CEFIC Long-range Research Initiative LRI-C4
Transcriptomics bioinformatics best practices in Toxicogenomics for regulatory application

C4 R-ODAF : Omics Data Analysis Framework for Regulatory Application

Marcha Verheijen[1], Weida Tong[2], Leming Shi[3], Tim Gant[4], Bruce Seligmann[5], Florian Caiment[1].

[1] Department of Toxicogenomics, Maastricht University, Maastricht, Netherlands

[2] U.S. Food and Drug Administration, Washington, USA

[3] Fudan University, Shanghai, China

[4] Public Health England & University of Surrey, Oxfordshire, UK
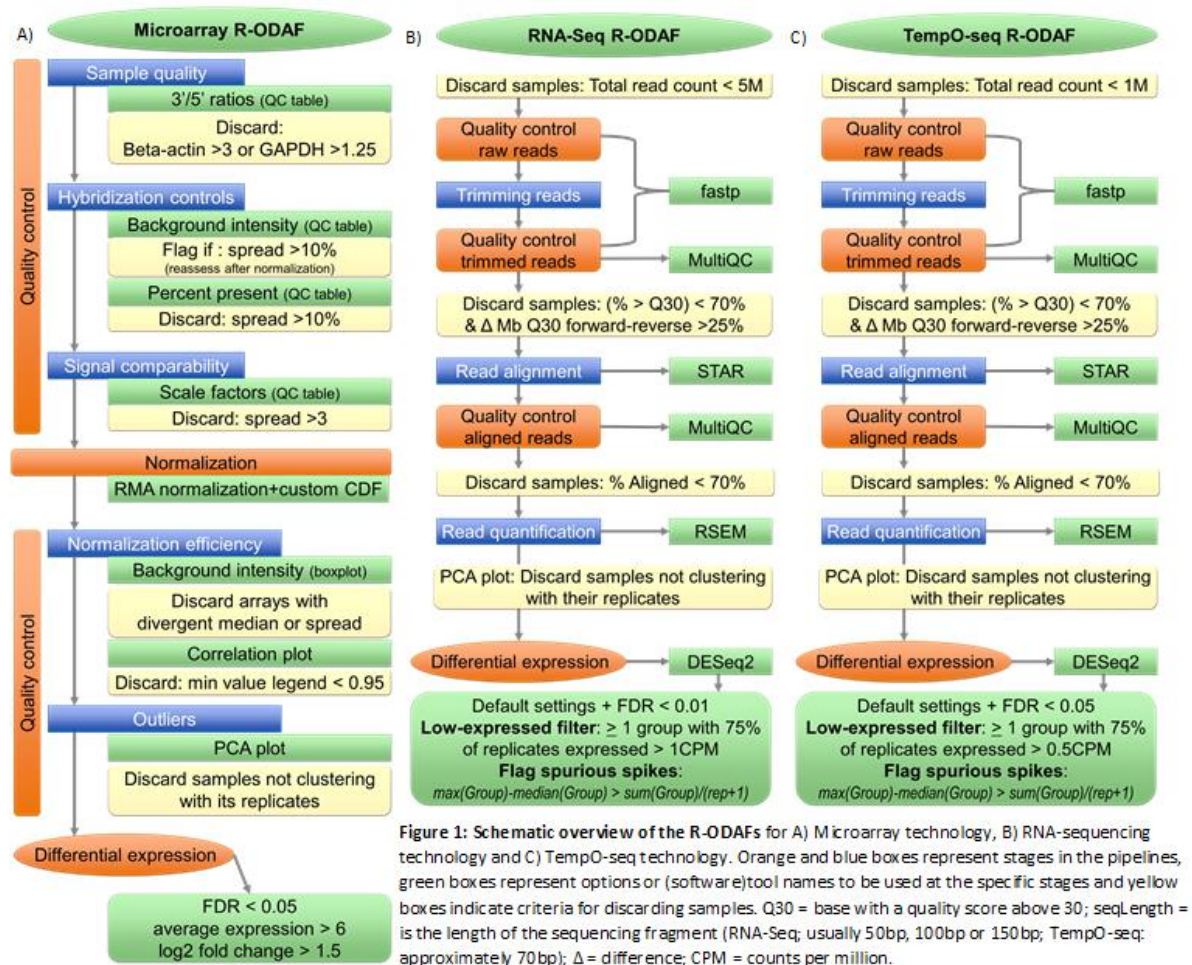
[5] Biospyder, Tucson, USA

Summary

The use of various omics techniques for scientific research is increasing. While toxicogenomics studies have already produced substantial data on diverse omics platforms, and despite the promises and excitement of 20 years ago when it was widely speculated that omics methods would reduce or even replace animal use and allow an enhanced understanding of hazard and susceptibility , to date there has been little routine application in regulatory toxicology. One of the reasons for this has been a trepidation about relying on the produced data. It has been argued that omics outputs might not be sufficiently reliable for regulatory application because the techniques, bioinformatics and interpretation can vary. For these reasons, the robustness of the obtained results is questioned. This reticence to trust omics data is further magnified by the lack of internationally agreed upon guidelines and protocols for both the generation and processing of omics data.

This C4 LRI project was established with the ambition to propose for the regulatory community an omics data analysis framework (R-ODAF) for the main transcriptomics platform available on the market: microarrays, RNA-sequencing or TempO-seq sequencing technologies. For this, in the first phase of the project, datasets of the three selected platforms were accumulated, and a complete review of all used data analysis pipeline were performed. After selecting of the most relevant tools and algorithms for each platform, the second phase of the project consisted in evaluating all possible parameters and design three

individual ODAF. The quality and reproducibility of the output was of course a primordial criterion for the selection of the pipeline, together with the accessibility of the tools, the user-friendliness, and the interoperability of the final product. In a final phase, the three different proposed ODAFs were assessed with several dataset to evaluate the efficiency of the proposed pipeline in calling a list of differentially expressed genes as accurately as possible.

The final R-ODAF is then composed of three individual pipelines for microarray, RNA-Sequencing and TempO-Seq platform, as summarize in figure 1:

Figure 1: Schematic overview of the R-ODAFs for A) Microarray technology, B) RNA-sequencing technology and C) TempO-seq technology. Orange and blue boxes represent stages in the pipelines, green boxes represent options or (software)tool names to be used at the specific stages and yellow boxes indicate criteria for discarding samples. Q30 = base with a quality score above 30; seqLength = is the length of the sequencing fragment (RNA-Seq; usually 50bp, 100bp or 150bp; TempO-seq: approximately 70bp); Δ = difference; CPM = counts per million.

Publication from C4:

1. Verheijen M, Tong W, Shi L, Gqnt TW, Seligmann B, Caiment F. 2020. Birch H, Kramer NI, Mayer P. 2019. Towards the development of an omics data analysis framework. Regul Toxicol Pharmacol Volume 112

A second manuscript presenting the final R-ODAF product is in preparation and will be submitted in the near future.