SALMON: signature analyzer for low mutation counts

User Guide

DongHyuk Lee[1], William Wheeler[2] and Bin Zhu[3]

[1] Department of Statistics, Pusan National University, Busan, Korea

[2] Information Management Services (IMS), Inc. Rockville, MD, USA

[3] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Version 0.0.5

October 27, 2022

Contents

1 Introduction

For the targeted sequenced tumors, applying the existing mutational signature analysis tools is challenging because they are developed for whole-exome or whole-genome sequencing (WES or WGS). Since the sequenced genomic sizes and contexts of targeted panels differ from WES or WGS, not all mutational signatures detected by WES or WGS can be found using the targeted sequencing panels. On the other hand, panel-based signatures may include *de novo* presence of signatures. To identify mutational signatures accurately, we developed a new method, SALMON (Signature Analyzer for Low Mutation cOuNt), for targeted sequenced tumors. In this guide, we present a step-by-step guide to detecting mutational signatures in targeted sequenced tumors, using the **R** package *SALMON*.

2 Installation

To install from Github directly, one can use the **R** package *devtools*:

```
if (!requireNamespace("devtools", quietly = TRUE))
        install.packages("devtools")
devtools::install_github("binzhulab/SALMON/source")
```

Alternatively, SALMON_0.0.5.tar.gz (for Unix) or SALMON_0.0.5.zip (for Windows, R version >= 4.0) from the Github page[1] can be installed using the following commands:

```
install.packages("SALMON_0.0.5.tar.gz", repos = NULL, type = "source")
install.packages("SALMON_0.0.5.zip", repos = NULL, type = "win.binary")
```

Once installed, *SALMON* can be loaded on **R** by calling

[1] https://github.com/binzhulab/SALMON

```
> library(SALMON)
```

3 Example data

For illustrative purposes, we simulated a 96 $\times$ 10027 mutation catalog matrix **V**, which contains 10027 targeted sequenced tumors for 96 single base substitution (SBS) types. Each element of the matrix is generated from the Poisson distribution (`rpois()` function) with the mean corresponding to each element of $\mathbf{L} \circ \mathbf{W}_T\mathbf{H}_T$, where $\circ$ is elementwise multiplication, **L** is a given panel size matrix with the same 96 $\times$ 10027 size of **V**, $\mathbf{W}_T$ is the profile matrix of the tumor mutation burden (TMB, in the unit of the number of mutations per million base pairs) based signatures of size 96$\times$6 for 6 SBS signatures (SBS1, SBS2/13, SBS4, SBS5, SBS40, SBS89) and $\mathbf{H}_T$ is the signature activity matrix of size 6$\times$10027. All 4 matrices are stored in `SimData` with corresponding names: **V** (`SimData$V`), **L** (`SimData$L`), $\mathbf{H}_T$ (`SimData$TrueH`), $\mathbf{W}_T$ (`SimData$TrueW_TMB`). In addition, the entire set of TMB-based catalog SBS signature profiles $\mathbf{W}_0$ (`SimData$W_TMB`) is also included (see Section 4.3).

```
> data(SimData, package = "SALMON")
> dim(SimData$V)
[1] 96 10027
> SimData$V[1:6, 1:6]
        Sample1 Sample2 Sample3 Sample4 Sample5 Sample6
A[C>A]A       0       0       0       0       0       0
A[C>A]C       0       0       0       0       0       1
A[C>A]G       0       0       0       0       0       0
A[C>A]T       0       0       0       0       0       0
C[C>A]A       2       0       0       0       1       0
C[C>A]C       1       0       0       0       0       0
```

```
> dim(SimData$L)
[1] 96 10027
> SimData$L[1:6, 1:6]
        Sample1  Sample2  Sample3  Sample4  Sample5  Sample6
```

```
A[C>A]A 0.051058 0.051058 0.101420 0.044490 0.051058 0.101420
A[C>A]C 0.039010 0.039010 0.074532 0.033841 0.039010 0.074532
A[C>A]G 0.013127 0.013127 0.023615 0.011521 0.013127 0.023615
A[C>A]T 0.042381 0.042381 0.084768 0.037095 0.042381 0.084768
C[C>A]A 0.062427 0.062427 0.115050 0.053886 0.062427 0.115050
C[C>A]C 0.051509 0.051509 0.091052 0.043542 0.051509 0.091052
```

```
> dim(SimData$TrueH)
[1]     6 10027
> SimData$TrueH[, 1:5]
            Sample1      Sample2      Sample3      Sample4      Sample5
SBS1    7.308622e+01 4.573251e-96 1.239838e-75 5.690435e-83 2.945896e-07
SBS2_13 1.795944e-19 4.129726e-47 2.046148e-60 2.191433e+01 2.741921e+01
SBS4    1.897719e+02 1.120712e+02 1.388643e-21 2.232536e-40 5.478792e-42
SBS5    2.695517e+02 3.070634e-11 2.803694e+01 6.568151e+01 1.006008e+02
SBS40   3.614360e-01 1.441568e+01 1.143864e-06 1.196264e-14 1.584391e-11
SBS89   7.538808e+00 1.783676e-08 4.931170e-17 1.832204e-19 6.009490e-13
```

```
> dim(SimData$TrueW_TMB)
[1] 96  6
> head(SimData$TrueW_TMB)
                SBS1       SBS2_13        SBS4         SBS5       SBS40       SBS89
A[C>A]A 1.214776e-04 0.0006839133 0.02363349 0.007220416 0.02148811 0.02136339
A[C>A]C 5.399604e-04 0.0005638725 0.03221248 0.009811094 0.01763674 0.02029637
A[C>A]G 1.936003e-04 0.0009479074 0.06969638 0.008880203 0.01775016 0.05123523
A[C>A]T 2.193298e-04 0.0002092971 0.02064726 0.004970584 0.01409405 0.01721664
C[C>A]A 4.659486e-05 0.0006579641 0.04922772 0.004869570 0.01726367 0.01784639
C[C>A]C 3.733196e-04 0.0006075963 0.06789517 0.005619718 0.01703851 0.01466225
```

4 SALMON procedure

The primary purpose of the SALMON algorithm is to conduct mutational signature analysis for targeted sequenced tumors, summarized in Figure 1. Somatic mutations (e.g., single base substitutions [SBS]) detected by the targeted sequencing are summarized into the mutation type matrix **V** based on the mutation and its genomic

context. Given the mutation type matrix **V**, *de novo* tumor mutation burden (TMB) signatures are discovered by SigneR (Rosales et al., 2017), adjusting the panel sizes (Section 4.2). Next, we map the *de novo* signature obtained by signeR to the TMB-based catalog signatures (`SimData$W_TMB`) using the penalized non-negative least squares (Tibshirani, 2011; Section 4.3). Then given the mutation type matrix (**V**), panel size of each tumor (**L**) and mapped TMB-based catalog signature profiles, signature activities are estimated for every tumor by the Expectation-Maximization (EM) algorithm simultaneously (Section 4.4). Finally, we computed the expected number of mutations due to each signature, called signature expectancy, for all tumors simultaneously (Section 4.5).

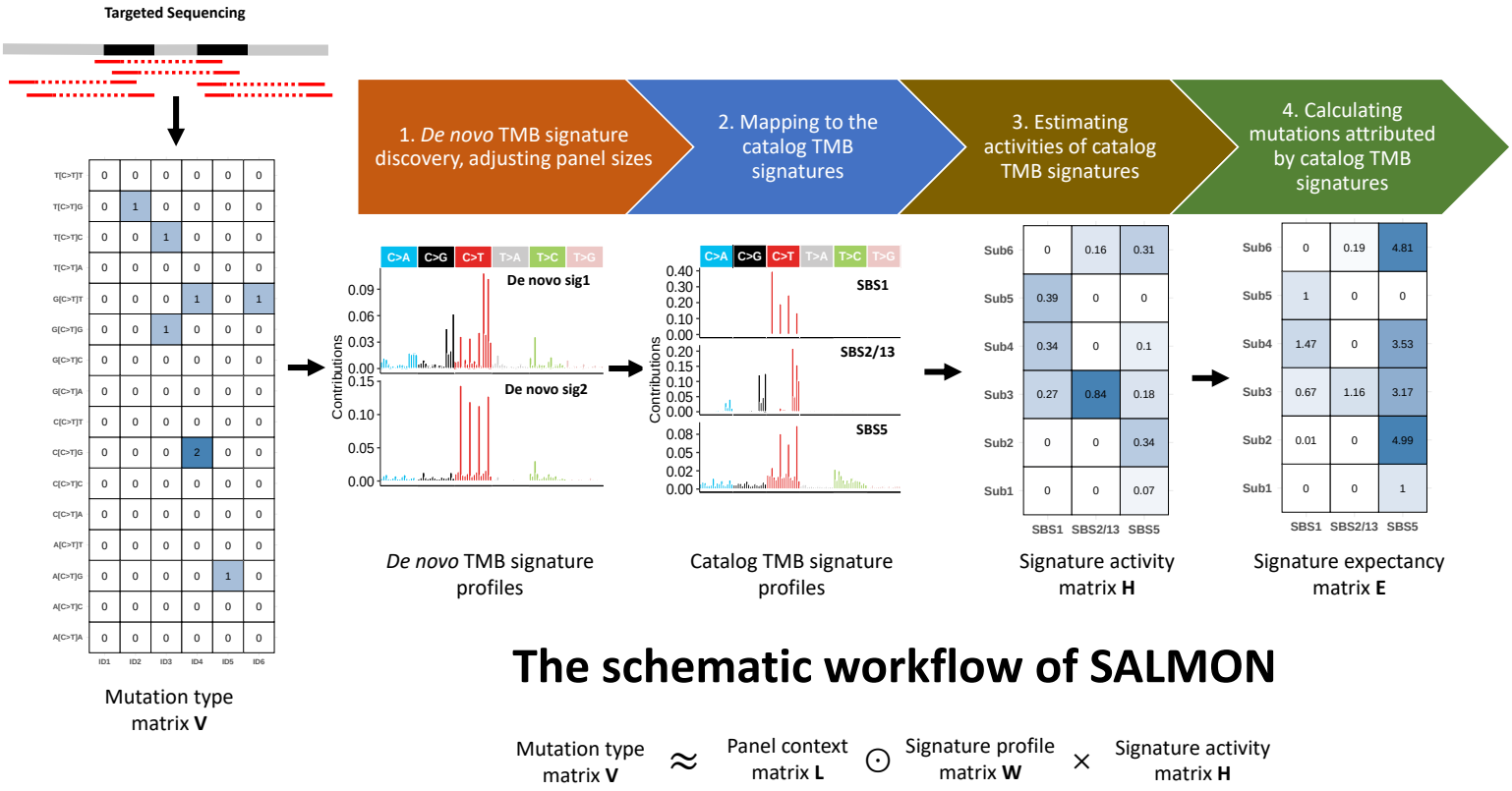

Figure 1: The schematic workflow of SALMON (Signature Analyzer for Low Mutation cOuNts).

4.1 Input data

For the mutation catalog matrix (**V**), it could be an **R** `dataframe` or `matrix` of which elements are non-negative counts. Each column of **V** corresponds to a tumor (or sample), while its row represents a mutation type. Although the selection of the number of signatures is independent of the order of mutation type, we specified the order of mutation type according to the COSMIC database (SBS signature[2]).

The element of the panel size matrix (**L**) represents the length of the trinucleotide contexts per million base pair to the corresponding sequencing panel. Similarly, each column of **L** corresponds to a tumor (or sample) while its row represents a mutation type.

4.2 Application of the signeR algorithm

To estimate *de novo* SBS signatures $\widehat{\mathbf{W}}$, we first apply signeR (need to reference) which finds an optimal solution $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{H}}$ such that $\mathbf{V} \approx \mathbf{L} \circ \widehat{\mathbf{W}}\widehat{\mathbf{H}}$. Since signeR uses a Bayesian approach with Markov chain Monte Carlo (MCMC), it is computationally intensive and time-consuming if the entire 96×10027 matrices **V** and **L** are used.

Therefore, we used pooled tumor of 100 combined samples, noting that pooling would not change the TMB-based signature profile. We are first interested in estimating **W**. In the codes below, `V_sum` and `L_sum` are combined mutation count matrix and the sum of panel size for these corresponding combined samples used to create `V_sum`, respectively. Then they are main arguments, `M` and `Opport` of `signeR()` function as shown below. Note that the resulting $\widehat{\mathbf{H}}$ has a different dimension from the original matrix **V** by pooling samples.

Another argument of `signeR()` function we used is `nlim` which specifies the range of possible number of optimal signatures. It needs much time to run the function if the upper bound of `nlim` is large. From our limited experience, the optimal number of signatures computed by `signeR()`, ranging between 2 to 5. Once `signeR()` is done, `BICboxplot()` can be used to visualize the number of

[2] https://cancer.sanger.ac.uk/signatures/sbs/

optimal signatures by signeR (Figure 2).

```
> V_group <- as.data.frame(t(V))
> V_group$srting <- as.character((1:dim(V_group)[1])%/%100)
> V_sum <- V_group %>% group_by(srting) %>%
        dplyr::select('A[C>A]A':'T[T>G]T')%>%summarise_all(sum)
Adding missing grouping variables: `srting`
> V_sum=as.data.frame(V_sum[,-1])
> rownames(V_sum) = paste0("G",rownames(V_sum))
>
> L_group <- as.data.frame(t(L))
> L_group$srting <- as.character((1:dim(L_group)[1])%/%100)
> L_sum <- L_group %>% group_by(srting) %>%
        dplyr::select('A[C>A]A':'T[T>G]T')%>%summarise_all(sum)
Adding missing grouping variables: `srting`
> L_sum=as.data.frame(L_sum[,-1])
> rownames(L_sum) = paste0("G",rownames(L_sum))
> ## Extract W hat
> set.seed(4022022)
> signeR_re <- signeR(M=V_sum, Opport=L_sum, nlim=c(1,5))
Evaluating models with the number of signatures ranging from 1 to 5,
please be patient.
Evaluating 1 signatures.
EM algorithm:
|==========================================================| 100%
Running  Gibbs sampler for 1 signature...Done.
Evaluating 2 signatures.
EM algorithm:
|==========================================================| 100%
Running  Gibbs sampler for 2 signatures...Done.
Evaluating 3 signatures.
EM algorithm:
|==========================================================| 100%
Running  Gibbs sampler for 3 signatures...Done.
Evaluating 4 signatures.
EM algorithm:
|==========================================================| 100%
Running  Gibbs sampler for 4 signatures...Done.
Evaluating 5 signatures.
EM algorithm:
|==========================================================| 100%
```

```
Running  Gibbs sampler for 5 signatures...Done.
The optimal number of signatures is 2.
Running  Gibbs sampler for 2 signatures...Done.
>BICboxplot(signeR_re)
```

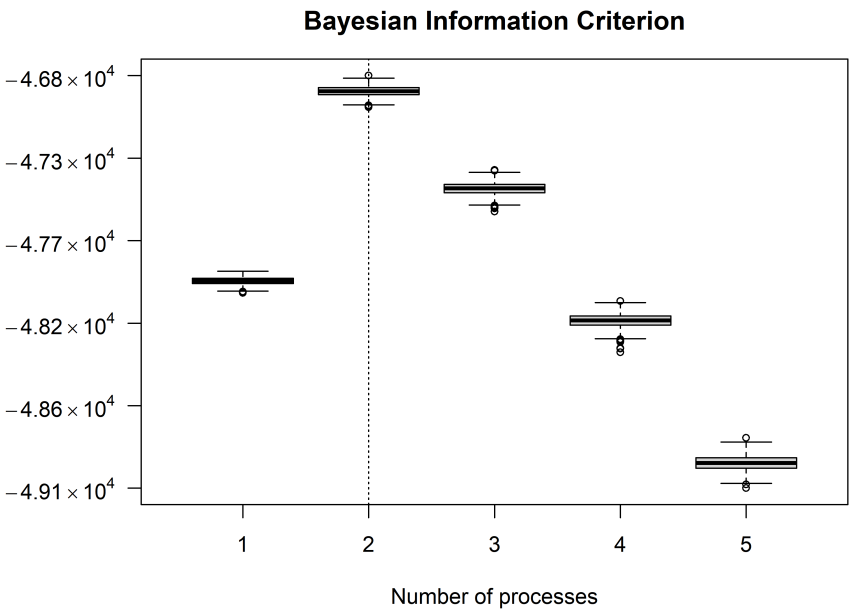

Figure 2: BIC (Bayesian Information Criterion) plot of signeR

4.3 Mapping *de novo* TMB-based signatures to TMB-based catalog signatures

Because the targeted sequencing panels have limited somatic mutations compared with WGS or WES, signeR identifies a few *de novo* TMB-based signatures (adjusted by panel size matrix **L**; Figure 3), which could be a linear combination of TMB-based catalog SBS signatures. Therefore we could consider mapping *de novo* signeR signature matrix $\widehat{\mathbf{W}}$ to TMB-based catalog signature $\mathbf{W}_0$ (stored in `SimData$W_TMB`).

To this end, we applied penalized non-negative least squares (pNNLS) for selecting the TMB-based catalog signatures. Specifically, we repeated pNNLS 100 times to reduce the randomness of cross-validation involved in pNNLS. Then TMB-based

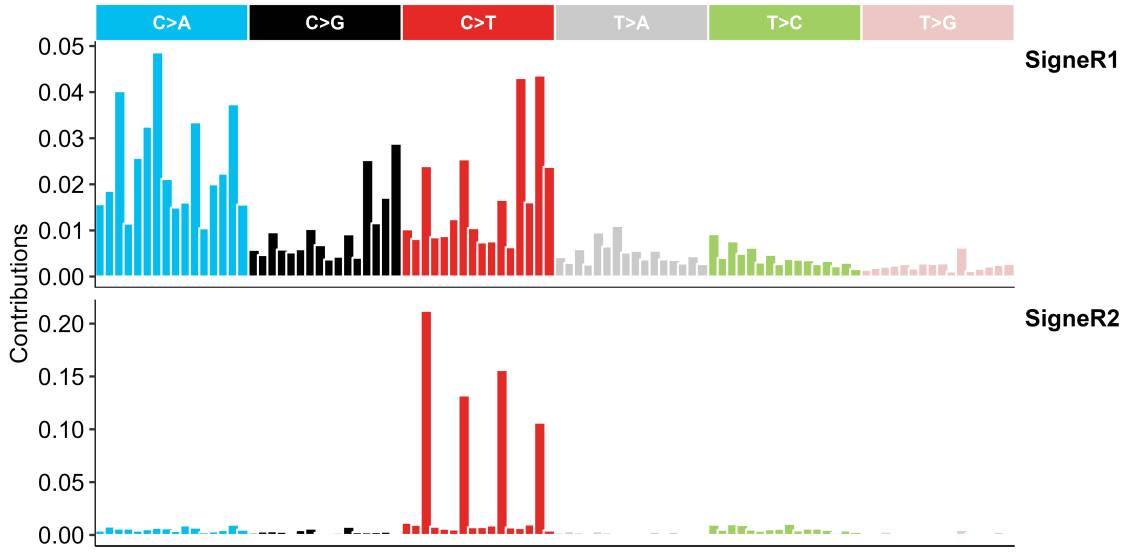

Figure 3: *De novo* TMB signature profiles identified by signeR (Section 4.2)

catalog signatures are selected with a coefficient greater than 0.1 in more than 80 repeats.

It can be done by calling `MappingSignature(W_hat, W_ref=NULL, niter=100, cutoff.I2=0.1, min.repeats=80)` in the *SALMON* package. `W_hat` is a *de novo* signatures from signeR (or any other signature analysis tool). `W_ref` is for the reference signature profiles, which will be mapped to. And the default is `NULL` with `SimData$W_TMB` used. The remaining `nIter`, `cutoff.I2` and `min.repeats` specifies the number of pNNLS repetitions (the default is 100), cutoff coefficient value to select signatures (the default is 0.1) and the minimum number of repetitions greater than cutoff coefficient value (the default is 80), respectively.

As a result, the mapped TMB-based catalog signatures are SBS1, SBS2/13, SBS4, SBS5, SBS40, and SBS89 with 100 frequencies. Their profiles are drawn in Figure 4. They are the same as signatures used to generate data (`colnames(SimData$TrueW_TMB)`).

```
> W_hat <- signeR_re$Phat
> MappedSig <- MappingSignature(W_hat = W_hat, W_ref = SimData$W_TMB)
> MappedSig
```

```
  Reference freq
1      SBS1  100
2   SBS2_13  100
3      SBS4  100
4     SBS40  100
5      SBS5  100
6     SBS89  100
> colnames(SimData$TrueW_TMB)
[1] "SBS1"    "SBS2_13" "SBS4"    "SBS5"    "SBS40"   "SBS89"
```

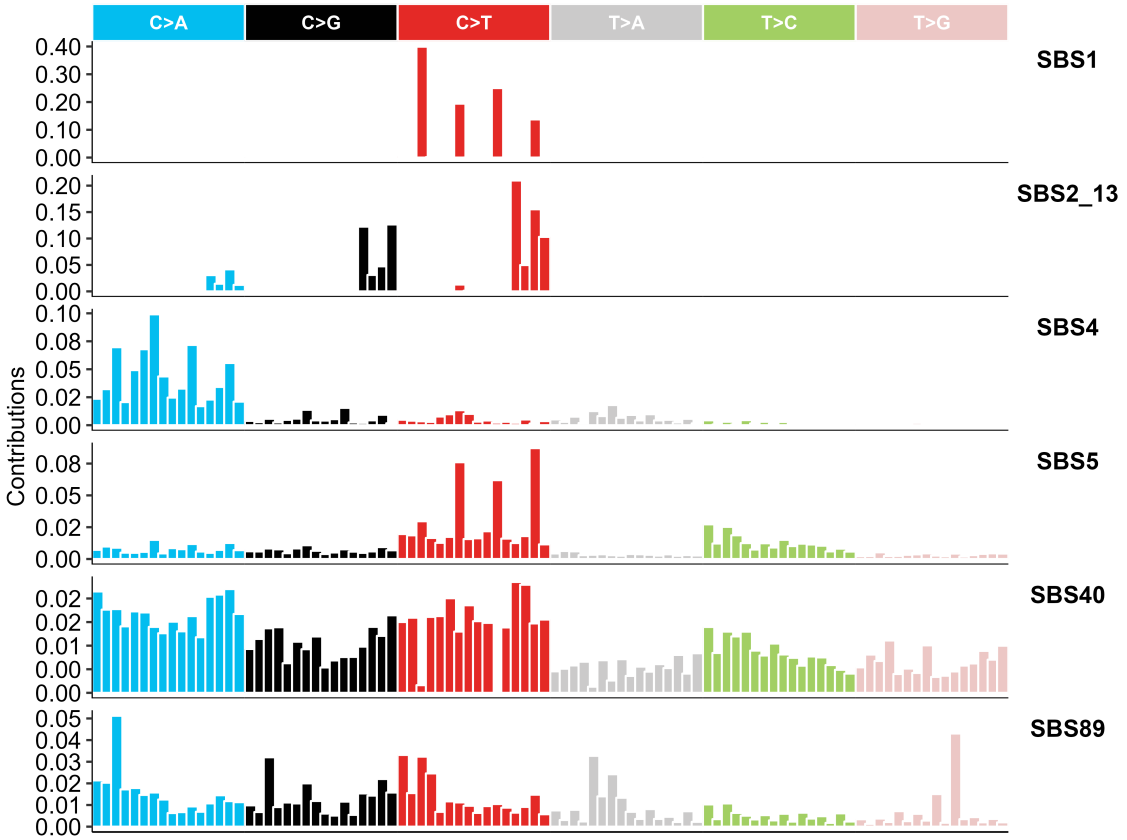

Figure 4: The mapped TMB-based catalog signatures obatined by `MappingSignature()` function

4.4 Estimation of signature activities by an expectation-maximization algorithm

Let $\mathbf{W}^*$ be the mapped TMB-based catalog signature profiles. Then given the original mutation catalog matrix **V**, the panel size matrix **L**, and $\mathbf{W}^*$, we can extract the signature activities $\widehat{\mathbf{H}}$ with the function `EstimateSigActivity(V, L, W, n.start=50, iter.max=5000, eps=1e-5)` in the *SALMON* package. The three main input matrices are **V**, **L**, and $\mathbf{W}^*$ (or any signature profiles).

Because the EM algorithm is used to estimate signature activities, `n.start`, `iter.max` and `eps` control EM part. Because the convergence to a local saddle point can be an issue of the EM algorithm, it would be good practice to try multiple initial values (`n.start`, the default is 50). For each initial value, the default value of the maximal iteration of the EM algorithm (`iter.max`) is 5000 and the stopping tolerance (`eps`) is set to 10^{-5}.

The function `EstimateSigActivity()` returns the signature activities matrix estimate $\widehat{\mathbf{H}}$, convergence status (1 for converged and 0 for not converged), and the largest log likelihood among `n.start` initial values.

```
> SBS.list <- mixedsort(unique(MappedSig[,"Reference"]))
> W_star <- as.matrix(SimData$W_TMB[,SBS.list])
> H_hat <- EstimateSigActivity(V = SimData$V, L = SimData$L, W = W_star)
> H_hat$H[, 1:5]
             Sample1      Sample2      Sample3      Sample4      Sample5
SBS1    2.954223e+02 1.269767e-90 0.000000e+00 2.808485e-69 3.176680e-39
SBS2_13 7.835971e-24 1.578359e-14 1.167664e+01 2.827798e+01 1.839790e+01
SBS4    1.809574e+02 4.490885e-04 4.085940e-67 3.033663e-03 2.052067e+01
SBS5    4.219347e+00 1.616008e-06 1.287058e-42 9.326036e+01 3.028428e+00
SBS40   4.463818e+01 8.555236e+01 2.433144e-40 2.167192e+01 7.239176e+01
SBS89   4.786104e-08 5.261826e+01 1.445489e-59 2.244540e-04 1.094364e-04
> H_hat$converged
[1] 1
> H_hat$loglike
[1] -0.2172882
```

4.5 Calculation of signature expectancy

The final step is to compute the expected number of mutations attributed by TMB-based catalog signatures (i.e., signature expectancy) given matrices **L**, $\mathbf{W}^*$ (the mapped TMB-based catalog signature profiles; see Section 4.3) and $\widehat{\mathbf{H}}$ (the signature activities matrix estimate; see Section 4.4). The signature expectancy can be computed via `CalculateSigExpectancy(L, W, H)` in the *SALMON* package. It decomposes mutations of a tumor into mutations due to the identified signatures. For example, Sample 1 has 14 mutations, about 4.8 caused by signature 1, 7 by signature 4, about 0.2 by signature 5 and two by signature 40.

```
> SigExp <- CalculateSigExpectancy(L = SimData$L, W = W_star, H = H_hat$H)
> round(SigExp[,1:5], 2)
        Sample1 Sample2 Sample3 Sample4 Sample5
SBS1       4.79    0.00       0    0.00    0.00
SBS2_13    0.00    0.00       1    1.08    0.82
SBS4       7.04    0.00       0    0.00    0.77
SBS5       0.21    0.00       0    2.93    0.07
SBS40      1.95    3.87       0    0.99    3.34
SBS89      0.00    2.13       0    0.00    0.00
> colSums(SimData$V[,1:5])
Sample1 Sample2 Sample3 Sample4 Sample5
     14       6       1       5       5
```

References

Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. and da Silva, I. T. (2017). signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc B* **73**, 273–282.