A pipeline for exploratory and pathway analysis of NGS data

Alejandra Cervera[1], Antonio Neme[2], Sampsa Hautaniemi[1]
[1]Research Programs Unit, Genome-Scale Biology, and Medicum, Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, Finland
[2]School of Medicine/ Institute of Biomedicine, University of Eastern Finland, Kuopio
alejandra.cervera@helsinki.fi, antonio.nemecastillo@uef.fi, sampsa.hautaniemi@helsinki.fi

We implemented a pipeline for processing and analysis of high-throughput sequencing and microarray level 3 data from ICGC Cancer Genome Consortium Challenges with the aim of finding the driver mechanisms in three cancers: Lung Adenocarcinoma (LUAD), Kidney Renal Clear Cell Carcinoma (KIRC), and Head and Neck Squamous Cell Carcinoma (HNSC).
We used a random forest to identify the genes that are most relevant for classifying the samples into normal tissue and tumor. The top 50 genes for each of the three cancers served as input for the SOMs showed in Figure 1. It can be observed that 50 genes are enough to achieve a fairly good separation of classes. DESeq2 was used to produce a list of differentially expressed genes (DEGs). Heatmaps using the genes with the greatest fold change from the DEGs are shown in Figure 2. It can be observed that the DEGs also achieve a fairly good separation of classes. The gene lists obtained from the random forest classification and the DEGs were used to identify possible relevant pathways for each cancer. From the top most represented pathways all genes belonging to those pathways were extracted. In the three cancers the Metabolic pathway (hsa01100) was enriched, and Salivary Secretion (has:04970), Cell Adhesion (hsa04514), and PI3K-Akt signaling pathways were enriched for HNSC, KIRC, and LUAD respectively. The new gene list obtained from the pathways was queried against the other levels of data: proteomics, miRNAs, copy number variation, and methylation. In the case of proteomics, we checked for indication of phosphorylation in any of the genes involved. Furthermore, DESeq was also used for obtaining a set of differentially expressed miRNAs and miRbase was used to find the known target genes. Figure 3 shows the miRNAs pathways in Cancer from where we found MIR17HG expressed.
The pipeline for preprocessing, analysing, and integrating all the datasets was implemented in Anduril; all steps are automated and it is available upon request (and soon made available in Anduril's website). Anduril is a framework for scientific data analysis that automates parallelization making it ideal for working with large datasets.

a) LUAD b) KIRC c) HNSC

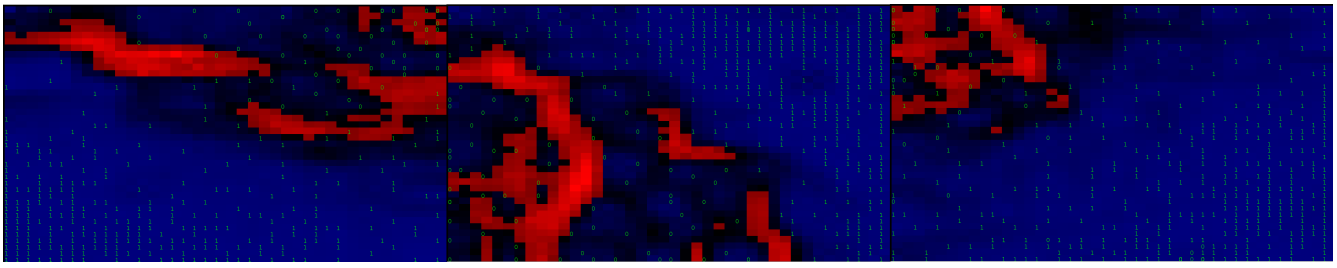

Figure 1. SOM using top50 most relevant genes from random forest classification of normal (0) vs tumor (1) samples. The color represents the distance between the points in the lattice (closer → blue, farther → red).

Figure 2. Heatmaps from gene expression of differentially expressed genes for each cancer.

a) LUAD b) KIRC c) HNSC

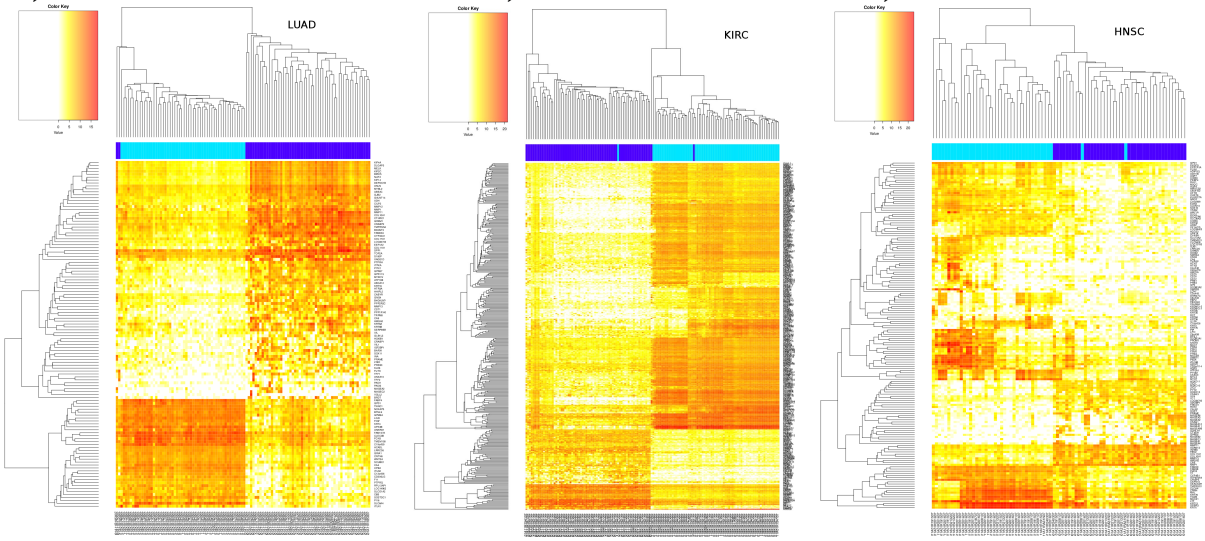

Figure 3. MiRNAs pathways in Cancer: miR-17~92 known as oncomiR-1 is known to be disregulated in cancer and we found MIR17HG (the primary transcript of the cluster) to be expressed in our samples.