Requirements document for a bioinformatic pipeline

Andrea Di Nenno - Alessandro Manfredi - Leonardo Tanzi

August 2018

1 Abstract

This document aims to outline which are the requirements involved in a generic bioinformatic tool for the analysis and the profiling of trascriptomes. For this purpose we analyzed two already deployed frameworks: on the one hand the RNA-seq sequencing technique, a next-generation sequencing technique for transcriptome profiling [3], while on the other hand the genome-wide miRNA expression data, which is used to study miRNA dysregulation comprehensively [2]. Moreover, we looked at a couple of web applications, [4] and [1], that already implement a bio-informatic pipeline, in order to gain some insights and requirements such a tool should have. We will list requirements and the pipeline's general structure in the next sections.

2 Functional requirements

These requirements are essential for the pipeline to work properly.

- **Interoperability between components:** one of the most crucial requirement for this domain. Indeed since many algorithms have been designed to perform different tasks, there is the need to make all different components work seamlessly in a pipeline. This task is hard to achieve, as these algorithms may have been written in different programming languages and for sure by different teams and organizations, such that the output from one cannot be used directly as output for others. Bridging scripts are then necessary.
- **Sample outliers detection:** RNA-seq is a complicated multistep process, so a mistake in any of the steps may end up with biased or unusable data. As a matter of facts, it is not uncommon that some samples have low quality and often substitute samples are not available, especially for RNA-seq of clinical specimen. To avoid that, it is necessary to define strict RNA-seq data quality metrics to identify outliers that should be excluded from further downstream data analysis.

- **Sample swapping and mislabeling detection:** especially with huge amounts of RNA samples to be sequenced and analyzed, it is likely that some of them will end up being mishandled and swapped. Such errors can become a serious problem for downstream data analyses and interpretation of results, especially for longitudinal sample analyses. By comparing genetic markers among samples, such as single nucleotide polymorphisms, we can tackle this problem down.

- **Interactive data visualization:** another crucial requirements, since all bio-informatics tools we have analyzed so far presents a very poor user friendly interface, making hard for simple users other than for experimental scientists to access output data and maybe reuse them for downstream analysis. A valid tool should be up to date to the newest web technologies (e.g. Javascript), in order to organize results in a user-friendly manner, make them fully accessible via a web interface and enable end users to interactively digest analysis results in a user friendly manner.

3 Non-functional requirements

These requirements don't interfere with the pipeline functioning, but are anyway important for the growth of the technology and more importantly to allow a wide adoption of it by end users.

- **Open-sourceness:** all components of the pipeline must be freely available in the public domain, opening the way for further improvements.

- **Simple parameter setup:** especially for researchers new to this field, the selection of the right tool (between many different) and the specification of the parameters may be non trivial. It's required both experience and a deep knowledge of the algorithms. Our ideal bioinformatic tool should then facilitate this task, by means of a friendly user interface, along with a manual that drives them through the configuration extensively.

- **Input consistency:** as we have seen that not all tools require the same gene annotation file format (e.g. GTF or BED), we believe that a valid pipeline should avoid any discrepancy or inconsistency between gene annotations formats. Already deployed scripts that convert files from GTF to BED or viceversa are available, and should be incorporated into the data input phase, making the pipeline more fluid.

- **Scalability:** ideally the pipeline should be designed to be independent from the bio-informatics tool itself. Although this requirement could be hard to achieve, it would lead to a major adoption in the industry, increasing so the amount of data publicly accessible by scientists for further analyses.

- **High speed processing:** this requirement can be met by introducing an high level of parallelism wherever possible. For instance in the very first step, where each sample is processed independently from other, parallelism may be used.

.

4 Pipeline

In this section we are going to illustrate all the operations involved in a bioinformatic pipeline, along with the challenges for each of them.

1. **Data collection:** The first and perhaps the most important step in every data analysis work is the quality control over the data taken as input. As we have mentioned above in the requirements, it's crucial that input data is filtered and cleaned properly, to avoid outliers other than mislabeling or swapped data.

2. **Align and Assemble:** A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity, thus making the definition of relationships between the sequences valuable. In general, the most widely used alignment tool is TopHat because it assures fast and high throughput alignment of sequencing reads from trascriptomes (e.g. RNA-seq). QuickRNASeq instead, uses STAR, an algorithm that aligns spliced sequences of any length with moderate error rates, providing scalability for possibly newer sequencing technologies, and generates output files ready for transcript and gene expression quantification.

3. **Computational Analysis:** In this step results coming from each individual RNA-seq sample are merged, in order to proceed with all the different across-sample metric evaluations that need to be calculated, such as correlation-based QC, SNP correlation matrix among samples.

4. **Data Visualization:** Thanks to modern Web 2.0 technologies, the project report can be visualized through the same entry point. This report includes interactive visualizations such as:

    - **QC Metrics:** read mapping summaries, read counting statistics, SNP correlations among samples, number of expressed genes at various RPKM cutoffs, and correlations among gene expression profiles;
    - **Parallel plot:** it offers an integrated view of linked QC measures for a single sample or group of samples. It is a common way to visualize high-dimensional data and it is used widely in multivariate data analysis.
    - **Expression table:** it provides links to raw read counts, a normalized RPKM table and gene expression levels.
    - **SNP correlation plots:** they help to verify whether samples are from the same subject or not.

References

[1] National Center for Biotechnology Information. Basic local alignment search tool, August 2018.

[2] Sarah Du Shanrong Zhao, William Gordon. Quickmirseq: a pipeline for quick and accurate quantification of both known mirnas and isomirs by jointly processing multiple samples from microrna sequencing. 2017.

[3] Sarah Du Shanrong Zhao, William Gordon. Quickrnaseq lifts large-scale rna-seq data analyses to the next level of automation and interactive visualization. 2017.

[4] Kengo Saton Yuki Kato. Ip based prediction of rna pseudo knots, March 2014.