cancer genomes (Nik-Zainal et al, Cell 149:979–993, 2012)
- 4 signatures

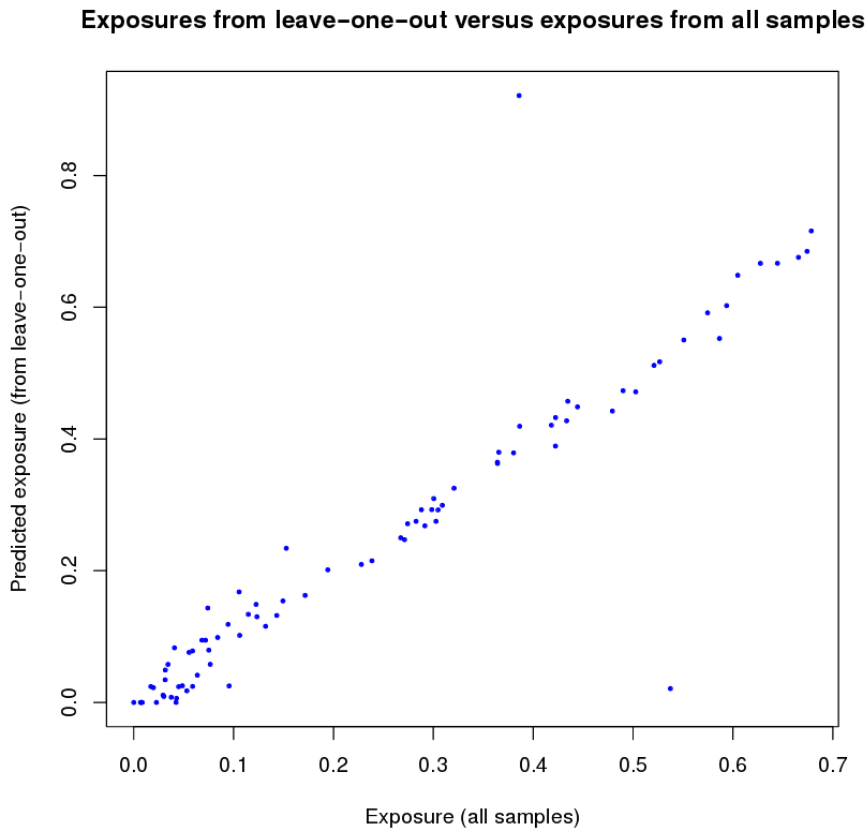
Dataset B:

- 435 tumor genomes from 10 tumor types (Alexandrov et al, Nature 500:415–421, 2013)
- 15 signatures

Procedure (simplified):

1. Compute "true" weights w_k using **all** tumors (`pmsignature`).
2. Leave out a single tumor (A) or 10% (B), compute signatures based on the remaining tumors (`pmsignature`).
3. Predict weights w'_k for each *single* left out tumor (`decompTumor2Sig`) using the signatures from the reduced dataset (step 2).
   - Note: this reflects a realistic case, e.g., in a clinical setting
4. Compare predicted weights (step 3) to "true" weights (step 1).

Results



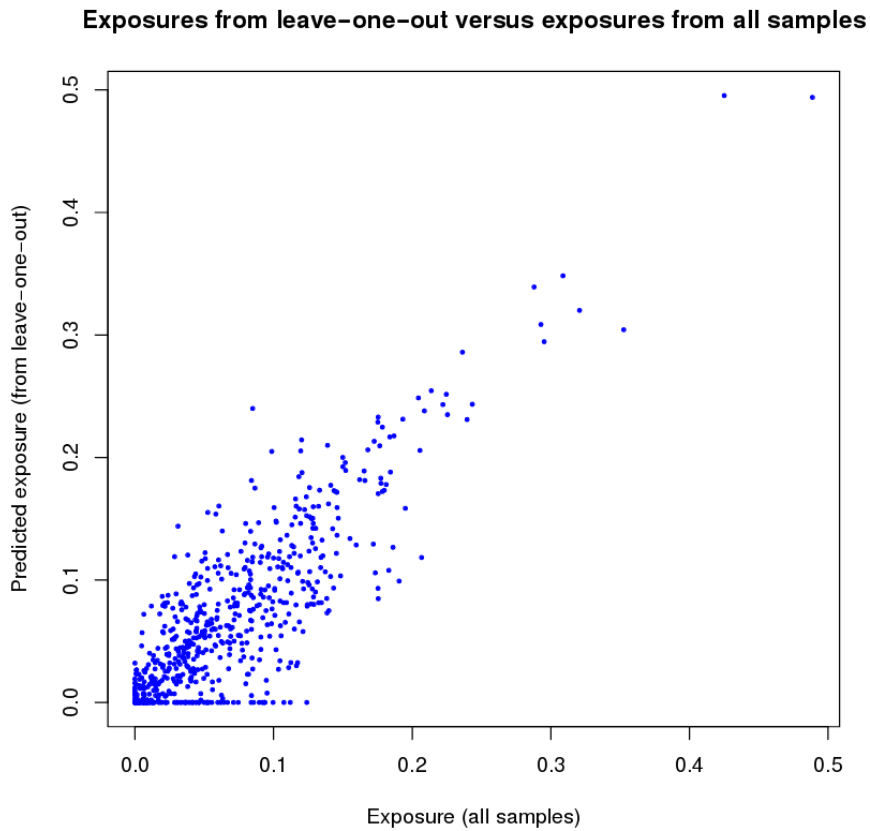


Figure 2: Comparison of contributions/weights ("exposures") predicted for individual tumors (`decompTumor2Sig`; y-axis) and collectively computed (`pmsignature`; x-axis). Left: leave-one-out test on 21 breast cancers ($r = 0.923$). Right: test set of 44 out of 435 tumors ($r = 0.807$).

Median weight differences ($|w_k - w'_k|$): (A) 0.018 and (B) 0.019

Summary

Contribution

- Development of the R package `decompTumor2Sig`
  - Determines the contribution of (given) Shiraishi signatures to the mutational load of an individual tumor
  - User friendly
- Evaluation of the accuracy of predicted contributions (weights w_k)

Outlook / further improvement

- Current version takes signatures from R objects produced by `pmsignature`
  - Allow to easily read from flat files
  - Allow to convert from Alexandrov signatures

Acknowledgments

- Freie Universtiät Berlin
  - Sandra Krüger *(*)*
- The University of Tokyo
  - Yuichi Shiraishi
- German Cancer Research Center (DKFZ), Heidelberg
  - Susanne Gröbner