Supplementary note:

Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer.

Jayendra Shinde, Quentin Bayard, Sandrine Imbeaud, Théo Z Hirsch, Feng Liu, Victor Renault, Jessica Zucman-Rossi and Eric Letouzé*

*Corresponding author. E-mail: eric.letouze@inserm.fr

In this supplementary note, we describe in more details the methodological details underlying the main functions of the *Palimpsest* package.

S1. Mutational signature analysis

Palimpsest allows both *de novo* analysis of mutational signatures and quantification of previously described signatures. In both cases, a set of base substitutions from a tumor series is first imported from a VCF file and each mutation is assigned to one of 96 mutation categories, defined by the 6 substitution types multiplied by 16 possible trinucleotide contexts (Alexandrov *et al.*, 2013). We can then represent the mutational catalog of the *n* tumors as a mutation matrix $M = \begin{pmatrix} m_1^1 & \dots & m_n^1 \\ \vdots & \ddots & \vdots \\ m_1^{96} & \dots & m_n^{96} \end{pmatrix}$ where m_i^j is the number of mutations of category *j* in tumor *i*. The goal of mutational signature analysis is to deconvolute this matrix *M* as the product of a matrix of mutational processes $P = \begin{pmatrix} p_1^1 & \dots & p_K^1 \\ \vdots & \ddots & \vdots \\ p_1^{96} & \dots & p_K^{96} \end{pmatrix}$ where p_i^j is the probability of the process *i* to cause a mutation of category *j*, and a matrix of exposures $E = \begin{pmatrix} e_1^1 & \dots & e_n^1 \\ \vdots & \ddots & \vdots \\ e_1^K & \dots & e_n^K \end{pmatrix}$ where e_i^j is the number of mutations attributed to process *i* in tumor *j*. To solve this problem, we use non-negative matrix factorization, as implemented in the *NMF* package (Gaujoux and Seoighe, 2010). For a *de novo* analysis, the number of processes *K* can be manually defined by the user or estimated automatically considering the cophenetic correlation coefficients and residual sum of squares (RSS) for each number of signatures. NMF then identifies the matrices *P* and *E* that verifies $M \approx P \times E$ and minimizes a Frobenius norm while maintaining non-negativity. For an analysis of previously described signatures, the user provides the matrix of processes P, and NMF is used only to reconstruct the exposure matrix *E*. The *P* matrix for the 30 signatures currently referenced in the COSMIC database is provided with the *Palimpsest* package.

S2. Association of mutational signatures with driver genes

Once a cancer genome has been deconvoluted as a combination of several mutational processes, we can estimate the probability of each somatic mutation being generated by each mutational process using simple Bayes statistics (Letouzé *et al.*, 2017). This key feature has been introduced in Palimpsest as follows:

Consider a mutation category c out of the 96 mutation categories, the number of mutations of category *c* in a tumor *t* can be expressed as:

$$m_t^c = \sum_{s=1}^{10} p_s^c \times e_t^s$$

where the product $p_s^c \times e_t^s$ represents the number of mutations of category *c* attributed to signature *s* in tumor *t*. The probability *P(m,s)* of a mutation *m* of category *c* in tumor *t* being due to signature *s* can then be estimated as:

$$P(m,s) = \frac{p_s^c \times e_t^s}{\sum_{s=1}^{10} p_s^c \times e_t^s}$$

This important feature of Palimpsest can be used to predict the mutational processes (point mutations or structural variants) at the origin of driver alterations in a cancer genome.

S3. Structural rearrangement signature analysis

Palimpsest performs structural rearrangement signature analysis by applying the same statistical framework used for mutational signature analysis. Somatic structural rearrangements from a series of tumors are first classified into 38 categories considering the type (deletion, tandem duplication, inversion, interchromosomal translocation) and size (<1kb, 1-10kb, 10-100kb, 100kb-1Mb, 1-10Mb, >10Mb) of rearrangements, as previously described (Nik-Zainal *et al.*, 2016). Clustered events, defined by the presence of ≥10 breakpoints within a 1Mb window, are identified using the *bedr* package (Haider *et al.*, 2016) and considered separately from non-clustered events. We then used non-negative matrix factorization, as implemented in the *NMF* package, to extract rearrangement signatures and their exposure in each tumor. Like mutational signature analysis, structural rearrangement signature analysis can be performed *de novo* or using a pre-defined set of known signatures to estimate the exposure matrix.

S4. Estimating the cancer cell fraction and clonality

Using the variant allele fraction (VAF), tumor cell content and absolute copy-number estimates, *Palimpsest* estimates the cancer cell fraction (CCF), i.e. the proportion of tumor cells harboring each mutation:

$$CCF = VAF \times \frac{\rho N_t + (1-\rho) N_n}{\rho n_{chr}}$$

where ρ is the tumor cell content, N_t and N_n the copy-number at the locus in tumor and normal cells, and n_{chr} the number of chromosomal copies harboring the mutation in tumor cells (also called multiplicity of the mutation). ρ and N_t can be estimated from copy-number data using various algorithms. The multiplicity of each mutation (n_{chr}) is set to the integer value closest to:

$$max\left(1,\ VAF\times\frac{\rho N_t+(1-\rho)N_n}{\rho}\right)$$

Finally, Palimpsest determines the 95% confidence of VAF using a binomial test and converts this interval to obtain the 95% confidence interval of CCF using the above formula. A mutation is then classified subclonal if the upper boundary of the 95% confidence interval is below a threshold set by the user (default 0.95), and clonal otherwise.

S5. Timing chromosomal duplications

When a chromosome is duplicated, mutations harbored by the chromosome are also duplicated and their VAF is increased as compared to mutations present on the other chromosome copy, or acquired after the duplication. The ratio of duplicated/non-duplicated mutations can thus be used to time the chromosome duplication event, early events having a low amount of duplicated mutations as compared to late events (Nik-Zainal *et al.*, 2012). *Palimpsest* uses previously described formulas (Letouzé *et al.*, 2017) to estimate the point mutation time of each chromosome duplication with a minimum of 30 mutations located on the duplicated segment.

Let us consider the simple case of a chromosome with absolute copy-number N_t=3. The molecular time at which the extra copy of the chromosome was gained can be estimated as:

$$T=\frac{N_{dup}}{N_{dup}+\frac{N_{ndup}-N_{dup}}{3}}\times 100$$

where N_{dup} and N_{ndup} are the number of duplicated and non-duplicated mutations, respectively.

We extrapolated this formula to chromosomes with N_t≥4. In this case, we timed the first duplication event using:

$$T=\frac{N_{dup}}{N_{dup}+\frac{N_{ndup}-N_{dup}}{(3+N_t)/2}}\times 100$$

where N_{dup} is the number of mutations at the maximal level of multiplicity and N_{ndup} the number of mutations at intermediate levels of multiplicity or non-duplicated.

For cases where the two parental chromosome copies were duplicated, e.g. N_t=4 with 2 copies of each chromosome copy, we adapted the formula as follows:

$$T=\frac{N_{dup}/2}{N_{dup}/2+\frac{N_{ndup}}{(3+N_t)/2}}\times 100$$

S6. Limitations of the package – number of mutations needed

Palimpsest can be applied to both whole genome and whole exome sequencing data. However, whole genome data are preferable for structural rearrangement signature analysis and timing chromosome duplications. Factors influencing extraction of mutational signatures have been extensively analyzed elsewhere (Alexandrov *et al.*, Cell Rep 2013) using simulated data with varying numbers of signatures, similarities between signatures, numbers of tumors and numbers of mutations per tumor. The authors concluded that, although both whole genome and whole exome sequencing data are suitable to accurately extract mutational signatures, the number of required mutations/samples increases with the number of signatures, and it is preferable to have more mutations/sample (e.g. whole genome sequences) than more samples with less mutations (e.g. large whole exome series). However, giving precise limitations regarding a required number of mutations per sample is delicate as it also depends on the type of mutational signatures analyzed. In our experience, very specific signatures caracterized by only a few mutation categories mutated at high frequency will be easily identified with relatively few mutations, whereas signatures with widespread mutation spectra will by harder to quantify.

References

Alexandrov,L.B. *et al.* (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.

Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.

Haider,S. *et al.* (2016) A bedr way of genomic interval processing. *Source Code Biol. Med.*, **11**, 14.

Letouzé,E. *et al.* (2017) Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.*, **8**, 1315.

Nik-Zainal,S. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.

Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.