Advancing transcriptome analysis in models of disease and ageing

de Jong, Tristan Vincent

DOI:
10.33612/diss.99203371

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 27-01-2020
Chapter 9

General Discussion

Tristan V. de Jong
RNA-sequencing technology is often utilized as an exploration or validation tool to identify genes which significantly change in expression under different experimental or natural conditions. A most popular method for interpretation of RNA-seq data is differential expression analysis that results in identification of genes and pathways that are most affected in the experiment. The results of such analyses are comparable to results obtained from expression microarray platforms and, in principle, underutilize the possibilities offered by RNA-sequencing approach.

In this general discussion I will highlight some of the less frequently utilized aspects of RNA-sequencing data analysis, that can help to uncover additional ‘layers’ of information such as alternative transcript isoforms, transcripts from non-coding and unannotated coding genes and inter-individual expression variability.

**tRNA, snoRNA and snRNA’s can be just as interesting as protein coding RNA.**

In chapter 3 we show an example of how RNA-sequencing data allows us to gain a more complete overview of factors at play than a gene-centered study. Through a forward genetic screen, a mutation causing a premature stop codon of LIR-3 was found to reduce the aggregation of proteins within body-wall muscle cells in *C. elegans*. A zinc-finger domain within MOAG-2/LIR-3 lead to the hypothesis that it can bind to DNA, which was verified through chromatin immunoprecipitation followed by deep-sequencing. As many as 678 unique MOAG-2/LIR-3 binding sites were identified, which often occurred in vicinity of the transcription start sites (TSS) of tRNA, snoRNA, rRNA and snRNA genes. Of the 678 binding sites more than half contained Box A and Box B binding sites, which constitute the promoter site recognized by polymerase (Pol) III complex. It was found through FLAG-tagged MOAG-2/LIR-3 co-immunoprecipitated with Pol III, suggesting the existence of a complex. Mutations in MOAG-2/LIR-3 were not observed to significantly impact the expression of protein coding genes but did significantly reduce the expression of many snRNAs, snoRNAs and tRNA genes. Yet, analysis of RNA-sequencing data showed that the partial deletion of MOAG-2/LIR-3 did not result in a significant change in the expression of
snRNAs, snoRNAs and tRNAs in the aggregation phenotype (Q40). Interestingly, we observed that the expression of snRNAs, snoRNAs and tRNAs was already strongly reduced in Q40 worms without the LIR3 deletion. This led to the conclusion that despite the knowledge that MOAG-2/LIR-3 co-regulates the expression of snRNAs, snoRNAs and tRNAs through Pol III, this characteristic is not the cause for the reduced aggregation. These fine nuances in the understanding of the intricacies of the inner workings of cellular regulation would not have been possible merely by investigating changes in the average expression of known protein coding genes.

**Unannotated genes and alternative exon usage might be left behind in a typical RNA-seq analysis**

Interestingly, when mapping the reads to the reference genome, a number of reads mapped to un-annotated regions. In total, we found 40 locations in the genome that contained unknown, potentially protein-coding transcripts (poly-adenylated) which were absent from gene annotation but appeared as prominently expressed in Q40 samples (Figure 1).

In addition, we detected hundreds of non-constitutive exons that were differently used between WT and Q40 worms (Figure 2). The insight into expression patterns of un-annotated genes and the alternative exons which are only used under certain conditions stress the value of exploring RNA-sequencing data beyond just a differential expression analysis.

By analyzing the genome coverage outside of the annotated regions new genes can be discovered and new protein-coding genes can be inferred, resulting in a further improvement of our understanding on transcriptional dynamics under stress (e.g. poly-glutamine toxicity).
**Figure 1.** Sashimi plot showing coverage of one of the 40 novel exonic regions where multiple reads map to an un-annotated part of the *C. elegans* genome. 3 wildtype samples (N2) and 3 samples with expanded poly-glutamine repeat (Q40) are mapped against the reference genome. The height of the bar shows relative coverage as compared to other samples. There is a 5.5 times fold change in Q40 samples, revealing the gene is upregulated under Q40 condition.

**Identification of genes, which change in expression with different temporal dynamics**

As acceptance of NGS-based technologies is improving and price is for RNA-seq analysis is constantly going down, researchers are increasing the number of conditions and perform time series analysis. This leads to more complex experimental designs which combine different model systems (or even organisms), experimental conditions and time points. Analysis of such experimental data typically involves multiple pairwise comparisons that need to be integrated and summarized or complex models that might limit the power of analysis.
Figure 2. Sashimi plot showing exon coverage and splice site usage of Pqn-52 gene. Here reads map to a non-constitutive exon 2 of Pqn-52. Reads from 3 wildtype samples (N2) and 3 samples with expanded poly-glutamine repeat (Q40) are mapped against the reference genome. The height of the bar shows relative read coverage. There is a 2.5-fold difference in usage of exon 2 in N2 samples compared to that in Q40 worms.

In the paper on which chapter 5 was based several RNA-sequencing experiments were explored to characterize genes that change their expression upon entering a senescent state. Due to the variable phenotype of the senescent state, and the lack of universal markers so far, an RNA-seq dataset was generated from fibroblasts (HCA-2), melanocytes, and keratinocytes at the proliferating stage as well as 2, 4, 10 and 20 days after ionizing radiation. Differential expression analysis with DE-seq, testing for genes which significantly change between 4, 10 and 20 days as compared to the proliferating state resulted in the identification of 61 genes that were shared among all cell types and time points, 34 of which were not shared with quiescent cells.
Our re-evaluation of this dataset (chapter 5) identified that the most difficult factor to account for in this experiment is the temporal dynamics among different cell types. The identification of genes which significantly change their average expression with a model which includes time and cell-type as factors would not directly account for a temporal delay in response between different cell types. If the time in days was set up as a linear factor it would mean that it would be difficult to detect significant change for genes that do not steadily increase or decrease their average expression. To this, genes which decrease their expression level after 20 days, after a strong increase at 2, 4, and 10 days would still be marked as increased, ignoring the observation that the gene expression returns to nominal values given enough time has passed. The differences in functions of the three cell-types, keratinocytes, melanocytes and fibroblasts could also cause an accelerated or delayed response to ionizing radiation among cell types. Inclusion of the timepoints between samples as factors would then result in significant changes being ignored due to a possible delay in response. The visualization of these data using Venn-diagram, would elude to the dynamics that underlie the cell type specific delay in response (Hernandez-Segura et al., 2017), yet combining both timepoints and cell type inside a single Venn-diagram would lead to a complex diagram that is difficult to interpret.

A method to gain a greater insight into the dynamics of the changes in expression is to cluster genes that follow the same patterns of expression in the different cell-types. Some genes might become up-regulated in fibroblasts after irradiation but might not respond as quickly in keratinocytes and melanocytes. When performing differential expression analysis these timepoints might not line up, wrongly assuming these genes do not change in expression after irradiation. Specific combinations of response dynamics might be of interest but evaluating all possible combinations of up- and down-regulation of genes across all timepoints would result in 243 patterns (3 directions on 4 timepoints in 3 cell-types). Grouping genes, such as k-means clustering could offer a solution in this case, yet such clustering tends to create large clusters of mildly similar patterns as a ‘the rest’ category unless many tens of clusters are created.
In order to reduce this complexity, we designed a series of qualitative temporal patterns (Chapter 5) representing generalizations of ways by which gene expression could respond to ionizing radiation. Only genes which significantly change their expression in the same direction after ionizing radiation between all three cell types were considered in this analysis. We included three patterns in the analysis: Maintained increase/decrease, delayed response and response with recovery. Genes which did not adhere to one of the three temporal patterns were ignored in this investigation. Consideration of the temporal dynamics resulted in the identification of 348 genes which were significantly up-regulated after ionizing radiation, while 339 genes were found to be significantly down-regulated.

Utilizing different cell types and time points when investigating a complex mechanism such as cellular senescence is of great importance to avoid calling genes as cellular senescence markers when observed in only one cell-type or at a single timepoint. The 687 genes which showed a significant change in expression and befitted one of the three profiles could lead to the discovery of robust senescence markers which might become useful in future experiments, regardless of cell type, time point of sampling after causing senescence. Pattern analysis has allowed for the discovery of additional information that could be of value for future in vivo experiments where senescent cells need to be detected in various tissue samples.

**Gene expression variability – the other dimension in transcriptome analysis**

A different paradigm of RNA-seq analysis is not only considering differential expression analysis, but also employing differential variability analysis. In chapter 2 of this thesis we elaborated on the variation in expression and proposed a method to quantify it in such a way that is independent from the average expression.

In chapter 7 we delved into the underlying causes of variation in expression. Here we found that a large part of the expression variability can be explained by genomic sequence composition around the TSS, is modulated epigenetic factors, as well as external factors such as tissue, diet and age. Currently several studies have shown the
positive effects of caloric restriction on lifespan for yeast, worms, mice and other rodents (Heilbronn & Ravussin, 2003). Experiments with rhesus macaque have not shown an extension of lifespan but did show a delay of age-related diseases (Colman et al., 2009).

The variability in expression might shed a light on the changes which occur with a change in diet and age, as a similar increase in health and lifespan was observed in chapter 8. Here, a C/EBPβ ΔuORF mutation, which mimics dietary restriction, was shown to be associated with an increase in lifespan, reduced tumor incidence in female mice and an overall reduction in the non-Poisson variability of a large number of genes with age.

To further highlight the merits of the variation in expression as a dimension of transcriptome analysis we elaborate on the variability observed in chapter 8. Using the methods proposed in chapter 2 I calculated the average expression and variability of gene expression under the different conditions. A total of 2,641 genes significantly changed in average expression with age under normal conditions, though only 569 genes significantly changed in average expression with age in the knock in (Kin) samples (figure 3b).

These are the genes which are usually investigated in further experiments. When observing the variation in expression, shown as non-Poisson variability there is an overall increase with age (figure 3a). When observing exclusively significant changes a reduction of variability is visible among Kin samples (figure 3c).

In total 414 genes were found to significantly increase in non-Poisson variability with an advanced age in the wildtype samples, whilst only a fraction of that significantly increased in variability in the Kin samples (Figure 3d). KEGG pathway analysis highlighted that several pathways, mostly involved with metabolism, showed an increase in expression variation with an advanced age.
Figure 3. Non-Poisson variability estimates for young and old samples with and without Knock-in (Kin) of C/EBPb-LIP (top left). We observe an increased overall expression variability with the C/EBPb-LIP knock-in among young samples. (bottom left) Here we see only the genes for which the non-Poisson variability significantly changes. (Top Right) The overlap of genes that significantly change in average expression with age for both the CEBPb-LIP knock-in and WT. (Bottom right) Overlap of genes that significantly change in variability with age for both the Kin and WT. The reduced overlap clearly shows the variability is decreased at an advanced age in the Kin mutants.

Interestingly, the Kin mice showed a strong rescue phenotype in the sense of expression variability for a multitude of these pathways (Figure 4).

A similar observation of an increased variation in expression with age was observed in chapter 7 among rats, in which older samples had more inter-individual variation in expression. These observations of an overall increase in variability with both an advanced age (chapter 7 and 8) and a high fat diet (chapter 7), as well as the reduction of variability within mice with a mutation mimicking a caloric restriction diet at an advanced age, lead us to hypothesize that there is a link between the observed stricter regulation of genes and the healthier phenotype under a calorie restricted diet. Previous research has shown that the intake of a high fat diet changes the chromatin
accessibility in mice, though these changes were mostly strain specific (Leung et al., 2014).

Figure 4. Boxplots of non-Poisson variability in genes belonging to different KEGG pathways. Some pathways show a clear reduction of variability with a knock-in of C/EBPβ, where other pathways show only an increased variation as a consequence of the C/EBPβ knock-in.
A greater understanding of changes in variation in expression upon ageing and dietary conditions will give an insight into factors influencing or influenced by histone modifications and will grant an insight into this chicken or egg conundrum. In this respect studying not only changes in the mean levels of transcripts, but also expression variability provides an extra dimension in studying age-related molecular changes.

**The possibility of revisiting old data sets**

RNA-sequencing is a method that leads to a treasure trove of information, one which is too often quickly discarded once the first pieces of value have been excavated. The analysis of non-protein coding transcripts can lead to insights on regulatory mechanisms not seen before. The mapping of un-annotated transcripts can, in turn, reveal new genes or exons that only surface under certain experimental conditions.

Beyond that, an increased variation of expression of certain genes can point at perturbations in regulation, which would be missed when only investigating changes in average expression. The methods proposed in chapter 2 can identify genes which influence biological systems in an impactful way, not by increasing or decreasing, but by deviating from the norm. These deviations can be measured and used as biomarkers for complex traits like development of diseases (Chapter 3), the onset of cellular senescence (chapter 5) or even susceptibility to obesity (chapter 7).

In certain complex diseases, such a chronic obstructive pulmonary disease (COPD) it is difficult to identify genes which universally increase or decrease to use as a diagnostic marker, yet an increase in variation can be observed amongst individuals with COPD due to the broad variety of causes and variations in the disease phenotype (De Vries et al., 2018). Currently, in trial experiments, the absence or presence of transcripts from several genes is utilized to determine whether a suspected tumor sample is malignant (Lin et al., 2013), yet not all types of cancer can be effectively diagnosed from molecular profiles due to the large genetic heterogeneity of tumors (Marusyk & Polyak, 2010). Searching for an abnormal increase in gene expression variation, stemming from a large heterogeneity of tumor cells or inter-individual
variation could circumvent this problem and give an insight into the likelihood of a sample containing cancerous cells. Instead of solely focusing on genes that universally change in expression, I suggest the identification of a baseline of variability in expression for complex multifactorial traits and diseases to provide an additional view on key genes.

A current limitation on the re-evaluation of samples is that only a few samples are required to perform differential expression analysis, often resulting in a common use if just several replicates (2-4) in datasets, making differential variation analysis difficult. However, future developments in more robust and cheap RNA-seq technologies will allow increased acceptance of differential variability analysis and will shed light on true importance of expression variability as an alternative molecular phenotype.
REFERENCES:

Colman, R. J., Anderson, R. M., Johnson, S. C., Kastman, E. K., Kosmatka, K. J., Beasley, T. M., ...


