Improving design, execution and analysis of transcriptomics experimentation

Bruning, O.

Citation for published version (APA):
Bruning, O. (2015). Improving design, execution and analysis of transcriptomics experimentation

General rights
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), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
General introduction
Background

The arrival of high-throughput, genome-wide omics technologies about two decades ago has had a major impact on life-sciences research (Box 1). Besides the sometimes paralyzing data explosion, the adoption of a non-reductionist approach surely has had a severe effect on life-sciences experimentation [1]. Since the introduction of omics technologies, a lot of effort has been put into adjusting the traditional ways in which experiments are set up, as to accommodate the requirements of these new techniques. Although many omics experiments have led to groundbreaking progress in life-sciences research, often the results from omics experiments remain mostly descriptive in nature. In contrast to the successful application of omics technologies in biomarker discovery and genome wide association studies (GWAS) [2–4], it has proven to be very difficult to directly derive new insights about complex biological mechanisms from these data.

There are a number of reasons for this, some of which will be discussed here. The first issues already arise at the beginning of omics experimentation. The design of an experimental setup may be suboptimal due to, for instance, budgetary or practical limitations. Although the basic statistical principles for experimental design regarding how to compare samples and numbers of required (biological) replicates have been established for decades, there are no strict rules on how to design omics experiments.

Box 1

What high–throughput genome-wide omics technologies are there?

- Genomics: The study of the genomic make-up of organisms using either microarrays or next generation sequencing
- Epigenomics: The study of the changes to the inheritable material in cells of organisms using either ChIP-Chip microarrays or ChIP-Seq next generation sequencing approaches
- Transcriptomics: The study of multiple RNA transcripts at once from biological samples using either microarrays or next generation sequencing
- Proteomics: The study of multiple proteins at once from biological samples using high–throughput mass spectrometry techniques
- Metabolomics: The study of multiple metabolites at once from biological samples using high–throughput mass spectrometry techniques
- Metagenomics: The study of the genomes from multiple organisms in a population using either microarrays or next generation sequencing
Moreover, the processing of a sample using omics techniques is, despite continuously decreasing prices, still quite expensive. Hence, each life science researcher faces difficult choices between the number of replicates versus the number of biological conditions. Unfortunately, this has caused many (expensive) projects to (partially) fail by lack of sufficient replicates to make justified statements about a population as a whole. This is basically an echo of the molecular biology lab practice at the last quarter of the 20th century, which falls short for omics experimentation due to the huge numbers of tested elements and the associated overwhelming complexity. Additionally, experiment designs are often plagued by the fact that no clear biological question is formulated upfront, to which the design can be tuned. Instead, a multitude of high-level biological questions are used as a starting point. This approach is called data-driven research, which quickly became popular due to the whole-genome approach of omics experimentation. Often, the observed phenomena in data-driven research cannot be interpreted, as the experimental design was not suited to investigate them.

Besides the design of omics experiments, there is also a need for optimizing the execution of omics experiments. During the last two decades, molecular biology has become a mainstream skill of life-sciences researchers. This inadvertently resulted in the fact that most researchers today are using pre-fab laboratory kits, often without knowing exactly what they are doing. This lack of molecular biology expertise has become a real problem in the omics era, where the detectors have become much more complex and sensitive. Hence, every, seemingly small, divergence in the experimental protocol will have a major impact on the results. One solution is to outsource parts of the experimentation procedure to professional omics service providers, but also generation and handling of the primary samples is extremely important. This is even more relevant, as the amount of starting material is sharply decreasing because experimentation in general is progressing towards the single cell level [5–10]. Altogether, the importance and sensitivity of sample isolation and omics technology execution are often underappreciated and should be a major topic of attention in life-sciences research.

One of the main issues in omics experimentation is the incredible amount of data generated, which put constrains on the data analysis of omics experimentation. Originally, a classical molecular-biological experiment consisted of studying around one to ten genes on DNA, RNA, or protein level. With the arrival of qPCR, comparing up to 384 genes in one experiment became feasible (Figure 1A) and with the introduction of the first microarrays (Figure 1B), thousands to tens of thousands of genes could be studied. Currently, the recent next-generation DNA sequencing (NGS) techniques have increased this data-burden even further as they are generating tens to hundreds of millions of sequencing reads and over 100 Gbases per run (Figure 1C). Recently, the first NGS platform that produces over 1 Tbases per run, which is the equivalent of 333 complete human genomes, has arrived. These huge numbers of microarray probes and NGS reads translate into data points, which in turn give rise to data with an ever increasing level of complexity (Figure 1D). Results from analyses of these complex data often do not rise above the level of description and confirmation of known biological processes. [11–14]
Chapter 1

Aim

The studies in this thesis research elements concerning experimental design, experiment execution, and data analysis from the whole chain of transcriptomics experimentation in order to improve them.

Approach

Our studies concerning experimental design focus on how one can tune an experimental setup to a specific biological question. This will be done by showing the value of range finding for transcriptomics experimentation. Our range finding approach aims to identify the "sweet spot" in the experiment design space that give the most information about the specific biological process under study.

Concerning experiment execution, we will look into improving experimental conditions in mouse embryonic fibroblasts (MEFs) experiments and optimizing the generation of total RNA from mouse skin biopsy samples.

In our studies of data analyses we will focus on trying to find new approaches to avoid

Figure 1. Evolution of high-throughput, genome-wide techniques output

Different types of high-throughput tools with increasing amounts of output A, Microtiter PCR plate with 384 wells capable of checking i.e. 384 genes in parallel. (http://www.thermoscientific.com) B, Spotted microarray slide capable of checking thousands to millions of transcripts in parallel. C, Sequencing chip capable of generating 60-80 million reads in parallel. D, Example of the complexity resulting from the ability of looking at up to tens of thousands of genes in parallel [26]
misinterpretation of transcriptomics data and gain truly new biological insights from these data.

**Biological example**

To support our studies, a biological use case was required. To this end we decided to study the mouse Trp53 gene, as p53 is one of the most studied proteins in biology. It plays a pivotal role in the suppression of tumors and is, as a transcription factor, involved in many biological processes, like cell cycle arrest, apoptosis, senescence, modulation of autophagy, DNA repair, and changes in metabolism. It is also an ancient protein that came into existence before the evolution of multicellular organisms and the arrival of cancer [15,16].

In the majority of tumors, the p53 function is somehow disrupted. Either it is directly inactivated via mutations or indirectly via increased inhibition, decreased activation or inactivation of its responsive elements [17,18].

Under normal circumstances, p53 prevents healthy cells from turning malignant by responding to various stresses on cells and activating counter measures. Whenever DNA damage, activation of oncogenes, or hypoxia are sensed, a number of steps are initiated that lead to the p53 response. In the first step, p53, which is normally a short-lived protein due to negative regulation by Mdm2 through degradation via ubiquitination, now becomes more stable as this interaction is disrupted. Next, it is posttranslationally modified by processes, like phosphorylation, methylation, acetylation, sumoylation or neddylation to further stabilize it and adjust the DNA binding properties to be able to react to the different types of DNA damage. Finally, after binding to the DNA it activates or represses specific target genes that will eventually determine the fate of a cell [19–21].

Next to this commonly known defense route through the p53-pathway, p53 is also involved in the regulation of various metabolic processes in the cell, like glycolysis, mitochondrial respiration, oxidative phosphorylation, autophagy, etc. These processes are often also affected in tumors [15,22,23].

Another way in which p53 contributes to the suppression of tumors is via regulation of microRNA's that, in turn regulate cell proliferation, differentiation and apoptosis. Again, processes that are commonly affected in cancerous cells [24,25].

Throughout this thesis, we will study P53-regulated cellular mechanisms in-vitro by means of MEFs and in-vivo by using mouse skin biopsies. Exposure to UV will be used as example of a typical biological perturbation in an attempt to unravel cellular processes at the molecular level.

**Results**

During the research described in this thesis, we have developed a number of approaches
Chapter 1

to improve several elements in the whole chain of transcriptomics experimentation.

Experimental design. We have set up a method for using small-scale and cost-effective range finding studies to pinpoint the experimental ‘sweet spot’ in a design space. Using this approach it is possible to tailor-make a larger zoomed-in experiment more geared towards the biological process in question. Our study showed the strong modularity of biological processes as each process turned out to have its own ‘sweet spot’ in the design space.

Experiment execution. To improve transcriptomics experiment execution in the laboratory, we have looked into the handling of MEFs sample material for the murine in-vitro system. Adjustments were made to the atmospheric oxygen concentrations under which the cells were grown, hereby removing perturbation-unrelated stress from the system. Furthermore, the cells were synchronized before used in the actual experiment. These steps strongly reduced the background noise found in the resulting material. For the murine in-vivo system, we implemented a method for efficient and cost-effective extraction of total RNA from skin biopsies. This was tested in both human and mouse skin material and yielded high amounts of good quality material.

Data analysis. We started by improving the analyses of transcriptomics data based on novel approaches that build on common-practice data-analyses methods at the time of our first experiment. After optimizing the experimental design and experiment execution, we found out that there are serious confounding factors that hamper proper analysis and interpretation of transcriptomics data.

Outline of the thesis

Our first study involved an extensive omics experiment, which we analyzed by a traditional approach augmented with novel elements. As biological used case, we aimed to identify the role of a p53 phosphorylation site (S389) via the transcriptome response of MEFs to UV exposure. For this, a transcriptomics experiment with a conventional setup was performed as described in Chapter Two. The effects on the basal gene expression levels of p53-dependent genes, the transcriptome response of wild-type MEFs over time, and the effect of the absence of p53.S389 phosphorylation on the UV response over time were analyzed. These three analyses revealed a complex response in the gene-expression pattern, with high numbers of differentially-expressed genes (DEGs).

However, when trying to extend and deepen the analysis of the data from this study, we uncovered multiple alarming issues caused by the original setup of this experiment. As described in Chapter Three, these issues could be traced back to the fact that the experiment was run from a non-optimal location in the design space for studying the biological processes of a UV specific response. In fact, most of the results could be attributed to the effects of generic-stress processes on the transcriptome. To circumvent these types of problems in future studies, we investigated the use of range finding studies preceding such transcriptomics experiments.
Before studying the optimal location in the design space, we first looked into optimizing the experiment execution. The results of transcriptomics experiments can for instance also be improved by optimizing the quality of the starting material. In Chapter Four, a contribution to this end is presented, when we introduce a protocol for improving the yield and quality of total RNA extraction from human and mouse skin biopsies.

The suggested use of range finding as part of the design for experimentation from Chapter three is further elaborated in Chapter Five. Here, for both in-vitro (MEFs) and in-vivo (skin biopsies) murine systems, wide-ranged dose and time range finding experiments were performed. This generic proof-of-concept approach and protocol showed clearly that different biological processes need to be studied using specific locations in the tested experimental design space.

Based on the results from Chapter Five, a more focused study on the specific transcriptome response to UV in mouse skin biopsies was performed in Chapter Six. This study focused on confounding factors in in-vivo transcriptome experimentation and showed the destructive effects of those factors on the outcome.

Finally, the concluding remarks in Chapter Seven put the lessons that were learned throughout the studies presented in this thesis in context. It shows that optimizing the experimental setup with respect to the design space, the experiment execution and the analysis for a specific biological process, is essential for answering the biological question under study.
Chapter 1

References


