Freie Universität Berlin

Identification of the mutational signatures active in individual tumors

Rosario Michael Piro

Freie Universität Berlin
Charité–Universitätsmedizin Berlin
German Cancer Consortium (DKTK)

Challenges in Data-driven Genomic Computing, Como, March 6-8, 2019

Outline

Background

Mutational processes and mutational signatures

De-novo inference of mutational signatures

Mutational signatures in individual tumors

Alexandrov signatures

Shiraishi signatures: `decompTumor2Sig`

Results

Evaluation

Decomposition of tumor genomes into signatures

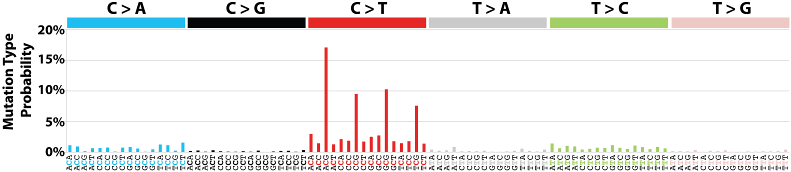
Mutational processes and somatic mutations

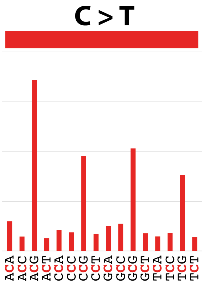
- Somatic mutations of individual tumors are caused by different mutational processes
- Mutational processes can significantly vary between tumors
  - between different cancer types
  - between individual tumors of the same cancer type
- The sequence context of mutated bases is important!
- Examples
  - 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:

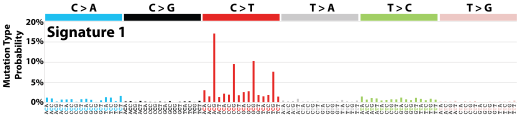
“Full” model (Alexandrov et al, Nature 500:415–421, 2013)



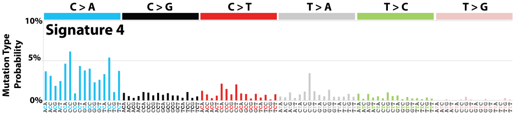


- Full dependency between mutated and adjacent bases
- For 3-base sequence contexts: $6 \times 4 \times 4 = 96$ parameters
- Can be described by a vector of 96 probabilities (i.e., sum is 1)

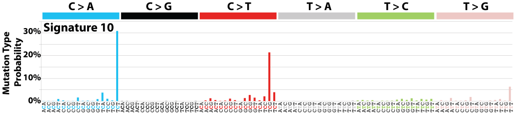
https://cancer.sanger.ac.uk/cosmic/signatures



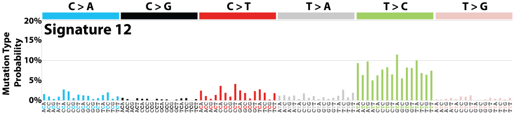
spontaneous deamination of 5-methylcytosine



associated with tobacco smoking



associated with recurrent POLE somatic mutations

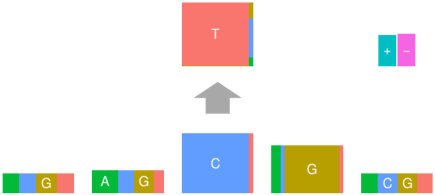


found in liver cancer (aetiology unknown)

“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 × 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

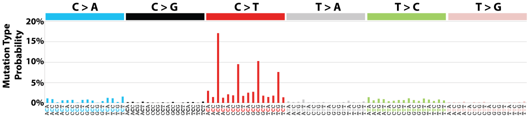
How are mutational signatures derived?

- 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



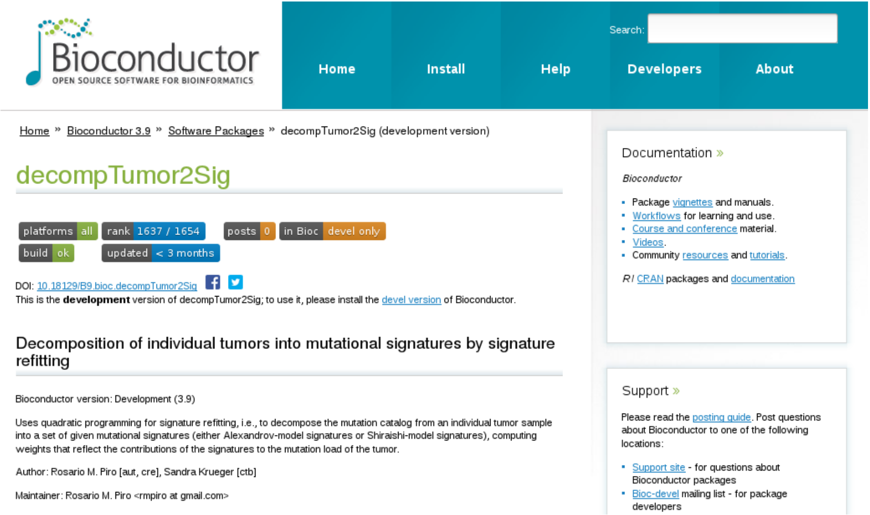
- Remember: the 96 mutation frequencies observed in *one* tumor genome $\vec{g}$ can be described as weighted sum of N signatures $\vec{s}_k$:

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

- Given: $\vec{g}$ (observed mutations), $\vec{s}_k$ (signatures)
- Goal: "Signature refitting"
  Compute weights w_k which minimize the error terms (ϵ_k)!
  → most likely 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

decompTumor2Sig



Home » Bioconductor 3.9 » Software Packages » decompTumor2Sig (development version)

decompTumor2Sig

platforms all · rank 1637 / 1654 · posts 0 · in Bioc devel only
build ok · updated < 3 months

DOI: 10.18129/B9.bioc.decompTumor2Sig

This is the **development** version of decompTumor2Sig; to use it, please install the devel version of Bioconductor.

Decomposition of individual tumors into mutational signatures by signature refitting

Bioconductor version: Development (3.9)

Uses quadratic programming for signature refitting, i.e., to decompose the mutation catalog from an individual tumor sample into a set of given mutational signatures (either Alexandrov-model signatures or Shiraishi-model signatures), computing weights that reflect the contributions of the signatures to the mutation load of the tumor.

Author: Rosario M. Piro [aut, cre], Sandra Krueger [ctb]

Maintainer: Rosario M. Piro <rmpiro at gmail.com>

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Quadratic programming approach

(Following Lynch, F1000Research 5:1253, 2016)

We want $\mathbf{S}\vec{w} \approx \vec{g}$, so we can solve the following (because $\vec{\epsilon} = \vec{g} - \mathbf{S}\vec{w}$)

Problem

$$\text{minimize} \qquad (\vec{g} - \mathbf{S}\vec{w})^T(\vec{g} - \mathbf{S}\vec{w})$$

$$\text{subject to} \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

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:

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

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

Accuracy of signature contributions/weights

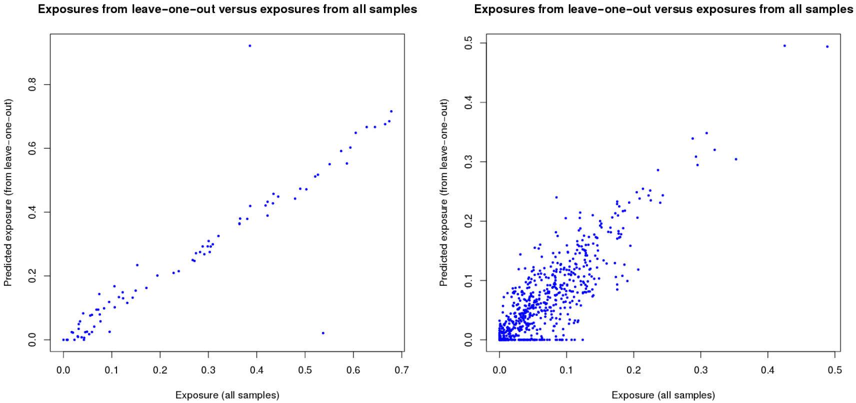
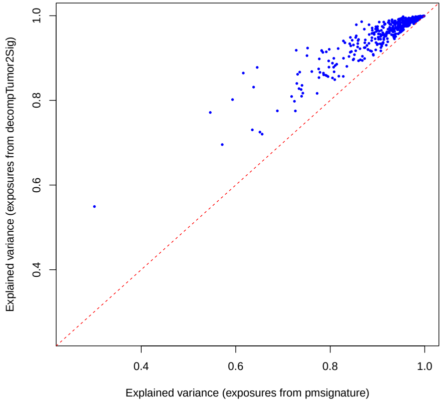


Figure 2: Comparison of contributions/weights (“exposures”) predicted for individual tumors (decompTumor2Sig; y-axis) and collectively computed (pmsignature; x-axis). Left: leave-one-out test on 21 breast cancers ($r = 0.923$). Right: test set of 44 out of 435 tumors ($r = 0.807$).

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

Signature refitting vs. de-novo inference

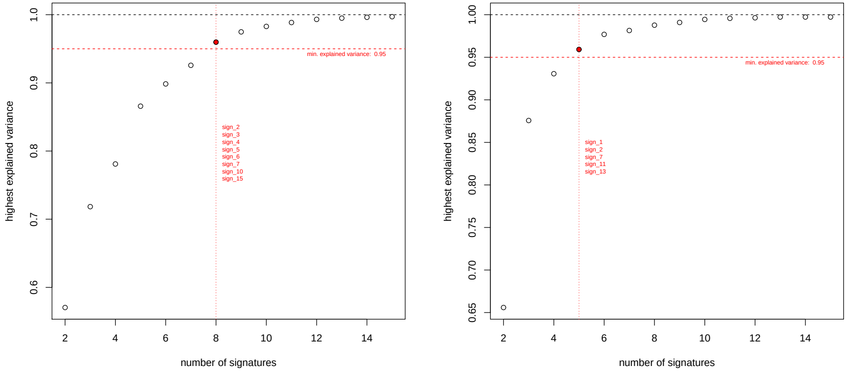


- Signature refitting better explains the observed variance of a tumor's mutation frequencies
- Explained variance R^2 defined as:

  $$R^2 = 1 - \frac{\mathrm{Var}(g - \hat{g})}{\mathrm{Var}(g)} = 1 - \frac{\sum_{i=1}^{P} (g_i - \hat{g}_i)^2}{\sum_{i=1}^{P} (g_i - g_i^*)^2}$$

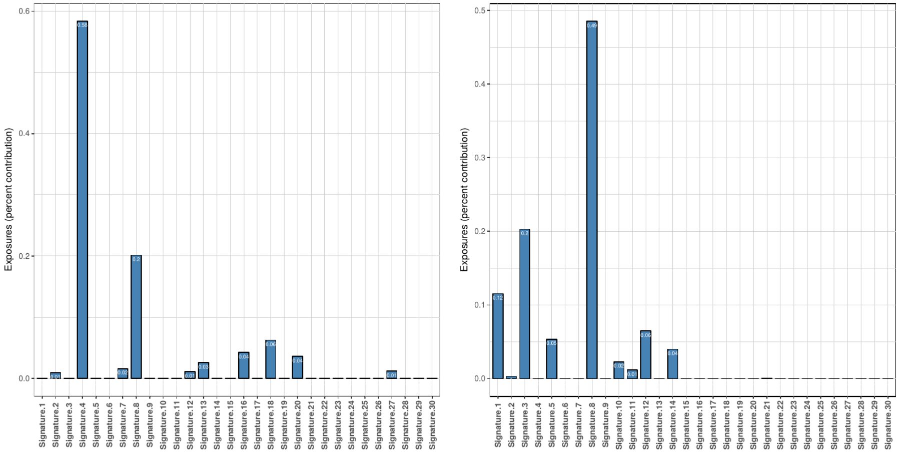
  where g is the observed, g^* a uniform, “flat” tumor genome, and $\hat{g}$ the genome predicted by the computed weights (hence $g - \hat{g}$ is the error).
- Much less parameters to be estimated ...

Subsets of signatures explain tumor genomes



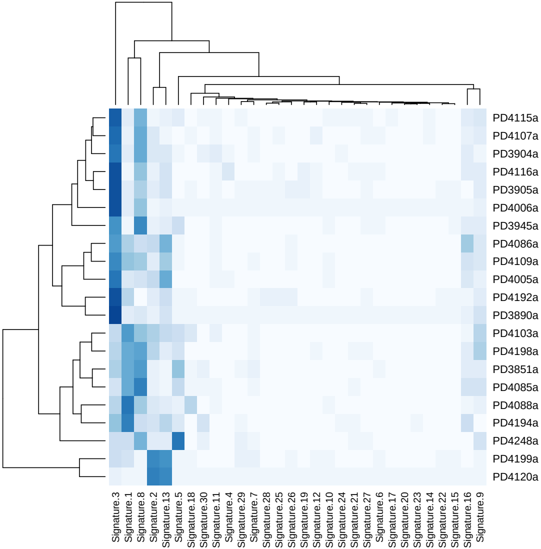
In many cases >95% of the variance in a tumor's mutation frequencies is determined by a subset of signatures.

Decomposition of example tumors



Contributions to lung adenocarcinoma (left) and medulloblastoma (right)

Decomposition of example tumors



Contributions to 21 breast cancer genomes
(data: Nik-Zainal et al., Cell, 2012)

Signature 3:

associated with failure of DNA double-strand break-repair (germline and somatic mutations in BRCA1 and BRCA2)

Signature 1:

spontaneous deamination of 5-methylcytosine (age-related)

Signatures 2 & 13:

attributed to activity of APOBEC family members

Acknowledgments

- Freie Universtiät Berlin
  - Sandra Krüger *(*)*
  - Annalisa Marsico
- The University of Tokyo
  - Yuichi Shiraishi
- German Cancer Research Center (DKFZ), Heidelberg
  - Susanne Gröbner
  - Marc Zapatka
  - Peter Lichter
  - Stefan Pfister
- Politecnico di Milano
  - Stefano Ceri
  - Eirini Stamoulakatou
  - Pietro Pinoli
- IBM Research - Zurich
  - Maria Rodriguez Martinez
  - An-phi Nguyen

Questions?