bioRxiv preprint doi: https://doi.org/10.1101/071761; this version posted August 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues

Francis Blokzijl[1], Roel Janssen[1], Ruben van Boxtel[1,*], Edwin Cuppen[1,*]

[1]Department of Genetics, Center for Molecular Medicine, Cancer Genomics Netherlands, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[*] Shared last author

Summary:
Mutational processes leave characteristic footprints in genomic DNA. The *MutationalPatterns* R package provides an easy-to-use toolset for the characterization and visualization of mutational patterns in base substitution catalogues of e.g. tumour samples or DNA-repair deficient cells. The package covers a wide range of patterns including: mutational signatures, transcriptional strand bias, genomic distribution and association with genomic features, which are collectively meaningful for studying the activity of mutational processes. The package provides functionalities for both extracting mutational signatures *de novo* and inferring the contribution of previously identified mutational signatures in a given sample. *MutationalPatterns* integrates with common R genomic analysis workflows and allows easy association with (publicly available) annotation data.
Availability and implementation: The *MutationalPatterns* R package is freely available for download at https://github.com/CuppenResearch/MutationalPatterns. The package documentation provides a detailed description of typical analysis workflows.
Contact: ecuppen@umcutrecht.nl

1. Introduction

Genomes of cells are constantly threatened by both endogenous and environmental sources of DNA damage, such as UV-light and spontaneous reactions. To safeguard their genomic integrity, cells employ various mechanisms that repair damaged DNA. When lesions are either incorrectly or not repaired prior to replication, these can lead to mutation incorporation into the genome (Iyama and Wilson, 2013). Each mutational process leaves a distinct genomic mark. For example, UV light preferentially induces CC>TT dinucleotide substitutions, whereas spontaneous deamination of 5-methylcytosines results in C>T substitutions at CpG sites. Mutational patterns can therefore be used to infer which mutational processes have been active in a cell during life (Helleday *et al.*, 2014).

In the past few years, large-scale analyses of tumour genome data across different human cancer types have revealed 30 mutational patterns. These so-called “mutational signatures” are characterized by a specific contribution of base substitution types with a certain sequence context (Alexandrov *et al.*, 2013; Helleday *et al.*, 2014). Each mutational signature is thought to reflect a single mutational mechanism. However, the aetiology of most mutational signatures remains currently unknown. In order to functionally link mutational signatures to biological processes, assessment of the contribution of these mutational signatures in e.g. cells that are exposed to specific mutagens or cells that are deficient for a certain DNA repair pathway will be essential.

The *MutationalPatterns* package provides an extensive toolset to explore and visualize a collection of mutational patterns that are relevant for deciphering which

bioRxiv preprint doi: https://doi.org/10.1101/071761; this version posted August 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mutational processes have been active in a sample. The package facilitates both (1) *de novo* mutational signature extraction and (2) quantification of the contribution of user-specified mutational signatures. While the first approach can be used to identify new mutational signatures, this is only meaningful for datasets with a large number of samples with diverse mutation spectra, as it relies on the dimensionality reduction method non-negative matrix factorization (NMF). The second approach can be used to study mutational processes in a single sample, and to further characterize previously-identified mutational signatures by assessing their contribution in different systems or under different conditions. Additionally, the package allows for exploration of other types of patterns such as transcriptional strand asymmetry, genomic distribution and associations with (publicly available) annotations such as chromatin organization. These features are useful for the identification of active mutation-inducing processes and the involvement of specific DNA repair pathways. For example, presence of a transcriptional strand bias in genic regions may indicate activity of transcription coupled repair (Haradhvala *et al.*, 2016; Pleasance *et al.*, 2010).

We conclude that the ability to assess combinations of mutational patterns, as facilitated by *MutationalPatterns*, is essential to identify the mutational processes that have been operative in a given sample.

2. Features

Any set of base substitution calls can be imported from a VCF file and represented as a GRanges object (Lawrence *et al.*, 2013). The sequence context of the base substitutions can be retrieved from a reference genome to construct a mutation matrix with counts for all 96 possible trinucleotide changes. In addition, other features such as transcriptional strand can be included, resulting in a 192 feature count matrix (96 trinucleotides * 2 strands). To this end, gene definitions - that can be retrieved from e.g. UCSC - are used to determine whether base substitutions in genes are located on the transcribed or untranscribed strand.

```
m = mut_matrix(vcfs, ref_genome)
m_s = mut_matrix_stranded(vcfs, ref_genome, genes)
```

Mutational signatures can be extracted *de novo* using NMF, where the number of signatures *n* is typically small compared to the number of samples in the mutation matrix *m* (Fig. 1A)

```
res = extract_signatures(m_s, n)
```

For mutational signatures with transcriptional strand features, the strand bias can be determined per base substitution type (Fig. 1B)

```
plot_signature_strand_bias(res$signatures)
```

Additionally, the non-negative linear combination of a set of user-specified mutational signatures that best describes the mutation profile of a sample *i,* can be determined by minimizing the Euclidean norm of the residual

$$\| S \cdot x_i - d_i \|$$

bioRxiv preprint doi: https://doi.org/10.1101/071761; this version posted August 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where *x* >= 0 and describes a linear combination of the signatures in matrix *S* for sample *i*, and *d* the original 96 mutation count vector of sample *i*. This well-studied non-negative least-squares constraints problem is solved using the pracma package (Borchers, 2016) for each sample in mutation matrix *m* (Fig. 1C)

```
fit_to_signatures(m, S)
```

To test whether base substitutions appear more or less frequently in specific genomic regions, enrichment or depletion can be visualized and tested for statistical significance (Fig. 1D-E). The analysis is corrected for the “surveyed area” of the genome; the positions at which base substitutions could be reliably called in that sample. The genomic regions are represented as a GRanges object (Lawrence *et al.*, 2013) and can be based on experimental data or publicly available annotation data retrieved via e.g. BiomaRt (Durinck *et al.*, 2005). The analysis can be performed over a number of samples collectively within user-specified groups, such as tissue type (Fig. 1E).

```
d = genomic_distribution(vcfs, surveyed, regions)
enrichment_depletion_test(d, by = tissue)
```

3. Discussion

Until now, somatic mutation catalogues have been mainly determined for tumour samples, owing to their clonal nature. Recent and ongoing advances in single cell sequencing (Gawad *et al.*, 2016), extremely deep sequencing of clonal patches of healthy tissue (Martincorena *et al.*, 2015; Xie *et al.*, 2014) and clonal cell cultures (Blokzijl *et al.*, 2016) will allow determination of somatic mutation catalogues of non-cancerous cells of various tissues. Furthermore, advances in gene-editing enables researchers to specifically knock-out a certain repair mechanism and evaluate this effect on mutational load (Meier *et al.*, 2014). *MutationalPatterns* aims to support the further dissection of mutational mechanisms by providing an extensive, easy-to-use toolset to characterize and visualize informative mutational patterns, not only for large collections of samples but also for single samples.

Funding

This work was financially supported by the NWO Gravitation Program Cancer Genomics.nl and the NWO/ZonMW Zenith project 93512003 to E.C.

bioRxiv preprint doi: https://doi.org/10.1101/071761; this version posted August 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

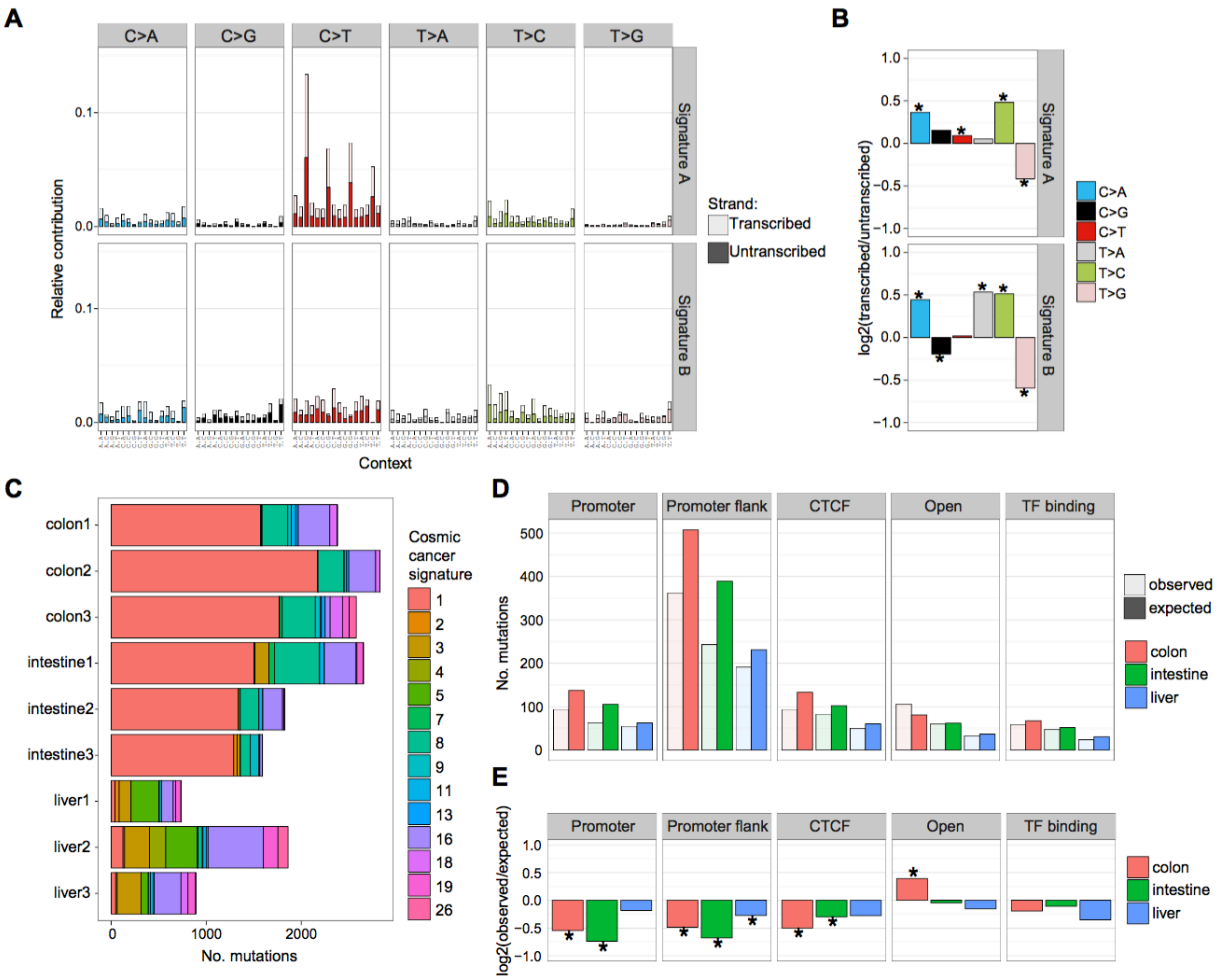

Fig. 1. MutationalPatterns-based analysis of somatic mutation catalogues of 9 normal human adult stem cells from 3 tissues (Blokzijl *et al.*, 2016). Two mutational signatures with transcriptional strand features were extracted *de novo* (panel A). The effect size (log2ratio) and significance (* P < 0.05, Poisson test) of the transcriptional strand bias were calculated per base substitution type per signature (panel B). Contribution of signatures of mutational processes in human cancer (Alexandrov *et al.*, 2013; Helleday *et al.*, 2014) was determined in the 9 normal adult stem cell samples (panel C). The expected and observed number of base substitutions in functionally annotated genomic regions from Ensembl regulatory features (Zerbino *et al.*, 2015) was determined for predicted promoters, promoter flanking regions, CTCF binding sites, open chromatin regions and transcription factor binding sites (panel D). The effect size (log2ratio) and statistical significance (* *P* < 0.05, binomial test) of the depletions/enrichments in the genomic regions was calculated (panel E).

bioRxiv preprint doi: https://doi.org/10.1101/071761; this version posted August 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Alexandrov,L.B. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–21.

Blokzijl,F. *et al.* (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, **in press**.

Borchers,H.W. (2016) pracma: Practical Numerical Math Functions.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Gawad,C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, **17**, 175–188.

Gehring,J.S. *et al.* (2015) SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, **31**, 3673–3675.

Haradhvala,N.J. *et al.* (2016) Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*, 1–12.

Helleday,T. *et al.* (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.

Iyama,T. and Wilson,D.M. (2013) DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst).*, **12**, 620–36.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

Martincorena,I. *et al.* (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. **348**, 880–886.

Meier,B. *et al.* (2014) C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.*, **1**, 1624–1636.

Pleasance,E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Xie,M. *et al.* (2014) Age-related cancer mutations associated with clonal hematopoietic expansion. *Nat. Med.*, **20**, 1472–1478.

Zerbino,D.R. *et al.* (2015) The ensembl regulatory build. *Genome Biol.*, **16**, 56.