FACULTY OF MEDICINE AND HEALTH SCIENCES

Dedication

To Pieter. It would have been impossible without your guidance, your clear and clever scientific perspective, your constant support and encouragement, and the calm you convey even when things get tough.
To Jo. You are also a big inspirator, no wonder you two make such a good working team. Scientific discussions with you were always stimulating and revealing.
I was really lucky to have the chance to work with both of you. It was a great honor.

To Pavel and Hua-Sheng. For those memorable times of hard work, taking advantage of the seven-hour difference between Gent and Houston. It was a pleasure working with you side by side and learning from your vast experience and clear insights.

To all the colleagues and scientists with whom I had the pleasure of collaborating and who contributed to making this work possible.

To the members of the examination board. Thank you for your time and effort in evaluating my thesis and for helping to improve the final result of this work.

To my family and friends, old and new, who are near or far.
You're always with me and I couldn't have done it without you.

GHENT UNIVERSITY

Examination Committee

Prof. dr. Vanessa Vermeirssen
Department of Biomolecular Medicine & Department of Biomedical molecular biology – Ghent University, Belgium.

Dr. Joost Kluiver
Department of Pathology and Medical Biology – University of Groningen, The Netherlands.

Prof. dr. Rory Johnson
Department of Biomedical Research (DBMR) – University of Bern, Switzerland.

Prof. dr. Lennart Martens
Department of Biomolecular Medicine – Ghent University, Belgium.

Prof. dr. Filip Van Nieuwerburgh
Department of Pharmaceutics – Ghent University, Belgium.

Prof. dr. Fransiska Malfait
Department of Biomolecular Medicine – Ghent University, Belgium.

Prof. dr. Jan Gettemans (chairman)
Department of Biomolecular Medicine – Ghent University, Belgium.

Link to the thesis:

CENTRUM MEDISCHE GENETICA GENT

CRIG

CHARTING REGULATORY RNAS IN THE HUMAN TRANSCRIPTOME USING RNA-SEQUENCING

Public PhD defense to obtain the degree of ‘Doctor in Health Sciences’

March 29, 2021

Lucía Lorenzi

Supervisor: Prof. dr. Pieter Mestdagh

Co-supervisor: Prof. dr. Jo Vandesompele

Summary

Technological advances in RNA-sequencing technologies have revolutionized *transcriptomics*, the field that studies RNA transcripts, by revealing that the wide majority of RNA produced in a cell do not code for proteins but exert diverse important regulatory functions. This new world of non-coding RNAs include a wide variety of molecules in different shapes and sizes, including short, long, linear and circular RNAs, all of which interact in complex regulatory networks that ultimately define how and when genetic information is expressed.

One of the most important and attractive applications of transcriptomics is its potential to aid in the understanding of human disease. The discovery and characterization of new RNA players can help better depict the molecular networks underlying disease and can lead to the identification of new therapeutic targets. For this reason, in Paper 1 we reviewed important current advancements, remaining challenges and future perspectives in the profiling of ncRNAs in cancer.

Next, we sought to generate a more comprehensive catalogue of the human transcriptome. Having a complete and high-quality gene annotation is of great importance for the functional interpretation of the genome and for genetic and transcriptomic studies that exploit these annotations as references. Several consortium-based efforts have generated large-scale RNA sequencing data sets that have proven extremely useful in enhancing and expanding human transcriptome annotation. Nevertheless, our picture of the transcriptome is still far from complete, in part due to traditional biases in RNA-sequencing studies. One of the main biases is that, in most datasets, RNA molecules are selected based on the presence of a poly(A) tail, by the use of *polyA RNA-sequencing*. Thus, non-polyadenlyated transcripts (those without a poly(A) tail) are not well represented in current transcriptome catalogues. Another limitation is that, despite its importance, the information on the DNA strand of origin of transcripts is usually lost during traditional sequencing experiments. This hinders the identification of *antisense RNAs*, a widespread class of regulatory RNAs. To tackle these limitations and capture a greater diversity of the human transcriptome, we initiated the *RNA Atlas project* (Paper 2). In this project, we applied complementary RNA-sequencing methods on a heterogenous collection of 300 human samples including 45 tissues, 162 cell types and 93 cell lines. For each of these samples, strand-specific total RNA, polyA, and small RNA-sequencing was performed. This unique dataset allowed us to reconstruct a comprehensive transcriptome including five major RNA biotypes (protein-coding RNAs, long intergenic non-coding RNAs, antisense RNAs, circular RNAs and micro RNAs) together with their matching expression profiles across all samples in our cohort.

To further refine and filter our assembled gene models and improve the overall quality of our transcriptome, we exploited publicly available large-scale data beyond RNA-sequencing. These included *chromatin states*, informative of transcriptional activity and enhancer regions, and cap analysis of gene expression (CAGE) sequencing, which helps to accurately identify transcription start sites (Paper 2).

Our matching polyA and total RNA-sequencing data allowed us to classify the polyadenylation status of all our transcripts and analyze its variability across tissues and cell types. These analyses revealed a large fraction of non-polyadenylated ncRNAs including a hitherto poorly catalogued class of single-exon lncRNAs (Paper 2).

As part of the data quality control in the RNA Atlas, we leveraged sample heterogeneity and tissue-specific expression in our dataset to directly estimate the magnitude of *read index hopping*, a previously reported cause of read-to-sample misassignment in RNA-sequencing. These analyses derived in a short technical report (Paper 3) describing the novel approach and demonstrating that the phenomenon was present, but at small rates, with likely negligible impact for most applications.

Finally, we exploited the unique aspects of our dataset to identify significant regulatory interactions mediated by all classes of non-coding RNAs. For this, we first integrated matching expression profiles for different types of RNAs with evidence for molecular interactions from public datasets and predicted thousands of regulatory relationships between all types of ncRNAs and protein-coding genes. We then used the broad exon and intron-coverage available from total RNA-sequencing data to provide mechanistic validation for a large fraction of the predicted interactions. These analyses allowed us to characterize the type of regulation (transcriptional or post-transcriptional) for thousands of non-coding RNAs and to derive a list of 316 miRNAs and 3,310 long non-coding RNAs with multiple lines of evidence for regulating protein-coding genes and pathways (Paper 2).

Publications

Paper 1

Long noncoding RNA expression profiling in cancer : challenges and opportunities.
Lucia Lorenzi, Francisco Avila Cobos, Anneleen Decock, Celine Everaert, Hetty Helsmoortel, Steve Lefever, Karen Verboom, Pieter-Jan Volders, Franki Speleman, Jo Vandesompele and Pieter Mestdagh. *Genes Chromosomes & Cancer. 2019*

Paper 2

The RNA Atlas, a single nucleotide resolution map of the human transcriptome.
Lucia Lorenzi*, Hua-Sheng Chiu*, et al. *Under revision in Nature Biotechnology.* Preprint: doi: https://doi.org/10.1101/807529
*contributed equally

Paper 3

Estimating sequencing read index hopping based on tissue-specific gene expression in RNA Atlas.
Lucia Lorenzi, Gary P. Schroth, Scott Kuersten, Jo Vandesompele and Pieter Mestdagh. *In preparation.*

Contact: lucia.lorenzi@ugent.be

lucialorenzi90@gmail.com