Proceedings of the LXII SIGA Annual Congress
Verona, Italy – 25/28 September, 2018
ISBN **978-88-904570-8-1**

Poster Communication Abstract – 6.12

VERSATILE AND EASY-TO-USE BIOINFORMATICS TOOL TO ANALYZE SINGLE CELL RNA SEQ DATA

SAERA-VILA A., SANSEVERINO W., AIESE CIGLIANO R.

Sequentia Biotech SL, Campus UAB, Edifici Eureka, Barcelona (Spain)

transcriptomics, single cell, RNA-Seq, bioinformatics, Chromium 10x Genomics

Next-generation sequencing (NGS) technologies have completely changed life sciences and biomedical research. Nowadays, sequencing empowers a broad range of fields from medical care to agrigenomics. In the transcriptomics field, RNA deep-sequencing (RNA-Seq) has revolutionized genome-wide scale gene expression studies becoming the gold standard. However, RNA-Seq measures the sample average gene expression so it is not sufficient to analyze heterogeneous systems. To address this weaknesses, in 2009, single cell RNA-Seq (scRNA-Seq) was developed but it remained almost forgotten until recently, mainly due to the lack of accessible protocols and high costs. In addition, non-bioinformatics researchers still need specific bioinformatics tools capable of dealing with the huge and complex data generated by scRNA-Seq experiments. Thus, the aim of this work was to design a bioinformatics tool accessible to any researcher and capable of analyzing all scRNA-Seq experiments: single/multiplexed experiments with/without UMIs (random molecular tags labeling each transcript) and/or internal control RNAs or “spike-ins”. Starting from the sequencing reads (FASTQ files), our pipeline can detect and extract real cell barcodes and UMIs. A quality control step filters out sequencing adapters and low quality ends. Filtered reads are then mapped to the reference genome so gene expression is calculated counting the number of reads per gene. When present, UMIs per gene are automatically condensed to correct for amplification biases. At this point, the expression matrix is filtered to remove low quality or outlier cells, and dropouts (excess zero or near zero expression values due to the low detection in each cell) are imputed. The expression matrix is then normalized and used to perform unsupervised clustering to identify the different cell groups in the original data set. Next, differential gene expression among groups is calculated and gene markers for each group are identified, including a Gene Ontology enrichment analysis. In conclusion, we have developed a bioinformatics tool for any sequenced species using the latest, peer-reviewed bioinformatics methods and algorithms in combination with thorough in-house benchmarking. We provide an almost automated solution to close the gap between scRNA-Seq data production and interpretation. It requires little bioinformatics knowledge with minimal interaction of the researcher so the errors associated to data/file managing, processing or storing are basically eliminated.