SomaticSignatures: Inferring Mutational Signatures from Single Nucleotide Variants

Julian S. Gehring [1,2], Bernd Fischer [1],
Michael Lawrence [2], and Wolfgang Huber [1]

October 24, 2014

Summary

Mutational signatures are patterns in the occurrence of somatic single nucleotide variants (SNVs) that can reflect underlying mutational processes. The *SomaticSignatures* package provides flexible, interoperable, and easy-to-use tools that identify such signatures in cancer sequencing studies. It facilitates large-scale, cross-dataset estimation of mutational signatures, implements existing methods for pattern decomposition, supports extension through user-defined methods and integrates with Bioconductor workflows.

The R package *SomaticSignatures* is available as part of the Bioconductor project (R Core Team, 2014; Gentleman *et al.*, 2004). Its documentation provides additional details on the methodology and demonstrates applications to biological datasets.

Contact

julian.gehring@embl.de, whuber@embl.de

Affiliations

[1]European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany
[2]Department of Bioinformatics and Computational Biology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080, USA

1 Introduction

Mutational signatures link observed somatic single nucleotide variants to mutation generating processes (Alexandrov *et al.*, 2013a). The identification of these signatures offers insights into the evolution, heterogeneity and developmental mechanisms of cancer (Alexandrov *et al.*, 2013b; Nik-Zainal *et al.*, 2012).

Existing implementations (Fischer *et al.*, 2013; Nik-Zainal *et al.*, 2012) are standalone packages with specialized functionality. Their reliance on non-standard data input and output formats limits integration into common workflows.

The *SomaticSignatures* package aims to encourage wider adoption of somatic signatures in tumor genome analysis by providing an accessible R implementation that supports multiple statistical approaches, scales to large datasets, and closely interacts with the data structures and tools of Bioconductor.

2 Approach

To detect the extent of sequence specific effects contributing to the set of observed somatic variants, the SNVs are analyzed with regard to their immediate sequence contexts, the flanking $3'$ and $5'$ bases (Alexandrov *et al.*, 2013a). This can capture characteristics of mutational mechanisms as well as technical biases (Nakamura *et al.*, 2011). As an example, the mutation of `A` to `G` in the sequence `TAC` defines the mutational motif `T[A>G]C`. Considering the frequency of the 96 possible motifs across all samples defines the mutational spectrum. It is represented by the matrix M_{ij}, with i enumerating the motifs and j the samples.

The observed mutational spectrum can be interpreted by decomposing M into two matrices of smaller size,

$$M_{ij} = \sum_{k=1}^{R} W_{ik} H_{kj} + \varepsilon_{ij}, \tag{1}$$

where the number of signatures R is typically small compared to the number of samples, and the elements of the residual matrix ε are minimized, such that WH is a useful approximation of the data. The columns of W describe the composition of a signature: W_{ik} is the relative frequency of somatic motif i in the k-th signature. In addition, the rows of H indicate the contribution of each signature to a particular sample j.

3 Methods

Several approaches exist for the decomposition (Eq. 1) that differ in their constraints and computational complexity. In principal component analysis (PCA), for a given k, W and H are chosen such that the norm $\sum_{ij} \varepsilon^2$ is minimal and the columns of W are orthonormal. Non-negative matrix factorization (NMF) (Brunet *et al.*, 2004) is motivated by the fact that the mutational spectrum fulfills $M_{ij} \geq 0$, and imposes the same requirement on the elements of W and H. Different NMF and PCA algorithms allow additional constraints on the results, such as sparsity. With unsupervised clustering, the elements of H are either 0 or 1, and each row contains exactly one entry of 1. In other words, the columns of W are the cluster representatives and H is the cluster membership matrix.

4 Results

SomaticSignatures is a flexible and efficient tool for inferring characteristics of mutational mechanisms. It integrates with the Bioconductor framework and its tools for importing, processing, and annotating genomic variants. An analysis starts with a set of SNV calls, typically imported from a VCF file and represented as a `VRanges` object (Obenchain *et al.*, 2014). Since the original calls do not contain information about the sequence context, we construct the mutational motifs first, based on the reference genome.

```
ctx = mutationContext(VRanges, ReferenceGenome)
```

Subsequently, we construct the mutational spectrum M. By default, its columns are defined by the samples in the data. Alternatively, users can specify a grouping covariate, for example drug response or tumor type.

```
m = motifMatrix(ctx, group)
```

Mutational signatures and their contribution to each sample's mutational spectrum are estimated with a chosen decomposition method for a defined number of signatures. We provide implementations for NMF and PCA, and users can specify their own functions that implement alternative decomposition methods.

```
sigs = identifySignatures(m, nSig, method)
```

The user interface and library of plotting functions facilitate subsequent analysis and presentation of results (Fig. 1). Accounting for technical biases is often essential, particularly when analyzing across multiple datasets. For this purpose, we provide methods to normalize for the background distribution of sequence motifs, and demonstrate how to identify batch effects.

In the documentation of the software, we illustrate a use case by analyzing 653,304 somatic SNV calls from 2,437 TCGA whole-exome sequenced samples (Gehring, 2014). The analysis, including NMF, PCA and hierarchical clustering, completes within minutes on a standard desktop computer. The different approaches yield a consistent and reproducible grouping of the cancer types according to the estimated signatures (Fig. 1).

We applied this approach to the characterization of kidney cancer and showed that classification of subtypes according to mutational signatures is consistent with classification based on RNA expression profiles and mutation rates (Durinck *et al.*, 2014).

Acknowledgment

We thank Leonard Goldstein and Oleg Mayba for their insights and suggestions.

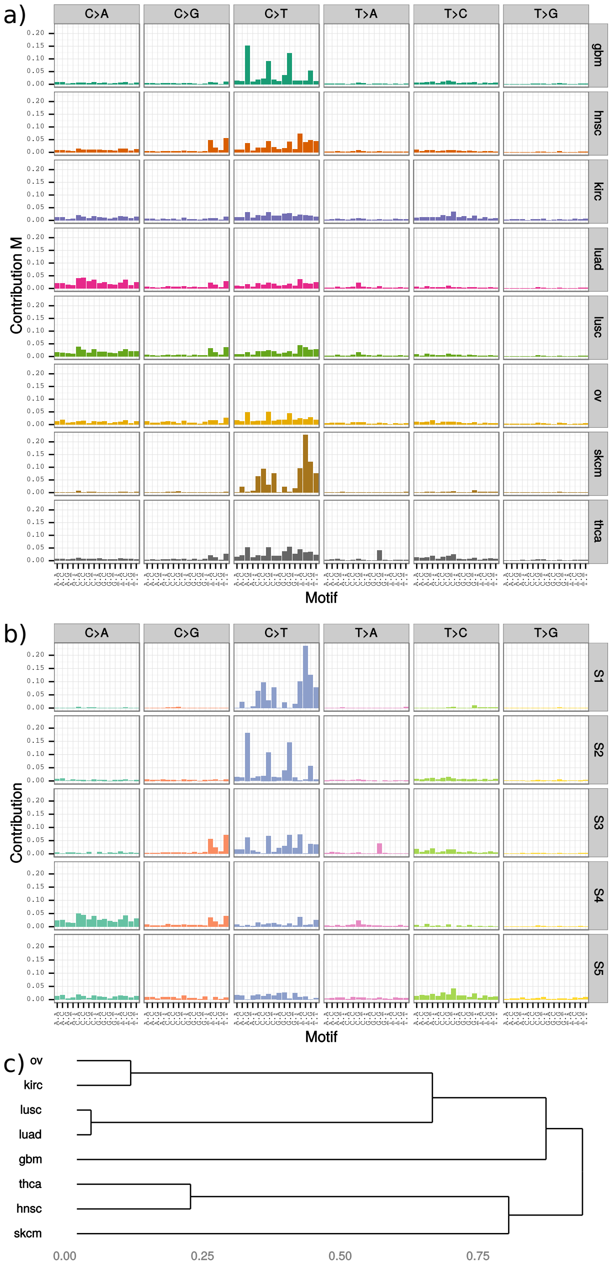

Figure 1: Analysis of mutational signatures for eight TCGA studies (Gehring, 2014). The observed mutational spectrum of each study (panel a, labels at the right side of the plot) was decomposed into 5 distinct mutational signatures S1 to S5 (panel b) with NMF. Hierarchical clustering (c) of the signatures based on cosine similarity confirms the similarities in mutational processes of biologically related cancer types. An annotated high-resolution version of this figure is included as Supplementary Figure S1.

Funding

This work was supported by European Molecular Biology Laboratory, the NSF award "BIGDATA: Mid-Scale: DA: ESCE: Collaborative Research: Scalable Statistical Computing for Emerging Omics Data Streams" and Genentech Inc.

References

Alexandrov *et al.* (2013) Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, doi:10.1016/j.celrep.2012.12.008

Alexandrov *et al.* (2013) Signatures of Mutational Processes in Human Cancer. *Nature*, doi:10.1038/nature12477

Brunet *et al.* (2004) Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *PNAS*, doi:10.1073/pnas.0308531101

Fischer *et al.* (2013) EMu: Probabilistic Inference of Mutational Processes and their Localization in the Cancer Genome. *Genome Biology*, doi:10.1186/gb-2013-14-4-r39

Gehring (2014). SomaticCancerAlterations. *Bioconductor package*, Version: 1.1.0, dx.doi.org/10.5281/zenodo.12279

Gentleman *et al.* (2004) Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology*, doi:10.1186/gb-2004-5-10-r80

Nik-Zainal *et al.* (2012) Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, doi:10.1371/journal.pgen.0030161

Obenchain *et al.* (2014) VariantAnnotation: A Bioconductor Package for Exploration and Annotation of Genetic Variants. *Bioinformatics*, doi:10.1093/bioinformatics/btu168

R Core Team. (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Durinck *et al.* Spectrum of Diverse Genomic Alterations Define Non-Clear Cell Renal Carcinoma Subtypes. *Nature Genetics*, in press

Nakamura *et al.* (2011) Sequence-Specific Error Profile of Illumina Sequencers. *Nucleic Acids Research*, doi:10.1093/nar/gkr344