14.08.2016 Vancouver

From Dr Groppa, The Rossi Lab
Subject: status of RNA Seq data analysis in collaboration with bioinformatics-units in Padua and Vancouver

Background:
Briefly, the RNA Seq data profiled four different cell populations collected from skeletal muscle tissue in steady state conditions (undamaged) or at different time points after notexin-injection. The cell subsets are: endothelial cells (EC, blood vessels), fibro-adipogenic mesenchymal progenitors (FAP), muscle progenitors, and inflammatory cells. Notexin-injection leads to an *acute muscle damage* in wild type mouse model, which is repaired by 14 days. As opposite to a regeneration process, we used CCR2 knock out mouse model where the notexin-injection causes a *transient fibrosis formation*, which dissolves and is replaced by normal regeneration by 21 days. Approximately, we can assume a time-shift of 7-10 days between the regeneration occurring in the wild type mouse model vs the knock out one.

Bioinformatics analysis-strategies:
The time course-gene profile of the different cell subsets is analyzed according two bioinformatics strategies (my apologies for the non-bioinformatics terminology!):

1) *"unbiased" analysis conducted by Farnush with the supervision of Sara*
   In this analysis Farnush is working with FPKM values, excluding genes with FPKM<1 and analyzing the different cell populations separately. For each cell subset she is identifying gene-clusters that is a group of genes that display the same expression-pattern through the time. The next step is to find similar clusters in two different cell subsets, for example FAP and EC, taking into account a *delta shift* that has to be defined (e.g. 1 hour?). Theoretically, we expect that we will find ligands in the cluster of one subset, for example ligands in FAP, whereas we will find the receptor-downstream genes in the respective cluster in the second cell subset, e.g. receptor-downstream genes in EC. I am talking about receptor-downstream genes because the majority of receptors probably do not display a time-dependent expression, rather activation (e.g. phosphorylation).
   <u>Current problems:</u>
   a) Farnush has encountered some problems because of the few time points to identify the clusters mentioned above, especially for inflammatory cells and muscle progenitors. I would suggest proceeding with the analysis on FAP and EC to develop bioinformatics tools. Later, the latter analyses/tools will be applied for studying the remaining populations;
   b) there is a general issue in regard of database-availability, in particular for the database of interaction ligand-receptor and receptor-downstream genes.

2) *TimeClip and database-supported analysis conducted by Paolo*

Paolo has been using the raw counts for his analyses excluding genes higher than 100 counts in at least the 40% of the time points. We should understand what is the relation between raw counts and FPKM (in term of numbers!) in order to understand the threshold we are using in the two different strategies 1) and 2). Taking advantage of the TimeClip, Paolo has assigned P value that allows us to study the gene-expression through the time, thereby identifying the most significantly regulated genes in the different cell subsets. Besides, Paolo has started a ligand-receptor analysis converting a previous ligand-receptor database from human to mouse and generated preliminary results (PMID 26198319). Here, Paolo has suggested different strategies to filter the outputs using 1) P value of ligand AND/OR 2) P value of receptor AND/OR 3) Spearman correlation. P value of ligand is a good parameter since ligands probably display time-dependent expression, however, P value of the receptor regulation and Spearman correlation might not be good filter-options, because of the poor time-dependent expression of receptors. In order to overcome this problem, the database of receptor-downstream genes is crucial such that we can associate the ligand-expression to receptor-activation (by looking at the receptor-downstream genes).

Final comment:

The general idea is that unbiased and biased analyses should converge to similar conclusions. Indeed, the unbiased analysis has a unique potential in terms of novel discoveries: for example, as Fabio suggested, we could use the cluster of genes generated by Farnush, and verify whether genes from the same cluster share sequence-specific DNA binding transcription factor. Such analysis might lead to the discovery of novel gene-regulation networks. This is a very interesting point, but I would suggest proceeding step by step.

Finally, I would suggest to focus the analysis on FAP and EC as their time course data collection is the most complete. Later, we can focus also on the other cell subsets when the bioinformatics tools have been established and the time course data collection of such cell subsets has be completed/refined.

What's next:

In the next days, Farnush and I will use the ligand-receptor mouse database used in the paper (PMID 27009580; *question: shall we expand this mouse-database using Gene Ontology inputs/BioMart/iRefWeb?*) to identify possible ligands in the clusters she has generated. Then, we will try to understand how to extrapolate the receptor-downstream genes from database as Kegg, which apparently is the most common used, and look for other data-sources.

When Paolo is back from vacation, I would suggest to repeat the last analyses he performed taking advantage of the mouse ligand-receptor database recently published (PMID 27009580), and perhaps help Farnush and me to find/generate the database receptor-downstream genes.

References

Here some papers that have been very useful for database source/general ideas/brainstorming discussion:

1) Intercellular network structure and regulatory motifs in the human hematopoietic system. PMID: 25028490

2) A draft network of ligand-receptor-mediated multicellular signalling in human. 26198319

3) Signaling Networks among Stem Cell Precursors, Transit-Amplifying Progenitors, and their Niche in Developing Hair Follicles. 27009580

4) timeClip: pathway analysis for time course data without replicates. 25077979

5) Loss of fibronectin from the aged stem cell niche affects the regenerative capacity of skeletal muscle in mice. 27376579