

MS.liverK

F. Petitprez

Introduction

Cancer is a highly heterogeneous disease, with marked differences between patients. Even within a specific type of cancer, such as hepatocellular carcinoma (HCC), heterogeneity persists on several points: diverse genetic mutations, resulting in diverse cancer phenotypes. Therefore, characterizing these differences within a cancer type is highly important to better understand the mechanisms that undergo cancer maintenance and progression. A more thorough classification of cancer samples into different subtypes also has tremendous benefits in terms of precision medicine, as it allows clinicians to target therapies to the specific disease of each patient. A striking example of targeted therapy is the treatment of resistant patients by anti-PD1 antibody immunotherapy. To better understand cancer heterogeneity and identify patients more likely to respond to a specific treatment, the notion of molecular subgroups has emerged.

In HCC, several teams have designed unsupervised or supervised classification strategies. MS.liverK provides easy-to-use functions to characterize HCC samples from transcriptomic data. It implements 6 different molecular subtypes classification and scores 45 genes signatures about molecular subtypes, prognosis or biological pathways.

Molecular classification provided in MS.liverK

Lee

In 2004, Lee and colleagues established a two-classes unsupervised classification of HCC (Lee et al., 2004). Their study was based on the gene expression profiles of 91 samples, which they divided into two classes highly associated to survival.

Boyault

Three years later, another team proposed a 6-groups unsupervised classification (Boyault et al., 2007). Their subgroups were found by unsupervised clustering of transcriptome data on 57 HCC samples and they have been linked to clinical and genomic information: G1 and G2 are linked to HBV virus' DNA presence and differ by the main mutations found. G3 is characterized by TP53 mutation and overexpression of cell cycle genes. G5 and G6 have mutations that activate Wnt pathway. G4 is an heterogeneous group.

Chiang

Chiang and colleagues, in 2008, derived a classification of HCC into 5 subgroups from the analysis of 91 hepatitis C-related HCC samples (Chiang et al., 2008). Their classes are labeled "CTNNB1", "Proliferation", "Inflammation", "Polysomy chr7" and an unannotated class. "CTNNB1" is enriched in mutations on the CTNNB1 gene, "Proliferation" class is enriched in IGF1 and RPS6, "Polysomy chr7" is defined by polysomy of chromosome 7.

Hoshida

In 2009, Hoshida and colleagues proposed a classification into 3 subgroups, from an analysis of genes expression profiles in nine cohorts, totalizing 603 patients. Their subgroups correlate with tumor size and cellular differentiation. Subgroup S1 contains samples with abnormal activation of Wnt pathway, S2 is characterized by proliferation and MYC and AKT activation, while S3 correlates with hepatocyte differentiation.

Roessler

In 2010, Roessler and colleagues (Roessler et al., 2010) presented a two subgroup classification to identify the risk of metastasis-related recurrence. They used 2 cohorts, totalizing 386 patients.

Example of use

Code

To give an example of use of MS.liverK, we will apply it to a cohort of 91 patients from GSE20238 (Minguez et al., 2011) that is provided with the package.

First, load MS.liverK

```
library("MS.liverK")
```

```
## Loading required package: pamr
```

```
## Loading required package: cluster
```

```
## Loading required package: survival
```

To load the GSE20238 data, enter :

```
data("GSE20238")
```

This dataset consists of two data frames. The first one is called **GEP** (for Gene Expression Profiles). It regroups the transcriptomic data from the 91 samples, at the probe set level. Data has previously been normalized and logged. The second data set, called **AnnotProbeset**, has two columns and one line per probe set. It contains the probe set's ID and the corresponding HUGO gene symbol.

We can now run the characterization of the samples, by using :

```
subtypes <- MS.liverK.subtypes(probesData=GSE20238$GEP, probesSymbols=GSE20238$AnnotProbeset)
```

subtypes now contains a list with 3 elements :

1. `subtypes$molecularSubtype` regroups all information concerning molecular subgroups classification :
 - `subtypes$molecularSubtype$prediction` contains the classes predicted by MS.liverK
 - `subtypes$molecularSubtype$signatures` contains the score of subtypes gene signatures
2. `subtypes$prognosis` regroups results for prognosis :

- `subtypes$prognosis$prediction` contains the 2 classes prognosis output from Nault's signature (Nault et Zucman Rossi, 2013)
 - `subtypes$prognosis$signatures` contains the score of prognosis gene signatures
3. `subtypes$biologicalPathwaysSignatures` contains the scores of biological pathways signatures.

MS.liverK provides a function to discretize continuous signatures scores. To use it, run :

```
subtypes_discretized <- MS.liverK.discretize(subtypes)
```

This transforms the continuous scores into a discrete score from 1 to 5. This discrete score is data-dependant : the lower scores of the series will be set to 1 and the higher to 5.

To visualize the result, you can then run :

```
x11(width = 8.3, height = 11.7)
MS.liverK.plot(subtypes)
```

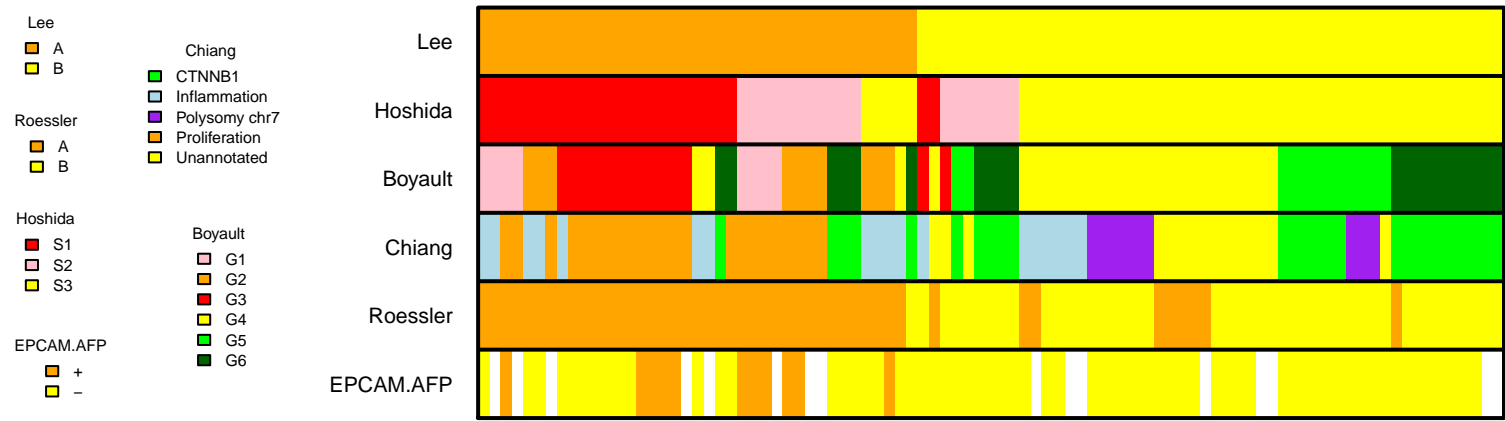
or

```
x11(width = 8.3, height = 11.7)
MS.liverK.plot(subtypes_discretized)
```

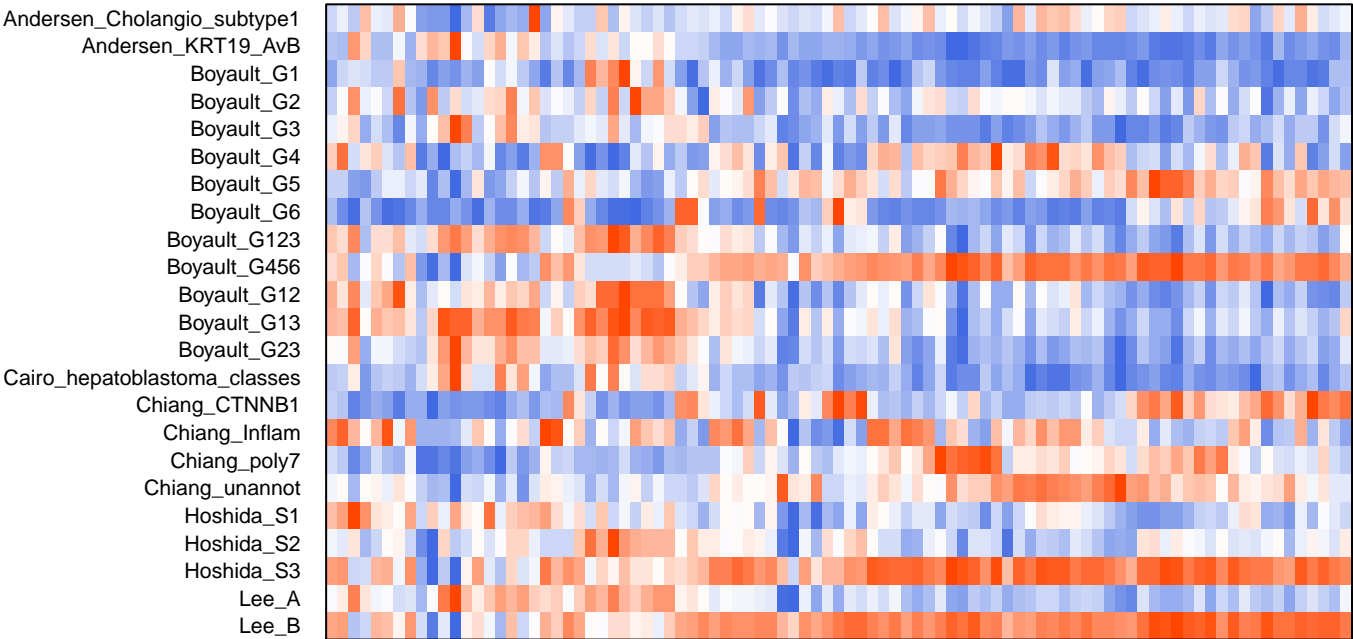
Graphical output

The two next pages present the graphical outputs obtained by the last two calls.

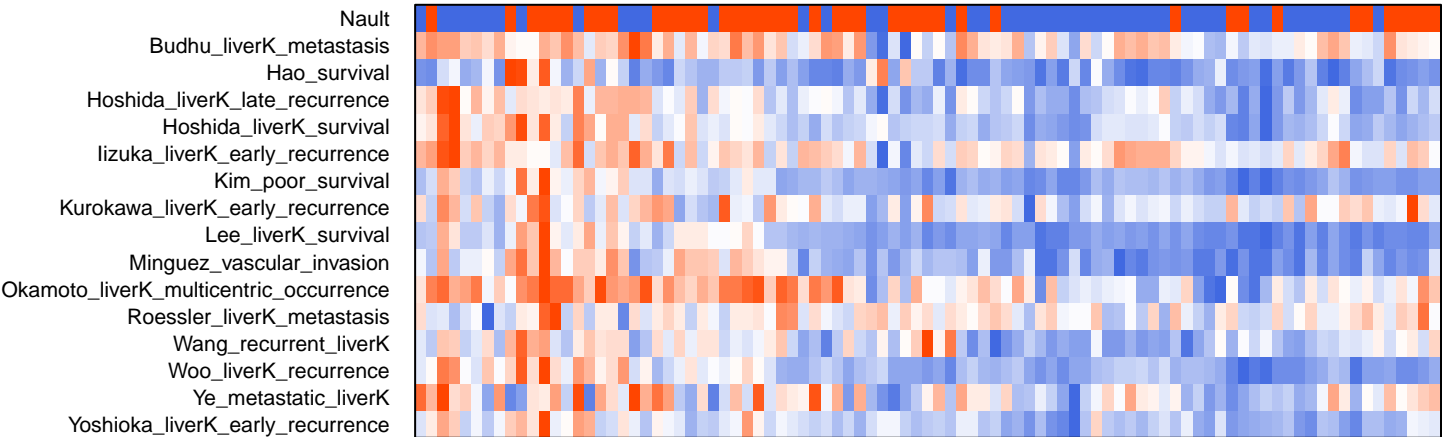
Molecular subtypes



Molecular subtypes signatures



Prognosis



Biological pathways signatures

