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 is 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 biologial 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 differenciation. Subgroup S1 contains samples with abnormal activation of Wnt pathway, S2 is characterized by proliferation and MYC and AKT activation, while S3 correlelates with hepatocyte differenciation.

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.

Prognosis and biological pathway signatures

MS.liverK implements a prognostic signature published by Nault et al. (2013). Several teams have published gene-signatures related to biological functions/pathways involved in liver cancer oncogenesis. These gene-signatures relate to TGFB1 signalling (Coulouarn et al., 2008), MET pathway (Kaposi-Novak et al., 2006), stemness (Oishi et al., 2012; Yamashita et al., 2008), EPCAM (Yamashita et al., 2008) and hypoxia (van Malenstein et al., 2010). These signatures are implemented in MS.liverK.

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). Is 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 aresults for prognosis :

- `subtypes$prognosis$prediction` contains the 2 classes prognosis output from Nault's signature (Nault et Zucman Rossi, 2013)
- `subtypes$prognosis$signatures` is NULL by default, itcontains the score of prognosis gene signatures if the parameter PrognosticSignatures has been set to TRUE (advanced users only)

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)
```

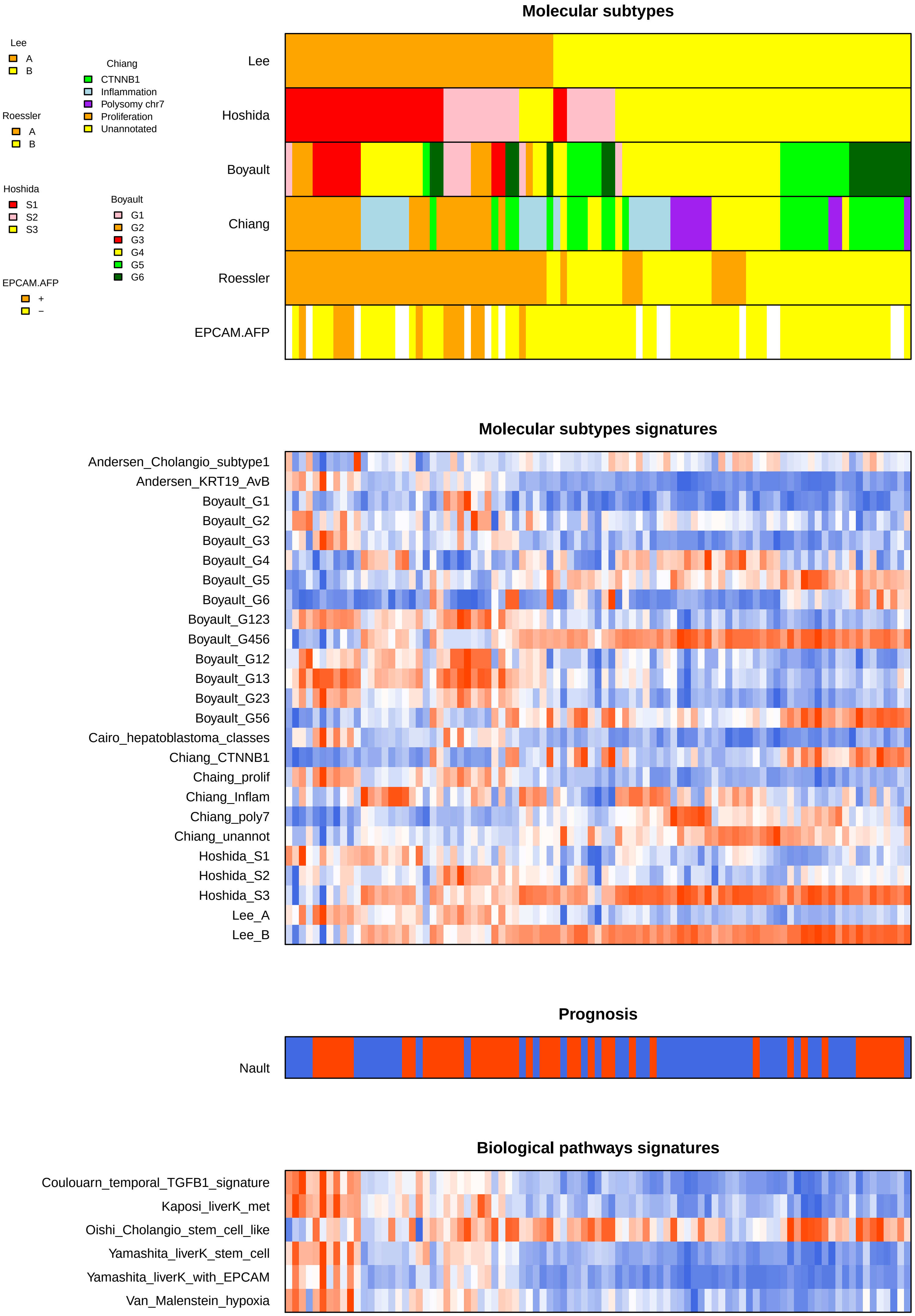
Dataset conversion

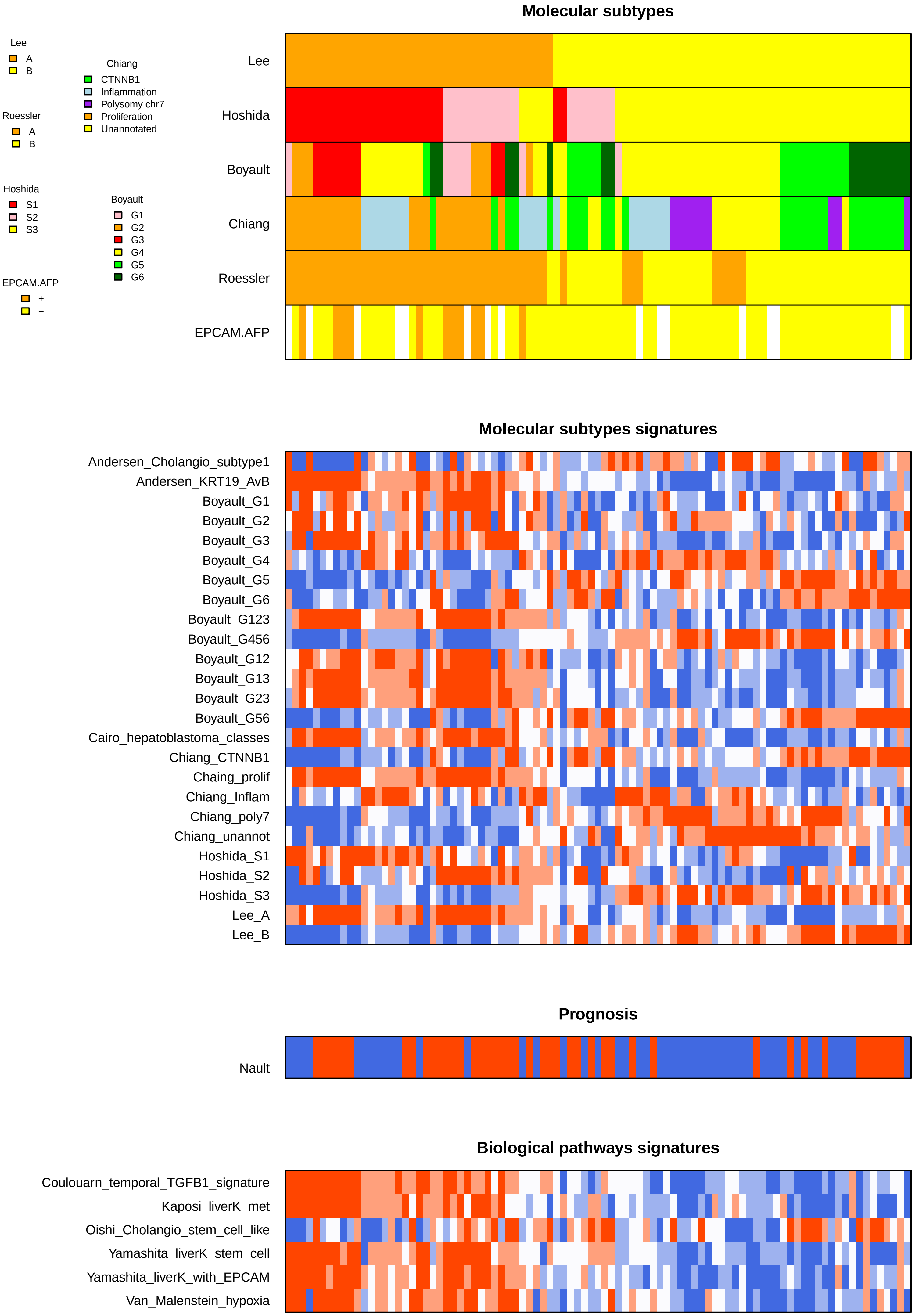
The predictors used by MS.liverK have been optimized for data aquired using Affmetrix Human Genome U133 Plus 2.0 Array. For data measured with other platforms, we encourage the users to convert it using the provided MS.liverK.convert function. To this end, the dataset must be already aggregated by gene, with HUGO gene symbols as row names, and samples in columns. For an optimal functionment, it is advised to have at least 50 samples. To convert your data set, simply replace data by the name of your dataset in the following code:

```
MS.liverK.convert(data)
```

Graphical output

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


Molecular subtypes
Lee
A
B
Roessler
A
B
Hoshida
S1
S2
S3
EPCAM.AFP
+
−
Chiang
CTNNB1
Inflammation
Polysomy chr7
Proliferation
Unannotated
Boyault
G1
G2
G3
G4
G5
G6
Lee
Hoshida
Boyault
Chiang
Roessler
EPCAM.AFP
Molecular subtypes signatures
Andersen_Cholangio_subtype1
Andersen_KRT19_AvB
Boyault_G1
Boyault_G2
Boyault_G3
Boyault_G4
Boyault_G5
Boyault_G6
Boyault_G123
Boyault_G456
Boyault_G12
Boyault_G13
Boyault_G23
Boyault_G56
Cairo_hepatoblastoma_classes
Chiang_CTNNB1
Chaing_prolif
Chiang_Inflam
Chiang_poly7
Chiang_unannot
Hoshida_S1
Hoshida_S2
Hoshida_S3
Lee_A
Lee_B
Prognosis
Nault
Biological pathways signatures
Coulouarn_temporal_TGFB1_signature
Kaposi_liverK_met
Oishi_Cholangio_stem_cell_like
Yamashita_liverK_stem_cell
Yamashita_liverK_with_EPCAM
Van_Malenstein_hypoxia



Molecular subtypes

Molecular subtypes signatures

Prognosis

Biological pathways signatures