Project summary for FK-132666

Year 1 - Outline

During the first year of the project, we focused on three tasks: 1) obtained the necessary resources for large-scale data processing, 2) benchmarked various software tools and set up a pipeline to reprocess public RNA-sequencing datasets, and 3) started developing R packages to characterize expression variance at the total gene expression and at the splicing isoform level.

Year 1 – Computing resources

Originally, we planned to use four different resources for data processing: a) our local server setup, b) the hungarian supercomputing network managed by "Kormányzati Informatikai Fejlesztési Ügynökség" (KIFÜ), c) a cloud computing resource provider available through the European Grid Infrastructure (EGI) and d) additional cloud computing resources using Oracle Cloud.

The initial local server was set up using independent funding from Semmelweis University. Additional storage, Vmware software for virtual machine management and installation costs were covered by this grant. We also obtained access to the hungarian supercomputing network. As we partially relied on processing the Cancer Genome Atlas (TCGA) dataset, that requires a Data Access Agreement between the National Institute of Health (NIH) in the USA and the institutes where the data is actually processed, we had to coordinate the signing of a Data Access Agreement between KIFÜ and NIH, besides Semmelweis University and NIH. This took much longer than expected, and the agreement with KIFÜ was signed only on 2020-10-22. While setting up our local server, and coordinating the NIH – KIFÜ Data Access Agreement, we also signed an agreement with EGI and got access to the SBG Horizon cloud computing system of CNRS - Institut national de physique nucléaire et de physique des particules. This required updating the Data Access Agreement between NIH and Semmelweis University, to include statements on cloud computing use. Complementing the above, we started a discussion with the hungarian branch of Oracle Cloud, to assess if they can provide computing and storage resources for a reasonable price. Finally, we decided to only use our local server, and the KIFÜ supercomputing resources, as EGI administration was unresponsive, virtual instances were hard to set up, while Oracle Cloud resources were not necessary.

Year 1 – TCGA pipeline development

We developed an initial pipeline for re-processing TCGA and other datasets. Even though already processed TCGA data is availabe, it only includes gene-level expression summaries, while we needed splicing isoform level expression estimates for a large part of our analysis. We originally aimed to use the kallisto tool for this analysis. However, considering recent papers and improvements on transcriptome pseudoalignment methods, we also tested the salmon tool in two different modes. This includes the mapping based mode, where we used a simple transcriptome pseudoalignment for quantification, and the alignment based mode, where we used RNA-sequencing reads aligned with the STAR splice-aware mapper for quantification. We benchmarked the three different methods (kallisto, salmon mapping, salmon alignment), using publicly available control samples from normal blood, myelodysplastic syndrome and secondary acute myeloid leukemia patients. Based on our

results, the three methods gave significantly different transcript isoform level expression estimates.

While developing the pipeline, we noticed that the offical, up-to-date data provided by TCGA at the Genomic Data Commons (GDC) repository contains already aligned RNA-sequencing data instead of the raw sequencing results in fastq format. However, this data contains a bug, and the pairing information from paired-end sequencing is lost, making it impossible to use with kallisto or salmon and get transcript level expression estimates. To circumvent this problem, we switched to the TCGA Legacy Database, that still contains the original fastq format sequencing results, although for a smaller number of samples. We tested our pipeline on the liver hepatocellular carcinoma (LIHC) dataset using the GENCODE v36 transcriptome annotation, and processed smaller cancer datasets locally.

Year 1 – R package development

As a third task, we started developing two R packages, for quantifying splicing and gene expression level variance within and across samples. The R package for quantifying splicing variance is available in Bioconductor, at the address: https://bioconductor.org/packages/release/bioc/html/SplicingFactory.html. The package uses transcript isoform level expression values to analyze splicing diversity based on various statistical measures. These include the Shannon-entropy, a pseudocount entropy corrected with Laplace's prior (pseudocount of 1), the Gini index, the Simpson index and the inverse Simpson index. To check for significant changes in splicing variance between conditions, we implemented a label shuffling and a Wilcoxon test based method in the package. The package accepts diverse input formats, including a matrix, data.frame, SummarizedExperiment, tximport, DGEList or ExpressionSet object. Input expression measures can be original read counts, RPKM, FPKM or TPM values at the transcript isoform level. The final output of a differential analysis is a data.frame, showing information on all genes of the dataset, including the fold-change of the diversity value between conditions, the p-value and the FDR corrected p-value. Using the package, we tested the effect of RNA-seq quantification tools (kallisto and salmon), quantification uncertainty, gene expression levels and isoform numbers on the isoform diversity calculation.

Together with the SplicingFactory package, we started developing an additional package, aimed at characterizing gene level expression variance changes, called ExpressionVariance. There are multiple measures for characterizing gene level variance, including standard deviation (sd), median absolute deviation (MAD) or the Fano-factor. However, experimental technologies, such as microarrays and RNA-sequencing used for quantifying gene expression, have known technical biases influencing variance measures and raw datasets need multiple filtering and normalization steps, before they can be analyzed in detail. Additionally, there are no well-documented, user-friendly tools for calculating and investigating gene expression variance. We defined the basic package structure, the necessary filtering, normalization and bias corrections steps based on various literature sources and implemented the basic functions. We used RNA modification enzyme KD, KO and overexpression cell lines to test the package and the expression variance measures, as we hypothesize that RNA modifications play a role in expression variance regulation in different conditions, including human disease.

Year 1 – Personnel changes

A candidate PhD student was selected to work full-time on the project (Aldo Sergi, from Rome, Italy) but he decided not to move, due to the COVID-19 pandemic. Several other students joined (Péter Szikora, Noam Makmal, András Udvarvölgyi, Alexa Szeifert, Benedek Dankó and Veronika Vraukó), besides a volunteer (Tamás Pór) and a part-time bioinformatician (Tibor Nagy).

Year 2 - Outline

During the second year of the project, we focused on the following tasks: 1) finalizing and publishing the SplicingFactory R package to characterize splicing isoform variance, 2) continuing the development of the ExpressionVariance package, 3) analyzing the effect of RNA modifications on gene expression variance using a number of public datasets, 4) de-novo transcriptome assembly and additional analysis of MDS RNA-sequencing samples, and 5) an additional project on central nervous system lymphoma miRNA expression patterns.

Year 2 – R package development

We finalized the SplicingFactory R package, publised it's updated version on Bioconductor, together with a full research paper in Bioinformatics. The package itself is open source, and freely available in the stable branch of Bioconductor. The paper is available as an open access publication. In the paper, we did an extensive benchmark of the effect of transcript isoform numbers, average gene expression values, upstream gene expression quantification methods and quantification uncertainty. We also benchmarked its performance, compared it to existing similar methods and tools carrying out a similar task, and analyzed a set of CD34+ hematopoietic stem cells and myelodysplastic syndrome samples. We found a set of genes, whose isoform diversity change is associated with the mutations of splicing factor SF3B1.

We continued the development of the ExpressionVariance R package. The package is currently in alpha version. It implements a number of data import, filtering, normalization and variance calculation steps. It can import transcript level expression data from the kallisto, salmon, sailfish and RSEM tools, and returns a SummarizedExperiment dataset. It is able to filter out genes with very low expression across samples, or expressed only in a small number of samples. The different expression normalization methods are the following: upperquartile, RLE, TMM, MRN, SizeFactor and CuffDiff. After initial filtering and normalization steps, it calculates a number of variance metrics, including the variance, standard deviation, median absolute deviation, Fano-factor and distance-to-median value. Finally, the package implements a LOESS regression to fit a model on the mean or median of expression values against a given variance metric and calculate the ratio of the predicted and observed variances.

Year 2 – Analysis of RNA modifications

We applied the ExpressionVariance package on 5 public datasets, including various cell lines and knock-down of the METTL3 or METTL14 RNA modifier protein coding genes. We calculated the available variance metrics and investigated changes in variance for various gene sets, including splicing factors, RNA modifiers, RNA binding proteins, transcription factors, kinases and histones. In parallel, we investigated the variance changes of known oncogenes or tumor suppressors. Finally, we selected a number of genes from the different gene set combinations (i.e. oncogenic transcription factors, tumor suppressor RNA binding

proteins) and analyzed the possible effect of the METTL3 or METTL14 knock downs on expression variance.

Year 2 – MDS RNA-sequencing sample analysis

To investigate the possible effect of splicing factor and epigenetic factor mutations on splicing diversity in more detail we collected a number of additional MDS RNA-sequencing datasets, including data from the following papers: (Im et al., 2018), (Pellagatti et al., 2018), (Dolatshad et al., 2015), (Madan et al., 2019), (Fernandez et al., 2019). In these datasets, we found 70 samples with SF3B1, SRSF2, U2AF1, or ZRSR2 splicing factor mutations, and 11 with ASXL1, BCOR, DNMT3A, EZH2 or TET2 epigenetic factor mutations. Using the Trinity de-novo assembly tool, we analyzed all samples at the Miskolc SGI UV 2000 supercomputer.

In the 81 samples we analyzed, the number of de-novo assembled transcripts varies significantly, minimum is 4847, maximum is 526711, median is 286980. When separating the two groups (splicing factor mutated or epigenetic factor mutated samples), the minimum, maximum and median values for splicing factor mutated samples are 156710, 526711 and 288396 respectively, while for epigenetic factor mutated samples they are 4847, 353670 and 192406. Even though the small sample size, possibly different experimental protocols, and quality differences between RNA-sequencing runs might all influence results, these intial values show, that epigenetic factor mutated samples have a smaller number of de-novo assembled transcripts, pointing to a significant change in the process of gene expression and/or splicing.

Additionally, using all collected MDS RNA-sequencing dataset and control CD34+ hematopoietic stem cell samples, we analyzed their difference with the iso-kTSP tool or LASSO logistic regression, and described MDS specific gene expression, transcript expression and splicing patterns. Using these results we built classification models, reaching good overall precision, sensitivity and specificity in classifying samples into control or MDS categories. Interestingly, the transcript expression quantification tool influenced the results significantly as splicing patterns performed well, when using salmon for transcript quantification, while gene expression performed well, when using kallisto for transcript quantification.

Year 2 – NanoString miRNA analysis

Finally, not included in the original research plan, we investigated the expression patterns of a large number of miRNAs in primary and secondary central nervous system lymphomas. Using the NanoString platform, we performed expression analysis of 798 miRNAs in 73 samples. We investigated how miRNAs vary between primary or secondary samples, between molecular subtypes or different mutation patterns. We identified a number of miRNAs differentially expressed between the above groups and defined a small set of samples with characteristically different miRNA expression patterns compared to the rest. Additionally, we identified differentially regulated pathways between groups and investigated the survival charactersitics of samples with specific miRNA expression patterns. The results were published in The Journal of Molecular Diagnostics as a regular article.

Year 2 – Personnel changes

Between December 2020 and February 2021 no personal change happened. However, the PI left Semmelweis University, starting with March 2021, and all part-time and volunteer contracts were cancelled.