FACULTY OF MEDICINE AND HEALTH SCIENCES

GHENT UNIVERSITY

ADDRESSING CHALLENGES IN TRANSCRIPTOME ANNOTATION AND CELL-TYPE HETEROGENEITY THROUGH INTEGRATION OF OMICS DATASETS AND COMPUTATIONAL DECONVOLUTION

Francisco Avila Cobos

Supervisor: Prof. dr. Katleen De Preter
Co-supervisor: Prof. dr. Pieter Mestdagh

Thesis submitted as fulfilment of the requirements for the degree of

Doctor in Health Sciences, Academic year 2019 - 2020

Center for Medical Genetics
Cancer Research Institute Ghent
Ghent University Hospital, Medical Research Building 1
Corneel Heymanslaan 10, 9000, Ghent, Belgium

Francisco.AvilaCobos@UGent.be

Thesis submitted to fulfill the requirements for the degree of Doctor in Health Sciences.

Supervisor

Prof. dr. Katleen De Preter

Department of Biomolecular Medicine, Ghent University (Ghent, Belgium)

Co-supervisor

Prof. dr. Pieter Mestdagh

Department of Biomolecular Medicine, Ghent University (Ghent, Belgium)

Members of the examination committee

Prof. dr. Andrei Zinovyev
Department of Bioinformatics of Cancer - Institut Curie (Paris, France)

Dr. Bram De Wilde, M.D., PhD
Department of Pediatric Hematology and Oncology, Ghent University Hospital (Ghent, Belgium)

Prof. dr. Frank Speleman (chairman)
Department of Biomolecular Medicine, Ghent University (Ghent, Belgium)

Prof. dr. Kathleen Marchal
Department of Plant Biotechnology and Bioinformatics, Ghent University (Ghent, Belgium)

Prof. dr. Vanessa Vermeirssen
Department of Biomedical Molecular Biology, Ghent University (Ghent, Belgium)

Dr. Volodimir Olexiouk
Department of Biomolecular Medicine, Ghent University (Ghent, Belgium)

Prof. dr. Yvan Saeys
Department of Applied Mathematics, Computer Science and Statistics, Ghent University (Ghent, Belgium)

The research in this thesis was conducted at the Center for Medical Genetics Ghent (Ghent University, Ghent, Belgium). This work was supported by a Concerted Research Action of Ghent University (BOF/GOA), a Special Research Fund scholarship of Ghent University (BOF.DOC.2017.0026.01) and a grant for a long stay abroad from Scientific Research Flanders (FWO - V440318N)

TABLE OF CONTENTS

ABBREVIATIONS

3P-seq	polyA-position profiling by sequencing
ALS	alternating least squares
asRNA	antisense RNA
ATAC-seq	assay for transposase-accessible chromatin sequencing
AUZ	area under the zipper
BSS	blind signal separation
CAGE-seq	cap analysis of gene expression sequencing
cDNA	complementary DNA
ChIP-seq	chromatin immunoprecipitation sequencing
circRNA	circular RNA
CN	condition number
DNA	deoxyribonucleic acid
DNase-seq	deoxyribonuclease sequencing
DSA	digital sorting algorithm
DWLS	dampened weighted least squares
FACS	fluorescence-activated cell sorting
FANTOM5	functional annotation of mammalian genomes 5
FDR	false discovery rate
FFPE	formalin-fixed paraffin-embedded
FPKM	fragments per kilobase million
GRO-seq	global run-on sequencing
GTEx	genotype-tissue expression
H3K4me1	histone H3 lysine 4 mono-methylation
H3K4me2	histone H3 lysine 4 di-methylation
H3K4me3	histone H3 lysine 4 tri-methylation
H4K20me1	histone H4 lysine 20 mono-methylation
H3K36me3	histone H3 lysine 36 tri-methylation
H3K79me2	histone H3 lysine 79 di-methylation
H3K9ac	histone H3 lysine 9 acetylation
H3K14ac	histone H3 lysine 14 acetylation
H3K27ac	histone H3 lysine 27 acetylation

HECS	highly expressed cell specific
ICA	independent component analysis
IHC	immunohistochemistry
ISH	in-situ hybridization
kb	kilobase
LCM	laser capture microdissection
lncRNA	long non-coding RNA
log	logarithm
LLS	linear least squares
LS	least squares
MAD	mean absolute difference
MAQC	microarray quality control
MAS5	microarray suite 5
MAST	model-based analysis of single-cell transcriptomics
MBEI	model based expression index
mRNA	messenger RNA
miRNA	micro RNA
OLS	ordinary least squares
NK	natural killer
NNLS	non-negative least squares
NMF	non-negative matrix factorization
nt	nucleotide
PCA	principal component analysis
PBMCs	peripheral blood mononuclear cells
PolyA	poly adenylated
PRO-seq	precision nuclear run-on sequencing
PSEA	population-specific expression analysis
QN	quantile normalization
QP	quadratic programming
RACE-seq	rapid amplification of cDNA ends sequencing
RLR	robust linear regression
RMA	robust multi-array analysis
RMSE	root-mean-square error
RNA	ribonucleic acid

RNA-seq	RNA sequencing
SA	simulated annealing
SEQC	sequencing quality control
sqrt	square-root
SNM	supervised normalization of microarray
SVD	singular value decomposition
SVM	support vector machine
SVR	support vector regression
TCGA	the cancer genome atlas
TES	transcription end site
TILs	tumor-infiltrating lymphocytes
TMM	trimmed mean of M-values
TPM	transcripts per million
tpm	tags per million mapped reads
TSS	transcription start site
UBERON	Uber-anatomy ontology
UHRR	universal human reference RNA
UMI	unique molecular identifier
UTR	untranslated region
UQ	upper-quartile
VST	variance-stabilizing transformation
ZH	zipper height

Summary

Researchers working on transcriptomics, and more specifically on long non-coding RNAs (lncRNAs), have to deal with the task of reconstructing transcript models from RNA-sequencing data, being particularly challenging for low abundant lncRNAs. Moreover, lncRNAs are often situated near protein coding genes and could be portions of untranslated regions (UTRs) instead of independent transcriptional units. This is especially the case for mono-exonic lncRNAs, which are numerous when assembling transcripts from RNA-sequencing data. Furthermore, the majority of tools for lncRNA annotation are mainly based on evolutionary conservation and may filter out lncRNA transcripts given their limited conservation, hereby leading to false negatives.

To address these issues I created the Zipper plot, a novel visualization and analysis method that combines publicly available CAGE, ChIP and DNase sequencing data across a very large collection of tissue and cell types from both FANTOM5 and Roadmap Epigenomics Project to obtain more reliable lncRNA transcript structures and annotation. We validated the Zipper plot using a set of well-characterized long non-coding RNAs and observed that fewer mono-exonic lncRNAs have CAGE peaks overlapping with their transcription start sites compared to multi-exonic lncRNAs.

In a second phase, I used the Zipper plot to contribute to the curation of the human transcriptome that was generated from three complementary RNA sequencing methods on 162 normal cell types and 45 tissues being part of the RNA Atlas dataset. Those transcripts with no evidence of independent transcription either at RNA or DNA level were discarded, leading to a stringent set of transcripts.

The heterogeneous nature of samples commonly used in research (e.g. cancer samples or tissues) has been largely overlooked and gene expression analyses of bulk tissues often neglect cell type composition as a key confounding factor in downstream analyses. Therefore, many computational approaches have been developed to infer cell type proportions and/or cell type-specific expression profiles in heterogeneous samples (computational deconvolution).

First, I thoroughly reviewed different computational deconvolution methodologies developed since 2001. Next, taking advantage of the plethora of different RNA biotypes present in the RNA Atlas and given that existing computational deconvolution methods have been tested on messenger RNAs only, I evaluated the performance of the different RNA fractions in a computational deconvolution framework, highlighting the importance

of including a comprehensive collection of cell types and high-quality markers in the reference matrix used in computational deconvolution of transcriptomics data, regardless of the RNA fraction being used.

Finally, even though few studies have addressed individual factors (other than the method) impacting the deconvolution results, an in-depth evaluation of their combined impact on the deconvolution results is still missing. Therefore, I assessed the combined impact of four data transformations, twenty scaling/normalization strategies, seven marker selection approaches and twenty different deconvolution methodologies on one thousand artificial pseudo-bulk mixtures from four different single-cell RNA-seq datasets. I also evaluated the impact of removing cell types from the reference matrix that were actually present in the mixtures. The findings from this benchmark study together with the set of general guidelines we have put forward will aid researchers to find the most appropriate computational deconvolution pipeline for their data and research question. Furthermore, it will allow them to obtain more accurate cell type proportion estimates of infiltrating immune cells and other relevant cell types from the tumor microenvironment, enhancing tumor subtype classification, immunotherapy response prediction and improving the sensitivity of survival analyses in cancer research.

Samenvatting

Omwille van de lage abundantie van lange niet-coderende RNAs (lncRNAs) is het voor onderzoekers niet evident om lncRNA transcriptmodellen te reconstrueren uit RNA-sequencing data. Bovendien liggen lncRNAs vaak dicht bij eiwitcoderende genen en zouden ze deel kunnen uitmaken van onvertaalde regios (UTRs) in plaats van transcriptionele entiteiten op zichzelf te zijn. Dit is zeker het geval voor de vele mono-exonische lncRNAs die aanwezig zijn tijdens het assembleren van transcripten uit RNA-sequencing data. Bovendien zijn de meeste tools voor lncRNA annotatie gebaseerd op evolutionaire conservatie en, gezien de beperkte conservatie van lncRNAs, kunnen bepaalde lncRNAs hierdoor weggefilterd worden, wat dan weer leidt tot vals negatieven.

Om tegemoet te komen aan deze problemen, heb ik de Zipper plot ontworpen. Deze nieuwe methode van visualisatie en analyse combineert publiek beschikbare CAGE-, ChIP- en DNase-sequencing data met een hele grote collectie van weefsel- en celtypes, zowel uit FANTOM5 als Roadmap Epigenomics Project. Dergelijke integratie leidt tot betrouwbaardere lncRNA transcriptstructuren en -annotatie. We hebben de Zipper plot gevalideerd aan de hand van een set van goed gekarakteriseerde lncRNAs. Hierbij viel op dat er in het geval van mono-exonische lncRNAs minder CAGE-pieken overlappen met de startpunten van transcriptie dan van multi-exonische lncRNAs.

In een tweede fase heb ik de Zipper plot ook gebruikt voor de curatie van het humaan transcriptoom dat gegenereerd was uit drie complementaire RNA sequeneringsmethodes op 162 normale celtypes en 45 weefsels in de RNA Atlas dataset. Transcripten waarvan we noch op het niveau van RNA noch op het niveau van DNA enig bewijs van onafhankelijke transcriptie konden vinden, werden weggelaten. Dit resulteerde uiteindelijk in een stringente set van transcripten.

Bij stalen die algemeen gebruikt worden in onderzoek (bv. kankerstalen of weefsels) wordt hun eventuele heterogene aard vaak grotendeels over het hoofd gezien. Ook voor genexpressie-analyses van bulk weefsels wordt de samenstelling van celtypes dikwijls genegeerd. Dit terwijl deze samenstelling nochtans een belangrijke verstorende factor voor verdere analyses kan zijn. Omwille van die reden zijn intussen vele computationele benaderingen ontwikkeld die in heterogene stalen de verhoudingen van celtypes en/of celtype-specifieke expressieprofielen proberen af te leiden (computationele deconvolutie).

Zelf heb ik eerst de verschillende computationele deconvolutiemethodes die ontwikkeld waren sinds 2001 grondig bestudeerd. Vervolgens maakte ik gebruik van de RNA Atlas,

waarin een overvloed aan verschillende RNA biotypes aanwezig is, om de prestatie op de verschillende fracties RNA in het kader van computationele deconvolutie te evalueren. Bestaande computationele deconvolutiemethodes waren immers enkel getest op messenger RNAs. Uit mijn evaluatie bleek telkens het belang van de referentiematrix voor de deconvolutie van transcriptomics data, onafhankelijk van welke RNA-fractie gebruikt werd. Een goede referentiematrix bevat een uitgebreide verzameling celtypes en merkers van hoge kwaliteit.

Tot slot, was de impact van individuele factoren (naast de methode) op resultaten van deconvolutie al aangekaart in enkele studies, maar een diepgaande evaluatie van hun gecombineerde impact ontbrak nog. Daarom heb ik de gecombineerde impact getest van vier datatransformaties, twintig herschalings-/normalisatiestrategieën, zeven methodes voor merkerselectie en twintig verschillende deconvolutiemethodologieën op duizend artificiële pseudo-bulk samenstellingen, afkomstig van vier verschillende single-cel RNA-seq datasets. Daarnaast ging ik ook de impact na van het verwijderen van celtypes uit de referentiematrix die ook werkelijk aanwezig weren in de artificiële samenstellingen. De resultaten van deze benchmark studie en de door ons voorgestelde set van algemene richtlijnen zullen onderzoekers helpen om voor hun specifieke data en onderzoeksvraag de meest geschikte computationele deconvolutie pipeline te selecteren. Bovendien zal het hen in staat stellen om een betere schatting te maken van de celtype fractie van infiltrerende immuuncellen en andere relevante celtypes in de micro-omgeving van de tumor. Dit zal uiteindelijk bijdragen tot de subtypeclassificatie van tumoren, het voorspellen van immunotherapierespons en het verbeteren van de sensitiviteit van overlevingsanalyses in kankeronderzoek.

Part I - Introduction

Living cells have their genetic information stored in their genomes. By a process known as transcription[1], this information is used to produce different types of ribonucleic acids (RNAs) and the collection of RNAs (also known as transcripts) present in a cell at a specific time constitutes its transcriptome[2]. Remarkably, a given cell can have different gene expression profiles (genes transcriptionally active) at different time points.

Furthermore, there are many different types of RNAs. Messenger RNAs (mRNAs) serve as template for the production of proteins in a process called translation[3] with the aid of the RNA polymerase II enzyme[4]. These mRNAs frequently contain introns (in eukaryotic organisms) that have to be removed by the spliceosome machinery in order to produce a mature mRNA[5]. Non-coding RNAs do not code for proteins, play regulatory roles during transcription and translation and are normally classified into small or long non-coding RNAs depending on their length. Small non-coding RNAs are shorter than 200 nucleotides and include microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), small interfering RNAs (siRNAs) and small nucleolar RNAs (snoRNAs), among others[6]. Importantly, snoRNAs participate in the processing of rRNAs and the assembly of ribosome sub-units[7].

Other non-coding RNAs being key components in the translation process are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). Of note, rRNAs are synthesized by RNA polymerase I whereas tRNAs and small non-coding RNAs are synthesized by the RNA polymerase III[4].

Long non-coding RNAs (lncRNAs) are longer than 200 nucleotides and, depending on its position with respect to other protein coding genes, can be further divided into sense, antisense, intergenic, bi-directional or intronic[8]. Furthermore, they have been found to be involved in different processes associated to cancer development (e.g. MALAT1 has been associated to cell proliferation in liver, breast and colon cancer[9]; over-expression of HOTAIR promotes metastasis of breast cancer cells[10]).

Over the years, the analysis of the transcriptome has substantially contributed to our understanding of the processes involved in human development[11] and disease[12]. Nevertheless, transcriptome analysis comes with a number of challenges.

RNA-sequencing (RNA-seq) is a high-throughput sequencing method that yields a precise estimation of gene expression levels[2] and has been routinely used worldwide for the past decade (there are more than 79,000 research articles in PubMed Central[13]). In contrast to microarrays, which require target sequences to be known *a priori* in order to design probes attached to their surface, RNA-seq does not require any prior genomic information[14] and has become the sequencing alternative of choice.

First, there are several technical factors affecting the RNA-seq itself. The choice of RNA purification kit[15] and starting material (e.g. fresh frozen tissue or formalin-fixed paraffin-embedded (FFPE) samples)[16,17] often lead to different amounts of RNA with different quality and degree of degradation. Next, both the reverse transcription efficiency converting RNA to cDNA during the library preparation step[18] and platform-specific differences and variations introduced during RNA-seq library construction[19] have also an impact.

There is also a plethora of tools to choose from regarding *de novo* transcriptome assembly[20], some of which they need a pre-existing genome assembly to guide the process. Since there are many different genome assemblies available[21,22] and each has a different number of genomic features, this also has an impact in the RNA-seq output.

Furthermore, there are diverse biological factors inherent to the samples being investigated. These typically include relevant differences due to clinical condition, age or gender, all of which should re-appear or be accounted for in downstream analyses such as differential gene expression. However, such analyses usually do not take into account cell type composition as a confounding factor, resulting in a loss of signal from less abundant cell types and limiting the conclusions that can be drawn from the experiments.

The combination of insufficiently documented or incorrect data processing practices[23,24], the technical factors affecting the RNA-sequencing described above and the sample heterogeneity issue can partially explain the problem of lack of reproducibility that the scientific community is currently facing[25].

Among all aforementioned issues, there are two big challenges in transcriptomic research I have addressed in this dissertation: 1) transcript annotation for novel RNAs found using RNA sequencing (RNA-seq); 2) the heterogeneous nature of samples and tissues used in research.

1) Novel RNAs: transcriptional noise or real biology?

The availability of thousands of RNA-seq profiles has enabled different efforts to obtain the most comprehensive reference of the human transcriptome, including CHESS[26], RefLnc[27] (focused on the long non-coding portion only) and most recently, the RNA Atlas dataset[28].

Despite of all these efforts, non-coding RNA annotation remains poor and thousands of novel transcripts identified during the past years still need experimental validation. First, long non-coding RNAs (lncRNAs) have tissue-specific expression patterns[29] and are generally lowly expressed[30]. Next, transcript reconstruction from RNA-seq data often gives rise to large numbers of single-exon transcripts. Furthermore, each lncRNA has on average two to four different isoforms per locus[31], varying in length and, sometimes, in the number of exons present. LncRNA transcripts are also less abundant than protein coding genes [30], often resulting in a lack of junction reads from which transcript models are inferred. Furthermore, lncRNAs are often located in the vicinity of protein-coding genes and could represent unannotated portions of untranslated regions (UTRs) rather than independent transcriptional units. Repetitive elements, which represent at least 50% of the human genome[32] and are known to be problematic in polymorphism identification and transcript reconstruction[33], have been also found in a high percentage in novel non-coding intergenic transcripts[34]. Given the limited evolutionary conservation of lncRNAs[35] and given that several tools for lncRNA annotation exclude transcripts partially or totally overlapping protein-coding genes, this may lead to large numbers of false negatives discarded throughout the process.

Sometimes, RNAs labelled as "non-coding" are actually found to have coding potential and generate micro-peptides (e.g. using tools such as CPAT[36]), highlighting how challenging the transcriptome annotation task is.

Therefore, even though RNA-seq provides an accurate picture of the transcriptome, some (or many) novel transcripts that are found can merely be transcriptional noise or DNA contamination.

The FANTOM5 project used Cap Analysis of Gene Expression (CAGE) to establish a comprehensive collection of transcription start sites (TSSs) across the majority of human cell types. LncRNA transcript models can be refined or discarded with the integration of complementary datasets marking transcription start sites (e.g. by CAGE-sequencing) and

chromatin states synonym of active transcription (e.g. by Chromatin Immunoprecipitation Sequencing (ChIP-seq).

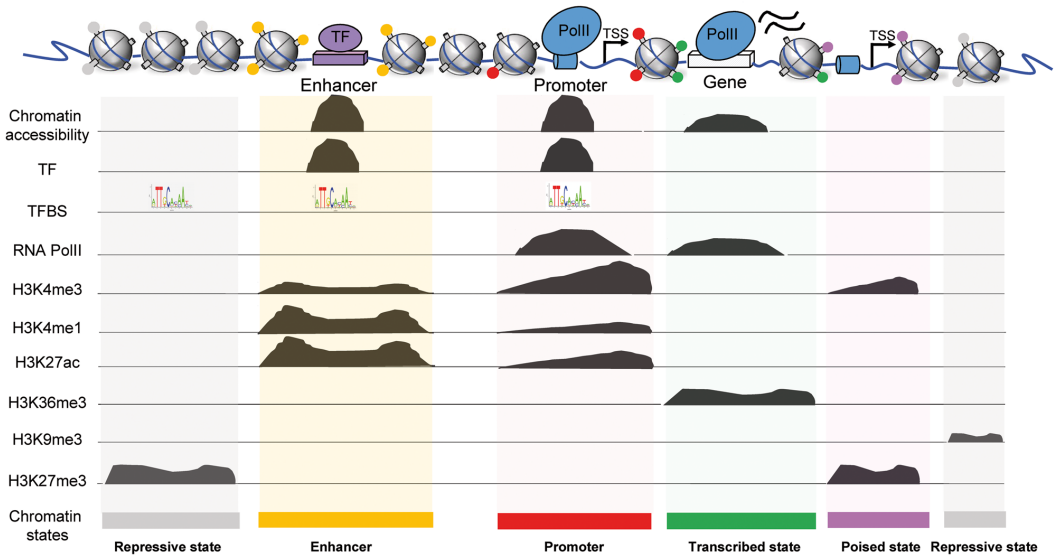


Figure 1 - Different combinations of histone modifications define diverse chromatin states. TF = transcription factor; Pol II = RNA polymerase II. Taken from Jiang and Mortazavi, 2018[37].

There are specific methylation and acetylation marks commonly found in actively transcribed promoters (H3K4me1, H3K4me2, H3K4me3, H4K20me1, H3K27ac, H3K9ac, and H3K14ac) and other relevant modifications that appear as consequence of transcription (H3K36me3, H3K79me2 at 5' end of gene bodies). Other relevant technology is DNase I hypersensitive sites sequencing (DNase-seq)[38], which enables the identification of genome-wide chromatin regions that are sensitive to cleavage by the DNase I enzyme, and thus, indicative of accessible ("open") chromatin regions, which are in turn related to transcriptional activity.

For instance, by combining the CAGE dataset from FANTOM5 together with several chromatin immunoprecipitation (ChIP) sequencing datasets, Andersson *et al.*[39] discovered the recurring pattern of bi-directional transcription of enhancers coincided with the co-occurrence of H3K27ac and H4K3me1 peaks, identifying more than 40,000 enhancers in the human genome in this manner. These analyses reveal the great potential of combining different sequencing datasets to find relevant signals of active transcription taking place.

2) Addressing sample heterogeneity through computational deconvolution

This section is built upon a review article I published in 2018[40] whose content has been updated to include the most recent developments that took place in 2019 regarding methodologies able to use single-cell RNA-sequencing data as input.

2.1 Challenges related to sample heterogeneity

As briefly mentioned before, the complex nature of samples and tissues used in transcriptomics research has been largely neglected. For instance, tumor samples are heterogeneous in nature, containing a variable portion of non-malignant cells that depends on the cancer type[41] (even when collected from the same patient) and include epithelial, stromal and infiltrating immune cells[42]. The expression level of each individual gene varies between different cell types and, when analysing bulk samples of heterogeneous tissues, only tissue-averaged expression levels are measured. As a result, the expression contribution of low abundant cell types could be masked by that of more abundant ones[43]. Therefore, observed changes in gene expression might be the result of underlying differences in cell type proportions between samples, genuine changes due to clinical condition or a combination of both.

Nevertheless, it is important to note that defining tumor heterogeneity is challenging at multiple levels: first, inter-tumor heterogeneity exists both between different tumor types and between samples within a given cancer (sub-)type (biological heterogeneity). Second, intra-tumor heterogeneity may also exist within a given sample (different tumor subclones).

The field of single-cell genomics has grown exponentially during the past few years, leading to the development of novel tools for the analysis of single cells within heterogeneous tissues[44–47]. Initially, single cells were isolated using Fluorescence-Activated Cell Sorting (FACS) or Laser Capture Microdissection (LCM) technologies. However, this upfront sorting required specific cell-surface markers to be known and appropriate antibodies to be available. To overcome this bottleneck, novel systems enable single cell isolation using other physical properties (e.g. cell size) by applying microfluidics or dielectrophoretic separation. Nevertheless, while promising, single-cell technologies have labour-intensive protocols and require expensive and specialized resources, currently hindering their establishment in a clinical setting. Moreover, some tissues are hard to dissociate, single-cell capture efficiencies still remain low[48] and single-cell technologies cannot be used to

analyse genetic material that is free in circulation (=not inside cells but present as circulating free RNAs (cfRNA) in liquid biopsies (=blood samples) or in other biofluids (e.g. saliva, urine)).

For all these reasons, and to exploit the wealth of publicly available bulk data that can be re-analysed, multiple computational approaches have been developed in the past years to infer abundance of different cell types and/or cell type-specific expression profiles in heterogeneous samples. This task is commonly known as computational deconvolution of expression data from mixed cell populations[40,49–51].

2.2) Defining the deconvolution problem

Eighteen years ago, Venet *et al.* [52] presented a method "*to infer the gene expression profile of the various cellular types […] directly from the measurements taken on the whole sample*". They framed this problem as a multiple linear regression model applied to microarray gene expression data. The generalization of this problem belongs to the category of blind signal separation (BSS) problems, with the "cocktail party problem" being the most known example[53]. It was formulated as the recognition of what a person says when others are speaking at the same time[54], which in turn can be translated into separating a set of observations into the constituent independent signals (sources).

The expression of a given gene in a heterogeneous sample can be modelled as the weighted sum (=linear combination) of the expression values from each cell type present in the mixture, assuming that every cell type has similar expression levels across different samples. Thus, the deconvolution problem can be formulated in matrix notation as follows (equation 1):

$$T = C \cdot P \qquad (1)$$

where T = measured expression values from heterogeneous tissue/tumor samples; C = cell type-specific average expression values and P = mixing proportions.

It can also be formulated algebraically as a latent variable model where the error term is not directly measurable (equation 2):

$$t_{ij=} \sum_{k=1}^{K} c_{ik} \cdot p_{kj} + e_{ij} \text{ ; i=1…M and j=1…N} \qquad (2)$$

where t_{ij} = observed expression value of gene i in heterogeneous sample j; c_{ik} = averaged expression value of gene i in cell type k; p_{kj} = proportion of cell type k in sample j; e_{ij} = error term; K = number of cell types; M = number of genes and N = number of samples.

The deconvolution can be performed if the system of linear equations has solution (the number of solutions of a system of linear equations can be determined using the Capelli-Fontené-Frobenius-Kronecker-Rouché theorem[55]). Depending on the information used as input, the deconvolution can have several definitions, as described in Figure 2. Specifically, unsupervised (=non-guided) scenarios where only T is available and both C and P are estimated, are known as **complete or full deconvolution** frameworks. In contrast, when *a priori* information (either matrix C or P) is available along with T, these supervised or guided approaches are also known as **partial deconvolution** frameworks.

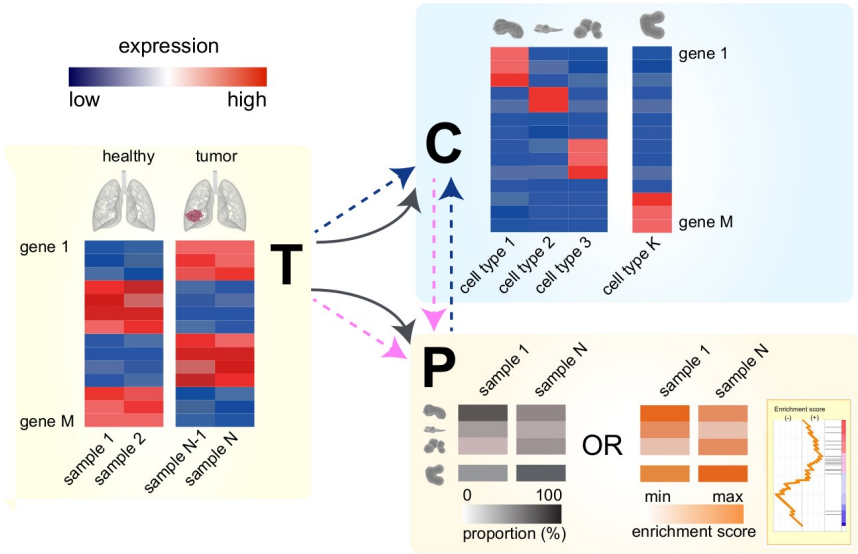


Figure 2 The deconvolution problem has multiple formulations depending on the available input data. T = matrix containing the observed (measured) expression values from heterogeneous (tissue/tumor) samples (M genes, N samples); C = matrix consisting of cell type-specific average expression values (M genes, K cell types); P = matrix containing the mixing proportions (=relative composition) (K cell types, N samples); min = minimum; max = maximum. **Case 1)** Only T is available, C and P are estimated (dark grey arrows). **Case 2)** Given T and C, P is estimated (dashed pink arrows; grey heatmap on the bottom-right corner). One variant of this formulation uses T and cell type signatures (lists of marker genes for each cell type) known from literature or obtained by supervised/unsupervised marker selection strategies to estimate relative measures of the tissue heterogeneity (=enrichment scores; orange heatmap on the bottom-right corner) instead of cell type proportions. (e.g. ESTIMATE[56,57]). Proportion values are strictly positive, bounded between 0 and 100 and with straightforward interpretation, whereas enrichment scores are unbounded and sometimes negative, making them harder to interpret.); **Case 3)** Given T and P, C is estimated (dashed blue arrows). See Supplementary Table 1 (available online in Avila Cobos *et al.* [40]) and "Mathematical approaches to solve the deconvolution problem" for more details.

Finally, some methods model the heterogeneous samples as a two-component system (e.g. tumor and non-tumor) [58–62], whereas others increase the complexity by including three or more cell types in the mixture[63–67], getting as far as 22[68] or 25 different cell types[69].

2.3) Mathematical approaches to solve the deconvolution problem

A detailed description of the particular deconvolution problem solved by each method, the necessary input data and their availability can be found online in the Supplementary Table 1 from Avila Cobos *et al.*[40]

2.3.1) Partial deconvolution approaches

The most commonly used group of methods is called ordinary least squares (OLS), linear least squares (LLS) or simply least squares (LS), whose goal is to **minimize the sum of squares** of the differences between fitted ($C \cdot P$) and observed values (T) (also known as minimization of the norm of the reconstruction error or minimization of the Euclidean distance) regardless of the distribution of the error term (equation 3; see Figure 2):

Given T and C (or T and P):

$$\min_{P \text{ (or C)}} ||C \cdot P - T||^2 \qquad (3)$$

This can be immediately seen as minimizing the residual sum of squares (RSS) of a standard linear regression ($\min_\beta || y - X\beta ||^2$).

Under the assumption that the error terms follow a normal distribution, a maximum likelihood estimation approach can also be applied to solve the minimization problem[70].

Optimization problems aim to minimize or maximize diverse objective functions with or without imposed constraints. Whilst the goal is always to find the global minimum or maximum of the objective function, some methods might get stuck in local minima or maxima. The sum of squared residuals can also be minimized using simulated annealing (SA)[71,72] (see BOX 1) or other non-convex optimization strategies. However, since only convex objective functions guarantee that a local solution corresponds to the global solution[73], these are not guaranteed to find the optimal solutions. Moreover, since gene expression matrices are typically not sparse, non-convex optimization strategies might result in high computational times and low rates of convergence (see BOX 1).

BOX 1- Glossary of terms

Gini index: in the context of marker selection, measure ranging from 0 to 1 used for the identification of tissue-enriched genes. The closer to 1, the higher the likelihood of a gene being exclusively expressed in one tissue.

Jensen-Shannon divergence: metric from information theory often used to discover cell-type specific genes. It quantifies the similarity between the expression of a given gene across tissues and that of a hypothetical gene whose expression is restricted to only one cell type.

Quadratic programming: optimization of a function that contains at least one quadratic term.

Simulated Annealing: Optimization of a function that allows worse solutions at some iterations with a probability that decreases as the solution space is explored. The worsening steps allow a broader search across the function domain.

Support Vector Regression (Support Vector Machine): supervised learning model used for regression or classification of linearly separable data into two categories.

Bayesian framework: statistical inference framework in which Bayes'theorem is used:

$$p(y|\theta) = \frac{p(\theta|y) * p(y)}{p(\theta)}$$

Therefore: $p(y|\theta)\ \alpha\ p(\theta|y) * p(y)$ (where α denotes proportionality). This is often translated into $posterior\ \alpha\ likelihood * prior$

In Bayesian inference, the **prior distribution** represents the knowledge we have about how the data was generated before its actual generation. The prior is combined with the probability distribution of the observed data to yield the **posterior distribution**. The **likelihood** function for the data represents how likely the data (y) is given the model specified by any value of θ. A **parameter** is the numerical characteristic of a statistical model and a **hyperparameter** is the parameter of a prior distribution.

Convex function: Function in which the midpoint of any segment between two points of the graph of the function is located above the graph or on the graph itself.

L2-norm function: $\sqrt{\sum_{j=1}^{p} \beta_j^2}$. In the context of the deconvolution, β_j can be the least squares coefficient estimate and p the number of predictors.

Condition number (CN) of a matrix: $||A||^*||A^{-1}||$; where $||.||$ is the matrix norm. For example (with L2-norm and matrix A):

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, A^{-1} = \begin{pmatrix} \frac{-1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{-1}{3} \end{pmatrix}$$

$$cond(A)_2 = \sqrt{1^2 + 2^2 + 2^2 + 1^2} * \sqrt{(\tfrac{-1}{3})^2 + (\tfrac{2}{3})^2 + (\tfrac{2}{3})^2 + (\tfrac{-1}{3})^2} = 3.33$$

Convergence: criterion used to evaluate the improvement of a solution found by an algorithm after each iteration. When a solution does not change more than a pre-specified threshold with respect to the last n iterations ($n >= 1$), it is said that the algorithm has converged and it halts.

K-dimensional polytope: geometric object of K dimensions with K flat sides.

Sparse matrix: matrix in which most elements are zero.

Fortunately, it can be shown that every LS problem can be formulated as a quadratic programming problem[74] (see BOX 1) and there are several implementations such as the quadprog function in MATLAB[75,76]) that guarantee a globally optimal solution. Other commonly used functions to solve this optimization problem are lsqlin from MATLAB[77,78]) or lsfit[79], limSolve[80,81] or the quadprog package[75,82] in R[83].

However, with this initial formulation of the problem (equation 3, unconstrained optimization problem), both positive or negative proportions (P) may arise and the sum of the proportions might be different than one (see BOX 2).

Box 2 - Dummy example for deconvolving the cell type proportions of 4 cell types (k = 1,...,4) present in 1 sample (j=1) assuming linear expression values for 8 genes (i = 1,...,8).

Assuming T and C are known (P is unknown): Each cell type proportion corresponds to the regression coefficient (β) of a linear model formulated as:

$$t_{ij} = c_{ik} * \beta_{kj}$$

$$\begin{pmatrix} 1 \\ 10 \\ 850 \\ 1000 \\ 50 \\ 6 \\ 1080 \\ 1300 \end{pmatrix} = \begin{pmatrix} 20000 & 1 & 1 & 1 \\ 10000 & 1 & 1 & 1 \\ 1 & 1000 & 1 & 1 \\ 1 & 1000 & 1 & 1 \\ 1000 & 1 & 1000 & 1 \\ 20000 & 1 & 1000 & 1 \\ 1 & 1000 & 1 & 1000 \\ 1 & 1000 & 1 & 500 \end{pmatrix} * \begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \beta_{3,1} \\ \beta_{4,1} \end{pmatrix}$$

When solving the above problem by linear least squares regression, the solution is: $\beta_{1,1}$ = -0.0005, $\beta_{2,1}$ = 0.9789, $\beta_{3,1}$ = 0.0320 and $\beta_{4,1}$ = 0.2092; with the total sum of proportions being 1.220.

Negative proportions are meaningless in the context of the deconvolution. When adding the *non-negativity* constraint by using the nnls function (R package), the new solutions are: $\beta_{1,1} = 0$, $\beta_{2,1} = 0.9789$, $\beta_{3,1} = 0.0268$ and $\beta_{4,1} = 0.2092$; with the total sum of proportions being 1.215. Finally, the *sum-to-one* constraint still has to be incorporated (during or after the optimization procedure) to obtain a definitive solution: $\beta_{1,1} = 0$, $\beta_{2,1} = 0.8057$, $\beta_{3,1} = 0.0221$ and $\beta_{4,1} = 0.1722$.

Since those scenarios are meaningless in the context of the deconvolution, two constraints are included into the optimization problem: the proportions must be strictly positive between 0 and 1 ("non-negativity" constraint) and the sum of proportions within each sample must add up to 1 ("sum-to-one" constraint). This approach is known as the non-negative least squares method (NNLS)[52,79,84–87]. The nnls or lsqnonneg functions in MATLAB or the 'nnls' package[88] in R are common functions implementing this approach.

Robust regression can also be used to solve the problem. Fast And Robust DEconvolution of Expression Profiles (FARDEEP)[89] is an adaptive least trimmed square approach that performs outlier removal before coefficient estimation based on the Bayesian Information Criterion (BIC) and assuming that the error term follows a log-normal distribution (instead of a normal distribution). Robust linear regression (RLR)[90] is an alternative to OLS to discard outliers or influential observations through iteratively reweighted least squares by assigning them lower weights.

Ridge regression (also known as Tikhonov regularization), lasso regression, elastic net and Digital Cell Quantifier (DCQ) are four different penalized regression approaches implemented using the glmnet[91] function that vary in the choice of parameters (ridge: alpha = 0; lasso: alpha = 1; elastic net: alpha = 0.2; DCQ: alpha = 0.05).

EPIC[92] is a weighted and constrained least square minimization to estimate the proportion of each cell type with a reference profile and another uncharacterized cell type. Marker genes with low variability are assigned bigger weights in the function to be minimized.

In the Digital Sorting Algorithm (DSA)[82], the reference matrix C is converted into a diagonal matrix constructed by averaging all genes that are highly expressed in each individual cell type. Using this new diagonal matrix and the bulk mixture as input, the cell type proportions are computed.

DeconRNASeq[93] solves the problem via quadratic programming using the lsei function (limSolve R package) with implicit non-negative and sum-to-one constraints. dtangle, contrarily to all other methods that require values on the linear scale as input, models the problem in the logarithmic scale.

A second group of methods are **support vector regression approaches** with linear kernel (ν-SVR) (see BOX 1), including CIBERSORT[68] and ImmuCC[69]. Support vectors are robust against noise introduced by unknown cell types present in the mixture and involve the minimization of both a linear loss function and a L2-norm function (see BOX 1), penalizing model complexity while minimizing the variance of the proportions assigned to highly correlated cell types, combating multicollinearity (see "*Multicollinearity: presence of correlated cell types in the mixture*").

2.3.2) Complete deconvolution approaches

These methods include different Bayesian and unsupervised non-negative matrix factorization (NMF or NNMF) approaches.

Regarding those with a **Bayesian framework** (see BOX 1), all attempt to maximize a likelihood function, but each method models the problem differently. They have a different type and number of parameters and hyper-parameters, with completely different *a priori* and *a posteriori* specifications (probability mass/density functions), leading to completely different likelihood functions. Since the joint estimation can be computationally intractable when the number of parameters is high, each method proposes different alternatives to make the problem tractable (e.g. approximating posterior distributions using Markov Chain Monte Carlo techniques, approximating expected values of parameters with Monte

Carlo integration or using expectation-maximization (EM) algorithms to iteratively maximize the likelihood of the observed data [58,59,61,94,95]. It is unfeasible to describe them individually here and I highly advise the reader to go to the original publications to get a detailed overview of the modelling approach of interest. CDSeq[96] is an example of a complete deconvolution methodology that uses a Bayesian framework.

The separation of heterogeneous samples into their constituent cell types can also be approached as an unsupervised (=non-guided) **dimensionality reduction problem**, with principal component analysis (PCA), Independent Component Analysis (ICA) or NMF being widely used for this goal[97,98]. The number of relevant components (=cell types present) can be established visually or by using diverse rules[99]. However, PCA-based approaches may not be the most appropriate since factors other than the cell type identity might be contributing to the proportion of variance explained.

ICA is another unsupervised statistical technique that identifies mutually independent non-gaussian components (dimensions) that are latent in the data[100]. In contrast to PCA (where the components are uncorrelated and ranked by the amount of variance they explain), ICA components might be correlated. ICA can be used in the context of deconvolution once the number of independent components present is indicated[101]. However, assigning components to specific biological processes, cell types and technical factors can be challenging. Typically, cell-types are associated with components through highest correlation[102] and R packages such as DeconICA (https://github.com/UrszulaCzerwinska/DeconICA) have been designed to use ICA in the deconvolution of transcriptomics data.

The **NMF formulation** factorizes T as the product of C and P and incorporates the non-negativity constraint for all elements of both C and P. As a first step, initial values for P or C have to be generated[103]. On one hand, these initial values can be easily implemented by sampling random numbers from a uniform distribution. However, multiple attempts with different initializations are needed to achieve a stable final solution. Moffitt *et al.*[104] successfully applied the NMF to pancreatic ductal adenocarcinoma with 20 random initializations, identifying different tumor subtypes with different tumor and stromal fractions. On the other hand, since the initialization process has a significant impact in the final results, singular value decomposition-based methods have been developed in an attempt to improve the initialization stage[105]. The most common algorithm used for NMF is called alternating least squares (ALS)[106] and consists of two iterative steps that are

repeated until convergence: first, P is fixed and, together with T, C is estimated by NNLS; secondly, C is fixed and, together with T, P is estimated.

ssKL and ssFrobenius[103] are modified versions of the original non-negative matrix factorization (NMF) algorithm using Kullback-Leibler divergence and Frobenius norm respectively. They are semi-supervised approaches that do not need a reference expression matrix as input (only the bulk mixtures and a list of markers (labels) to be used).

Linseed[107] is another complete deconvolution method that first identifies a subset of mutually linear genes, followed by a determination of a putative number of cell types present in a mixture using singular value decomposition (SVD). Finally, it performs the deconvolution in a DSA-like manner.

2.4) Deconvolution methods using single-cell expression data as reference

Various methods capable of using single-cell RNA-sequencing data for deconvolution have emerged in the past year.

deconvSeq[108], which also allows deconvolution of bisulphite sequencing data, requires the input to be in linear scale and un-normalized. It uses the scRNA-seq data to perform an internal differential gene expression step using edgeR[109] to find markers and to obtain the reference matrix for the deconvolution.

MuSiC[110], DWLS[111], SCDC[112] and Bisque[113] are four different variants of the non-negative least squares methodology. MuSiC computes cell-type-specific library sizes and cross-subject mean and variance for each gene. Instead of pre-selecting marker genes based only on mean expression, it assigns weights to all genes, prioritizing low cross-subject variance across subjects while simultaneously paying attention to the cross-cell variability in gene expression. Importantly, it is also able to account for gene-specific protocol bias (bulk and single-cell expression data are likely to come from different protocols).

DWLS (dampened weighted least squares) is a weighted least squares approach tweaked to properly adjust the contribution of each gene (e.g. avoid minimal contribution of good markers only due to low mean expression levels). The weights are a function of an initial solution found through ordinary least squares and includes a dampening constant that avoids infinite weights coming from low gene expression levels or low proportions. It includes an internal differential gene expression analysis step using MAST[114] where the data is log2 transformed and genes with adjusted p-value <= 0.01 and a log fold-change

>= 0.5 between each cell type and all the others are considered markers. The final number of marker genes to be kept is the one which minimizes the condition number of the reference matrix. Eventually, DWLS creates a reference matrix consisting of the mean expression values across all cells from each cell type which is used in the deconvolution.

Bisque internally converts single-cell counts to counts per million (thus, only non-normalized data can be used as input), filters zero-variance genes and generates a reference profile from the single-cell data by averaging read counts across each cell type. Based on the fact that there is a strong (but not perfect) positive correlation between bulk and single-cell reconstituted pseudo-bulk expression data (by adding up counts from individual cells belonging to the same cell type), it transforms the bulk data to maximize the global linear relationship with the pseudo-bulk data across all genes. Finally, using the reference profile and the transformed bulk data, cell type proportions are computed via NNLS.

The previous three methods are only able to use one single-cell expression dataset at a time as reference matrix for bulk gene expression deconvolution. SCDC aims to improve the robustness and accuracy of the deconvolution by integrating multiple single-cell datasets at once while accounting for batch effects. Compared to MuSiC, SCDC estimates cell-type proportions with a different weighted NNLS framework where the contribution of each subject to the reference matrix varies according to the data quality (higher weights to scRNA-seq datasets more closely related to the bulk).

2.5) Selection of cell-type specific markers or expression profiles

For the second formulation of the deconvolution problem (see "Case 2" described in the legend of Figure 2), cell type-specific markers or cell-type specific expression profiles are needed. This section describes several useful approaches for this endeavour.

Importantly, I have focused on expression values at gene level. However, scenarios where underlying differences in alternative transcript expression among different samples are masked at gene level may very well arise. Moreover, the usage of transcript expression might result in more candidate biomarkers and even higher cell-type specificities, potentially increasing the accuracy of the deconvolution results. Therefore, a deconvolution using expression values at both gene and transcript level could be considered if possible.

Ideally, a cell type-specific marker is a gene whose expression is restricted to one cell type and is robustly expressed across different biological replicates from the same cell type[64]. However, since the deconvolution can only be solved if the number of marker genes is greater or equal than the number of cell types present in the mixture[55] and the presence

of closely related cell types (= with only subtle differences in their transcriptome) is a very frequent scenario, the original restrictive definition of a cell type-specific marker is changed to a gene predominantly expressed in one cell type and to a lesser extent expressed in others[52].

A first approach to select marker genes consists of finding genes whose average expression value in one cell type is several times greater than the median expression value across all cell types. The "highly expressed, cell specific" (HECS) gene database[115] is an example of this approach and contains lists of cell type-specific genes from microarray data across 84 human and 96 murine tissues and cell types. The previous approach can be refined by statistically assessing differential gene expression between every cell type against all other cell types and setting arbitrary fold-change (e.g. >=3) and p-value (e.g. <0.05) thresholds[69,103,116,117]. Of note, several authors recommend the use of medium-to-high expressed genes as robust markers, instead of the most expressed ones[43,85].

Some methods go one step further and rank the markers based on signal-to-noise ratios[67,72] or include an extra feature selection strategy to remove poorly discriminating marker genes[43,68,80]. The F-statistic (measure of their fit in the multiple linear regression model)[65], the Gini index[118], the Jensen-Shannon divergence[29] (see BOX 1) or the components from PCA, ICA or NMF analyses[104,119] can be also used to identify marker genes.

More advanced methodologies include CellMapper[120], Nanodissection 1.0[121], UNDO[62] and CAM[87]. Assuming that marker genes for a given cell type should correlate with each other and starting with as little as one cell type- specific marker gene, CellMapper (developed and validated using microarray data but potentially applicable to RNA-seq data) uses thousands of publicly available expression profile datasets (pre-loaded as objects in "CellMapperData" Bioconductor package) or custom datasets to find other marker genes with similar expression patterns and specifically expressed in a cell type of interest.

By selecting among 28 different human tissues and uploading a set of at least ten candidate marker genes ("positive standard") and ten genes expressed in other lineages ("negative standard"), Nanodissection 1.0 estimates the probability that a gene is cell-type specific using an iterative linear support vector machine (SVM) approach.

Both UNDO and CAM are completely unsupervised approaches that allow novel marker identification without any prior information by geometrically identifying the vertices and resident genes of a K-dimensional polytope (see BOX 1) built from a gene expression matrix, with K being the number of cell types present in the mixture.

In conclusion, the generation of cell type-specific expression matrices is not trivial, varies from method to method and is a determinant factor to consider when approaching the deconvolution strategy.

2.6) Factors affecting the deconvolution efficiency

Several studies have shown that the detection of differentially expressed genes after the deconvolution of bulk expression data is less prone to the identification of false positives and false negatives[72], resulting in more accurate[82], specific and sensitive results[122] when compared to those obtained from bulk heterogeneous (tumor) samples.

Of note, there are multiple factors affecting the performance of the deconvolution, which are discussed below.

2.6.1) Effect of pre-processing and normalization

As Hoffmann *et al.*[64], Clarke *et al.*[123] and Repsilber *et al.*[85] pointed out, the data normalization procedure has an impact on the estimation of cell type proportions, cell type-specific expression profiles and thus, the power to detect differential expression. Moreover, Newman *et al.*[124] highlighted the need of accounting for normalization differences in order to perform meaningful comparisons between different deconvolution methods. Most methods presented in this introduction assume that the data is appropriately pre-processed and normalized prior to the deconvolution (see Supplementary Table 1 from Avila Cobos *et al.*[40]). Some methods applied to data coming from different platforms incorporate a batch effect correction using Combat[57,125] or the supervised normalization of microarray (SNM) method[126].

Controversially, some methods propose background correction[69,80] whereas others recommend not to apply it [127]. On the one hand, Hoffmann *et al.*[64] finds the Microarray Suite 5 (MAS5) to provide a more robust estimation of the proportions compared to the robust multi-array analysis (RMA) and model based expression index (MBEI). On the other hand, Ahn *et al.*[82] discuss a deviation from the linearity assumption when applying MAS5 scale normalization, which was not observed when using RMA together with quantile normalization. Interestingly, Irizarry et al.[128] shows that background correction rather than the normalization method, is the main factor explaining differences between different pre-processing alternatives for Affymetrix GeneChip systems. Thus, a quantitative evaluation of the impact on the deconvolution results would be relevant for the field. A detailed summary about the normalization strategies can be found in Supplementary Table 1 from Avila Cobos *et al.*[40].

2.6.2) Logarithmic vs linear space

Statistical tests used to assess differential gene expression typically assume an underlying normal distribution of the data being analysed. For this reason, since the log-normal distribution is considered as a good approximation for microarray expression data[129] and stabilizes the variance[130], the data is often transformed into logarithmic scale.

However, Zhong and Liu[131] showed that log transformed microarray data violated the linearity assumption of equation 2 (see "Defining the deconvolution problem"), leading to a consistent over-estimation of the cell-type proportions. On the other hand, when the data was transformed back to linear scale, it resulted in an accurate deconvolution. The linearity assumption was also confirmed by[43,132] on non-log transformed microarray data. Zhong *et al.*[82] alleged that the linearity assumption also holds true for RNA-seq data and recently, Jin *et al.*[133] performed a thorough assessment of the linearity assumption of transcript abundance from RNA-seq data. They showed the need of normalizing the data prior to the deconvolution and concluded that when using RNA-seq data, TPM values from Salmon[134], RSEM[135] or Kallisto[136] provided the most accurate reconstruction of cell type proportions present in a mixture.

In line with this argument, the vast majority of methods agreed on transforming the data into log scale for pre-processing and data normalization followed by a conversion back to linear scale (using the anti-log transformation) prior to the deconvolution[59,87,137]. Although the linearity assumption is valid for most genes, a more accurate deconvolution might be achieved by detecting and excluding genes affected by non-linear amplification[122], excluding noisy genes with little biological signal[79] or removing outliers (= trimmed robust regression)[64] before applying the least squares method.

However, others claimed that it is possible to apply the deconvolution to both log-transformed and non-log transformed data[85,95] or claimed more accurate results when using quantile normalized and log2-transformed data[80]. Furthermore, Clarke *et al.*[123] requires log-transformed data to find accurate estimates of the proportion of a cell type in a mixture.

A counterintuitive statement comes from Repsilber *et al.*[85], claiming optimal deconvolution of cell type-specific gene expression using log-transformed data whereas cell type-specific differential expression is optimal when using non-log-transformed data.

2.6.3) Multicollinearity: presence of correlated cell types in the mixture

Significant correlation between two or more cell types (also known as multicollinearity in the context of linear regression) might result in an increase of the estimation errors and the impossibility of separating the contribution from individual cell types[43].

Even though some authors assume gene expression profiles between different cell fractions to be uncorrelated[52], this might be an unrealistic scenario with important consequences. As Newman *et al.*[68] pointed out, the deconvolution results can be negatively affected when many related cell types were present, which may result in higher proportions being assigned to the cell type whose expression profile is most similar to the mixture. One possible solution to tackle this problem is the support vector regression (SVR) methodology implemented by CIBERSORT[68], which minimizes the variance of weights assigned to highly correlated predictors. CIBERSORT was able to deal with five highly collinear cell types and has been successfully applied to more than 18,000 expression profiles to analyse overall survival across 25 cancer types and abundance of diverse tumor-associated leukocyte subsets[125]. Mohammadi *et al.*[51] found that using the L2 loss function together with an R2 regularizer gave the best results and they reasoned that the regularization of the objective function can improve the performance in cases where highly correlated cell types are present in a mixture.

2.6.4) Condition number of a matrix

It is known that the condition number ("CN"; see BOX 1) has an impact when solving systems of linear equations (equation 1)[138]. Abbas *et al.*[79] and Newman *et al.*[68] stated that reference expression profiles (matrix C in Figure 2) could become more robust by minimizing the CN. Abbas *et al.*[79] found the CN to be high for matrices containing small or large number of genes whereas the CN was minimum for moderate numbers. Newman *et al.*[68] calculated the CN value for all candidate signature matrices for 22 cell types and kept the one with lowest CN. Glass and Dozmorov[139] discovered that a high CN of the matrix containing the cell proportions (matrix P in Figure 2) negatively affected the sensitivity of the deconvolution. Interestingly, Gentles *et al.*[125] noticed that the exclusion of cell types with the lowest proportion mean resulted in a noticeable improvement in sensitivity and in a considerable reduction of the CN. Interestingly, Teschendorff and Zheng[140] also emphasize the importance of optimizing the CN when selecting CpGs to deconvolute DNA methylation data. For all these reasons, this factor should not be overlooked when building the necessary matrices for a deconvolution problem, aiming at the smallest CN values as possible.

2.6.5) Cell cycle

Cells are dynamic systems, reflected by continuous changes in their transcriptome. Each sample has a mixture of cells in different phases of the cell cycle. When working with cultured cells, the cell cycle can be synchronized by chemical arrest or nutrient starvation[141]. However, this is not possible when tissue samples are profiled. Lu *et al.* [71] pioneered the estimation of the proportions of cells in different phases of the cell cycle using microarray expression data. They proposed the use of phase-specific markers (such as cyclin *CLN2* for phase G1 or *CLB4* for phase G2) to establish different time points of the cell cycle. Even though the vast majority of methods do not include this complex aspect when modelling the deconvolution problem, this must be ideally taken into account when developing new tools.

2.7) Minimum cell type proportions that can be detected

Zhong *et al.*[82] were able to accurately estimate cell types present at more than 10%, with a substantial decrease in accuracy if the percentage was smaller than that threshold. PERT[126] and DeconRNAseq[93] were able to retrieve proportions as small as 2% whereas CIBERSORT[68] detected fractions down to 0.5% in mixtures containing <50% of tumor content.

2.8) Cell type proportions as output: RNA content vs number of cells

Even though different cells have different sizes and RNA content, most deconvolution methods assume an equal amount of RNA in each cell, regardless of the cell type. When this assumption does not hold, the cell type contribution deviates from the cell abundance and the results are in fact mRNA proportions rather than absolute cell type proportions. EPIC[142] and Linseed[107] are two deconvolution methods that address this issue by re-normalizing the proportions based on cell-type-specific mRNA content or by including an extra cell size coefficient in the optimization problem, respectively.

2.9) Assessment of the deconvolution results

Multiple empirical approaches have been proposed to assess the validity of the estimations generated by the deconvolution methods: 1) in-situ hybridization (ISH)[43,132] or immunohistochemistry (IHC) stainings from the Human Protein Atlas[121] to validate cell type-specific gene expression; 2) comparison of predicted proportions with those measured by flow cytometry[126]; 3) combination of microscopy and FACS analysis to evaluate the

estimated proportion of yeast cells in different stages of the cell cycle[87]; 4) correlation with immune-fluorescence cell estimates or cell fractions inferred from DNA methylation[57,66] or DNA copy number data[57].

Moreover, when single-cell RNA-sequencing data is available, pseudo-bulk mixtures can be made to validate the performance of deconvolution methods. First, the number of cell types to be present and their identities have to be chosen, followed by choosing the proportion assigned to each cell type while enforcing a sum-to-one constraint. Finally, once the amount of cells to be picked up from specific cell types is determined, cells can be randomly selected and the pseudo-bulk mixtures can be made by adding up count values from individual cells. Since the proportions used in each mixture are known, they can be compared with the output from the deconvolution (e.g. using Pearson correlation and root-mean square-error (RMSE) values).

2.10) Potential issues with traditional linear modelling

There are four important aspects that need to be taken into account when modelling gene expression data as the weighted sum of gene expression profiles of pure populations:

1) **There should be reference profiles for all populations present in the mixture or at least one marker for each cell type**. This might be problematic for some cell types that cannot be isolated easily (mostly the less abundant ones) and might not have been analysed or sequenced yet. Since reference profiles are assumed to accurately represent the actual cell types present in heterogeneous samples[126], they should be carefully obtained. Moreover, the existence of a sufficient number of marker genes to perform the deconvolution is crucial[64]. Some methods need as little as one marker per cell type[52] but most of them recommend a higher number (5-10) to avoid the potential influence of outliers[60].

2) **Since the true composition is unknown, some cell types may be ignored**. Some methods require precise knowledge of either the constituent cell types[43] or the cell type proportions present in the heterogeneous sample[61] (e.g. assessment from a pathologist or estimated by FACS) for solving the deconvolution problem. However, it is possible that there are no surface markers available yet[143] for sorting unknown populations. Moreover, since the assessment of a pathologist provides information about cell type proportions but not on the amount of mRNA present, the estimates might not be accurate[123]. Even though I have stated that some unsupervised methods take advantage of *a priori* information whenever this is available, other authors are against this practice. For example, Chikina

et al.[116] argues that Coulter counter measurements can have an error >= 5% for lowly abundant cell types, advising not to use them as input for the deconvolution. Furthermore, Gong *et al.*[77] showed that Erkkila's Bayesian approach could not find any solutions when seeded with random estimates (= absence of prior information). Therefore, although *a priori* information can be efficiently exploited (e.g. in a Bayesian framework), the use of incorrect proportion estimates can negatively affect the deconvolution. Finally, an incorrect model specification (e.g. ignoring a cell type that is actually present) might result in incorrect estimates of cell type-specific expression levels for some methods[43,86].

3) **Some methods designed to infer the cell type composition from expression data assume a stable cell type composition within a given heterogeneous tissue[60].** Marker genes are not guaranteed to be expressed at the same levels across different cells[82], even in a tumor from the same patient. Furthermore, the expression profiles are platform-specific, which might result in markers not being present in all platforms and in varying expression levels for a given marker across different platforms[61,144].

Assuming that the expression of a marker gene in one cell type is independent from other cell types present in the mixture is often unrealistic due to potential paracrine signalling effects. This can be tackled by including an extra coefficient in the linear model accounting for the cross-product between different cell types: Kuhn *et al.*[43] excluded all those genes likely to be expressed by a cell type that was not included in the model and Stuart *et al.*[63] observed many transcripts with high cross-product values, suggesting that the expression levels in one cell type are affected by the presence and abundance of other cell types.

4) **The majority of the methods do not take into account the fact that the reference expression profiles are often perturbed by microenvironment or developmental effects or were simply obtained under different conditions or with different technologies or platforms**. To address this issue, PERT[126] estimates a shared perturbation factor across all cell types to account for transcriptional variation between the reference and constituent expression profiles. ISOLATE[145] uses a multinomial model to measure noise in gene expression data and assumes that there is a new population not represented by the available reference profiles. Finally, ISOpure[146] (ISOpureR[137]) is similar to ISOLATE in the estimation of tumor purities and a reference cancer profile but assumes that each healthy profile is the weighted sum of the available healthy tissue profiles and imposes non-negative and sum-to-one constraints.

2.11) Deconvolution methods readily available as webtools

The column "Availability/GUI" on Supplementary Table 1 from Avila Cobos *et al.*[40]) contains detailed information about how to get access to the different reviewed methods. Most of them are accessible as pre-built packages or raw code from different programming languages (e.g. R, Python, Java…). For scientists lacking bioinformatics skills, I highlight seven tools readily accessible for everyone with an internet connection, with little or no bioinformatics background required:

- CellPred[65]: Allows estimation of cell type proportions using Affymetrix microarray data as input. Available at http://webarraydb.org/webarray/index.html (CellPred tab).
- TIMER[66]: A great resource containing the proportions of B cells, CD4+ and CD8+ T cells, macrophages, neutrophils and dendritic cells across 11,509 samples corresponding to 32 cancer types from The Cancer Genome Atlas (TCGA). Available at https://cistrome.shinyapps.io/timer/. Users can download the TIMER method from https://github.com/hanfeisun/TIMER to run it on their own samples.
- DSection[95]: Estimation of cell type-specific expression profiles, corrected cell type proportions and differential gene expression using microarray data. Available at: http://informatics.systemsbiology.net/DSection/
- DCQ[143] and CoD[147] are two tools from the Irit Gat-Viks lab allowing the estimation of cell type quantities to identify disease-relevant cell types using microarray or RNA-seq data. Available at: http://www.dcq.tau.ac.il/ (detailed information: http://dcq.tau.ac.il/application.html) and http://www.csgi.tau.ac.il/CoD/ (detailed information: http://www.csgi.tau.ac.il/CoD/application.html)
- ESTIMATE[56]: Allows quick access to relative stromal and immune cell type composition across all samples available at TCGA (microarray and RNA-seq data).

Available at: http://bioinformatics.mdanderson.org/estimate/

- CIBERSORT[68]: Given microarray or RNA-seq data from heterogeneous samples and selecting pre-built or custom-made matrices with cell type-specific expression profiles, it generates proportions of up to 22 cell types. Available at: https://cibersort.stanford.edu/runcibersort.php

2.12) Alternative data types to perform the deconvolution

Apart from transcriptomics data, other omics data is currently being used as input for the deconvolution problem. Teschendorff and Zheng[140] recently highlighted the impact of cell-type heterogeneity in DNA methylation data and provided a detailed overview of algorithms for correcting cell-type composition in the context of Illumina Infinium Methylation Beadarrays. Titus *et al.*[148] have also recently published a review about cell-type deconvolution from DNA methylation. EpiDISH[149] infers cell-type composition using DNA methylation data and cell-type specific DNase hypersensitive sites. Other tools such as MeDeCom[150] and eFORGE[151] have been designed to estimate cell type-specific signal and account for tumor purity in heterogeneous methylomes. Onuchic *et al.*[152] proposed EDec, a two-step approach in which cell-type proportions in each sample and cell type-specific methylation and gene expression profiles are retrieved. Importantly, as Teschendorff and Zheng[140] pointed out, a direct comparison between expression-based and DNA methylation-based cell type composition estimates has not been performed yet.

Several methods have been proposed to detect copy number aberrations from DNA profiling of heterogeneous samples: BACOM 2.0[153], ABSOLUTE[154] and CloneCNA[155]. Finally, Aran *et al.*[156] created the Consensus measurement of Purity Estimation (CPE), a robust value for tumor purity obtained from combining gene expression, somatic copy number, methylation and immunohistochemistry data that they successfully applied to more than 10,000 samples from The Cancer Genome Atlas (TCGA).

2.13) Applications of computational deconvolution in cancer research

It has already been shown that taking tumor heterogeneity into account led to an increase in the sensitivity of relapse-free survival analyses and more accurate tumor subtype predictions. Specifically, Elloumi *et al.* [157] showed that there was an under-estimation of patient risk induced by the non-tumorous portion present in breast tumor samples and relapse-free survival was more sensitive after accounting for such non-tumorous proportion. A link has been shown between the spontaneous regression of the paediatric cancer neuroblastoma and the participation of tumor-infiltrating lymphocytes, anti-neuroblastoma antibodies or natural killer cells[158], highlighting the relevance of retrieving the proportions of those cell types in the tumor samples.

Computational deconvolution has been successfully used in immune-oncology to infer cell type proportions which have been linked to relevant clinical parameters. Cancer and immune cell content in tumor tissue has already been assessed through computational

deconvolution[92,159] and the presence of tumor-infiltrating lymphocytes (TILs) and other immune cells in the tumor microenvironment is currently a very active field of research[160–162] (e.g. in the context of immunotherapy). For instance, Şenbabaoğlu et al.[57] used computational deconvolution to obtain a T-cell infiltration score and an overall immune infiltration score across 19 different malignancies, where clear cell renal cell carcinoma tumors showed the highest T-cell infiltration. Li et al. [66] analyzed tumor-infiltrating cells over 10,000 RNA-seq samples across 23 cancer types from The Cancer Genome Atlas (TCGA) and showed that immune infiltrates obtained through computational deconvolution were strongly associated to patient clinical features. Specifically, they found the infiltration levels of B-cells to significantly predict patient survival in glioblastoma multiforme, lung adenocarcinoma and bladder carcinoma and CD8+ T-cell infiltration to predict tumor relapse in skin cutaneous melanoma, colon adenocarcinoma, rectum adenocarcinoma and cervical squamous carcinoma.

Importantly, there are other applications of the computational deconvolution oriented towards the retrieval of tumor-specific expression profiles or the development of sensitive differential gene expression frameworks that account for sample heterogeneity.

As stated in section 2.1, tumor samples have varying proportions of non-malignant cells. ISOpure is able to account for differences in tumor purity and generates, for each tumor sample, the "purified" expression profile coming only from the tumor fraction. Using the purified version of the tumors versus the non-purified counterparts showed led to improved prognostic signatures for lung adenocarcinoma and an increase of 8% in prediction accuracy when predicting extra-prostatic extension in prostate tumors[163], which is a strong predictor for recurrence.

Next, csSAM[122] and contamDE[164] are two differential gene expression analysis frameworks that account for sample heterogeneity and were successfully used with microarray and RNA-seq data, respectively. Using whole-blood expression data from two groups of kidney transplant recipients ("stable patients post-transplant" versus "experiencing acute transplant rejection"), cell type–specific significance analysis of microarrays (csSAM) was compared against significance analysis of microarrays (SAM)[165], a traditional differential gene expression between groups that do not account for cell type heterogeneity. While SAM was unable to find a single gene as differentially expressed using a permissive false discovery rate (FDR <= 0.3), csSAM detected more than 300 differentially expressed genes. Using RNA-seq data from 14 patients with matching prostate cancer and normal adjacent tissue, contamDE was tested against edgeR and DESeq2. contamDE accounted

for differences in tumor purity across the prostate cancer samples and outperformed the other two in terms of power and false discovery rates. Moreover, contamDE uniquely identified 85 additional differentially expressed genes that in turn were found to be associated with biological functions directly involved in tumor progression.

In conclusion, failing to account for sample heterogeneity hampers the identification of biologically relevant differentially expressed genes and potential candidate biomarkers for specific cancer types might go unnoticed.

REFERENCES

1. Wissink, E. M., Vihervaara, A., Tippens, N. D. & Lis, J. T. Nascent RNA analyses: tracking transcription and its regulation. Nat. Rev. Genet. 20, 705–723 (2019).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63 (2009).
3. Clancy, S. & Brown, W. Translation: DNA to mRNA to Protein. Nat. Educ. 1, 101 (2008).
4. Clancy, S. RNA transcription by RNA polymerase: prokaryotes vs eukaryotes. Nature 1, 125 (2008).
5. Clancy, S. RNA splicing: introns, exons and spliceosome. Nat. Educ. 1, 31 (2008).
6. Hombach, S. & Kretz, M. Non-coding RNAs: Classification, Biology and Functioning. in Non-coding RNAs in Colorectal Cancer (eds. Slaby, O. & Calin, G. A.) 3–17 (Springer International Publishing, 2016). doi:10.1007/978-3-319-42059-2_1.
7. Maxwell, E. & Fournier, M. THE SMALL NUCLEOLAR RNAs. Annu. Rev. Biochem. 64, 897–934 (1995).
8. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. RNA Biol. 10, 924–933 (2013).
9. Gutschner, T., Hämmerle, M. & Diederichs, S. MALAT1 — a paradigm for long noncoding RNA function in cancer. J. Mol. Med. 91, 791–801 (2013).
10. Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076 (2010).
11. Liu, L. et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat. Commun. 10, 1–11 (2019).
12. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int. J. Mol. Sci. 18, (2017).
13. Maloney, C., Sequeira, E., Kelly, C., Orris, R. & Beck, J. PubMed Central. (National Center for Biotechnology Information (US), 2013).
14. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE 9, e78644 (2014).
15. Sellin Jeffries, M. K., Kiss, A. J., Smith, A. W. & Oris, J. T. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. BMC Biotechnol. 14, (2014).
16. Abdueva, D., Wing, M., Schaub, B., Triche, T. & Davicioni, E. Quantitative Expression Profiling in Formalin-Fixed Paraffin-Embedded Samples by Affymetrix Microarrays. J. Mol. Diagn. 12, 409–417 (2010).
17. Esteve-Codina, A. et al. A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. PLOS ONE 12, e0170632 (2017).
18. Zucha, D., Androvic, P., Kubista, M. & Valihrach, L. Performance Comparison of Reverse Transcriptases for Single-Cell Studies. Clin. Chem. (2019) doi:10.1373/clinchem.2019.307835.
19. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotech 32, 903–914 (2014).
20. Hölzer, M. & Marz, M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 8, (2019).

21. Church, D. M. et al. Extending reference assembly models. Genome Biol. 16, 13 (2015).
22. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? Genome Biol. 20, 159 (2019).
23. Consortium, M. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat. Biotechnol. 28, 827–838 (2010).
24. Simoneau, J., Dumontier, S., Gosselin, R. & Scott, M. S. Current RNA-seq methodology reporting limits reproducibility. Brief. Bioinform. doi:10.1093/bib/bbz124.
25. Baker, M. 1,500 scientists lift the lid on reproducibility. Nat. News 533, 452 (2016).
26. Pertea, M. et al. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. bioRxiv 332825 (2018) doi:10.1101/332825.
27. Jiang, S. et al. An expanded landscape of human long noncoding RNA. Nucleic Acids Res. 47, 7842–7856 (2019).
28. Lorenzi, L. et al. The RNA Atlas, a single nucleotide resolution map of the human transcriptome. bioRxiv 807529 (2019) doi:10.1101/807529.
29. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25, 1915–1927 (2011).
30. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–1789 (2012).
31. Ziegler, C. & Kretz, M. The More the Merrier—Complexity in Long Non-Coding RNA Loci. Front. Endocrinol. 8, (2017).
32. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
33. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46 (2012).
34. Weirick, T. et al. The identification and characterization of novel transcripts from RNA-seq data. Brief. Bioinform. 17, 678–685 (2016).
35. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22, 1–5 (2006).
36. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, e74–e74 (2013).
37. Jiang, S. & Mortazavi, A. Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics 17, 104–115 (2018).
38. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. 2010, pdb.prot5384 (2010).
39. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
40. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics 34, 1969–1979 (2018).
41. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. Nat. Rev. Cancer 12, 298–306 (2012).
42. Egeblad, M., Nakasone, E. S. & Werb, Z. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. Dev. Cell 18, 884–901 (2010).

43. Kuhn, A. et al. Cell population-specific expression analysis of human cerebellum. BMC Genomics 13, 610 (2012).
44. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat. Commun. 10, 1–11 (2019).
45. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 20, 194 (2019).
46. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. Nat. Biotechnol. 37, 547–554 (2019).
47. Dal Molin, A., Baruzzo, G. & Di Camillo, B. Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. Front. Genet. 8, (2017).
48. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. Nat. Commun. 11, 1–14 (2020).
49. Shen-Orr, S. S. & Gaujoux, R. Computational Deconvolution: Extracting Cell Type-Specific Information from Heterogeneous Samples. Curr. Opin. Immunol. 25, (2013).
50. Yadav, V. K. & De, S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Brief. Bioinform. 16, 232–241 (2015).
51. Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc. IEEE 105, 340–366 (2017).
52. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. Bioinformatics 17, S279–S287 (2001).
53. Cherry, E. C. Some Experiments on the Recognition of Speech, with One and with Two Ears. J. Acoust. Soc. Am. 25, 975–979 (1953).
54. Bronkhorst, A. W. The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten. Percept. Psychophys. 77, 1465 (2015).
55. Gorodentsev, A. L. Algebra I: Textbook for Students of Mathematics. (Springer, 2016).
56. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. 4, ncomms3612 (2013).
57. Şenbabaoğlu, Y. et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol. 17, 231 (2016).
58. Ghosh, D. Mixture models for assessing differential expression in complex tissues using microarray data. Bioinforma. Oxf. Engl. 20, 1663–1669 (2004).
59. Lähdesmäki, H., Shmulevich, L., Dunmire, V., Yli-Harja, O. & Zhang, W. In silico microdissection of microarray data from heterogeneous cell populations. BMC Bioinformatics 6, 54 (2005).
60. Ahn, J. et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinforma. Oxf. Engl. 29, 1865–1871 (2013).
61. Li, Y. & Xie, X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. BMC Bioinformatics 14 Suppl 5, S11 (2013).
62. Wang, N. et al. UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. Bioinforma. Oxf. Engl. 31, 137–139 (2015).

63. Stuart, R. O. et al. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc. Natl. Acad. Sci. U. S. A. 101, 615–620 (2004).
64. Hoffmann, M. et al. Robust computational reconstitution - a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. BMC Bioinformatics 7, 369 (2006).
65. Wang, Y. et al. In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res. 70, 6448–6455 (2010).
66. Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 17, 174 (2016).
67. Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 17, 218 (2016).
68. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015).
69. Chen, Z. et al. Inference of immune cell composition on the expression profiles of mouse tissue. Sci. Rep. 7, 40508 (2017).
70. Berkson, J. Estimation by Least Squares and by Maximum Likelihood. in (The Regents of the University of California, 1956).
71. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proc. Natl. Acad. Sci. 100, 10370–10375 (2003).
72. Wang, M., Master, S. R. & Chodosh, L. A. Computational expression deconvolution in a complex mammalian organ. BMC Bioinformatics 7, 328 (2006).
73. Belzer, J., Holzman, A. G. & Kent, A. Encyclopedia of Computer Science and Technology: Volume 11 - Minicomputers to PASCAL. (CRC Press, 1979).
74. Hu, H. Positive definite constrained least-squares estimation of matrices. Linear Algebra Its Appl. 229, 167–174 (1995).
75. Weingessel, S. original by B. A. T. R. port by A. quadprog: Functions to solve Quadratic Programming Problems. (2013).
76. Quadratic Programming Algorithms - MATLAB. https://nl.mathworks.com/help/optim/ug/quadratic-programming-algorithms.html#brnox76.
77. Gong, T. et al. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PloS One 6, e27156 (2011).
78. Solve constrained linear least-squares problems - MATLAB lsqlin. https://nl.mathworks.com/help/optim/ug/lsqlin.html?s_tid=gn_loc_drop.
79. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS One 4, e6098 (2009).
80. Shannon, C. P. et al. Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. PloS One 9, e95224 (2014).
81. Soetaert, K., Meersche, K. V. den, Oevelen, D. van & authors, L. limSolve: Solving Linear Inverse Models. (2017).
82. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics 14, 89 (2013).
83. R: The R Project for Statistical Computing. https://www.r-project.org/.

84. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126 (1994).
85. Repsilber, D. et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics 11, 27 (2010).
86. Zuckerman, N. S., Noam, Y., Goldsmith, A. J. & Lee, P. P. A self-directed method for cell-type identification and separation of gene expression microarrays. PLoS Comput. Biol. 9, e1003189 (2013).
87. Wang, N. et al. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. Sci. Rep. 6, 18909 (2016).
88. Stokkum, K. M. M. and I. H. M. van. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). (2012).
89. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLOS Comput. Biol. 15, e1006976 (2019).
90. Ripley, B. et al. MASS: Support Functions and Datasets for Venables and Ripley's MASS. (2002).
91. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–22 (2010).
92. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife 6, e26476 (2017).
93. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinforma. Oxf. Engl. 29, 1083–1085 (2013).
94. Roy, S., Lane, T., Allen, C., Aragon, A. D. & Werner-Washburne, M. A hidden-state Markov model for cell population deconvolution. J. Comput. Biol. J. Comput. Mol. Cell Biol. 13, 1749–1774 (2006).
95. Erkkilä, T. et al. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinforma. Oxf. Engl. 26, 2571–2577 (2010).
96. Kang, K. et al. CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. PLOS Comput. Biol. 15, e1007510 (2019).
97. Kassambara, A. et al. GenomicScape: An Easy-to-Use Web Tool for Gene Expression Data Analysis. Application to Investigate the Molecular Events in the Differentiation of B Cells into Plasma Cells. PLOS Comput. Biol. 11, e1004077 (2015).
98. Lenz, M., Müller, F.-J., Zenke, M. & Schuppert, A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Sci. Rep. 6, srep25696 (2016).
99. Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput. Stat. Data Anal. 49, 974–997 (2005).
100. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430 (2000).
101. Sompairac, N. et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int. J. Mol. Sci. 20, 4414 (2019).

102. Nazarov, P. V. et al. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. BMC Med. Genomics 12, 132 (2019).
103. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 12, 913–921 (2012).
104. Moffitt, R. A. et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat. Genet. 47, 1168–1178 (2015).
105. Boutsidis, C. & Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognit. 41, 1350–1362 (2008).
106. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52, 155–173 (2007).
107. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat. Commun. 10, 2209 (2019).
108. Du, R., Carey, V. & Weiss, S. T. deconvSeq: deconvolution of cell mixture distribution in sequencing data. Bioinformatics doi:10.1093/bioinformatics/btz444.
109. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).
110. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. 10, 1–9 (2019).
111. Tsoucas, D. et al. Accurate estimation of cell-type composition from gene expression data. Nat. Commun. 10, 1–9 (2019).
112. Dong, M. et al. SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. bioRxiv 743591 (2019) doi:10.1101/743591.
113. Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. bioRxiv 669911 (2019) doi:10.1101/669911.
114. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 16, 278 (2015).
115. Shoemaker, J. E. et al. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. BMC Genomics 13, 460 (2012).
116. Chikina, M., Zaslavsky, E. & Sealfon, S. C. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinforma. Oxf. Engl. 31, 1584–1591 (2015).
117. Reinartz, S. et al. A transcriptome-based global map of signaling pathways in the ovarian cancer microenvironment associated with clinical outcome. Genome Biol. 17, 108 (2016).
118. Zhang, J. D. et al. Detect tissue heterogeneity in gene expression data with BioQC. BMC Genomics 18, 277 (2017).
119. Zinovyev, A., Kairov, U., Karpenyuk, T. & Ramanculov, E. Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem. Biophys. Res. Commun. 430, 1182–1187 (2013).
120. Nelms, B. D. et al. CellMapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types. Genome Biol. 17, 201 (2016).
121. Ju, W. et al. Defining cell-type specificity at the transcriptional level in human disease. Genome Res. 23, 1862–1873 (2013).

122. Shen-Orr, S. S. et al. Cell type-specific gene expression differences in complex tissues. Nat. Methods 7, 287–289 (2010).
123. Clarke, J., Seo, P. & Clarke, B. Statistical expression deconvolution from mixed tissue samples. Bioinforma. Oxf. Engl. 26, 1043–1049 (2010).
124. Newman, A. M., Gentles, A. J., Liu, C. L., Diehn, M. & Alizadeh, A. A. Data normalization considerations for digital tumor dissection. Genome Biol. 18, 128 (2017).
125. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat. Med. 21, 938–945 (2015).
126. Qiao, W. et al. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput. Biol. 8, e1002838 (2012).
127. Liebner, D. A., Huang, K. & Parvin, J. D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinforma. Oxf. Engl. 30, 682–689 (2014).
128. Irizarry, R. A., Wu, Z. & Jaffee, H. A. Comparison of Affymetrix GeneChip expression measures. Bioinformatics 22, 789–794 (2006).
129. Hoyle, D. C., Rattray, M., Jupp, R. & Brass, A. Making sense of microarray data distributions. Bioinformatics 18, 576–584 (2002).
130. Tsai, C.-A., Chen, Y.-J. & Chen, J. J. Testing for differentially expressed genes with microarray data. Nucleic Acids Res. 31, e52 (2003).
131. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. Nat. Methods 9, 8–9 (2012).
132. Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L. M. & Luthi-Carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat. Methods 8, 945–947 (2011).
133. Jin, H., Wan, Y.-W. & Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. BMC Bioinformatics 18, 117 (2017).
134. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).
135. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).
136. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).
137. Anghel, C. V. et al. ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. BMC Bioinformatics 16, 156 (2015).
138. Fang, Q. A note on the condition number of a matrix. J. Comput. Appl. Math. 157, 231–234 (2003).
139. Glass, E. R. & Dozmorov, M. G. Improving sensitivity of linear regression-based cell type-specific differential expression deconvolution with per-gene vs. global significance threshold. BMC Bioinformatics 17, 334 (2016).
140. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics 9, 757–768 (2017).
141. Bar-Joseph, Z. et al. Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. Proc. Natl. Acad. Sci. 105, 955–960 (2008).

142. Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife 6, e26476 (2017).
143. Altboum, Z. et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol. Syst. Biol. 10, 720 (2014).
144. Shannon, C. P. et al. Enumerateblood - an R package to estimate the cellular composition of whole blood from Affymetrix Gene ST gene expression profiles. BMC Genomics 18, 43 (2017).
145. Quon, G. & Morris, Q. ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. Bioinforma. Oxf. Engl. 25, 2882–2889 (2009).
146. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. 5, 29 (2013).
147. Frishberg, A., Steuerman, Y. & Gat-Viks, I. CoD: inferring immune-cell quantities related to disease states. Bioinforma. Oxf. Engl. 31, 3961–3969 (2015).
148. Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. Cell-type deconvolution from DNA methylation: a review of recent applications. Hum. Mol. Genet. 26, R216–R224 (2017).
149. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics 18, (2017).
150. Lutsik, P. et al. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. 18, 55 (2017).
151. Breeze, C. E. et al. eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. Cell Rep. 17, 2137–2150 (2016).
152. Onuchic, V. et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. Cell Rep. 17, 2075–2086 (2016).
153. Fu, Y. et al. BACOM2.0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. Sci. Rep. 5, srep13955 (2015).
154. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 30, 413–421 (2012).
155. Yu, Z., Li, A. & Wang, M. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. BMC Bioinformatics 17, 310 (2016).
156. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. Nat. Commun. 6, ncomms9971 (2015).
157. Elloumi, F. et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med. Genomics 4, 54 (2011).
158. Brodeur, G. M. Spontaneous regression of neuroblastoma. Cell Tissue Res. 372, 277–286 (2018).
159. Schelker, M. et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat. Commun. 8, 2032 (2017).
160. Sharma, A. et al. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. bioRxiv 698845 (2019) doi:10.1101/698845.
161. Hendry, S. et al. Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group. Adv. Anat. Pathol. 24, 235–251 (2017).

162. Research, A. A. for C. Low-Heterogeneity Melanomas Are More Immunogenic and Less Aggressive. Cancer Discov. (2019) doi:10.1158/2159-8290.CD-RW2019-144.
163. Stephenson, A. J. et al. Postoperative Nomogram Predicting the 10-Year Probability of Prostate Cancer Recurrence After Radical Prostatectomy. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 23, 7005–7012 (2005).
164. Shen, Q. et al. contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples. Bioinforma. Oxf. Engl. 32, 705–712 (2016).
165. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. 98, 5116–5121 (2001).

Part II - Research Objectives

The analysis of the transcriptome has substantially contributed to our knowledge of the processes involved in development and disease. Moreover, since the introduction of the RNA-sequencing technology (RNA-seq) a decade ago, research has shown that up to three quarters of the human genome can be transcribed. Non-coding RNAs represent the majority of this transcriptome and reconstructing accurate transcript models for these is a complex endeavour when processing RNA-seq data.

Furthermore, the heterogeneous nature of samples typically used in research (e.g. cancer samples or tissues) has been largely overlooked. Gene expression analyses of bulk tissues often neglect cell type composition as an important confounding factor in downstream analyses, masking the signal coming from lowly abundant cell types. Therefore, many computational approaches have been developed to infer cell type proportions and/or cell type-specific expression profiles in heterogeneous samples (computational deconvolution).

Aim 1 - Development of a new bioinformatics tool: the Zipper plot

Discriminating between truly independent transcriptional units and untranslated regions of an upstream protein coding gene or DNA contamination is far from straightforward. Many functional experiments aiming at discovering the function of novel long non-coding RNAs (lncRNAs) occasionally fail even when their expression values are highly correlated with a protein coding gene in the vicinity. Most of the time, this is due to the absence of marks indicative of putative transcription.

For this reason, I developed the Zipper plot (**Avila Cobos *et al.*, 2017. BMC Bioinformatics**), a novel visualization and analysis method that enables users to simultaneously interrogate thousands of human putative transcription start sites (TSSs) in relation to various features that are indicative for transcriptional activity: CAGE-sequencing, ChIP-sequencing and DNase-sequencing.

Aim 2 – Use the Zipper plot tool to refine the human transcriptome

The human transcriptome assembly from the RNA Atlas project (Lorenzi *et al.*, 2019) resulted in more than 40,000 genes, including thousands of novel lncRNAs, microRNAs, circular RNAs and protein-coding genes. However, a portion of these novel genes were likely to be false positives. I have refined and filtered these new transcript models with the Zipper plot, by integrating complementary datasets on chromatin states associated with transcription or with enhancer activity from the Roadmap Epigenomics project) and transcript boundaries (i.e. CAGE-seq to mark the TSS).

Aim 3 - Assessing the biological signal of different RNA biotypes through computational deconvolution of healthy tissues

Multiple computational deconvolution approaches have been developed to infer cell type proportions in heterogeneous samples. However, current methods have been only used with messenger RNAs (mRNAs) as input. Using expression data across 162 normal cell types and 45 tissues from the RNA Atlas project, I investigated the performance of additional RNA fractions in the computational deconvolution of healthy tissues **(Avila Cobos *et al.*, 2019. In preparation)**.

Aim 4 – Comprehensive benchmarking of computational deconvolution methods

The deconvolution of the RNA Atlas was assessed with only one choice of data transformation, normalization and deconvolution method. I extensively reviewed many deconvolution methods developed since 2001 (**Avila Cobos *et al.*, 2018. Bioinformatics**) and performed a quantitative evaluation of the combined impact of data transformation, scaling/normalization, marker selection, cell type composition and choice of methodology on the deconvolution results **(Avila Cobos *et al.*, 2019. Under review).**

Part III – Results

Paper 1

Zipper plot: visualizing transcriptional activity of genomic regions

Francisco Avila Cobos, Jasper Anckaert, Pieter-Jan Volders, Celine Everaert, Dries Rombaut, Jo Vandesompele, Katleen De Preter* and Pieter Mestdagh*

(*) These authors contributed equally to this work.

Contribution: Conceptualization of the manuscript; wrote all the R code; developed the visualization method with PJV; validation of the method; made all the figures and wrote the manuscript.

Published in BMC Bioinformatics (2017) 18:231
DOI 10.1186/s12859-017-1651-7

Zipper plot: visualizing transcriptional activity of genomic regions

Francisco Avila Cobos[1,2,3], Jasper Anckaert[1,2,3], Pieter-Jan Volders[1,2,3], Celine Everaert[1,2,3], Dries Rombaut[1,2,3], Jo Vandesompele[1,2,3], Katleen De Preter[1,2,3,†] and Pieter Mestdagh[1,2,3,†]

[†] Equal contribution. [1] Center for Medical Genetics, Ghent University, De Pintelaan 185, Ghent, Belgium. [2] Cancer Research Institute Ghent, De Pintelaan 185, Ghent, Belgium. [3] Bioinformatics Institute Ghent from Nucleotides to Networks, De Pintelaan 185, Ghent, Belgium. Corresponding author: Pieter.Mestdagh@UGent.be

Abstract

Reconstructing transcript models from RNA-sequencing (RNA-seq) data and establishing these as independent transcriptional units can be a challenging task. Current state-of-the-art tools for long non-coding RNA (lncRNA) annotation are mainly based on evolutionary constraints, which may result in false negatives due to the overall limited conservation of lncRNAs.

To tackle this problem we have developed the Zipper plot, a novel visualization and analysis method that enables users to simultaneously interrogate thousands of human putative transcription start sites (TSSs) in relation to various features that are indicative for transcriptional activity. These include publicly available CAGE-sequencing, ChIP-sequencing and DNase-sequencing datasets. Our method only requires three tab-separated fields (chromosome, genomic coordinate of the TSS and strand) as input and generates a report that includes a detailed summary table, a Zipper plot and several statistics derived from this plot.

Using the Zipper plot, we found evidence of transcription for a set of well-characterized lncRNAs and observed that fewer mono-exonic lncRNAs have CAGE peaks overlapping with their TSSs compared to multi-exonic lncRNAs. Using publicly available RNA-seq data, we found more than one hundred cases where junction reads connected protein-coding gene exons with a downstream mono-exonic lncRNA, revealing the need for a careful evaluation of lncRNA 5'-boundaries. Our method is implemented using the statistical programming language R and is freely available as a webtool.

Background

The introduction of RNA-sequencing (RNA-seq) has revolutionized the field of molecular biology, revealing that up to 75% of the human genome is actively transcribed [1]. The majority of this transcriptome consists of so-called long non-coding RNAs (lncRNAs). Reconstructing accurate transcript models for these lncRNAs is a major challenge when processing RNA-seq data. In general, lncRNA transcripts are less abundant compared to protein coding genes [2], often resulting in a lack of junction reads from which transcript models are inferred. In addition, lncRNAs are frequently located in the vicinity of protein coding genes and could therefore represent unannotated extensions of untranslated regions (UTRs) rather than independent transcriptional units. Finally, transcript reconstruction from RNA-seq data often gives rise to large numbers of single-exon transcripts. Distinguishing single-exon fragments that represent independent transcriptional units from those that result from genomic DNA contamination or incomplete transcript assembly is not straightforward.

State-of-the-art tools for lncRNA annotation based on evolutionary constraints such as PLAR (pipeline for lncRNA annotation from RNA-seq data) [3] and slncky [4], might filter out some putative lncRNA transcripts depending on stringent conservation criteria. PLAR removes transcripts that are short (< 2 kb) and lowly expressed (FPKM < 5) and focuses on the annotation of syntenic lncRNAs. Given the limited conservation of lncRNAs [5] and given that both tools exclude any transcript that partially or totally overlaps protein-coding genes, such approaches may result in a large number of false negatives.

LncRNA transcript models can be refined and filtered by integrating complementary datasets on chromatin state (i.e. ChIP sequencing (ChIP-seq) for histone marks or DNase sequencing (DNase-seq)) and transcript boundaries (i.e. CAGE sequencing (CAGE-seq) to mark the transcription start site (TSS) or 3P-seq to mark the 3' end of poly-adenylated transcripts)[6]. Transcripts for which the transcription start site coincides with a CAGE-peak and is in close proximity to a H3K4me3 or H3K27ac mark are more likely to be independent transcriptional units compared to transcripts that lack these features.

GRIT [7] is a command line-based tool that uses CAGE in conjunction with RNA-seq data but does not take advantage of other important layers of genomic information such as open chromatin (DNase-seq) and histone marks (ChIP-seq data) typically associated with active transcription.

To tackle the challenge of establishing lncRNAs as independent transcriptional units we have created the Zipper plot, a novel visualization and analysis method available as a quick and user-friendly webtool [8] that employs publicly available CAGE-seq, ChIP-seq and DNase-seq data across a large collection of tissue and cell types. The user only needs to provide a list of genomic features (one per line), each consisting of three tab-separated fields: chromosome, human genomic coordinate (hg19) of the TSS and strand. Our webtool will retrieve the closest CAGE-seq/DNase-seq/ChIP-seq peak to each TSS for thousands of genomic features at the same time. The closer these peaks are, the higher the evidence of independent transcriptional activity for the set of genomic features.

Results and discussion

Implementing the Zipper plot as a webtool

The Zipper plot is freely available as a webtool (front-end) at [8] and has been implemented using the JavaScript library jQuery, PHP and HTML5. The back-end (server) contains a peak-based database (see Methods) and the necessary code to retrieve and sort the closest CAGE-seq/ChIP-seq/DNase-seq peak to each TSS, to create the plot (see "Zipper plot construction") and to compute several statistics to assess the TSS-peak associations (see "Summary statistics and generation of html summary reports"). This code was written using the R statistical programming language [9] along with the data.table[10], ggplot2[11], knitr[12], R.utils[13], grid [9] and gridExtra [14] packages. The communication between the web interface and our server is established using PHP.

Due to memory constraints on our server, we limited the number of genomic features per input file to 20,000. However, to allow users to integrate our tool as part of bigger pipelines, we have made our scripts available at Github [15].

Database querying

To start using the webtool, the user only needs to upload a list of genomic features (one per line), each consisting of three tab-separated fields: chromosome, human genomic coordinate (hg19) of the TSS and strand. Optionally, users can provide an additional fourth column containing labels for the genomic features being studied.

If the user has a file from another genomic build (e.g. hg38), we propose two alternatives to convert it to hg19: 1) hgLiftOver [16]: a webtool where users can upload a file with "chrN:start-end" or BED format and select the new genomic build of interest; 2) CrossMap

[17]: a tool that supports more file types as input, including BAM, SAM and BigWig among others. Detailed information about its usage and download can be found at [18].

Importantly, hgLiftOver can also be installed locally on unix-based systems by downloading the executable [19] and appropriate chain files [20].

In a second step, the user has to select the data type of interest among the ones available in our database (CAGE-seq, ChIP-seq or DNase-seq peaks; see Methods) and has the option to run the analysis in one sample type of interest or across all available sample types. In the first option, the user knows in advance in which tissue the set of genomic features are more likely to be expressed; with the second option, each individual genomic feature is analyzed across all samples and the sample in which the peak is most closely associated to the genomic feature is retained for further analysis. Importantly, all CAGE-peaks are used by default but the user can set a more stringent threshold if desired (tags per million mapped reads (tpm) > 0). A detailed user guide can be found at [21].

Zipper plot construction

Once the user's input is uploaded to our website and the data type of interest has been selected, the data.table package [10] is used to sort TSSs from the user's input in a chromosome-wise manner and to perform a fast binary search (O(log n) time) in compiled C to retrieve the closest ChIP-seq/DNase-seq/CAGE-seq peak to each TSS. It retrieves the "start" and "end" genomic coordinates of the closest peak, always considering the "start" as the part of the peak closest to the TSS. The supplementary methods ("Definition of the distance between a TSS and the closest peak" section) contain three different examples on how these coordinates are determined.

The peaks are then ranked based on the distance from the TSS to the "start" of the closest peak and a Zipper plot is generated with the aid of the ggplot2 package [11]: peaks overlapping with the TSS are placed at the top of the plot and the zipper starts to open as the peaks are located further away from the TSSs. By default, the Zipper plot is visualized in a +/- 5 kilobase (kb) window around the TSS but the window size can be adjusted by the user. Fig 1 shows in detail how the Zipper plot is built.

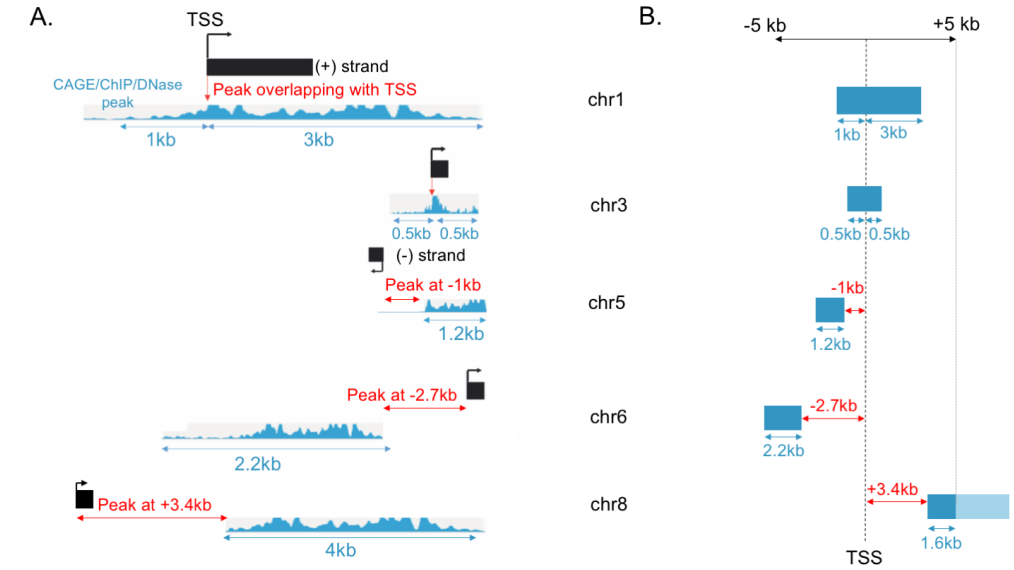


Fig 1. The closest CAGE-seq/ChIP-seq/DNase-seq peak to each TSS is rapidly retrieved using a binary search. A) The process of finding the closest CAGE peak takes into account the strand information supplied by the user (ChIP-seq and DNase-seq data are unstranded). If a TSS is located on the positive DNA strand (TSSs on chromosomes 1, 3, 6 and 8), peaks with a genomic coordinate greater than the TSS are considered downstream (=positive distance) of the genomic feature. If a TSS is located on the negative DNA strand (third TSS on chromosome 5), peaks with a genomic coordinate greater than the TSS are considered upstream (=negative distance) of the genomic feature. Peak widths and overall peak enrichment for each region (signalValue for ChIP-seq and DNase-seq data; tpm expression values for CAGE-seq) are simultaneously retrieved. B) Once the distances to the closest peaks have been retrieved they are ordered and placed on top of a vertical axis representing the TSS. Since the Zipper plot is visualized (by default) in a 5kb window, peaks that are wider than 5kb or are further away from the TSS will not appear (i.e. TSS on chromosome 8; darker region will appear whereas the faded region exists but it is not displayed).

Summary statistics and generation of html summary reports

In parallel with the construction of every Zipper plot, two statistics named Zipper Height (ZH) and Area Under the Zipper (AUZ) are calculated. ZH corresponds to the quotient between the number of genomic features with a peak overlapping with the TSS and the total number of genomic features being studied (ZH $\in$ [0,1]). The AUZ_global is computed as the sum of all the areas between the closest peak and the TSS of each genomic feature (see "Definition of the sum of all areas between the closest peak and the TSS" and "Small AUZ values, areas in the plot and how AUZ_window is calculated (Fig 3D)" in the supplementary methods for detailed explanation).

However, since the distribution of peaks upstream or downstream of the TSSs can be asymmetric, AUZleft (sum of all the areas for cases where the closest peak was found

upstream the TSS) and AUZright (sum of all the areas for cases where the closest peak was found downstream the TSS) are considered independently (see "Rationale for calculating both positive and negative distances between closest peaks and TSSs" in supplementary methods for more details).

The closer the peaks are distributed around the TSSs, the smaller the AUZ and the higher the evidence of independent transcriptional activity for the set of genomic features. A "closed zipper" (AUZ=0) indicates an overlap between the closest peak and TSS for all the genomic features being studied. We have also incorporated the AUZ_window, which depends on the window size choice (by default +/- 5kb) and is computed using only those peaks that lie within the window. The method virtually sets to 5kb (or other value if the user changes the default window size) all those distances that are located more than 5kb away from the TSS. This allows a quick visual comparison between two Zipper plots built using the same window size. Following the same reasoning as the paragraph above, we have incorporated both AUZ_window_right and AUZ_window_left separately. Of note, ZH and AUZ are negatively correlated.

A one-sided p-value (AUZ_pval) is calculated by comparing the AUZ of the Zipper plot built with the user's input to 100 (by default) or 1000 random Zipper plots created by selecting as many random locations as the number of genomic features supplied by the user while maintaining the same distribution of TSSs per chromosome. Since truly random locations picked uniformly along the length of each chromosome are not representative of possible lncRNA TSSs, we have excluded from the selection those genomic regions containing gaps, centromeres, telomeres, heterochromatin and repetitive regions from [22,23] using the BEDTools suite [24]. The p-value is computed dividing the number of random cases with AUZ values smaller than or equal to the AUZ for the user case by the total number of repetitions. The p-value represents the chance of finding a random Zipper plot with an AUZ_global smaller than or equal to the AUZ_global of the actual use case or, in other words, whether it is likely that the set of TSSs chosen by the user was randomly selected or not. Therefore, the smaller the p-value, the higher the likelihood your set of genomic features are truly independent transcriptional units.

When evaluating genomic features in one sample type, the closest peaks in that sample type are retrieved for both the random TSSs and the user input. Optionally, the closest peak in each sample type can be retrieved for each TSS and, for each TSS, a TSS p-value is calculated comparing how many tissues have a peak as close (or closer) to the TSS than the one found in the tissue chosen by the user.

On the other hand, if the user selects all sample types, the closest peaks among all possible sample types are retrieved for both the random TSS and the user input. AUZs are calculated and a p-value is calculated similarly to the case where the user selects one sample type.

Eventually, the knitr package [12] is used to generate an html report containing 1) the Zipper plot; 2) all the aforementioned parameters/statistics; 3) a summary table listing closest peaks, peak widths and overall peak enrichment information.

Validation and applications of the Zipper plot

To assess the usefulness of our webtool, we first investigated a set of 36 well-characterized lncRNAs proposed by [4]. The Zipper plot created using only the FANTOM5 (CAGE-seq) data showed that 26 out of 36 lncRNAs have a CAGE peak within +/- 5kb from their TSSs in at least one of the sample types present in our database (Fig 2A; detailed output available in S1 Table). Moreover, when also including H3K4me3 and DNaseI (marks for active transcription and open chromatin) together with H3K4me1 and H3K27ac (marks for active enhancer RNAs), 32 out of 36 lncRNAs have peaks within +/5 kb from their TSSs (Fig 2B). These results demonstrate that, while most of the well-characterized lncRNAs have evidence for transcription initiation at or near their presumed TSS, some may be incompletely annotated with respect to their TSS. This is especially apparent from the CAGE-seq Zipper plot (Fig 2A).

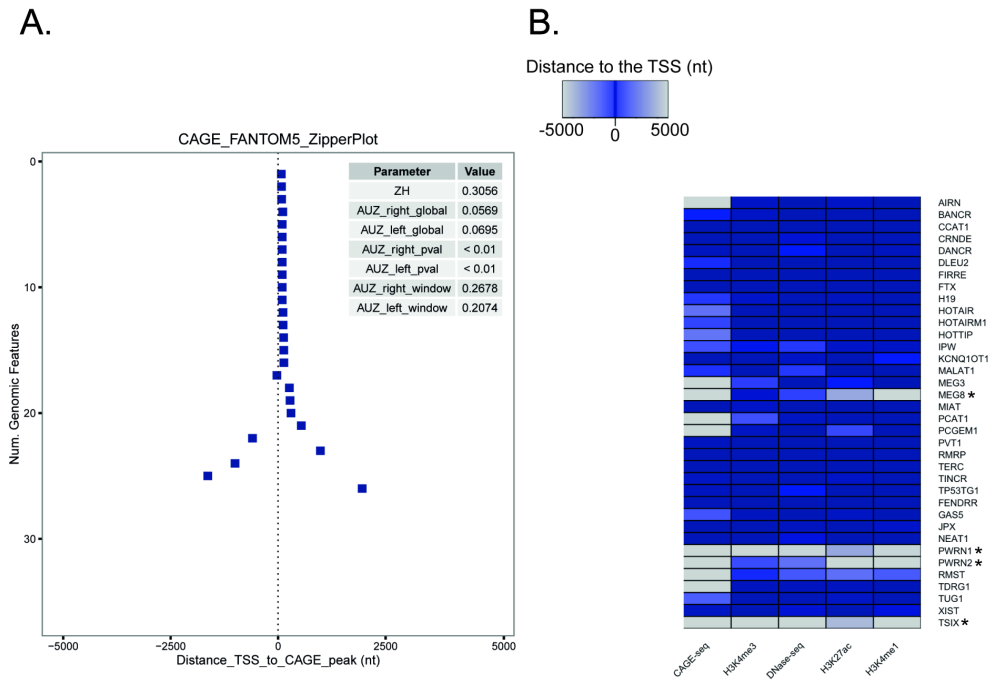


Fig 2. There is evidence of transcriptional activity for 32 out of 36 well-characterized lncRNAs using the Zipper plot. A) Zipper plot and associated statistics for the set of 36 well-characterized lncRNAs proposed by [4] using CAGE-seq data. Even though the visualization contains a +/- 5kb window, it is clear that the closest CAGE peaks for 26 lncRNAs are within +/- 2.5kb from the TSS. Both AUZ_right_pval and AUZ_left_pval are smaller than 0.01, suggesting that the set of TSSs are more closely associated with CAGE peaks compared to random regions in the genome. B) Heatmap showing the distance between TSSs and CAGE-seq, DNase-seq, H3K4me1, H3K4me3 and H3K27ac peaks. Darker colours represent peaks that are closer to the TSSs. LncRNAs marked with an asterisk do not have enough evidence of transcriptional activity. (nt = nucleotides).

As a second example application of the Zipper plot, we evaluated the transcriptional independence of all human lncRNAs listed in Lncipedia 3.1 [25]. We studied the distribution of the closest CAGE-seq peaks (FANTOM5 data) around the TSSs of all mono-exonic and all multi-exonic human lncRNA transcripts (21,102 and 90,508 respectively) (Fig 3A-C) and found that 589 mono-exonic lncRNAs (2.8 %) presented a CAGE-peak overlapping with the TSS and 6,256 (29.7 %) had a peak within a +/- 5kb window. On the other hand, 14.419 multi-exonic lncRNAs (15.9 %) presented a CAGE-peak overlapping with the TSS and 45.878 (50.7 %) had a peak within a +/- 5kb window (Fig 3D). These differences, also reflected in greater AUZ_global values in the former case, suggest that numerous mono-exonic lncRNAs might not be truly independent transcriptional units.

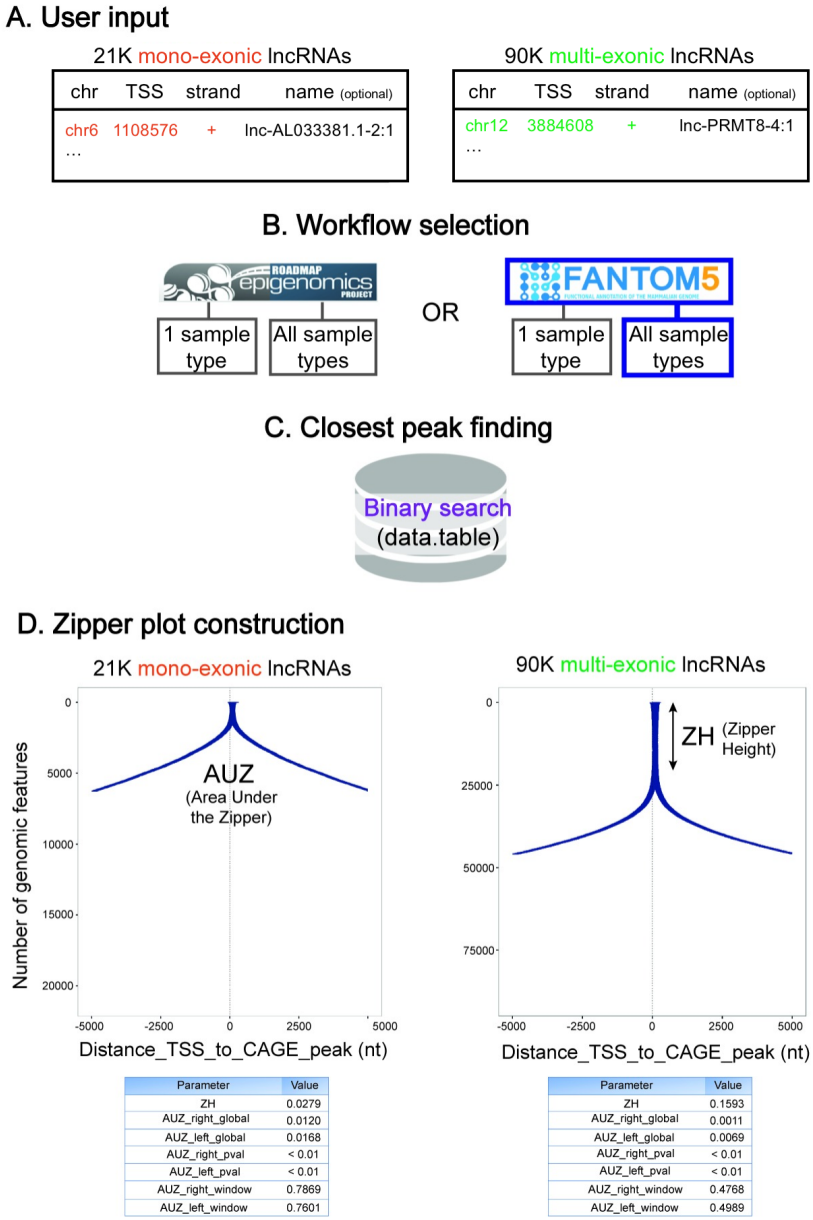


Fig 3. Fewer mono-exonic lncRNAs have CAGE-seq peaks overlapping with their TSSs compared to multi-exonic lncRNAs. This is reflected in smaller Zipper Height (ZH) and higher Area Under the Zipper (AUZ) values. FANTOM5 data (CAGE-seq) and "All sample types" workflow was selected. As described in the "Database querying" section, users may provide an additional fourth column in the input file with labels for each TSS (optional). Since both plots are visualized in a +/- 5 kb window, AUZ_window values can be directly compared: smaller values (multi-exonic lncRNAs) represent higher evidence of independent transcriptional activity for the set of genomic features being studied. This conclusion can also be made looking at the ZH values: a bigger ZH value means a higher proportion of lncRNAs with a CAGE peak overlapping with the TSS. Finally, both AUZ_right_pval and AUZ_left_pval are smaller than 0.01, so it is unlikely that the set of TSSs from mono and multi-exonic lncRNAs were randomly selected.

We hypothesized that at least a fraction of mono-exonic lncRNAs were actually extensions of UTRs from upstream protein coding genes or genomic DNA contamination. To further investigate this hypothesis, we first retrieved the intron lengths for all RefSeq protein coding genes (hg19; using the UCSC Table Browser data retrieval tool) [26,27] and found that 80% of them are smaller than or equal to 5,827 nucleotides. In a second step, we artificially "stitched" mono-exonic lncRNAs that do not have a CAGE peak within 500 nucleotides from their TSSs to the 3' end of any protein coding gene located within 5,827 nucleotides on the same strand. This process led to 536 mono-exonic lncRNAs stitched to upstream protein coding genes.

If these lncRNAs were actually unannotated portions of upstream coding genes (and therefore, not lncRNAs but false positives), we should find junction reads spanning one exon from a protein coding gene and another exon from a lncRNA. To evaluate this, we used RNA-seq data from The Cancer Genome Atlas (TCGA) [28,29] and Universal Human Reference RNA (UHRR) samples [30,31] (see Methods). Since junction reads that are shared between exons of overlapping lncRNAs and protein coding genes cannot be assigned unambiguously, they were excluded from the analyses. Next, we established a minimum of at least 1 junction read linking a lncRNA to an upstream protein coding gene and a minimum overlap of 2 nucleotides between the junction read and the protein coding gene exon and a minimum overlap of 2 nucleotides between the junction read and the lncRNA exon (see "Table 1" and "S2 Table").

Strikingly, we found spanning reads for 135 out of the 536 cases (25.19%) based on the TCGA RNA-seq data and 35 (6.53%) based on UHRR RNA-seq data (S2 Table).

We also tried to stitch multi-exonic lncRNAs that do not have a CAGE peak within 500 nucleotides from their TSSs in the same manner as we did for mono-exonic lncRNAs, resulting in 675 multi-exonic lncRNAs stitched to upstream protein coding genes. We found spanning reads for 127 out of the 675 cases (18.81%) based on the TCGA RNA-seq data and 33 (4.89%) based on UHRR RNA-seq data (S2 Table). 92.59% of the junction reads from the TCGA RNA-seq data entirely overlap with protein coding gene exons and 88.15% of them entirely overlap with lncRNA exons. On the other hand, 89.31% of the junction reads from UHRR RNA-seq data entirely overlap with a protein coding gene exons and 91.91% of the junction reads entirely overlap with lncRNA exons.

Both TCGA and UHRR samples shared junction reads for 34 mono-exonic and 31 multi-exonic lncRNAs stitched to an upstream protein coding gene. Table 1 shows the

distribution of junction reads spanning a protein coding gene and downstream lncRNA based on the TCGA RNA-seq data.

Table 1. Distribution of junction reads (JR) from 1,460 TCGA samples connecting protein-coding gene exons with a downstream exon previously annotated as part of a mono or multi-exonic lncRNA. These junction reads suggest that the latter is actually an extension of a UTR from an upstream protein coding gene rather than a truthful lncRNA. Detailed information for each individual case can be found on S2 Table.

	1 <= JR <= 10	11 <= JR <= 100	JR > 100	Total
Protein coding gene + mono-exonic lncRNA	86	37	12	135
Protein coding gene + multi-exonic lncRNA	81	31	15	127

These results support our hypothesis and reveal the need for a careful evaluation of lncRNA 5'-boundaries using CAGE-seq data and histone marks as demonstrated here or alternative procedures such as 5'-RACE(-seq) [32].

To further expand the applicability of our tool, we plan to extend the number of samples when new data becomes available, to allow users to work with their own data and to integrate publicly available data from i) 25 chromatin states across 127 epigenomes reflecting the interaction between two or more histone marks in their spatial context, ii) a plethora of publicly available data from methods that detect nascent RNAs (GRO- and PRO-sequencing) and iii) open chromatin regions (ATAC-sequencing).

Conclusion

We have created the Zipper plot, a novel visualization and analysis method available as a webtool [8] that allows researchers to quickly evaluate the reliability of the annotation of thousands of novel transcripts and lncRNAs at the same time. Using the Zipper plot we found evidence of transcription for a set of well-characterized lncRNAs and observed that fewer mono-exonic lncRNAs have CAGE peaks overlapping with their TSSs compared to multi-exonic lncRNAs. Using publicly available RNA-seq data, we discovered more than one hundred cases where junction reads connected protein-coding gene exons with a

downstream mono-exonic lncRNA, revealing the need for a careful evaluation of lncRNA boundaries.

We also recognize a limitation in our webtool: the presence of a CAGE-peak and activating histone marks at the TSS is indicative of independent transcription, but the absence of such features does not imply the opposite. Low abundant transcripts may not show up in the CAGE-seq data because of too low sequencing depth or the expression of the lncRNA may be restricted to a tissue of cell type not (yet) included in the CAGE-seq, ChIP-seq and DNase-seq database. Importantly, TSSs of RNA transcripts reconstructed from RNA-seq data might appear several nucleotides downstream of a CAGE-seq peak. Particularly for low abundant RNA transcripts, this inconsistency may be the result of an incomplete transcript assembly due to non-uniformity of read coverage towards 5' ends and should be carefully examined.

Methods

Establishing a peak-based database using publicly available datasets

ChIP-seq & DNase-seq from 127 consolidated human epigenomes already processed in the context of the Roadmap Epigenomics Project (111 from NIH Roadmap Epigenomics Mapping Consortium (Release 9 of the Human Epigenome Atlas) [33] and 16 cell line epigenomes from the ENCODE Project Consortium [34,35] were retrieved from the "Peak Calling" section at [36].

DNase-seq and ChIP-seq data consists of ENCODE narrowPeak, broadPeak and gappedPeak files. Detailed information about these formats can be found at [37].

These files contain lists of peaks that were obtained by a peak caller algorithm in the context of the Roadmap Epigenomics Project. The peak calling process identified regions in the genome that were enriched with aligned reads ("peaks") as a consequence of the ChIP or DNase-seq experiment.

We focused our filtering approach on the qValue, being a measurement of statistical significance for the signal enrichment of each peak using the false discovery rate (FDR). We set a FDR <= 0.05, implying that only those peaks with qValue <= 0.05 were retained in our database for downstream applications.

The following activating marks [38] were used to construct the database: marks for open chromatin (DNaseI); acetylation marks commonly found in actively transcribed promoters (H3K27ac, H3K9ac, and H3K14ac), methylation marks found in actively transcribed

promoters (H3K4me1, H3K4me2, H3K4me3 and H4K20me1) and modifications added as consequence of transcription (H3K36me3, H3K79me2 at 5' end of gene bodies) adding up to more than 134 million peaks. (S3 Table).

CAGE-seq expression data (RLE normalized) for human samples was retrieved from the Functional Annotation of the Mammalian Genome (FANTOM5) project [39,40]. CAGE-seq measures expression by means of sequencing from the 5' end (transcription start site (TSS)) of capped molecules. In case of multiple replicates per sample type, only one replicate was retained, bringing the total number of samples to 649 with a total of 200,737 peaks. (S4 Table).

Obtaining junction reads from publicly available RNA-seq data

1,460 RNA-seq samples from TCGA across different cancer types [28,29] (See S5 Table for detailed information on cancer type and TCGA barcodes) and 80 UHRR samples from the Sequencing Quality Control (SEQC) project publicly available at the Gene Expression Omnibus (GEO) database with accession number GSE47774 (Sample A: Replicates 1-4; Beijing Genomics Institute) [30,31] were mapped to the human genome (GRCh37) using TopHat2 [41] with default parameters, resulting in 279,507,060 and 12,679,075 junction reads respectively.

Abbreviations

3P-seq: Poly(A)-Position Profiling by Sequencing; **AUZ**: Area Under the Zipper; **CAGE-seq**: Cap Analysis of Gene Expression sequencing; **ChIP-seq**: Chromatin Immunoprecipitation sequencing; **DNase-seq**: DNase sequencing; **FANTOM5**: Functional Annotation of Mammalian Genomes 5; **FDR**: False Discovery Rate; **FPKM**: Fragments Per Kilobase Million; **GRO-seq**: Global Run-On Sequencing; **H3K4me1**: Histone H3 lysine 4 monomethylation; **H3K4me2**: Histone H3 lysine 4 dimethylation; **H3K4me3**: Histone H3 lysine 4 trimethylation; **H4K20me1**: Histone H4 lysine 20 monomethylation; **H3K36me3**: Histone H3 lysine 36 trimethylation; **H3K79me2:** Histone H3 lysine 79 dimethylation; **H3K9ac**: Histone H3 lysine 9 acetylation; **H3K14ac**: Histone H3 lysine 14 acetylation; **H3K27ac**: Histone H3 lysine 27 acetylation; **lncRNA**: Long non-coding RNA; **PRO-seq**: Precision nuclear Run-On sequencing; **RACE-seq**: Rapid amplification of cDNA ends sequencing; **RNA-seq**: RNA sequencing; **SEQC**: Sequencing Quality Control; **TCGA**: The Cancer Genome Atlas; **tpm**: tags per million mapped reads; **TSS**: Transcription Start Site; **UHRR**: Universal Human Reference RNA; **ZH**: Zipper Height.

Declarations

Acknowledgements

We thank Tom Sante for his technical advice and fruitful discussions throughout the development of this webtool and Karen Verboom and Annelynn Wallaert for their suggestions to improve the website.

Funding

This work was supported by the Concerted Research Action of Ghent University (BOF/GOA) to FAC; P.M. and C.E. are supported by the Fund for Scientific Research Flanders (FWO); D.R. is supported by the Agency for Innovation by Science & Technology (IWT).

Author Contributions

PM, KP, FAC, JV and DR contributed to the conceptualization of the manuscript. FAC prepared the original draft under the supervision of PM, KP and JV. FAC and PJV developed the visualization method. FAC wrote the R code and JA contributed to its curation and implementation as a website. FAC, PM, KP, contributed to the validation of the method and CE processed the publicly available RNA-seq data. All authors read and approved the final manuscript.

The funders had no role in the study design, data collection and analysis, interpretation of the data, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The original curated list of 36 functionally characterized lncRNAs can be found in the additional file 2 from[4]. S6 Table contains the genomic coordinate of the TSS for these 36 lncRNAs.

We acknowledge the TCGA Consortium and its members for their project initiatives.

The results shown here are in part based upon data generated by the TCGA Research Network [28,29], FANTOM5 Project [39], NIH Roadmap Epigenomics Mapping Consortium [33] and ENCODE Project Consortium [34,35].

FANTOM5 data by RIKEN is licensed under a Creative Commons Attribution 4.0 International License [40,42]. Processed data from NIH Roadmap Epigenomics Mapping Consortium and ENCODE Project Consortium are freely available at [36].

The UHRR RNA-seq data (GSE47774) is available from the NCBI website [30].

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
2. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22:1775–89.
3. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 2015;11:1110–22.
4. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 2016;17:19.
5. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 2006;22:1–5.
6. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011;147:1537–50.
7. Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ, et al. Genome-guided transcript assembly by integrative analysis of RNA sequence data. Nat. Biotechnol. 2014;32:341–6.
8. ZipperPlot [Internet]. [cited 2017 Feb 1]. Available from: http://zipperplot.cmgg.be/
9. R: The R Project for Statistical Computing [Internet]. [cited 2017 Feb 2]. Available from: https://www.r-project.org/
10. data.table.pdf [Internet]. [cited 2017 Feb 2]. Available from: https://cran.r-project.org/web/packages/data.table/data.table.pdf
11. Wickham H. ggplot2 [Internet]. New York, NY: Springer New York; 2009 [cited 2017 Feb 2]. Available from: http://link.springer.com/10.1007/978-0-387-98141-3
12. Xie Y, Vogt A, Andrew A, Zvoleff A, http://www.andre-simon.de) AS (the C files under inst/themes/ were derived from the H package, Atkins A, et al. knitr: A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2016 [cited 2017 Feb 2]. Available from: https://cran.r-project.org/web/packages/knitr/index.html
13. Bengtsson H. R.utils: Various Programming Utilities [Internet]. 2016 [cited 2017 Mar 31]. Available from: https://cran.r-project.org/web/packages/R.utils/index.html
14. Auguie B, Antonov A. gridExtra: Miscellaneous Functions for "Grid" Graphics [Internet]. 2016 [cited 2017 Mar 31]. Available from: https://cran.r-project.org/web/packages/gridExtra/index.html
15. favilaco/Zipper_plot [Internet]. GitHub. [cited 2017 Mar 31]. Available from: https://github.com/favilaco/Zipper_plot
16. Lift Genome Annotations [Internet]. [cited 2017 Mar 31]. Available from: https://genome.ucsc.edu/cgi-bin/hgLiftOver
17. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30:1006–7.
18. What is CrossMap ? — CrossMap documentation [Internet]. [cited 2017 Mar 31]. Available from: http://crossmap.sourceforge.net/
19. UCSC Genome Browser Store [Internet]. [cited 2017 Mar 31]. Available from: https://genome-store.ucsc.edu/
20. UCSC Genome Browser: Downloads [Internet]. [cited 2017 Mar 31]. Available from: http://hgdownload.cse.ucsc.edu/downloads.html#source_downloads

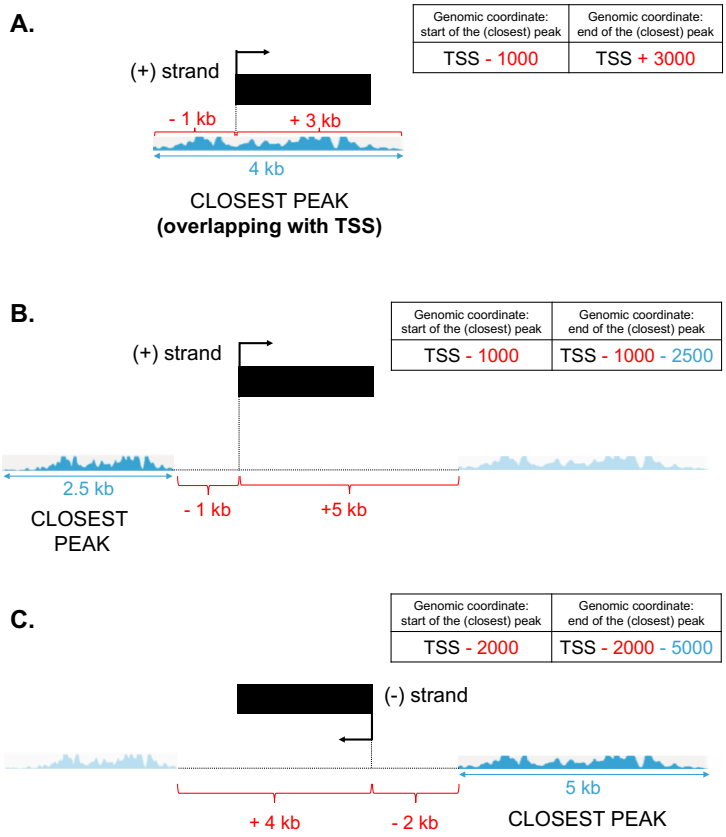
21. Microsoft Word - ZP_Documentation_FINAL.docx - manual_1.2.pdf [Internet]. [cited 2017 Feb 1]. Available from: http://zipperplot.cmgg.be/manual_1.2.pdf
22. Index of /goldenPath/hg19/database [Internet]. [cited 2017 Mar 31]. Available from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gap.txt.gz
23. Index of /goldenPath/hg19/bigZips [Internet]. [cited 2017 Mar 31]. Available from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromOut.tar.gz and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromTrf.tar.gz
24. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
25. Volders P-J, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, et al. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res. 2015;43:D174-180.
26. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
27. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32:493D–496.
28. The Cancer Genome Atlas Home Page [Internet]. Cancer Genome Atlas - Natl. Cancer Inst. [cited 2017 Feb 1]. Available from: https://cancergenome.nih.gov/
29. Welcome to The Genomic Data Commons Data Portal | GDC [Internet]. [cited 2017 Feb 2]. Available from: https://gdc-portal.nci.nih.gov/
30. GEO Accession viewer [Internet]. [cited 2017 Feb 1]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47774
31. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.
32. Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM, et al. Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). Nat. Commun. 2016;7:12339.
33. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
34. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
35. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
36. Roadmap Epigenomics [Internet]. [cited 2017 Feb 1]. Available from: http://egg2.wustl.edu/roadmap/web_portal/processed_data.html#ChipSeq_DNaseSeq
37. Genome Browser FAQ [Internet]. [cited 2017 Mar 31]. Available from: https://genome.ucsc.edu/FAQ/FAQformat
38. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.
39. (dgt) TFC and the RP and C. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
40. FANTOM - Data Summary [Internet]. [cited 2017 Feb 1]. Available from: http://fantom.gsc.riken.jp/5/data/

41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
42. Creative Commons — Attribution 4.0 International — CC BY 4.0 [Internet]. [cited 2017 Feb 2]. Available from: https://creativecommons.org/licenses/by/4.0/

Supplementary methods

Definition of the distance between a TSS and the closest peak

Each peak consists of two genomic coordinates delimiting its start and end positions. Regardless of whether there is a peak overlapping with the TSS or not, the data.table package retrieves the "start" and "end" genomic coordinates of the closest ChIP-seq/DNase-seq/CAGE-seq peak as depicted in this figure, **with the "start" always being the part of the peak closest to the TSS**:



"A" represents a case where the closest peak overlaps with the TSS whereas "B" and "C" are cases where it does not. In "B", the closest peak is the one on the left hand side, since a distance of 1kb is smaller than 5kb. In "C", the closest peak is the one on the right hand side, since a distance of 2kb is smaller than 4kb.

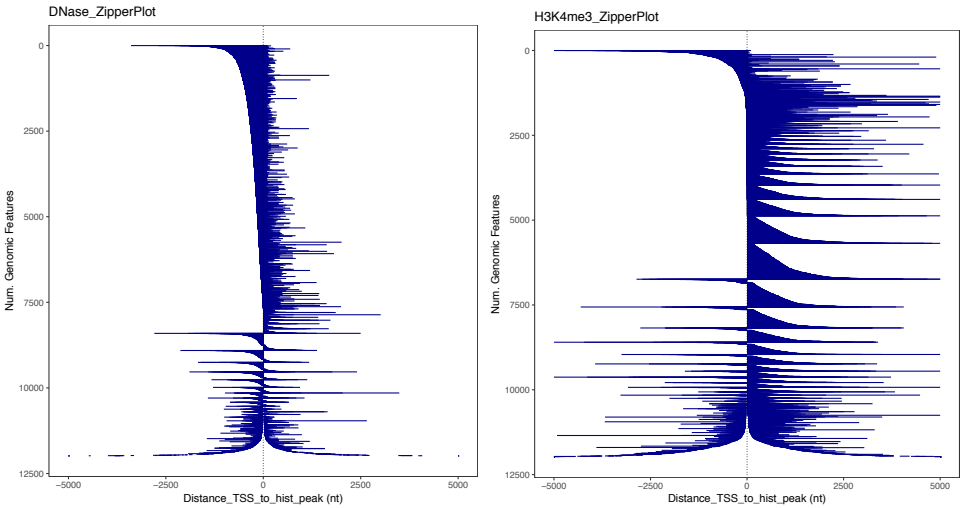
Finally, **peaks are ranked based on the distance from the TSS to the "start" of the closest peak** and a Zipper plot is generated as described in the main manuscript.

Rationale for calculating both positive and negative distances between closest peaks and TSSs

Thurman and colleagues [1] demonstrated that, for 56 different cell types, a common pattern for the distribution of the H3K4me3 and DNase marks around the TSS can been observed: DNase marks were located a few nucleotides upstream the TSS whereas the H3K4me3 appeared several nucleotides downstream the TSS (Fig. 3a-b from [1*]).

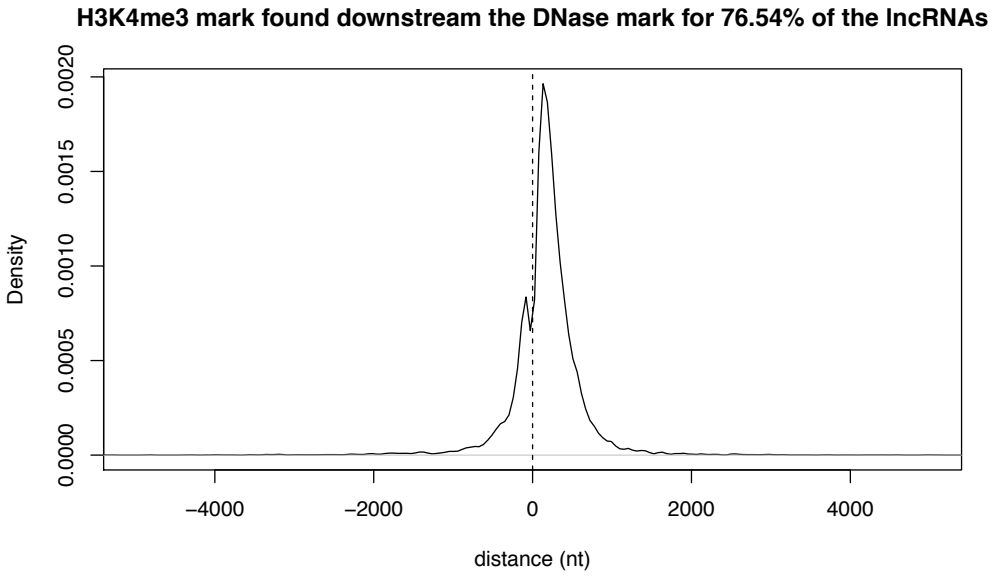
We have investigated this hypothesis as an argument to support the need of calculating AUZleft (negative distances between TSSs and closest peak) and AUZright (positive distances between TSSs and closest peak).

We used all lncRNAs (mono and multi-exonic) from Lncipedia 3.1 and retrieved 11,989 lncRNAs (unique TSSs) with a CAGE peak overlapping with the TSS. Next, we retrieved the closest H3K4me3 and DNase peaks (narrowPeak; across all sample types) for those and generated the following Zipper plots:



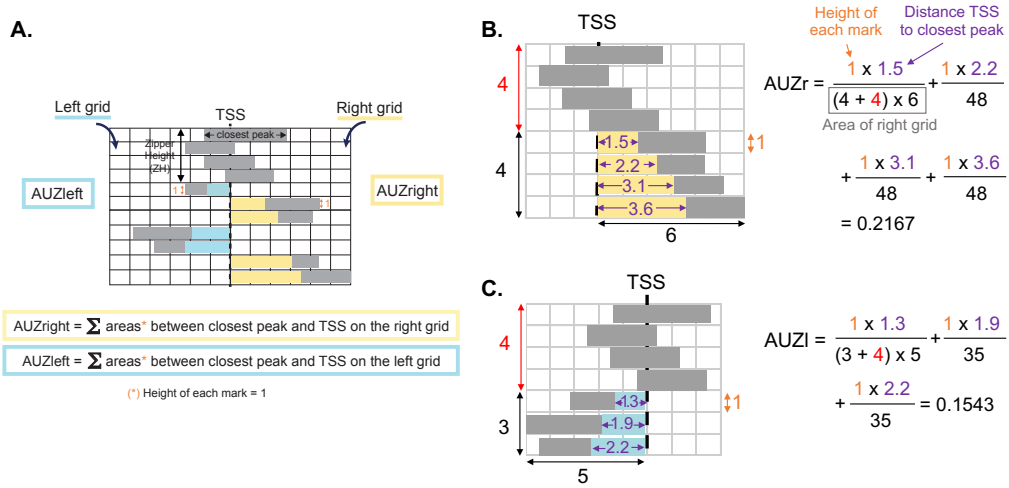
From these plots is clear that the DNase marks (left) were mostly found a few nucleotides upstream the TSS (=negative distances in the plot; OX axis) whereas the H3K4me3 (right) were mostly found several nucleotides downstream the TSS (=positive distances in the plot; OX axis), in agreement with what was observed in [1*].

Finally, we plotted the distribution of distances between these two marks and found that for 9,177 lncRNAs (76.54%) the H3K4me3 mark was found downstream the DNase mark (=distances greater than 0):



This result is a useful example where differences between positive and negative distances are clear.

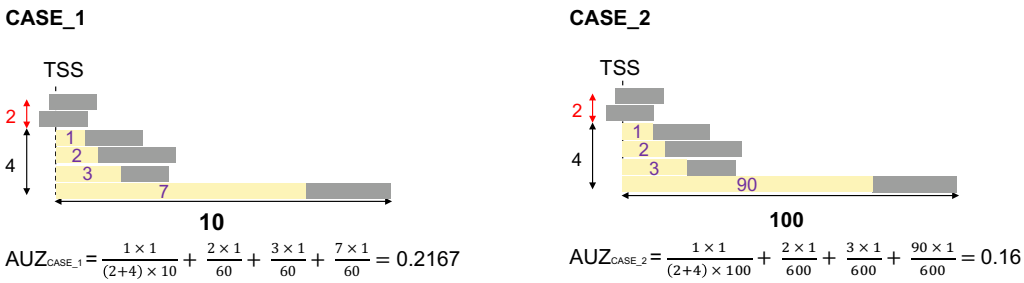
Definition of the sum of all areas between the closest peak and the TSS



A) AUZ is computed as the sum of all the areas between the closest peak and the TSS of each genomic feature. Since the distribution of peaks upstream or downstream of the TSSs can be asymmetric, AUZleft and AUZright are considered independently. B) shows how the AUZ_right is computed; C) shows how AUZ_left is computed.

Importantly, the width of the grid for each Zipper plot is determined by the (closest) peak furthest away from the TSS among all retrieved ones. Looking at the image below,

AUZ_{CASE_1} seems bigger than AUZ_{CASE_2} while it should be smaller (marks in CASE_1 are closer to TSS than marks in CASE_2):



This issue is due to the difference in grid areas: to compare two AUZ values directly, both need to come from a grid with the same width. This can be achieved by re-scaling the AUZ from the case (or cases, if we are comparing more than 2 Zipper plots) with the smallest width by the ratio between both grids (and maintaining the AUZ from the case with biggest width unchanged):

$AUZ_{CASE\_1\_rescaled}$ = $AUZ_{CASE\_1\_original}$ x ($Grid_{CASE\_1}$ / $Grid_{CASE\_2}$) = 0.2167 * (60/600) = 0.02167

The resulting AUZ values are: **$AUZ_{CASE\_1\_rescaled}$ = 0.02167 < $AUZ_{CASE\_2\_unchanged}$ = 0.16, as expected.**

This re-scaling formula is used during the computation of the AUZ values of Zipper plots constructed from random regions. More specifically, to compute the AUZ_pval.

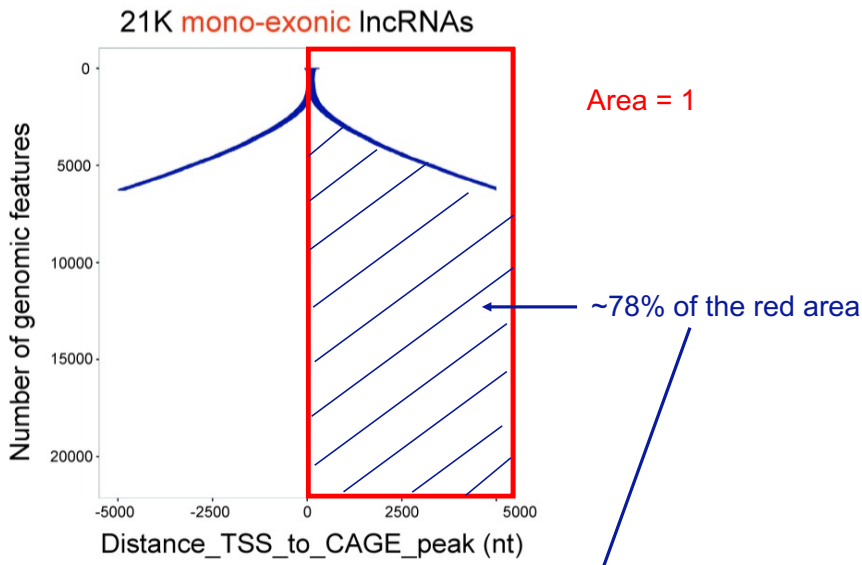
Small AUZ values, areas in the plot and how AUZ_window is calculated (Fig 3D)

As explained in the previous section, the width of the grid for each Zipper plot is determined by the (closest) peak furthest away from the TSS among all retrieved ones. Regarding the Zipper plot for the 21,000 mono-exonic lncRNAs, we found that there are few cases where the closest CAGE-seq peak is several Mb away (x-axis) from the TSS. If we also take into account that we are studying thousands of lncRNAs simultaneously (y-axis), this results in a grid area of the order of 10^9. Since 76.24% of the mono-exonic lncRNAs have a CAGE peak within 50kb from the TSS (very small value compared to 1Mb), this resulted in very small AUZ_global values.

The "actual areas in the plot" correspond to the AUZ_window values. By default, the Zipper plot is visualized in a +/- 5kb window. In the case of the plot for the 21,000 mono-

exonic lncRNAs, the method virtually sets to 5kb (or other value if the user changes the default window size) all those distances that are located more than 5kb away from the TSS. Therefore, the y-axis extends down to 21,000.

AUZ_window values correspond to the AUZ value that users can "visually" infer from the plot visualized in the pre-defined (5kb) window.



Parameter	Value
ZH	0.0279
AUZ_right_global	0.0120
AUZ_left_global	0.0168
AUZ_right_pval	< 0.01
AUZ_left_pval	< 0.01
AUZ_right_window	0.7869
AUZ_left_window	0.7601

References

1*. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82. doi:10.1038/nature11232.

Supplementary material (available online at BMC Bioinformatics, too big to be included)

S1 Table.xls Summary table for the set of 36 well-characterized lncRNAs using CAGE-seq data.
S2 Table.xls Junction reads between protein coding genes and mono/multi-exonic lncRNAs based on RNA-seq data from TCGA and UHRR; nucleotides of junction read overlapping with lncRNA and protein coding gene exons.
S3 Table.xls Correspondence between Roadmap Epigenomics names and actual sample types; number of peaks and number of epigenomes available for each case; peak width and peak enrichment distributions across chromosomes for narrow, broad and gapped peaks (for each mark).
S4 Table.xls Correspondence between FANTOM5 names and actual sample types; number of CAGE-seq peaks per chromosome; peak width and tpm distributions across chromosomes.
S5 Table.xls Cancer type and barcode for each sample from TCGA**.**
S6 Table.xls HGNC, Ensembl ID, PMID, chromosome location, TSS and strand information for the set of 36 well-characterized lncRNAs.

Paper 2

Refining the human transcriptome and assessing the biological signal of its different RNA biotypes through computational deconvolution

Francisco Avila Cobos, Lucía Lorenzi, Jo Vandesompele, Gary Schroth, Katleen De Preter and Pieter Mestdagh

Contribution: filtering of the initial transcriptome assembled in the RNA atlas project and performed the computational deconvolution analyses; made all the figures and wrote the manuscript.

In preparation

Refining the human transcriptome and assessing the biological signal of its different RNA biotypes through computational deconvolution

Francisco Avila Cobos[1,2], Lucia Lorenzi[1,2], Jo Vandesompele[1,2], Gary Schroth[3], Katleen De Preter[1,2] and Pieter Mestdagh[1,2]

1. Center for Medical Genetics , Ghent University, Ghent, Belgium
2. Cancer Research Institute Ghent, CRIG, Ghent, Belgium
3. Illumina, San Diego, California, USA

Abstract

The RNA Atlas, which integrates data from three complementary RNA-sequencing methods, provides a comprehensive reference of the human transcriptome. Building this reference required stringent filtering of newly assembled transcripts. We used publicly available data as proxy for independent transcription, to filter and refine the assembled transcriptome.

Furthermore, taking advantage of the plethora of different RNA biotypes present in the RNA Atlas, we investigated the performance of the different RNA fractions in a computational deconvolution framework. Our analysis highlighted the importance of including a comprehensive collection of cell types and high-quality markers in the reference matrix used in computational deconvolution of transcriptomics data, regardless of the RNA fraction being used.

Introduction

Since the introduction of RNA-sequencing (RNA-seq) technology more than a decade ago[1], many different efforts to generate a comprehensive annotation of the human transcriptome have been launched. For instance, Pertea *et al.*[2] used un-stranded poly-adenylated (polyA) RNA-seq from the Genotype - Tissue Expression (GTEx) project to build the CHESS database, which expanded the traditional number of ~20,500[3] protein coding genes to 21,306 and also included 21,856 non-coding genes. Lncipedia[4] (currently in its fifth edition) gathers a set of ~50,000 long non-coding RNAs (lncRNAs) coming from Ensembl, RefSeq and the FANTOM CAT projects. MicroRNAs (miRNAs), antisense RNAs (asRNAs) and circular RNAs (circRNAs) are also part of the plethora of non-coding RNAs and have been shown to have regulatory functions and to be relevant in human development and disease[5,6], making the aforementioned resources very valuable for the scientific community.

To date, the most complete attempt to map the entire human transcriptome comes from the RNA Atlas project[7], which applied three complementary RNA-sequencing methods (small, polyA and Total RNA-seq) on 162 normal cell types, 45 tissues and 93 cancer cell lines.

Previous efforts such as GTEx[8] (~1000 samples across 54 non-diseased human tissues), the Cancer Genome Atlas (TCGA[9]; both cancer and matched normal tissues) or the International Cancer Genome Consortium (ICGC[10]; both cancer and matched normal tissues) only included unstranded libraries for polyA and small RNA-sequencing. In contrast, the novelty of the RNA Atlas project lies in the introduction of the Total RNA sequencing layer, the use of stranded libraries and the inclusion of both tissues and its constituent cell types.

Thousands of transcripts belonging to five different RNA biotypes (asRNAs, circRNAs, lncRNAs, miRNAs and messenger RNAs (mRNAs)) were identified in the RNA Atlas dataset through transcriptome assembly, but an extra filtering step was needed in order to distinguish truly independent transcriptional units from transcriptional noise or contamination.

We used the Zipper plot tool[11] to refine the initial transcriptome assembly in the RNA Atlas project and retained features with evidence of independent transcription at DNA level, RNA level or both. As an additional quality control step, and to exploit the potential of the various RNA biotypes, we performed computational deconvolution on each individual RNA fraction using the tissue and cell type expression profiles. Since the tissue and cell type samples came from different and independent sources, finding biological signal of the expected cell types composing a tissue can highlight data quality and consistency of expression profiles for different RNA biotypes.

Methods

Building a stringent version of the RNA Atlas transcriptome

Independent evidence for transcription of genomic features in the RNA Atlas transcriptome is obtained from two different resources:

1. Cap Analysis of Gene Expression (CAGE) sequencing data from the FANTOM5 project, which maps transcription start sites (TSSs) in promoters.

2. Chromatin states from the Roadmap Epigenomics Project (25 chromatin state model across 127 epigenomes using 12 marks; https://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp): DNase; active

transcription start site (1_TssA); transcription (5_Tx5; 6_Tx; 7_Tx3); transcribed and regulatory (9_TxReg); transcribed and enhancer (10_TxEnh5; 11_TxEnh3); active_enhancer (13_EnhA1; 14_EnhA2) and bivalent_promoter (23_PromBiv) (Figure 1; Supplementary Table 1).

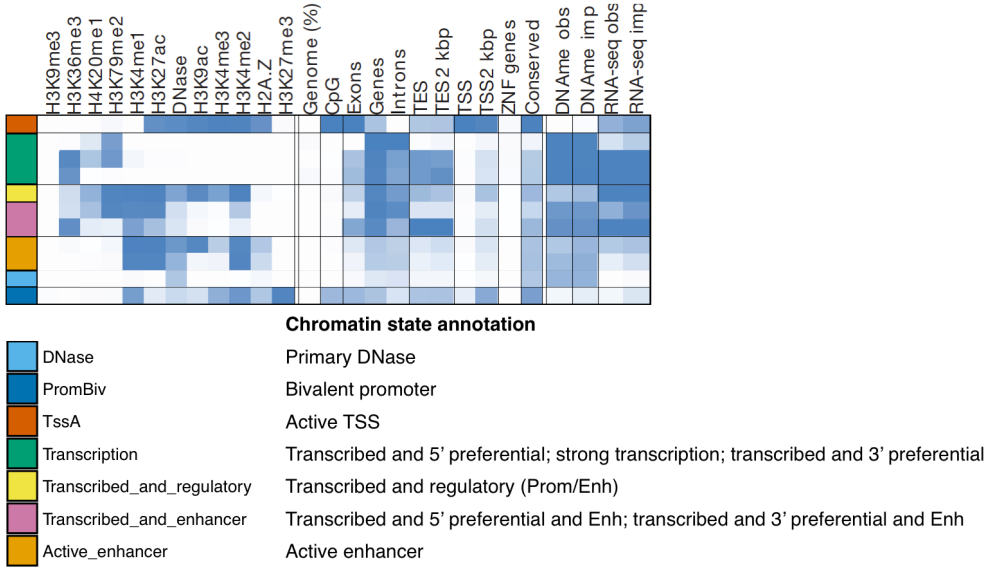


Figure 1 (Adapted from Ernst and Kellis, 2015[12]) - Chromatin states from the Roadmap Epigenomics Project used to annotate the genomic features from the RNA Atlas project (evidence at DNA level). TES = Transcription End Site; obs = observed; imp = imputed; Prom = promoter; Enh = enhancer; DNase = DNase I hypersensitivity regions lacking enhancer/promoter marks.

For each TSS of genes with Total RNA-seq expression values greater or equal to 0.1 transcripts per million (TPM) in at least one tissue from the RNA Atlas and not being part of chromosome Y (chromatin states were not computed for that chromosome in the original article), we used the Zipper plot[11] approach to retrieve the closest CAGE-seq and chromatin state peaks across all samples from the FANTOM5[13] and Roadmap Epigenomics project[14], respectively. We defined "strong evidence for independent transcription" as presence of those peaks within 500 nucleotides upstream or downstream of the TSSs and assigned the genes to one of the three following categories: 1) evidence at both DNA and RNA level; 2) evidence only at RNA level; 3) evidence only at DNA level.

Genes not belonging to any of those three categories were excluded from the final (stringent) version of the RNA Atlas transcriptome. More information about the original transcriptome assembly and how miRNAs and circRNAs were selected can be found in the original RNA Atlas manuscript[7].

Assessing the biological signal of different RNA fractions through computational deconvolution of healthy tissues

Multiple approaches have been developed to infer abundance of different cell types in heterogeneous samples (= computational deconvolution). Albeit potentially applicable to different RNA fractions, current methods have been designed and tested only with messenger RNAs (mRNAs) as input. Using expression data of long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), microRNAs (miRNAs) and mRNAs from the RNA Atlas project, we investigated the impact of using different RNA fractions as input in a computational deconvolution framework.

Tissues and cell types in the RNA-Atlas were matched based on Uber-anatomy ontology (UBERON) from EMBL-EBI[15], resulting in pairs encompassing 28 tissues and 102 cell types (Supplementary Table 2). For each cell type and using VST-normalized[16] expression matrices from Total RNA-sequencing data, we selected cell-type specific markers (matrix C in Figure 2) from each RNA fraction, namely asRNAs, circRNAs, lncRNAs, miRNA and mRNAs.

Following the recommendation from Zhong and Liu[17], RNA expression values were converted into linear scale using the anti-logarithmic function and ranked across cell types. Next, we computed the fold change between the cell type with the highest expression value and the second highest. RNAs with a fold-change greater than or equal to 5 were considered markers, and the maximum number of markers per cell type was capped at 10.

Together with the expression data from the tissues in the RNA Atlas (matrix T), these markers and their expression in the different cell types (matrix C) were used to determine the proportion (matrix P) of each cell type in each of the tissues through computational deconvolution. To that end, we used the Lawson-Hanson implementation of a non-negative least squares framework (using the nnls package[18] from the R statistical programming language), meaning that all proportions must be greater or equal than zero and must sum to one (Figure 2; see Avila Cobos *et al.*[19] for a detailed review). For any given tissue, we defined the "signal" as the sum of the proportions of all its constituent cell types and this signal was computed for mRNA, miRNA, lncRNA and circRNA markers separately.

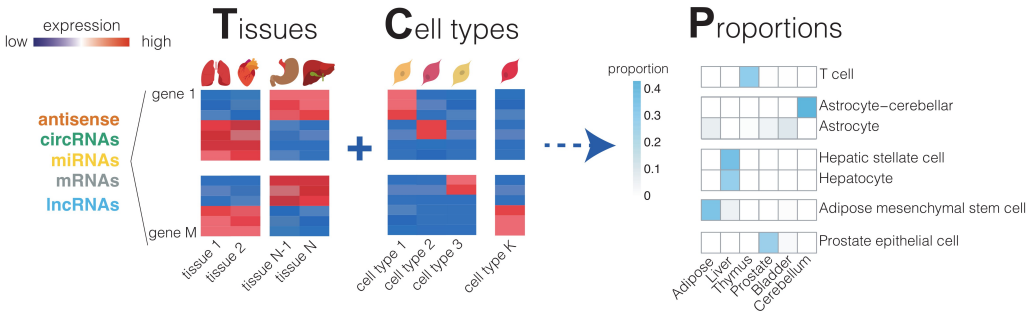


Figure 2 – Scheme for deconvolution of RNA Atlas tissues using cell-type specific expression profiles from Total RNA-sequencing. For instance, the signal for liver tissue in the picture will be the sum of the proportions from the hepatic stellate and hepatocyte cells.

Validation of tissue-specific RNAs from the Tissue Atlas

Expression data was retrieved from 23 tissues from the Tissue Atlas dataset (part of the Human Protein Atlas[20]; http://www.proteinatlas.org) with matching tissues in the RNA atlas (Supplementary Table 3). Within the Tissue Atlas dataset, 1,320 tissue-specific genes with an expression value of at least 5 TPM and a fold change of at least 10 between the first and the second tissue with highest expression values were selected.

Next, we moved to the RNA Atlas dataset and computed the log2 fold-change between the expression in the matching tissue and the highest expression among the remaining tissues for those 1,320 marker genes.

A marker was considered validated in the RNA atlas if it had the highest expression value in the same tissue (see Figure 6).

RESULTS

The stringent version of the RNA Atlas transcriptome contains thousands of novel genes

Using publicly available CAGE-seq data from the FANTOM5 and various chromatin states associated with transcription or with enhancer activity from the Roadmap Epigenomics project, we obtained a comprehensive human transcriptome consisting of 19,107 mRNA genes (of which 188 are novel with respect to the Ensembl transcriptome (v86[21])), 18,387 lncRNAs (of which 13,175 are novel) and 7,309 asRNAs (of which 2,519 are novel) with evidence at RNA level, DNA level or both (Figure 3). For information on circRNAs and miRNAs, see Lorenzi *et al.*[7]

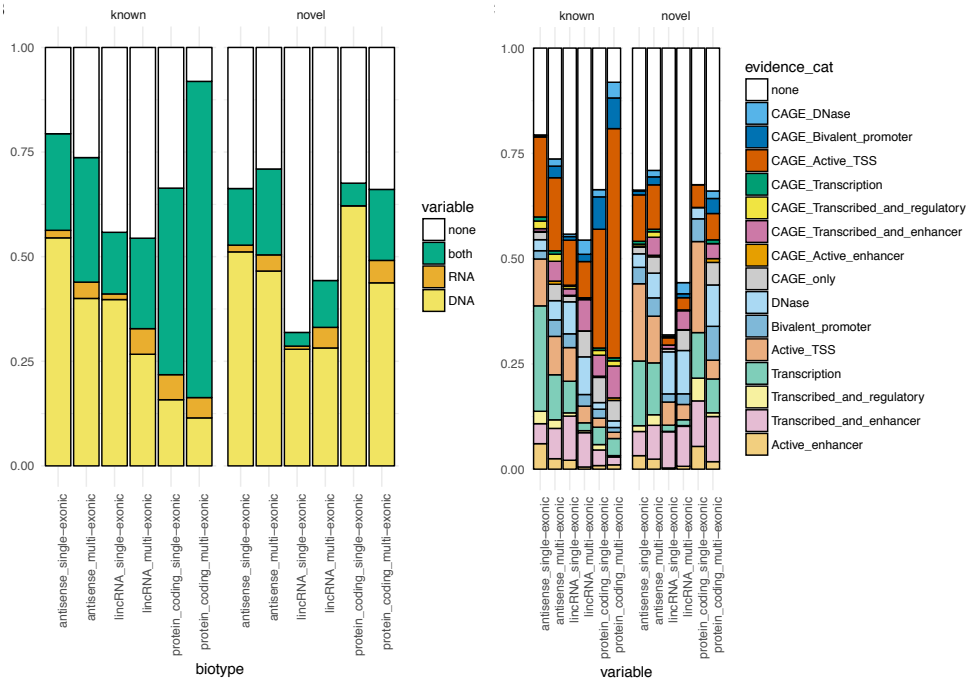


Figure 3 – Fraction of known and novel genes in the RNA Atlas with evidence for independent transcriptional activity. A) Fraction of genomic features with evidence at DNA level only (yellow), at RNA level only (orange) or both DNA and RNA level (green) across the different RNA biotypes. No evidence is shown in white. B) Detailed version of A) where the RNA evidence is represented by all categories containing the word "CAGE" whereas the DNA evidence is represented by 7 different coloured categories.

Of note, we performed a correlation analysis between the Total RNA-seq expression from the RNA Atlas and matching CAGE-seq expression from the FANTOM5 project, showing that there is a positive correlation between both features (Supplementary Figure 1), adding further evidence for the use of CAGE-seq as a mark for transcriptional activity in the RNA Atlas.

Computational deconvolution showed variable biological signal across tissues and RNA fractions

For those tissues for which we found markers in at least one constituent cell type (Figure 4A-B), we defined the signal as the sum of the proportions of all its constituent cell types. A higher signal represents a better deconvolution performance (Figure 4C). We detected a high biological signal (>= 0.5, meaning a proportion of at least 50% for those cell types composing such tissue, based on the UBERON ontology) for five tissues (thyroid, liver, brain, placenta and spleen) using mRNAs, for three tissues (liver, bladder and kidney)

using lncRNAs and for two tissues using either circRNAs (brain, liver) or asRNAs (spleen, placenta).

Moreover, we further investigated the usability of single-exon lncRNAs, non-polyadenylated lncRNAs and novel miRNAs in the deconvolution of healthy tissues present in the RNA Atlas. Single-exon RNAs are typically discarded during de novo transcriptome assembly and non-polyadenylated RNAs cannot be detected with polyA(+) RNA-seq. However, a fraction of these were found to have evidence of independent transcriptional activity at DNA and RNA level (Figure 3; Figure 5) and were successfully used in computational deconvolution of several healthy tissues (Figure 4D).

Figure 4. Computational deconvolution results using VST-normalized counts from Total RNA-Seq data.

A) heatmap showing, for each RNA fraction, whether markers were found for the different cell types present in the RNA Atlas (orange = presence of marker).

B) barplots and boxplots depicting the number of cell types for which markers were found (left), the number of markers found per cell type within each RNA fraction (middle) and the log2 fold-change of those markers (right).

C) heatmap showing the signal values for each tissue across antisense, circRNAs, lncRNAs, miRNAs and mRNAs. The boxplot on the right hand side depicts the signal distribution across the different RNA fractions (each dot represents one tissue type).

D) deconvolution results using single-exonic (SE) vs multi-exonic (ME) lncRNAs (left panel); non-polyA vs polyA lncRNAs (middle panel) and novel vs known miRNAs.

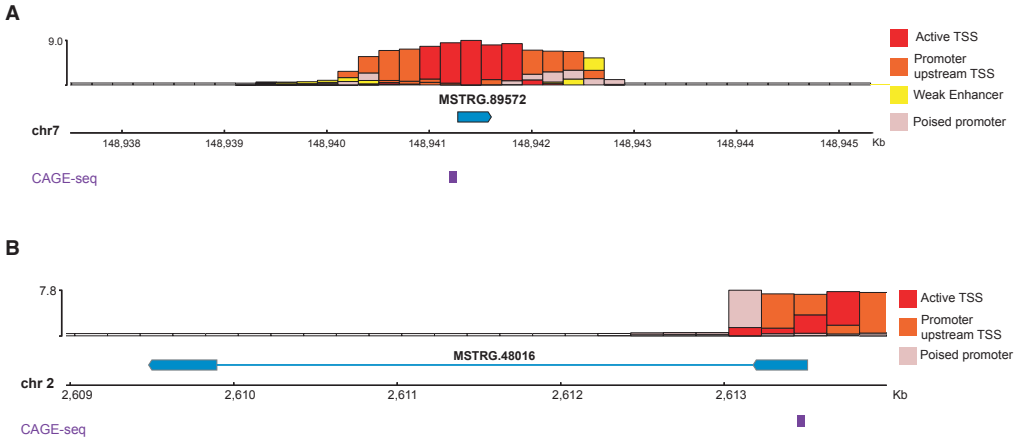


Figure 5 - Examples of a novel mono (A) and novel multi-exonic (B) lncRNA with evidence at both DNA (represented with the chromatin states depicted on the top track) and RNA level (depicted as a purple track named "CAGE-seq").

Failure to include a relevant cell type in the reference matrix has a dramatic impact on the deconvolution results

Including human hepatocytes in the reference matrix used to deconvolve human liver tissue resulted in a high biological signal (sum of the proportions of human hepatocytes, hepatic stellate cells, hepatic mesenchymal stem cells and hepatic sinusoidal endothelial cells (constituent cell types in liver according to UBERON; Supplementary Table 2)) (Figure 6, left barplot). However, removing this cell type from the reference matrix leads to re-distributed and distorted cell type proportions, leading to a low biological signal (Figure 6, right barplot).

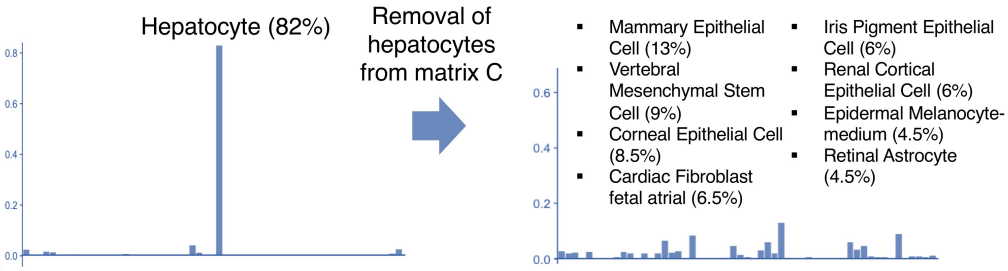


Figure 6 – Failure to include hepatocytes in the reference has dramatic effects on the deconvolution of liver tissue.

Tissue-specific mRNA markers from the Tissue Atlas dataset were validated in the RNA Atlas

A second way of assessing the quality of the RNA Atlas dataset is by analysing external markers coming from an independent dataset. Out of 1,320 mRNA markers selected from the Tissue Atlas (part of the HPA), 1,269 (96.1%) were cross-validated in 19/23 tissues (82.6%) of the RNA Atlas dataset (Figure 7), confirming the quality and relevance of the RNA-Atlas data.

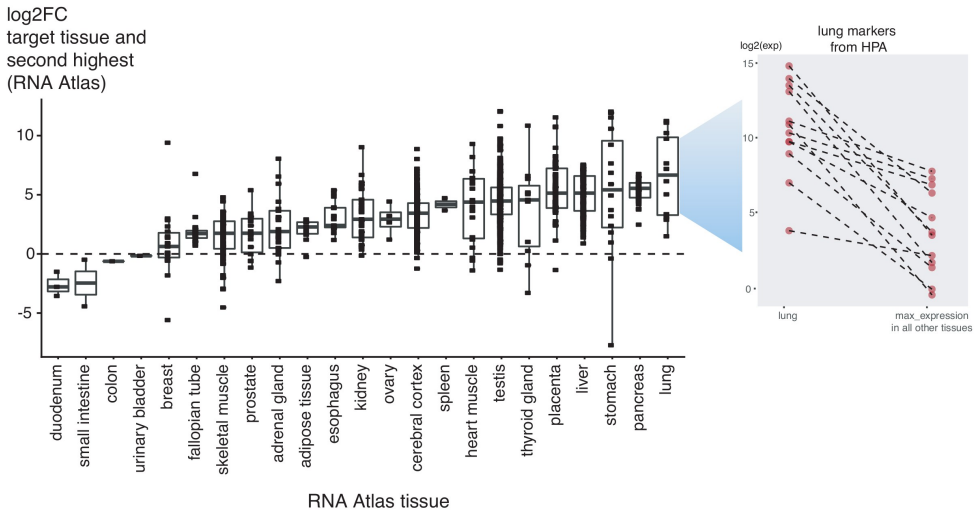


Figure 7 – Log2 fold-changes between the expression values in the tissue were a marker was found (target tissue) and the tissue having the second highest expression value. This analysis was done for mRNA markers found in tissues from the Tissue Atlas and for tissues that are present both in the Tissue Atlas and the RNA Atlas datasets. The grey scatter plot on the right hand side shows the underlying log2(expression) values in the RNA Atlas for all lung markers found in the Tissue Atlas: all lung markers were found to have the highest expression values in the lung tissue from the RNA Atlas compared to the tissue having the second highest expression value for the marker (each marker is depicted as a pair of red circles connected by a black dashed line).

mRNA markers found at tissue level do not always overlap with those found at cell type level

Focusing on the case of heart tissue, we performed the reciprocal analysis to the one depicted in Figure 7 and confirmed that mRNA marker genes for heart tissue from the RNA Atlas were also validated as markers of heart muscle in the Tissue Atlas dataset (Figure 8). However, the majority of mRNA markers for different cardiac cell types from the RNA Atlas are not found back as markers of heart tissue, neither in the RNA Atlas nor the Tissue Atlas datasets (Figure 9).

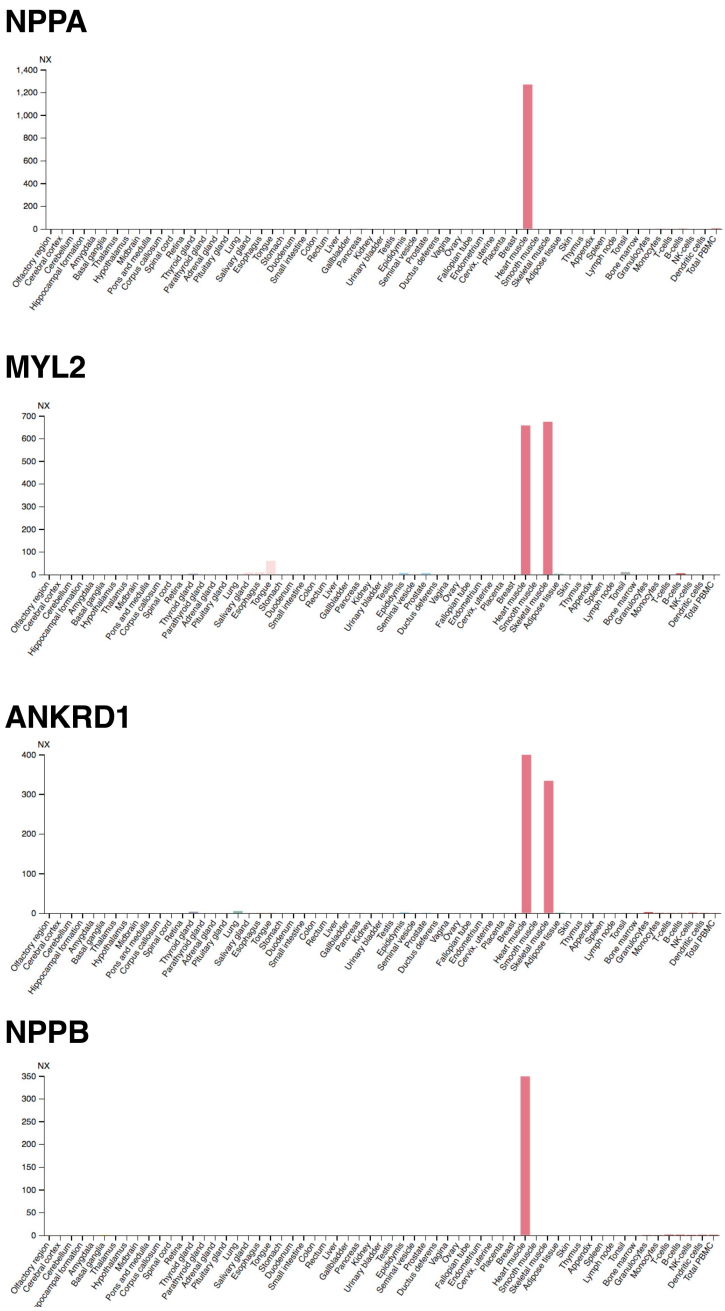


Figure 8 – mRNA markers for heart tissue from the RNA Atlas were validated as markers for heart muscle in the Tissue Atlas dataset. NX = Consensus Normalized eXpression. Image credit: Image credit: Human Protein Atlas. Available from: https://www.proteinatlas.org/ENSG00000175206-NPPA/tissue; https://www.proteinatlas.org/ENSG00000111245-MYL2/tissue; https://www.proteinatlas.org/ENSG00000148677-ANKRD1/tissue; https://www.proteinatlas.org/ENSG00000120937-NPPB/tissue

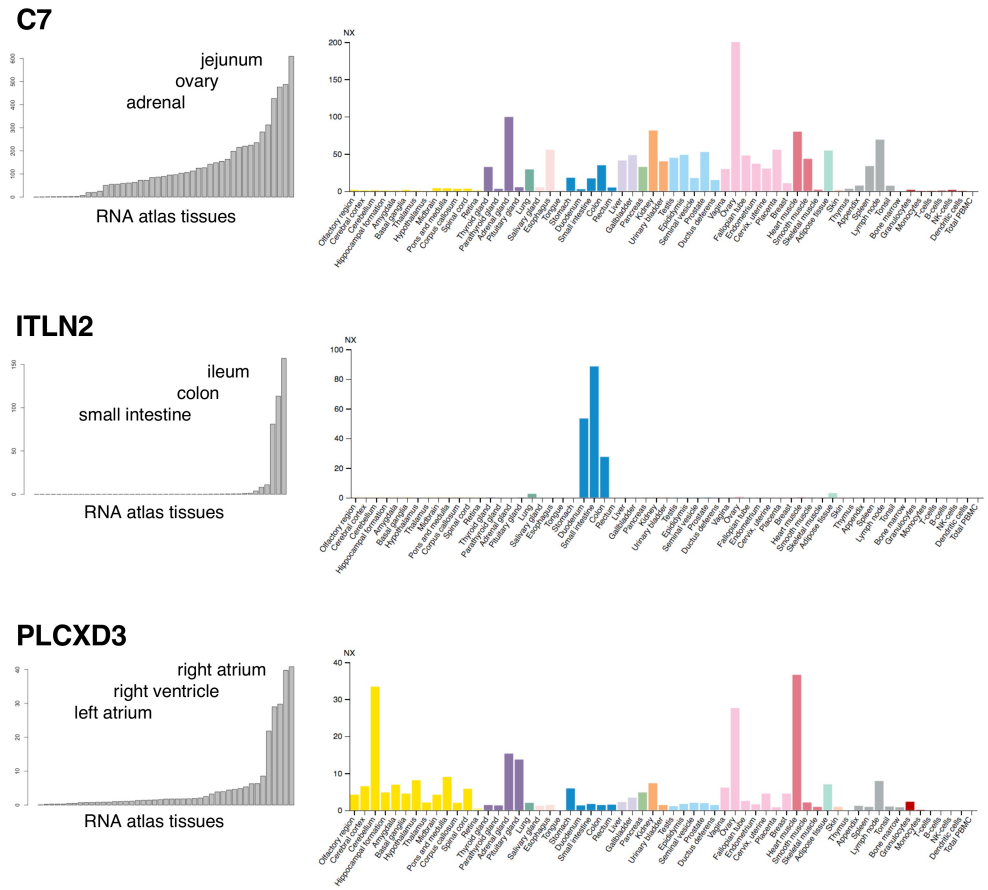


Figure 9 – the majority of mRNA markers for different cardiac cell types (C7, ITLN2 and PLCXD3) are not found back as markers for heart tissue (neither in RNA Atlas nor Tissue Atlas datasets). Grey barplots (left) show the ranked expression values across all tissues in the RNA Atlas, with labels representing the top three tissue were the highest expression values were found. Coloured barplots (right) show NX (= Consensus Normalized eXpression) levels across 55 tissue types and 6 blood cell types, created by combining the data from three independent transcriptomics datasets (HPA, GTEx and FANTOM5). Image credit: Human Protein Atlas. Available from: https://www.proteinatlas.org/ENSG00000112936-C7/tissue; https://www.proteinatlas.org/ENSG00000158764-ITLN2/tissue; https://www.proteinatlas.org/ENSG00000182836-PLCXD3/tissue

Discussion

Here we used the Zipper plot tool to refine the human transcriptome assembly from the RNA Atlas project and, using expression data across 162 normal cell types and 45 tissues from the RNA Atlas project, we investigated the performance of several non-coding RNA fractions (in addition to mRNAs) in the computational deconvolution of healthy tissues.

Surprisingly, Figure 4C revealed a sub-optimal deconvolution performance, apparently being useful only for a subset of the tissues present in the RNA Atlas, and with miRNAs

seemingly carrying no biological signal in any tissue. A plausible explanations for this is that, although comprehensive, the RNA Atlas and UBERON are not complete collections of all tissues and cell types present in the human body. Failure to include just one cell type in the reference matrix being used to deconvolute a tissue can lead to completely distorted results (Figure 6). Furthermore, having only one biological replicate per cell type and tissue in the RNA Atlas is a second limitation. This prevented us from having inherent inter-sample biological variability and thus, the markers that were selected and employed may not be robust in a second dataset.

In order to find out why the deconvolution of heart using mRNAs showed null biological signal (Figure 4C), we used the RNA Atlas to retrieve mRNA markers for cardiac microvascular endothelial cells, cardiac myocytes and cardiac fibroblasts (=cell types found to be present in heart tissue based on UBERON). We ranked those markers by fold change and looked at their expression levels both in the RNA Atlas tissues and in an external dataset (Tissue Atlas). Only one marker contained cardiac tissues among the "top three" tissues with the highest expression (PLCXD3; top 3 in RNA Atlas: right atrium, right ventricle and left atrium; top 3 in Tissue Atlas: heart muscle, cerebellum and ovary) while the other markers did not (C7: jejunum, ovary and adrenal; ITLN2: ileum, colon and small intestine) (Figure 9). However, when doing the marker selection at tissue level rather than at cell type level, results were completely different. We were able to validate heart (mRNA) markers from the RNA Atlas as markers for heart muscle in the Tissue Atlas dataset (Figure 8), complementing the results from Figure 7.

Finally, we also tried to discover the reason why not a single tissue was successfully deconvolved using miRNAs as input. We investigated the fold-change distribution for miRNA markers and found out that using a threshold of at least 5-fold difference resulted in discarding the majority of miRNA markers. When lowering the fold-change threshold from 5 to 2, we obtained a signal of 0.94 for placenta (tissue). The set of ten (ranked) miRNA markers used for this second deconvolution was: hsa-miR-518b, hsa-miR-519d-3p, hsa-miR-520g-3p, hsa-miR-518f-3p, hsa-miR-523-3p, hsa-miR-520d-3p, hsa-miR-373-3p, hsa-miR-371a-5p, hsa-miR-520f-3p, hsa-miR-518d-3p. The first two (hsa-miR-518b and hsa-miR-519d-3p) are well-known placenta-specific microRNAs[22] and showed a fold-change greater than three in placental cell types (villous trophoblast and villous mesenchymal fibroblast, respectively) with respect to the other cell types present in the RNA Atlas dataset.

These analyses highlight a second problem in the deconvolution: not only is it important to include a comprehensive collection of cell types in the reference matrix but also a high quality set of markers.

Conclusion and future perspectives

We refined the initial transcriptome assembled as part of the RNA Atlas and investigated the performance of additional RNA fractions in a computational deconvolution workflow using 162 different normal cell types and 45 tissues from the RNA Atlas project.

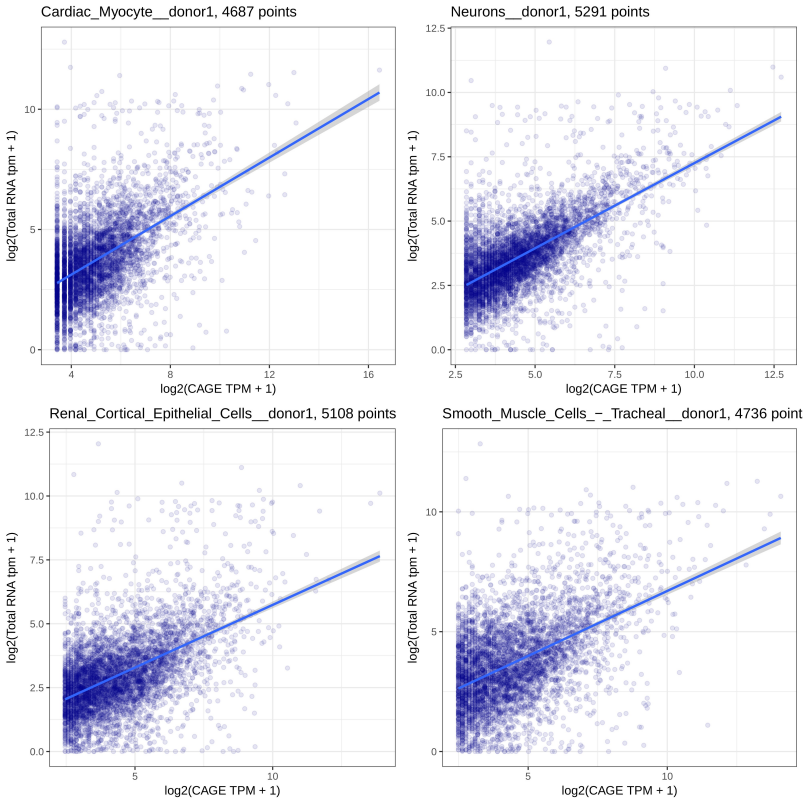
Even though the RNA Atlas is a very comprehensive and very valuable resource for the community, a higher number of biological replicates (it only contains one biological replicate per tissue and cell type in its current form) will improve the results and accuracy of downstream analyses such as computational deconvolution.

Furthermore, only one deconvolution framework has been assessed here: VST-normalized counts transformed into linear scale with the anti-logarithmic function together with the non-negative least squares methodology. However, different combinations of data transformation, normalization and deconvolution method together with different markers and cell types present in the reference matrix have not been yet tested. Such comprehensive evaluation would reveal which combinations lead to optimal results and which ones should not be used.

REFERENCES

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, nrg2484 (2009).
2. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 332825 (2018) doi:10.1101/332825.
3. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19428–19433 (2007).
4. Volders, P.-J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **47**, D135–D139 (2019).
5. Vidigal, J. A. & Ventura, A. The biological functions of miRNAs: lessons from in vivo studies. *Trends Cell Biol.* **25**, 137–147 (2015).
6. Yu, C.-Y. & Kuo, H.-C. The emerging roles and functions of circular RNAs and their generation. *J. Biomed. Sci.* **26**, 29 (2019).
7. Lorenzi, L. *et al.* The RNA Atlas, a single nucleotide resolution map of the human transcriptome. *bioRxiv* 807529 (2019) doi:10.1101/807529.
8. GTEx Portal. https://gtexportal.org/home/.
9. The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas - National Cancer Institute* https://cancergenome.nih.gov/.
10. International Cancer Genome Consortium. https://icgc.org/.
11. Avila Cobos, F. *et al.* Zipper plot: visualizing transcriptional activity of genomic regions. *BMC Bioinformatics* **18**, 231 (2017).
12. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).
13. The FANTOM Consortium and the RIKEN PMI and Clst (dgt) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
14. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
15. S Jupp. A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) Proceedings of SWAT4LS 2015 - Semantic Web Applications and Tools for Life Sciences. in.
16. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
17. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8–9 (2012).
18. Mullen, K. M. & van Stokkum, I. H. M. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4. https://CRAN.R-project.org/package=nnls.
19. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
20. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).
21. Homo sapiens - Ensembl genome browser 86. http://oct2016.archive.ensembl.org/Homo_sapiens/Info/Index.
22. Higashijima, A. *et al.* Characterization of placenta-specific microRNAs in fetal growth restriction pregnancy. *Prenat. Diagn.* **33**, 214–222 (2013).

SUPPLEMENTARY MATERIAL



Supplementary Figure 1 – correlation analysis between Total RNA-seq expression from the RNA Atlas and matching CAGE-seq expression from the FANTOM5 project.

Supplementary Table 1 – description for the selected chromatin states indicative of transcriptional activity

chromatin state number	**abbreviation**	**description**
1	TssA	Active TSS
5	Tx5’	Transcribed - 5' preferential
6	Tx	Strong transcription
7	Tx3’	Transcribed - 3' preferential
9	TxReg	Transcribed & regulatory (Prom/Enh)
10	TxEnh5’	Transcribed 5' preferential and Enhancer
11	TxEnh3’	Transcribed 3' preferential and Enhancer
13	EnhA1	Active Enhancer 1
14	EnhA2	Active Enhancer 2
23	PromBiv	Bivalent Promoter

Supplementary Table 2 – tissues and cell types in the RNA Atlas dataset were matched based on Uber-anatomy ontology (UBERON). (*) Organs containing "immune component" as part of the ontology.

cell type	tissue
human adipose microvascular endothelial cell	adipose
human mesenchymal stem cell-adipose	
human preadipocyte-subcutaneous	
human adrenal microvascular endothelial cell	adrenal
human adrenal cortical cell	
human adrenal fibroblasts	
human bladder microvascular endothelial cell	bladder
human bladder smooth muscle cell	
human urothelial cell	
human neuron	brain
human Schwann cell	
human oligodendrocyte precursor cell	
human astrocyte	
human astrocytes-mid brain	
human brain vascular smooth muscle cell "	
human brain vascular adventitial fibroblasts"	
human brain vascular pericyte	
human choroid plexus epithelial cell	
human choroid plexus fibroblast	
human astrocytes-brain stem cell	total_rna_brain_stem_human_25ug
human astrocyte-cerebellar	total_rna_brain_cerebellum_human_25ug
human astrocyte-hippocampal	total_rna_brain_frontal_cortex_human_25ug
human mammary endothelial cell	breast
human mammary epithelial cell	
human mammary fibroblast	
human colonic smooth muscle cell	colon
human colonic epithelial cell total RNA	
human colonic microvascular endothelial cell	
human esophageal epithelial cell	esophagus

human esophageal fibroblasts	
human esophageal smooth muscle cell	
human cardiac microvascular endothelial cell	heart
human cardiac myocyte	
human cardiac myocyte-adult	
human cardiac fibroblast	
human cardiac fibroblast-adult ventrical	
human cardiac fibroblast-adult atrial	
human cardiac fibroblast adult	
human cardiac fibroblast fetal atrial	
human renal glomerular endothelial cell	kidney
human renal proximal tubular epithelial cell	
human renal cortical epithelial cell	
human renal epithelial cell	
human renal mesangial cell	
human hepatocyte	liver
human hepatic stellate cell	
human mesenchymal stem cell-hepatic	
human hepatic sinusoidal endothelial cell	
human pulmonary microvascular endothelial cell	lung
human pulmonary alveolar epithelial cell	
human pulmonary fibroblast	
human pulmonary fibroblast-adult	
human pulmonary mesenchymal stem cell	
human bronchial epithelial cell	
human bronchial smooth muscle cell	
human pulmonary artery endothelial cell	
human pulmonary artery fibroblast	
human pulmonary artery smooth muscle cell	
human lymphatic endothelial cell	lymph node
human lymphatic fibroblast	
human ovarian surface epithelial cell	ovary
human ovarian fibroblast	
human ovarian microvascular endothelial cell	

human pancreatic stellate cell	pancreas
human pericardial fibroblast	pericardium
human villous trophoblast	placenta
human villous mesenchymal fibroblast	
human amniotic mesenchymal stromal cell	
human amniotic epithelial cell	
human chorionic mesenchymal stromal cell	
human prostate microvascular endothelial cell	prostate
human prostate epithelial cell	
human prostate fibroblast	
human skeletal muscle cell	skeletal muscle
human skeletal muscle satellite cell	
human skeletal muscle myoblast	
human intestinal smooth muscle cell	small intestine
human intestinal fibroblasts	
human splenic endothelial cell	spleen
human splenic fibroblast	
human seminal vesicle microvascular endothelial cell	testicle
human seminal vesicle fibroblast	
human seminal vesicle epithelial cell	
human thymus fibroblasts	thymus
human thyroid fibroblasts	thyroid
human tracheal epithelial cell	trachea
human tracheal smooth muscle cell	
human myometrial microvascular endothelial cell	uterus
human myometrial smooth muscle cell	
human endometrial microvascular endothelial cell	
alveolar macrophage 4	lymph node/spleen/thymus*
alveolar macrophage 7	
monocytes	
iDCs_d6	
mDCs_d7_lps	

cd34+lin- 2x10e6c/via (hematopoietic stem cell)	
cd3+ 2.5x10e6c/vial (T cell)	
cd14+ 2x10e6c/vial (monocyte)	
cd56+ 1.6x10e6c/vial (natural killer)	
cd19+ 1x10e6c/vial (B cell)	
cd10+ 1x10e6c/vial (granulocyte)	
cd14-cd15+ 2x10e6c/vial (granulocyte)	

Supplementary Table 3 – 20 matching tissues across Roadmap, FANTOM5 and RNA Atlas

Roadmap	**FANTOM5**	**RNA Atlas**
Adipose_Nuclei	adipose	adipose
Fetal_Adrenal_Gland	adrenal gland, adult	adrenal
Brain_Angular_Gyrus	brain, adult	brain
Colon_Smooth_Muscle, Colonic_Mucosa, Sigmoid_Colon	colon, adult colon, fetal	colon distal colon proximal colon
Duodenum_Mucosa	duodenum, fetal	duodenum
Esophagus	esophagus, adult	esophagus
Fetal_Heart	heart, adult	heart
Fetal_Kidney	kidney, adult	kidney
Right_Atrium	left atrium, adult	right atrium
Left_Ventricle	left ventricle, adult	left ventricle
Liver	liver, adult	liver
Lung	lung, adult	lung
Ovary	ovary, adult	ovary
Pancreas	pancreas, adult	pancreas
Placenta	placenta, adult	placenta
Skeletal_Muscle_Female	skeletal muscle, adult	skeletal muscle
Small_Intestine	small intestine, adult	small intestine
Stomach_Smooth_Muscle	stomach, fetal	stomach
Spleen	spleen, adult	spleen
Thymus	thymus, adult	thymus

Paper 3

Comprehensive benchmarking of computational deconvolution of transcriptomics data

Francisco Avila Cobos, José Alquicira-Hernandez, Joseph Powell*, Pieter Mestdagh* and Katleen De Preter*

(*) These authors contributed equally to this work.

Contribution: Conceptualization of the manuscript; conducted all the analyses and wrote the manuscript.

Under review

Comprehensive benchmarking of computational deconvolution of transcriptomics data

Francisco Avila Cobos[1,2,3], José Alquicira-Hernandez[3,4], Joseph Powell[3,4],* Pieter Mestdagh[1,2,*] and Katleen De Preter[1,2,*]

* These authors contributed equally to this work. Corresponding author: Katleen.DePreter@UGent.be

[1] Center for Medical Genetics Ghent, Department of Biomolecular Medicine, Ghent University, Ghent, Belgium. [2] Cancer Research Institute Ghent (CRIG), Ghent, Belgium. [3] Garvan Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, Australia. [4] Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

Abstract

Many computational methods to infer cell type proportions from bulk transcriptomics data have been developed. Attempts comparing these methods revealed that the choice of reference marker signatures is far more important than the method itself. However, a thorough evaluation of the combined impact of data transformation, pre-processing, marker selection, cell type composition and choice of methodology on the results is still lacking.

Using different single-cell RNA-sequencing (scRNA-seq) datasets, we generated hundreds of pseudo-bulk mixtures to evaluate the combined impact of these factors on the deconvolution results. Along with methods to perform deconvolution of bulk RNA-seq data we also included five methods specifically designed to infer the cell type composition of bulk data using scRNA-seq data as reference.

Both bulk and single-cell deconvolution methods perform best when applied to data in linear scale and the choice of normalization can have a dramatic impact on the performance of some, but not all methods. Overall, single-cell methods have comparable performance to the best performing bulk methods and bulk methods based on semi-supervised approaches showed higher error and lower correlation values between the computed and the expected proportions. Moreover, failure to include cell types in the reference that are present in a mixture always led to substantially worse results, regardless of any of the previous choices. Taken together, we provide a thorough evaluation of the combined impact of the different factors affecting the computational deconvolution task across different datasets and propose general guidelines to maximize its performance.

Introduction

Since bulk samples of heterogeneous mixtures only represent averaged expression levels (rather than individual measures for each gene across different cell types present in such mixture), many relevant analyses such as differential gene expression are typically confounded by differences in cell type proportions**.** Moreover, understanding differences in cell type composition in diseases such as cancer may enable scientists to identify potentially interesting cellular populations to be targeted therapeutically. For instance, the abundance of tumor infiltrating lymphocytes and other immune cells in solid tumors (also known as the tumor microenvironment) is currently a very active field of research[1–3] (e.g. in the context of immunotherapy) and it has already been shown that accounting for the tumor heterogeneity resulted in more sensitive survival analyses and more accurate tumor subtype predictions[4]. For these reasons, many methodologies to infer proportions of individual cell types (= computational deconvolution) from bulk transcriptomics data have been developed during the last two decades[5] and various methods able to use single-cell RNA-sequencing data have emerged in the past year alone.

Several studies have addressed different factors affecting the deconvolution results but only focused on one or two individual aspects at a time. For instance, Zhong and Liu[6] showed that applying the logarithmic transformation to microarray data led to a consistent under-estimation of cell-type specific expression profiles. Hoffmann *et al.*[7] showed that four different normalization strategies had an impact on the estimation of cell type proportions from microarray data and Newman *et al.*[8] highlighted the importance of accounting for differences in normalization procedures when comparing the results from CIBERSORT[9] and TIMER[10]. Furthermore, Vallania *et al.*[11] observed highly concordant results across different deconvolution methods in both blood and tissue samples, suggesting that the reference matrix was more important than the methodology being used.

Sturm *et al.*[12] already investigated scenarios where reported cell type proportions were higher than expected (spillover effect) or different from zero when a cell type was not present in a mixture (background prediction), possibly caused by related cell types sharing similar signatures or marker genes not being sufficiently cell-type specific. Moreover, they provided a guideline for method selection depending on which cell type of interest needs to be deconvolved. However, each method evaluated in Sturm *et al.* was accompanied by its own reference signature for the different immune cell types, implying that differences may be marker-dependent and not method-dependent. Moreover, they did not evaluate

the effect of data transformation and normalization in these analyses and only focused on immune cell types.

Here we provide a comprehensive and quantitative evaluation of the combined impact of data transformation, scaling/normalization, marker selection, cell type composition and choice of methodology on the deconvolution results. In this study we evaluated the performance of 20 deconvolution methods aimed at computing cell type proportions, including five recently developed methods that use single-cell RNA-sequencing data as reference. The performance is assessed by means of Pearson correlation and root-mean-square error (RMSE) values between the cell type proportions computed by the different deconvolution methods (P_C; computed proportions; Figure 1) and known compositions (P_E; expected proportions) of a thousand pseudo-bulk mixtures from each of four different single cell RNA-sequencing datasets (three from human pancreas and one from peripheral blood mononuclear cells (PBMCs)). Furthermore, to evaluate the robustness of our conclusions, different number of cells (cell pool sizes) were used to build the pseudo-bulk mixtures.

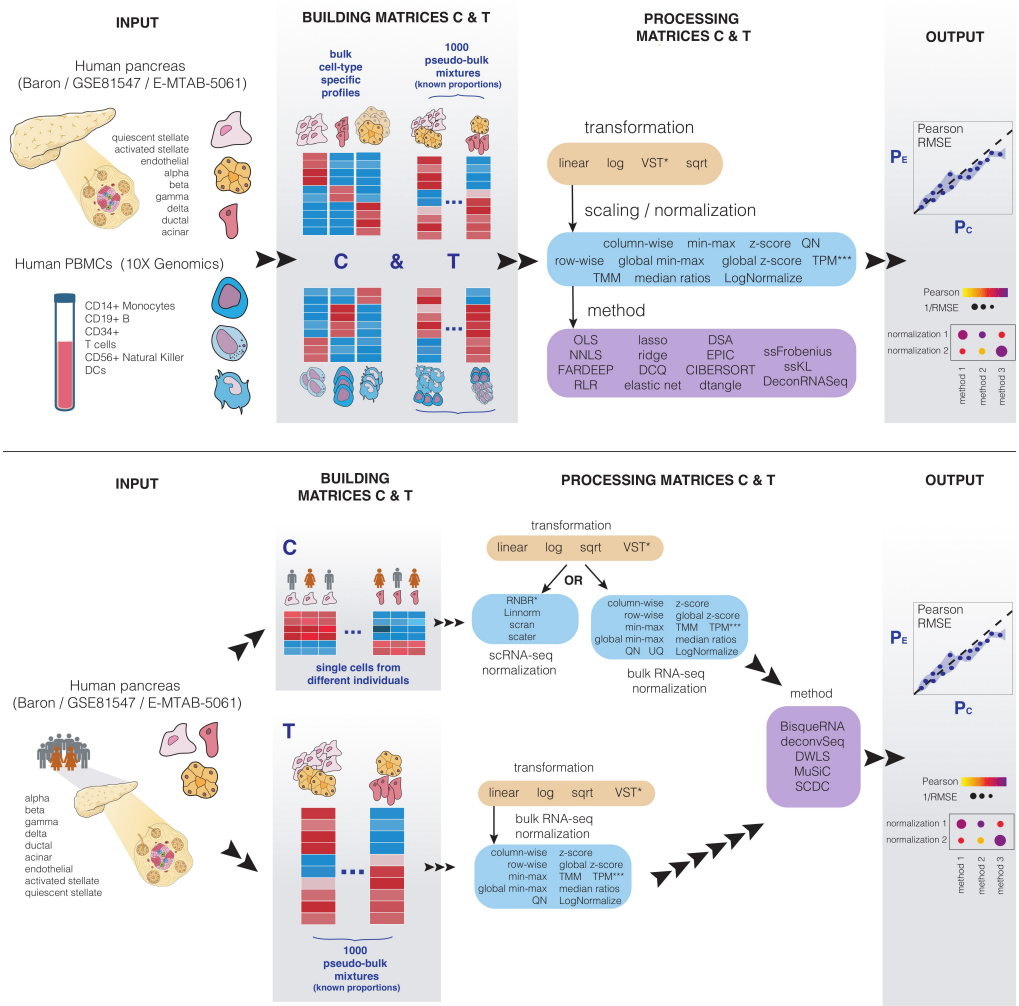


Figure 1 – Schematic representation of the benchmarking study**.** Top panel: workflow for bulk deconvolution methods. Bottom panel: workflow for single-cell methods. In both cases the deconvolution performance is assessed by means of Pearson correlation and root-mean-square error (RMSE). PBMCs = peripheral blood mononuclear cells; log = logarithmic; sqrt = square-root; VST = Variance stabilization transformation. P_E = Expected proportions; P_c = Computed proportions.

Results

Different normalization and methodology combinations have different memory requirements and time consumption

Even though computational resources keep on growing exponentially, memory requirements and time consumption can become important bottlenecks for non-experienced users that may be constrained to limited resources on a personal laptop or for

implementations in clinical settings where short processing times are required. While simple logarithmic (log) and square-root (sqrt) data transformations were performed almost instantaneously in R (between 1 and 5 seconds; see Table 1 for information about the number of cells subject to transformation in each single-cell RNA-seq dataset), the variance stabilization transformation (VST) performed using DESeq2[13] applied to the single-cell RNA-sequencing datasets had high memory requirements and took several minutes to complete (time increasing linearly with respect to the number of cells) (Supplementary Figure 1). Importantly, we used the developer version of DESeq2 v1.25.9, which reduced the running time from quadratic (Suppl. Fig 27 from Soneson *et al.*[14]) to linear with respect of the number of cells.

We further evaluated the impact of different scaling and normalization strategies as well as the choice of deconvolution method. Although the different scaling/normalization strategies consistently have similar memory requirements, RNBR[15] and scran[16] (two single-cell RNA-sequencing specific normalization methods) required up to seven minutes to complete, a 14 fold difference with the other methods, which finished under 30s (Supplementary Figure 2).

The bulk deconvolution methods DSA[17], ssFrobenius and ssKL[18] (all implemented as part of the CellMix[19] R package) had the highest RAM memory requirements, followed by DeconRNASeq[20]. Not surprisingly, the ordinary least squares (OLS[21]) and non-negative least squares (nnls[22]) were the fastest, as they have the simplest optimization problem to solve. For single-cell methods, Dampened Weighted Least Squares (DWLS[23]), which includes an internal marker selection step, resulted in the longest time consumption (6 to 12 hours to complete) whereas MuSiC[24] and SCDC[25] finished in 5 to 10 minutes.

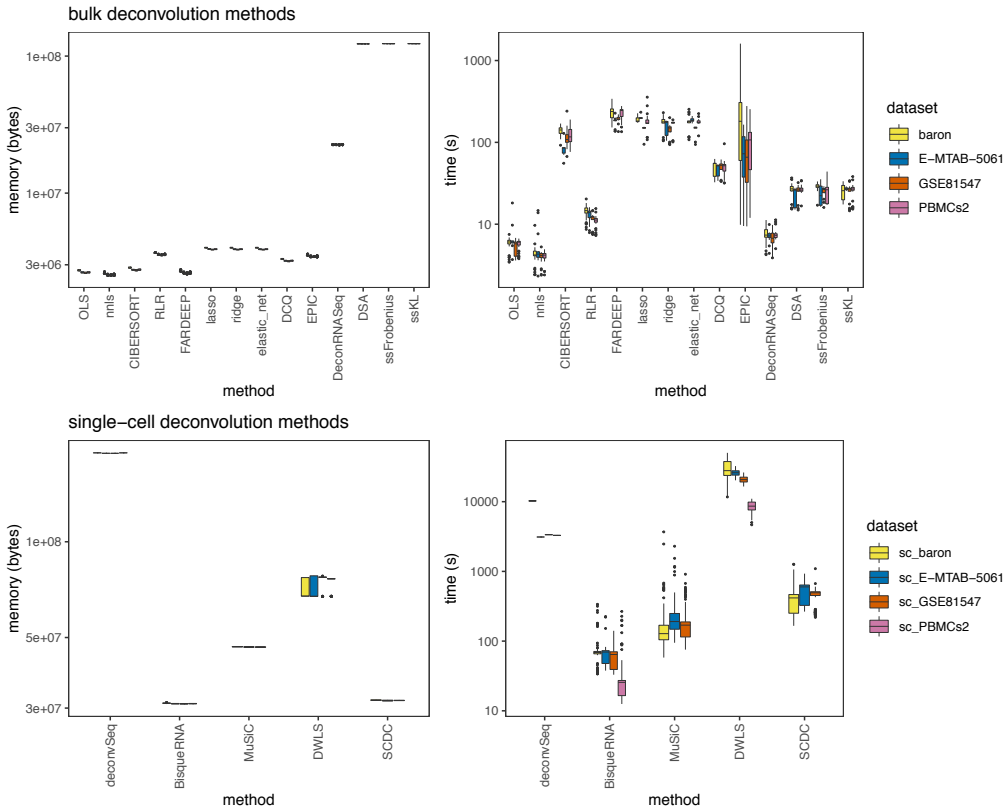


Figure 2 – RAM memory (bytes) and time (seconds) requirements for the different bulk (top panel) and single-cell (bottom panel) deconvolution methodologies across datasets with expression values in linear scale (boxplots depict all scaling/normalization strategies across all pseudo-bulk cell pool sizes).

Data transformation has a dramatic impact on the deconvolution results

While logarithmic and variance-stabilizing (VST) transformations are often used during the pre-processing of omics datasets in the context of differential gene expression analyses [26,27], Zhong and Liu[6] argued against log-transforming the data when performing computational deconvolution. Therefore, we investigated the overall performance of each individual deconvolution method across four different data transformations and all normalization strategies (Figure 3; Supplementary Figures 3-4). Maintaining the data in linear scale ("none" transformation, in grey) consistently showed the best results (lowest RMSE values) whereas the logarithmic (in orange) and VST (in green; which also performs an internal complex logarithmic transformation) scale led to a poorer performance, with two to four-fold higher median RMSE values. For a detailed explanation concerning several bulk and single-cell deconvolution methods that could only be applied with a specific data transformation or dataset, please see Supplementary Methods.

The choice of normalization strategy has a substantial impact on the deconvolution for EPIC[28], DeconRNASeq[20] and DSA[17]. For the remaining methods, the choice of transformation has a higher impact on the deconvolution results than the choice of normalization strategy. These conclusions also hold when repeating the analysis with different pseudo-bulk pool sizes in all datasets tested (collapsing all scaling/normalization strategies and all bulk (Supplementary Figure 5) or single-cell (Supplementary Figure 6) deconvolution methods together). For these reasons, all downstream analyses were performed on datasets in linear scale. Interestingly, five bulk (OLS, nnls, RLR, FARDEEP and CIBERSORT) and three single-cell deconvolution methods (DWLS, MuSiC, SCDC) are able to achieve very accurate cell type proportions in linear scale (median RMSE values lower than 0.05) and there is up to 20 and 19 fold-change difference in RMSE values between the best and worst case for bulk (best: TPM + CIBERSORT; worst: column z-score + EPIC) and single-cell (best: TMM scalingT + no scalingC + DWLS; worst: row scalingT + no scalingC + MuSiC) deconvolution methods, respectively.

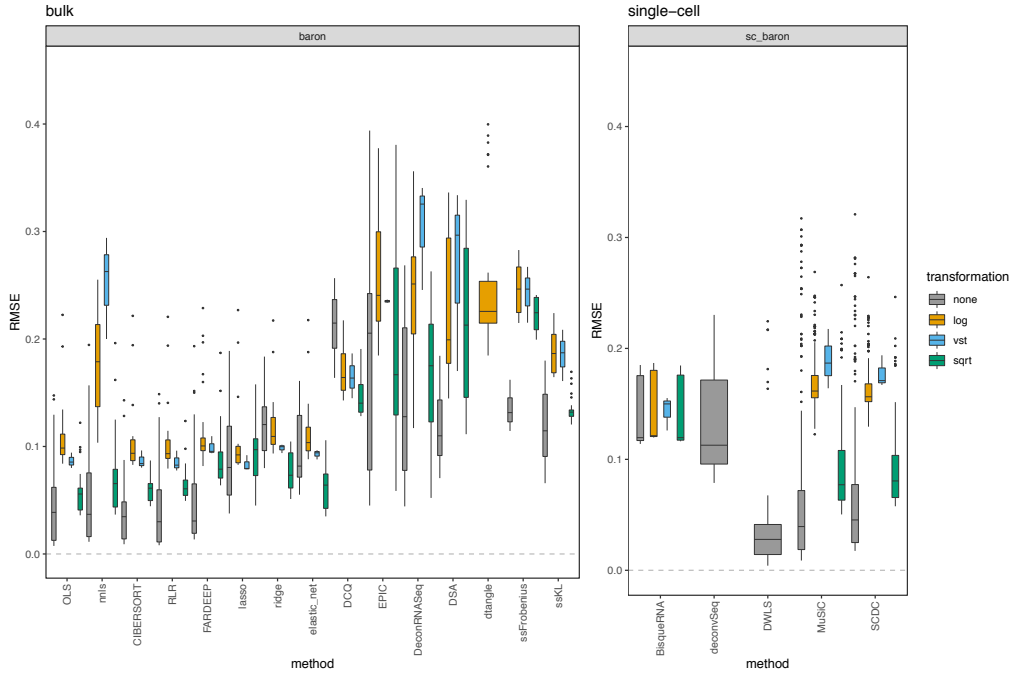


Figure 3 – RMSE values between the known proportions in 1000 pseudo-bulk tissue mixtures from the baron dataset (pool size = 100 cells per mixture) and the predicted proportions from the different bulk (left) and single-cell (right) deconvolution methods. Each boxplot contains all normalization strategies that were tested in combination with a given method.

Different combinations of normalization and deconvolution methodologies reveal important differences in performance

From Figure 3 it is clear that different combinations of normalizations and methodologies lead to substantial differences in performance. Focusing on the data in linear scale, Figure 4 delves into the specific method and normalization combinations evaluated in this manuscript. Among the bulk deconvolution methods, least-squares (OLS, nnls), support-vector (CIBERSORT) and robust regression approaches (RLR/FARDEEP) gave the best results across different datasets and pseudo-bulk cell pool sizes (median RMSE values $<$ 0.05; Figure 4a, Supplementary Fig 7). Regarding the choice of normalization/scaling strategy, column min-max and column z-score consistently led to the worst performance. In all other situations, the choice of normalization/scaling strategy had a minimal impact on the deconvolution results for these methods. Of note, quantile normalization always resulted in sub-optimal results in any of the tested bulk deconvolution methods (Figure 4a,c).

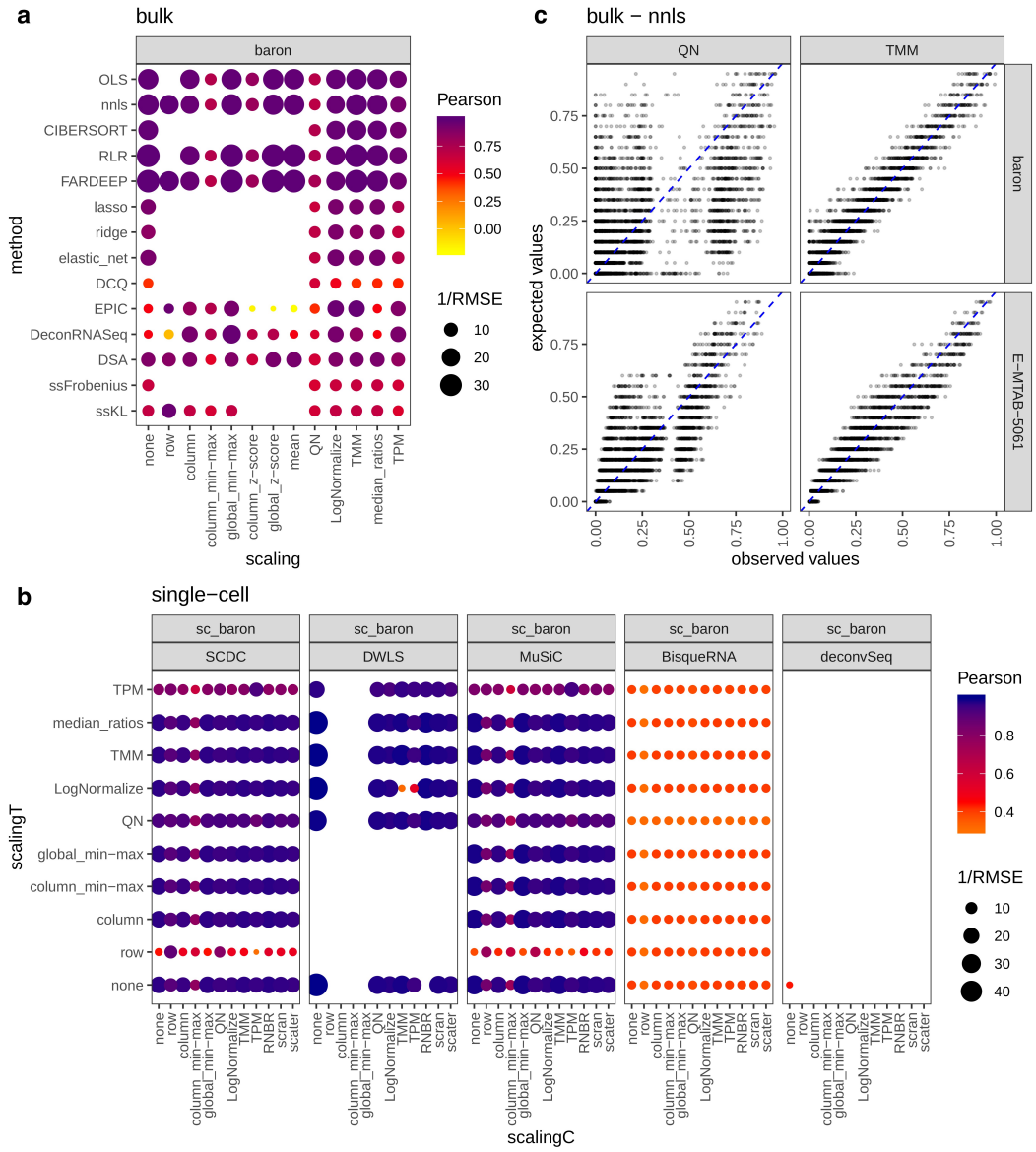


Figure 4 – Pearson correlation values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures in linear scale (pool size = 100 cells per mixture) and the output proportions from the different bulk (a) and single-cell (b) deconvolution methods. The darker the blue and the higher the area of the circle (depicting 1/RMSE) represents higher Pearson and lower RMSE values, respectively. c) Scatter plot showing the impact of the normalization strategy (TMM versus quantile normalization (QN)) comparing the expected proportions (y-axis) and the results obtained through computational deconvolution using nnls (x-axis) for baron and E-MTAB-5061 datasets. Empty locations represent combinations that were not feasible (see Supplementary methods).

Penalized regression approaches including lasso, ridge, elastic net regression and DCQ performed slightly worse than the ones described above (median RMSE ~ 0.1). As stated in its original publication, EPIC assumes transcripts per million (TPM) normalized expression values as input. We indeed observed that the choice of scaling/normalization has a big impact on the performance of EPIC, with TPM giving the best results.

Quadratic programming (DeconRNASeq), Digital Sorting Algorithm (DSA) and the semi-supervised approaches ssKL and ssFrobenius (using only sets of marker genes, in contrast to the supervised counterparts which use a reference matrix with expression values for the markers) showed the poorest performances with the highest root-mean-square errors and lower Pearson correlation values.

For single-cell deconvolution methods (Figure 4c), we evaluated the different combinations of normalization strategies of both the pseudo-bulk mixtures ("scalingT", y-axis) and the single-cell expression matrices ("scalingC", x-axis). DWLS, MuSiC and SCDC consistently showed the highest performance (comparable to the top-performers from the bulk methods, see also Figure 3) across the different choices of normalization strategy (with the exception of row-normalization, column min-max and TPM). While these results are consistent for deconvSeq, MuSiC, DWLS and SDCD regardless of the dataset and pseudo-bulk cell pool size, we observed a substantial performance improvement in BisqueRNA when the pool size increased or when the dataset contained single-cell RNA-sequencing from more individuals (E-MTAB-5061 and GSE81547, with n=6 and 8 respectively) (Supplementary Figure 8). Note that it was not feasible to evaluate all combinations (empty locations in the grid), see Supplementary methods for a detailed explanation.

The set of markers used in bulk deconvolution methods impacts deconvolution results

Based on the previous results, we wanted to evaluate whether different marker selection strategies had an impact on the deconvolution results starting from bulk expression data in linear scale. To that end we assessed the impact of eight different marker selection strategies (see Methods) on the deconvolution results using bulk deconvolution methods (Figure 5, Supplementary Figure 9). This analysis was not done with the single-cell methods because they do not require marker genes to be known prior to performing the deconvolution.

The use of all possible markers ("all" strategy) showed the best performance overall, followed by positive fold-change markers ("pos_fc"; negative fold-change markers are those

with small expression values in the cell type of interest and high values in all the others) or those on the top 50% of average expression values ("top_50p_AveExpr") or log fold-changes ("top_50p_logFC"). As expected, the use of random sets of 5 markers per cell type ("random5"; negative control in our setting) was consistently the worst choice across all datasets regardless of the deconvolution method. Using the bottom 50% of the markers per cell type based on average expression levels ("bottom_50p_AveExpr") or log fold changes ("bottom_50p_logFC) also led to sub-optimal results. Specifically in the baron and PBMC datasets, the use of the top 2 markers per cell type ("top_n2") led to a) optimal results when used with DSA; b) similar results as using the bottom_50p_AveExpr or bottom_50p_logFC with ordinary linear regression strategies; c) worse results than random when used with penalized regression strategies (lasso, ridge, elastic_net, DCQ) and CIBERSORT.

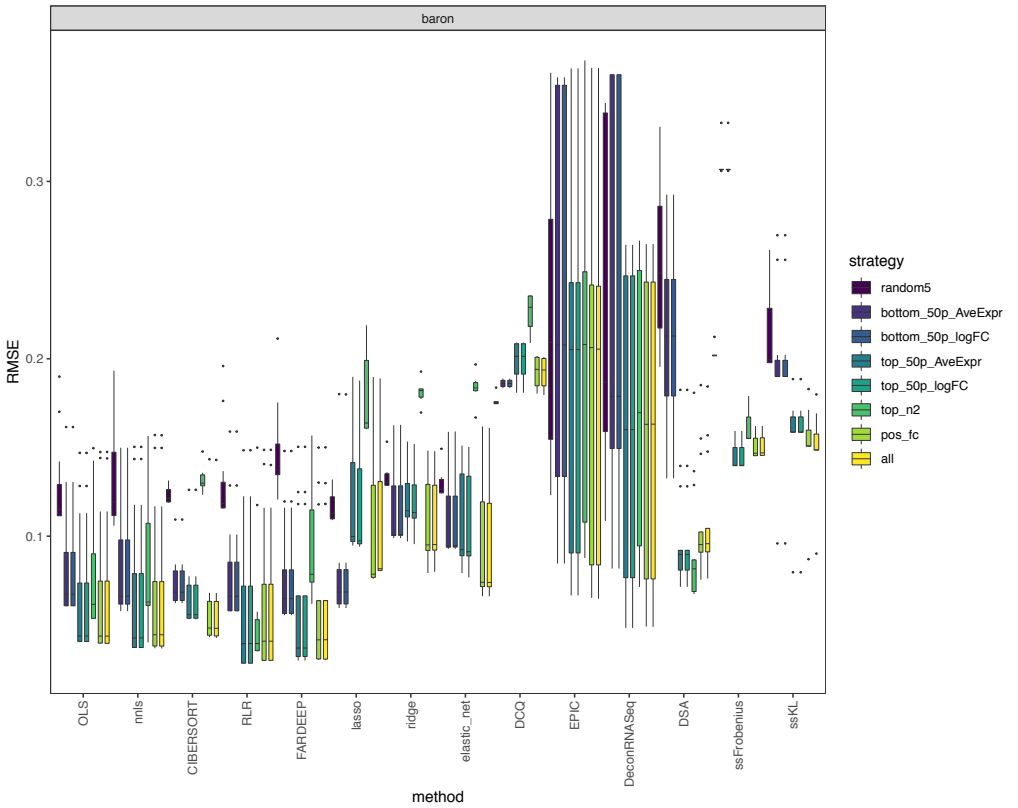


Figure 5 – RMSE values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures (linear scale; pool size = 100 cells per mixture) and the output proportions from the baron dataset, using eight different marker selection strategies. Each boxplot contains all normalization strategies that were tested in combination with a given marker strategy across the different bulk deconvolution methods.

Removing cell types from the reference matrix results in substantially worse deconvolution results compared to reference matrices composed of all cell types present in the mixtures

Based on the results from all the analyses thus far, we decided to evaluate the impact of removing cell types with the data in linear scale and using all available markers ("all" marker selection strategy). Furthermore, we selected nnls and CIBERSORT as representative top-performing bulk deconvolution methods and DWLS and MuSiC as top-performing single-cell methods. To also be able to evaluate the impact of the normalization strategy, we included a representative sample of normalization strategies that result in small RMSE and high Pearson correlation values (see Figure 4 and Supplementary Figures 7-8): column, median ratios, none, TMM and TPM for nnls and CIBERSORT; column, scater, scran, none, TMM and TPM for DWLS and MuSiC.

We assessed the impact of removing a specific cell type by comparing the absolute RMSE values between the ideal scenario where the reference matrix contains all the cell types present in the pseudo-bulk mixtures (leftmost column in Figures 6a-b and 7a-b, with grey label "none") and the RMSE values obtained after removing one cell type at a time from the reference (all other grey labels).

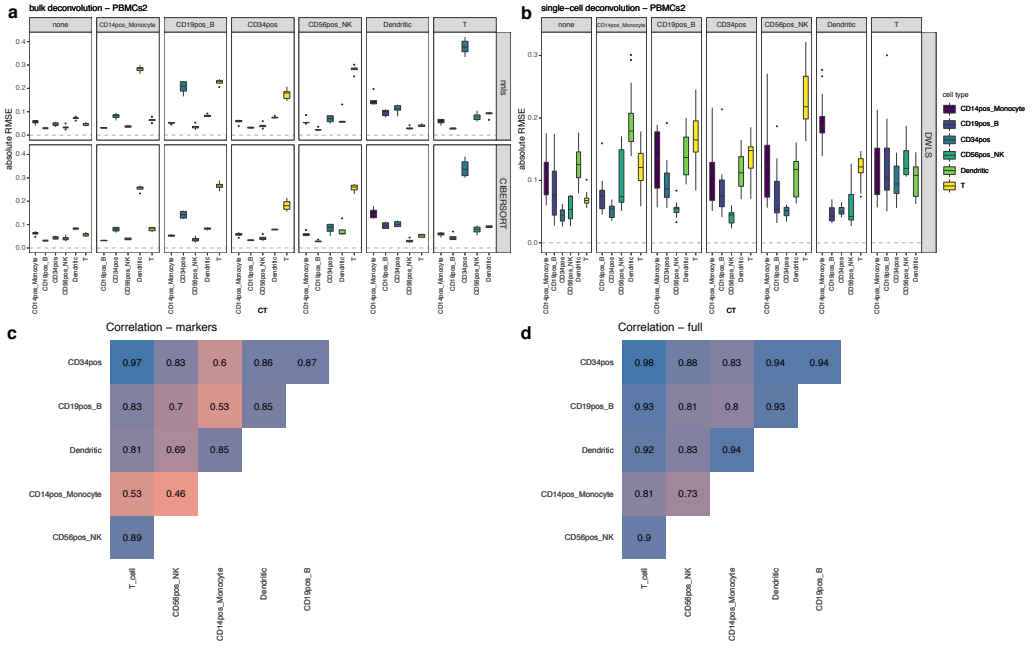


Figure 6 – Effect of cell type removal on the deconvolution results using the PBMCs dataset [100-cell pseudo-bulk mixtures in linear scale]. a) results using bulk deconvolution methods (nnls and CIBERSORT); b) results using single-cell deconvolution methods (only DWLS because the scRNA-seq data comes from only one individual); c) pairwise Pearson correlation values between expression profiles for the different cell types, using a subset of the reference matrix containing only the markers used in the bulk deconvolution; d) pairwise Pearson correlation values between complete expression profiles for the different cell types. In a) and b), each grey column represents a specific cell type removed. Each data point conforming a boxplot represents a different scaling/normalization strategy used.

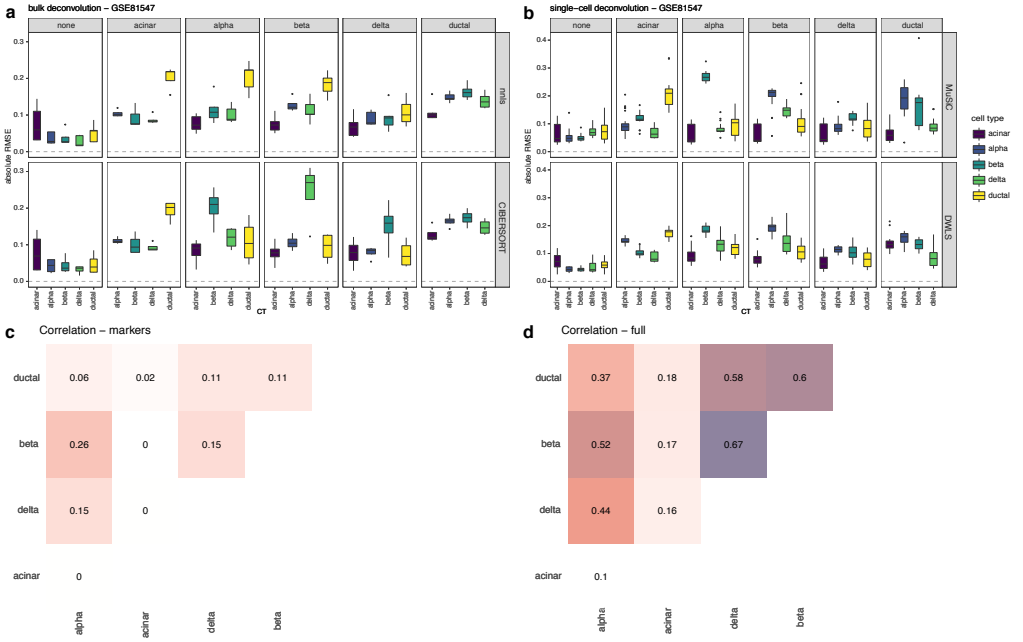


Figure 7 – Effect of cell type removal on the deconvolution results using the GSE81547 dataset [100-cell pseudo-bulk mixtures in linear scale]. a) results using bulk deconvolution methods (nnls and CIBERSORT); d) results using single-cell deconvolution methods (MuSiC and DWLS); c) pairwise Pearson correlation values between expression profiles for the different cell types, using a subset of the reference matrix containing only the markers used in the bulk deconvolution; d) pairwise Pearson correlation values between complete expression profiles for the different cell types. In a) and b), each grey column represents a specific cell type removed. Each data point conforming a boxplot represents a different scaling/normalization strategy used.

We then focussed on those cases where the median absolute RMSE values between the results using the complete reference matrix (depicted as "none" in Figures 6a-b and 7a-b) and all other scenarios where a cell type was removed, increased at least 2-fold. In the PBMC dataset (Fig 6a-b), removing CD19+, CD34+, CD14+ or NK cells had an impact on the computed T-cell proportions (between a three and six-fold increase in the median absolute RMSE values, both in bulk and single-cell deconvolution methods). The GSE81547 dataset (Figure 7a-b) shows that removing acinar cells has a dramatic impact in all other cell type proportions. Supplementary Figures 10 and 11 showed the results for baron and E-MTAB-5061 datasets, respectively. Remarkably, no method and normalization combination was able to provide accurate cell type proportion estimates when the reference missed a cell type.

To investigate whether the proportion of the omitted cell type was re-distributed equally among all remaining cell types or only among those that are transcriptionally most similar, we computed pairwise Pearson correlation values between the expression profiles of the

different cell types (Figure 6c-d and Figure 7c-d). Figure 6c-d shows that CD14+ monocytes were mostly correlated with dendritic cells (Pearson = 0.85 when computing pairwise correlations on the reference matrix containing only marker genes and 0.94 when using the complete expression profiles from all cell types, respectively) and Figure 6a-b shows that, when removing CD14+ monocytes, the highest RMSE value was found in dendritic cells. Figure 7c-d shows that acinar cells are not correlated with any other cell type (Pearson values close to zero with all other cell types) and Figure 7a-b shows that, when removing acinar cells, all cell type proportions estimates have higher RMSE values compared to the case where no cell type is missing ("none", leftmost panel).

For the baron dataset (Supplementary Figure 10): the removal of ductal cells (highest correlation with quiescent stellate and endothelial cells) led to highest RMSE values for both quiescent stellate and endothelial cells while the removal of endothelial cells (mostly correlated with quiescent stellate, beta and ductal cells) led to the highest RMSE values for quiescent, ductal and beta cells. For the E-MTAB-5061 dataset (Supplementary Figure 11): no cell type is correlated to one another and removing any cell type from the reference matrix led to distorted proportions for all other cell types.

Discussion

Using both Pearson correlation and RMSE values as measures of the deconvolution performance, we comprehensively evaluated the combined impact of four data transformations, twenty scaling/normalization strategies, seven marker selection approaches and twenty different deconvolution methodologies on four different single-cell RNA-seq datasets. These datasets encompass two different biological sample types (human pancreas and peripheral blood mononuclear cells) and three different sequencing protocols (CEL-Seq, Smart-Seq 2 and GemCode Single-Cell 3′). Additionally, we assessed the impact of using different number of cells when making the pseudo-bulk mixtures and the impact of removing cell types from the reference matrix that were actually present in the mixtures. Remarkably, each dataset was split into train and testing fractions in a "sample-agnostic" manner that took into account the cell number distribution across cell types, generated cell type pools including cells coming from different individuals and prevented cells to be simultaneously present in both train and test fractions. By doing so, realistic intra-cell type and inter-sample variability were retained in both train and test fractions.

Even though the four datasets used throughout this manuscript encompass different sequencing protocols that led to hundred-fold differences in the number of reads sequenced

per cell (Table 1), our findings were consistent regardless of the dataset being evaluated or the number of cells used to make the pseudo-bulk mixtures.

The logarithmic transformation is routinely included as a part of the pre-processing of omics data in the context of differential gene expression analysis [26,27], but Zhong and Liu[6] showed that it led to worse results than performing computational deconvolution in the linear (un-transformed) scale. Silverman *et al.*[29] showed that using log counts per million with sparse data strongly distorts the difference between zero and non-zero values and Townes *et al.*[30] showed the same when log-normalizing UMIs. Tsoucas *et al.*[23] showed that when the data was kept in the linear scale, all combinations of three deconvolution methods (DWLS, QP or SVR) and three normalization approaches (LogNormalize from Seurat, Scran or SCnorm) led to a good performance, which was not the case when the data was log-transformed. Here, we assessed the impact of the log transformation on both full-length and tag-based scRNA-seq quantification methods and confirmed that the computational deconvolution should be performed on linear scale to achieve the best performance.

Data scaling or normalization is a key pre-processing step when analysing gene expression data. Data scaling approaches transform the data into bounded intervals such as [0, 1] or [-1, +1]. While being relatively easy and fast to compute, scaling is sensitive to extreme values. Therefore, other normalization strategies that do not result in bounded intervals may be preferred. In the context of transcriptomics, the term "normalization" refers to removing biases that may have been introduced in the data while being generated and is needed to only keep true differences in expression. Normalizations such as TPM aim at removing differences in sequencing depth among the samples. We refer the reader to Evans *et al.*[31], for an in-depth analysis of RNA-seq normalization methods. Vallania *et al.* [11] assessed the impact of standardizing (= substracting the mean and dividing by the standard deviation) both the bulk and reference expression profiles into z-scores prior to deconvolution, which is performed by CIBERSORT but not in other methods. They observed high pairwise correlations between the estimated cell type proportions with and without standardizing the data, suggesting a neglectable effect. However, a high Pearson correlation value is not always synonym of a good performance. As already pointed out by Hao *et al.*[32], high Pearson correlation values can arise when the proportion estimations are accurate (low RMSE values) but also when the proportions differ substantially (high RMSE values), making the correlation metric alone not sufficient to assess the deconvolution performance. Both for bulk and single-cell deconvolution methods, our analyses show that the normalization strategy had little impact (except for EPIC, DeconRNASeq and DSA bulk methods). Of note, quantile normalization (QN), an

approach used by default in several deconvolution methods (e.g. FARDEEP, CIBERSORT), consistently showed sub-optimal performance regardless of the method.

Schelker *et al.*[33] and Racle *et al.*[28] showed that the origin of the expression profiles had also a dramatic impact on the results, revealing the need of using appropriate cell types coming from niches similar to the bulk being investigated.

Hunt *et al.*[34] showed that a good deconvolution performance was achieved if the markers being used were predominantly expressed in only one cell type and with the expression in other cell types being in the bottom 25%. Monaco *et al.*[35] showed similar conclusions when the reference matrix was pre-filtered by removing markers with small log fold change between the first and second cell types with highest expression. In our analyses, markers were selected based on the fold change with respect to the cell type with the second highest expression. Therefore, the pre-filtering proposed by Hunt *et al.* and Monaco *et al.* was already implicitly done. Furthermore, when sub-setting the markers based on their average gene expression or fold changes, those in the top fifty percent led to smaller RMSEs compared to those in the bottom fifty percent (Figure 5).

Wang *et al.*[24] explored the effect of removing one immune cell type at a time from the reference matrix on the estimation accuracy using artificial bulk expression of six pancreatic cell types (alpha, beta, delta, gamma, acinar and ductal) and removing one cell type from the single-cell expression dataset. They observed that, when a cell type was missing in the reference matrix, MuSiC, NNLS and CIBERSORT did not produce accurate proportions for the remaining cell types. Gong and Szustakowski[20] also investigated this issue by performing a first deconvolution using DeconRNASeq, then removing the least abundant cell population from the reference/basis matrix, and finally repeating the deconvolution with the new matrix. They observed an uneven redistribution of the signal and observed that some initial proportions became smaller. Moreover, Schelker *et al.*[33] investigated this phenomenon by looking at the correlation coefficient between the results obtained with the complete reference matrix and the results removing one cell type at a time.

We performed similar analyses for four deconvolution methods (two bulk and two single-cell) and eleven normalization strategies (five for bulk, six for single-cell) on three single-cell human pancreas and one PBMC dataset, keeping the data in linear scale. We observed both cases where the choice of normalization strategy had no impact and other cases where it did. Interestingly, the removal of specific cell types did not affect all other cell types equally. Both bulk and single-cell deconvolution methods showed similar trends when

removing specific cell types. However, there were some discrepancies in the RMSE values (e.g. removal of beta cells had a substantial impact on the proportions of delta cells but CIBERSORT showed three times higher RMSE values compared to either nnls, MuSiC or DWLS). This may be explained by the fact that for bulk deconvolution methods, we removed both the cell type expression profile and its marker genes from the reference matrix whereas for the single-cell methods, only the cells from the specific cell type were excluded, without applying extra filtering on the genes (MuSiC, SCDC) or because a different signature was internally built (DWLS).

Schelker *et al.* found that B cell and dendritic cell proportions were affected by removing macrophages or monocytes whereas NK cell proportions were affected by removing T cells. Sturm *et al.*, also reported the impact of removing CD8+ T cells on NK cell proportions. Our results on the PBMC dataset agree with those from Schelker *et al.* and Sturm *et al.* but also include novel insights: removing CD19+ B-cells, CD34+, CD14+ monocytes or NK cells had an impact on the computed T-cell proportions and removing CD19+ B-cells, CD56+ NK or T cells had an impact on CD34+ cell proportions.

Furthermore, we found a direct association between the correlation values among the cell types present in the mixtures and the effect of removing a cell type from the reference matrices. Specifically, we hypothesize that: a) removing a cell type that is barely or completely uncorrelated (Pearson < 0.2) to all other cell types remaining in the reference matrix has a dramatic impact in the cell type proportions of all other cell types; b) removing a cell type that was strongly positively correlated (Pearson > 0.6) with one or more cell types still present in the reference matrix leads to distorted estimates for the most correlated cell type(s).

EPIC[28] shows a first attempt in alleviating this problem by considering an unknown cell type present in the mixture. Nevertheless this is currently restricted to a cancer setting, using markers of non-malignant cells that are not expressed in cancer cells.

Conclusion and future perspectives

The three most relevant factors affecting the deconvolution results are: i) the data transformation, ii) all cell types being part of the mixtures must be represented in the reference matrix and, for bulk deconvolution methods, iii) a sensible marker selection strategy.

When performing a deconvolution task, we advise users to: a) keep their input data in linear scale; b) select any of the scaling/normalization approaches described here with

exception of row scaling, column min-max, column z-score or quantile normalization; c) choose a regression-based bulk deconvolution method (e.g. nnls, CIBERSORT or FARDEEP) and also perform the same task in parallel with DWLS, MuSiC or SCDC if single-cell data is available; d) use a stringent marker selection strategy that focuses on differences between the first and second cell types with highest expression values; e) use a comprehensive reference matrix that include all relevant cell types present in the mixtures.

Finally, as more scRNA-seq datasets become available in the near future, its aggregation (while carefully removing batch effects) will increase the robustness of the reference matrices being used in the deconvolution and will fuel the development of methodologies similar to SCDC, which allows direct usage of more than one scRNA-seq dataset at a time.

Methods

Dataset selection and quality control

Four different datasets coming from different single-cell isolation techniques (FACS and droplet-based microfluidics) and encompassing both full-length (Smart-Seq2) and tag-based library preparation protocols (3'-end with UMIs) were used throughout this article (see Table 1). After removing all genes (rows) full of zeroes or with zero variance, those cells (columns) with library size, mitochondrial content or ribosomal content further than three median absolute deviations (MADs) away were discarded. Next, only genes with at least 5% of all cells (regardless of the cell type) with a UMI or read count greater than 1 were kept. Finally, we retained cell types with at least 50 cells passing the quality control step and, by setting a fixed seed and taking into account the number of cells across the different cell types, each dataset was further split into "training" and "testing" datasets with a similar distribution of cells per cell type.

Regarding E-MTAB-5061: cells with "not_applicable", "unclassified" and "co-expression_cell" labels were excluded and only cells coming from six healthy patients (non-diabetic) were kept. After quality control, we made two-dimensional t-SNE plots for each dataset. When adding coloured labels both by cell type and donor (Suppl. Fig 12), the plots showed consistent clustering by cell type rather than by donor, indicating an absence of batch effects.

Table 1 – Details of the four datasets used. (*) Since this dataset originally contained six closely related T-cell subtypes (and other people have failed in their attempts of distinguishing them[36,37]) we re-labelled all cells from these sub-types as "T cells". Moreover, to reduce the memory and time requirements needed to run all combinations of data transformation, normalization and methodology, we randomly selected 10,000 cells out of the original 68,000. (**) 10X genomics data is not in a public repository but available at: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a

Dataset	Biological sample type	Sequencing protocol	Number of individual samples	Number of cell types	Number of cells after QC	Number of genomic features after QC	Median total counts per cell after QC	Median number of non-zero features per cell after QC	Ref
Baron (GSE84133)	Human pancreatic islands	inDrop platform + CEL-Seq protocol	4 (2 male, 2 female)	10	7692	8386	4856	1723	[38]
E-MTAB-5061	Human pancreatic tissue and islets	FACS sorting into 384-well plates + Smart-Seq2	6 (5 male, 1 female)	6	908	13899	329217	5521	[39]
GSE81547	Human pancreatic tissue	FACS sorting into 96-well plates + Smart-Seq2	8 (6 male, 2 female)	5	2068	11694	481825	3072	[40]
PBMCs**	Human fresh peripheral blood mononuclear cells	Chromium GemCode Single-Cell Instrument + GemCode Single-Cell 3′ Gel Bead and Library Kit (10x Genomics)	1	6*	10000*	2175	1142	401	[41]

Generation of reference matrices for the deconvolution

Using the "training" splits from the previous section, the mean count across all individual cells from each cell type was computed for each gene, constituting the original (un-transformed and un-normalized) reference matrix (C in equation (I) from section "Computational deconvolution: formulation and methodologies") and were used as input for the bulk deconvolution methods described in that section. Importantly, the "training" splits without applying the mean collapsing step were used by the single-cell deconvolution methods and for the marker selection step.

Cell-type specific marker selection

TMM normalization (edgeR package[42]) was applied to the original (linear) scRNA-seq expression datasets and limma-voom[43] was used to find out marker genes. Only genes with positive count values in at least 30% of the cells of at least one cell type were retained. Among the retained ones, those with absolute fold changes greater or equal to 2 between the first and second cell types with highest expression and BH adj p-value < 0.05 were kept as markers in all three pancreatic datasets. Since the PBMCs contained more closely related cell types, the fold-change threshold was lowered to 1.5.

Once the set of markers was retrieved, the following approaches were evaluated: i) "all": use of all markers found following the procedure described in the previous paragraph; ii) "pos_fc": using only markers with positive fold-change (=over-expressed in cell type of interest; negative fold-change markers are those with small expression values in the cell type of interest and high values in all the others); iii) "top_n2": using the top 2 genes per cell type with the highest log fold-change; iv) "top_50p_logFC": top 50% of markers (per CT) based on log fold-change; v) "bottom_50p_logFC": bottom 50% of markers based on log fold-change; vi) "top_50p_AveExpr": top 50% of markers based on average gene expression (baseline expression); vii) "bottom_50p_AveExpr": low 50% based on average gene expression; viii) "random5": for each cell type present in the reference, five genes that passed quality control and filtering were randomly selected as markers.

Generation of thousands of artificial pseudo-bulk mixtures

Using the "testing" datasets from the quality control step, we generated matrices containing 1,000 pseudo-bulk mixtures (matrix T in equation (I) from **"**Computational deconvolution: formulation and methodologies") by adding up count values from the randomly selected individual cells. The minimum number of cells used to create the pseudo-bulk mixtures (pool size) was 100 and the maximum was determined by the second most abundant cell type (rounded down to the closest hundred, to avoid non-integer numbers of cells) in each of the four datasets. When the difference between the minimum and maximum values was greater than or equal to 200, three different pool sizes were created by rounding up the mean value between both extremes to the closest hundred (n = 100, 700 and 1200 for Baron; n = 100, 300 and 400 for PBMCs). Due to this constraint, only two pool sizes were feasible for GSE81547 (n = 100 and 200) and one for E-MTAB-5061 (n = 100). Each (feasible) pseudo-bulk mixture was created by randomly selecting the number of cell types to be present (between 3, 4 and 5) and their identities, followed by choosing the cell type proportion assigned to each cell type (enforcing a sum-to-one

constraint) among all possible proportions between 0.05 and 1, in increasing intervals of 0.05. Finally, once the amount of cells to be picked up from specific cell types was determined, the cells were randomly selected (without replacement).

Data transformation and normalization

The next step is applying four different data transformations to: i) the un-transformed and un-normalized reference matrix C; ii) the un-transformed and un-normalized single-cell "training" splits and iii) the un-transformed and un-normalized matrix T containing the 1000 pseudo-bulk mixtures.

Since count data from both bulk and single-cell RNA-seq show the phenomenon of over-dispersion[42,44], the following data transformations were chosen: a) leave the data in the original (linear) scale; b) use the natural logarithmic transformation (with the log1p function in R[45]) ; c) use the square-root transformation; d) variance-stabilizing transformation (VST). The second and third are simple and commonly used transformations aiming at reducing the skewness in the data due to the presence of extreme values[27] and stabilizing the variance of Poisson-distributed counts[46], respectively. VST (using the varianceStabilizingTransformation function from DESeq2) removes the dependence of the variance on the mean, especially important for low count values, while simultaneously normalizing with respect to library size[13].

Each transformed output file was further scaled/normalized with the approaches listed on Table 2. The mathematical implementation can be found at the original publications ("Ref" column) and in our GitHub repository (http://github.com/favilaco/deconv_benchmark). Due to the sparsity of the single-cell RNA-seq matrices (most genes with zero counts), the UQ normalization failed (all normalization factors were infinite or NA values) and thus was eventually not included in downstream analyses. TMM includes an additional step that uses the normalization factors to obtain normalized counts per million. LogNormalize and Linnorm include an additional exponentiation scale after normalization in order to transform the output data back into linear scale. Median of ratios can only be applied to integer counts in linear scale.

Table 2 – Detailed description of different scaling/normalization approaches used in the benchmarking

Scaling/normalization method	Single-cell specific	Output containing negative values	Output bounded in [0,1] interval	Reference
Column-wise (="Total count" or library size normalization)	no	no	yes	[47]
Column min-max	no	no	yes	[48]
Column z-score	no	yes	no	[49]
Row-wise	no	no	yes	[50]
Global min-max	no	no	yes	[48]
Global z-score	no	yes	no	[49]
Quantile normalization (QN)	no	no	no	[51]
Upper quartile (UQ)	no	no	no	[52]
Transcripts per million (TPM)	no	no	no	[53]
Trimmed mean of M-values (TMM)	no	no	no	[54]
LogNormalize	no	no	no	[55]
Median of ratios	no	no	no	[13]
Scran	yes	no	no	[16]
Scater	yes	no	no	[56]
Linnorm	yes	no	no	[57]
RNBR	yes	no	no	[15]

Computational deconvolution: formulation and methodologies

The deconvolution problem can be formulated as $T = C \cdot P$ (I) [5] , where T = measured expression values from bulk heterogeneous samples; C = cell type-specific expression values and P = cell-type proportions. Specifically, T represents the 1000 pseudo-bulk mixtures from "Generation of thousands of artificial pseudo-bulk mixtures" and C is the reference matrix from "Cell-type specific marker selection and generation of reference matrices for the deconvolution". In the context of this article, the goal is to obtain P using T and C as input.

Fifteen bulk deconvolution methods a have been evaluated, including two traditional (ordinary least squares (OLS[21]) and non-negative least squares (NNLS[22])) and one weighted least squares method (EPIC[28]); two robust regression (FARDEEP[58], RLR[59]), one support-vector regression (CIBERSORT[9]) and four penalized regression (ridge, lasso, elastic net[60] and Digital Cell Quantifier (DCQ[61])) approaches; one quadratic programming (DeconRNASeq[20]), one method that models the problem in logarithmic scale (dtangle[34]) and three methods included in the CellMix R package[19]: Digital Sorting Algorithm (DSA[17]) and two semi-supervised non-negative matrix factorization methods (ssKL and ssFrobenius[18]). Furthermore, five single-cell deconvolution methods have been evaluated: deconvSeq[62], MuSiC[24], DWLS[23], Bisque[63] and SCDC[25]. We refer the reader the original publications and our Github repository (http://github.com/favilaco/deconv_benchmark) for details about their implementation.

Measures of deconvolution performance

Changes in memory were assessed with the mem_change function from the pryr package[64] and the elapsed time was measured with the proc.time function (both functions executed in R v.3.6.0).

We computed both the Pearson correlation values and the root-mean-square error (RMSE) between cell type proportions from thousands of pseudo-bulk mixtures with known composition and the output from different deconvolution methods for each combination of data transformation, scaling/normalization choice and deconvolution method. Higher Pearson correlation and low RMSE values correspond to a better deconvolution performance.

Evaluation of missing cell types in the reference matrix C

For every cell type removed, the deconvolution was applied only to mixtures where the missing cell type was originally present. For bulk deconvolution methods, the marker genes of the cell type that was removed from the reference were also excluded (single-cell methods did not require a priori marker information).

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to acknowledge Evan Benn, Derrick Lin, Vikkitharan Gnanasambandapillai and Manuel Sopena Ballesteros for their IT support. This work was supported by the Concerted Research Actions from Ghent University (BOF.DOC.2017.0026.01) and a scholarship for a long stay abroad (V440318N) from the Fund for Scientific Research Flanders (FWO) to FAC.

Author contributions

FAC conducted all the analyses and wrote the manuscript. JAH contributed to the quality control assessment of the single-cell RNA-seq data. KDP, PM and JAH reviewed and edited the manuscript. All authors read and approved the final manuscript.

Author information

FAC is a PhD student in the Department of Pediatrics and Medical Genetics, Faculty of Medicine and Health Sciences, Ghent University (Belgium). PM and KDP are Professors in the Department of Pediatrics and Medical Genetics, Faculty of Medicine and Health Sciences, Ghent University (Belgium). JAH is a PhD student at the Institute for Molecular Bioscience, University of Queensland. JP is the head of the Garvan-Weizmann Centre for Cellular Genomics and head of the Computational Genomics Group at the Garvan Institute of Medical Research in Sydney (Australia).

REFERENCES

1. Sharma, A. *et al.* Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. *bioRxiv* 698845 (2019) doi:10.1101/698845.
2. Hendry, S. *et al.* Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group. *Adv. Anat. Pathol.* **24**, 235–251 (2017).
3. Research, A. A. for C. Low-Heterogeneity Melanomas Are More Immunogenic and Less Aggressive. *Cancer Discov.* (2019) doi:10.1158/2159-8290.CD-RW2019-144.
4. Elloumi, F. *et al.* Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med. Genomics* **4**, 54 (2011).
5. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
6. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8–9 (2012).
7. Hoffmann, M. *et al.* Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics* **7**, 369 (2006).
8. Newman, A. M., Gentles, A. J., Liu, C. L., Diehn, M. & Alizadeh, A. A. Data normalization considerations for digital tumor dissection. *Genome Biol.* **18**, 128 (2017).
9. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
10. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
11. Vallania, F. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat. Commun.* **9**, 4735 (2018).
12. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).
13. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
14. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
15. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv* 576827 (2019) doi:10.1101/576827.
16. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
17. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
18. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12**, 913–921 (2012).

19. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211–2212 (2013).
20. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinforma. Oxf. Engl.* **29**, 1083–1085 (2013).
21. Chambers, J., Hastie, T. & Pregibon, D. Statistical Models in S. in *Compstat* (eds. Momirović, K. & Mildner, V.) 317–321 (Physica-Verlag HD, 1990). doi:10.1007/978-3-642-50096-1_48.
22. Mullen, K. M. & van Stokkum, I. H. M. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4. https://CRAN.R-project.org/package=nnls.
23. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10**, 1–9 (2019).
24. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
25. Dong, M. *et al.* SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. *bioRxiv* 743591 (2019) doi:10.1101/743591.
26. Gene Expression Studies Using Affymetrix Microarrays. *CRC Press* https://www.crcpress.com/Gene-Expression-Studies-Using-Affymetrix-Microarrays/Gohlmann-Talloen/p/book/9781138112315.
27. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures. *PLOS ONE* **9**, e85150 (2014).
28. Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).
29. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *bioRxiv* 477794 (2018) doi:10.1101/477794.
30. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *bioRxiv* 574574 (2019) doi:10.1101/574574.
31. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).
32. Hao, Y., Yan, M., Lei, Y. L. & Xie, Y. Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles using Least Trimmed Squares. *bioRxiv* 358366 (2018) doi:10.1101/358366.
33. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).
34. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).
35. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627-1640.e7 (2019).
36. Wagner, F. Straightforward clustering of single-cell RNA-Seq data with t-SNE and DBSCAN. *bioRxiv* 770388 (2019) doi:10.1101/770388.
37. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).

38. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).
39. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).
40. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321-330.e14 (2017).
41. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
42. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
43. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
44. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 1–17 (2018).
45. Becker, R. A., Chambers, J. M. & Wilks, A. R. *The New s Language: A Programming Environment for Data Analysis and Graphics*. (Chapman & Hall, 1988).
46. Lun, A. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *bioRxiv* 404962 (2018) doi:10.1101/404962.
47. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
48. The min-max scaling method - Feature Engineering Made Easy [Book]. https://www.oreilly.com/library/view/feature-engineering-made/9781787287600/aa5580ee-6fb7-4ac2-a1fe-369d95b70168.xhtml.
49. Clark-Carter, D. z Scores. in *Wiley StatsRef: Statistics Reference Online* (American Cancer Society, 2014). doi:10.1002/9781118445112.stat06236.
50. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* **10**, 2209 (2019).
51. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
52. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
53. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
54. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
55. LogNormalize function | R Documentation. https://www.rdocumentation.org/packages/Seurat/versions/3.1.1/topics/LogNormalize.
56. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

57. Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. & Wang, J. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* **45**, e179–e179 (2017).
58. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLOS Comput. Biol.* **15**, e1006976 (2019).
59. Ripley, B. *et al. MASS: Support Functions and Datasets for Venables and Ripley's MASS*. (2002).
60. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
61. Altboum, Z. *et al.* Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10**, 720 (2014).
62. Du, R., Carey, V. & Weiss, S. T. deconvSeq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics* doi:10.1093/bioinformatics/btz444.
63. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *bioRxiv* 669911 (2019) doi:10.1101/669911.
64. Wickham, H. & R), R. C. team (Some code extracted from base. *pryr: Tools for Computing on the Language*. (2018).

SUPPLEMENTARY MATERIAL

Supplementary Methods

Incompatible data transformations or normalizations with several deconvolution methods

Global and column z-score normalizations generated negative values, making it incompatible with the single-cell deconvolution methods and with bulk deconvolution methods such as DeconRNASeq, ssKL, ssFrob and DSA. Quantile normalization is used by default in FARDEEP but we disabled it to observe the impact of other scaling/normalization strategies. Row scaling led to singular matrices (several rows were identical; determinant = 0) and thus methods such as robust linear regression (RLR) failed. Linnorm normalization performs an internal logarithmic transformation step, so it is not compatible with logarithmic, square-root and VST transformed input data. CIBERSORT performs and internal z-score standardization of the input matrices prior to fit the support vector regression. The glmnet function used in penalized regression approaches such as ridge, lasso, elastic net and DCQ, includes an internal standardization step (=predictors to be scaled as z-scores) to ensure that the penalty affects each coefficient equally. DSA, ssFrobenius and ssKL can only be applied to data in linear scale [http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix/gedAlgorithm.ssKL.html] whereas dtangle only accepts input matrices in logarithmic scale. ssFrobenius performs an internal mean-centering step of each signature separately whereas in ssKL no re-scaling is performed at all.

MuSiC and SCDC could not be tested using PBMCs because n=1 (they are "multi-subject" methods). deconvSeq is formulated as a generalized linear model that accounts for the quadratic relationship between the mean and the variance in RNA-seq count data using the log link function for a negative binomial distribution, so it is not compatible with logarithmic, square-root and vst transformed input data, and it requires the input to be un-normalized. DWLS includes an internal log2 transformation step followed by differential gene expression analysis (internal marker selection step) with Model-based Analysis of Single-cell Transcriptomics (MAST)[1]. For these reasons, only single-cell input data in linear scale and normalization strategies not generating negative or bounded values were compatible with DWLS.

Explicit versus implicit non-negativity and sum-to-one constraints

For some methods, the output needed to be explicitly ("E" in the table below) modified after the deconvolution to enforce only positive proportions (non-negativity constraint: negative proportions were set to 0) and that they sum to one. For others, these constraints were implicitly ("I" in the table below) included and the output was left unchanged.

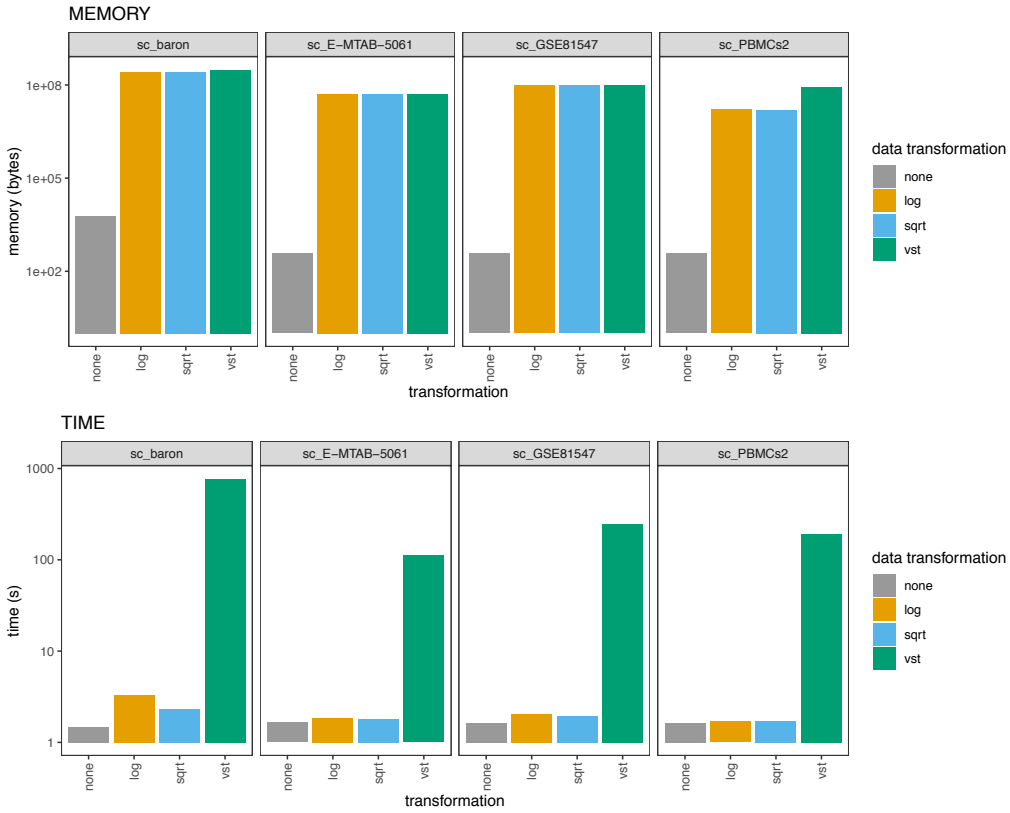
Supplementary Table 1 – Explicit (E) versus implicit (I) non-negativity and sum-to-one constraints

deconvolution method	non-negativity	sum-to-one
OLS	E	E
NNLS	I	E
FARDEEP	I	E
RLR	E	E
lasso	E	E
ridge	E	E
elastic net	E	E
DCQ	E	E
DSA	E	E
EPIC	I	I
dtangle	I	I
DeconRNASeq	I	I
CIBERSORT	I	I*
ssFrobenius	I	I
ssKL	I	I
deconvSeq	I	I
MuSiC	I	I
SCDC	I	I
Bisque	I	I
DWLS	E	E**

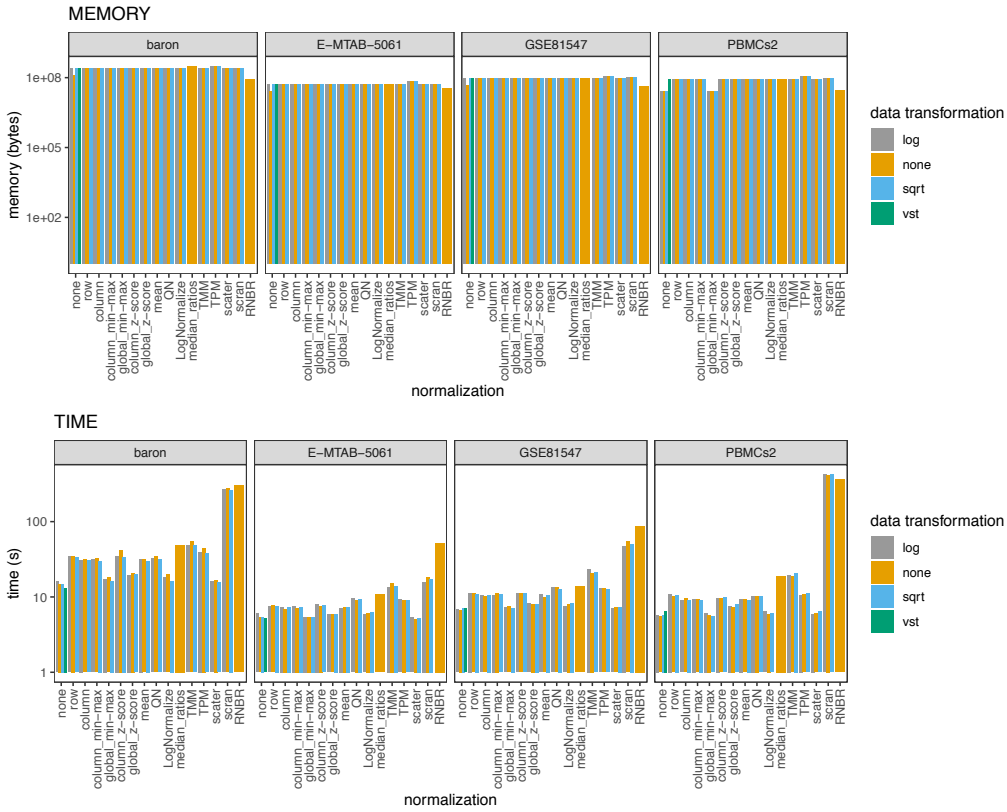
(*) Users can select "absolute" mode to remove the sum-to-one constraint or to use a signature score as output.

(**) It only included the implicit sum-to-one constraint. Thus, to enforce both constraints, we artificially enforced the non-negativity constraint followed by sum-to-one.

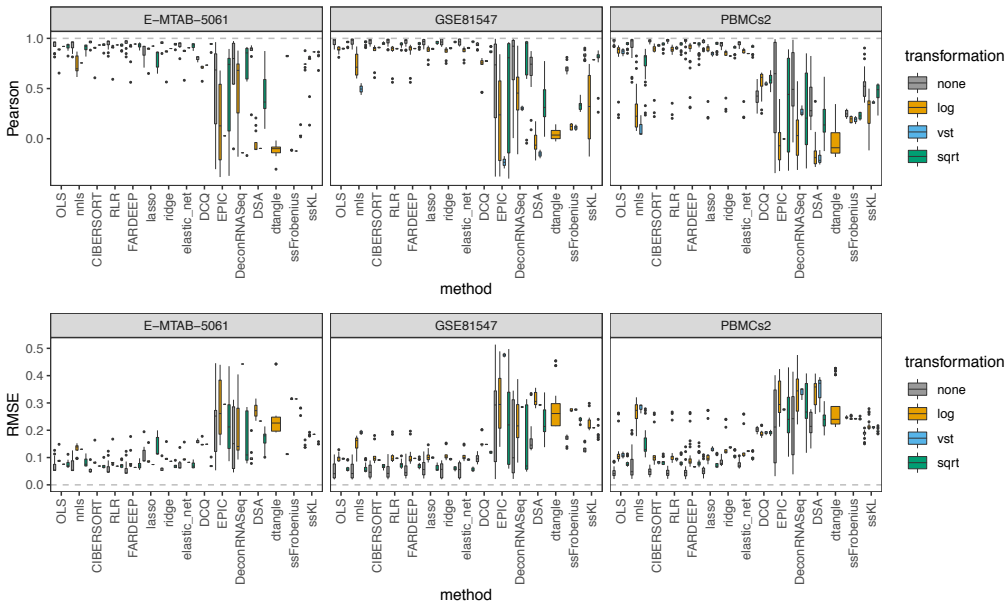
Supplementary Figures



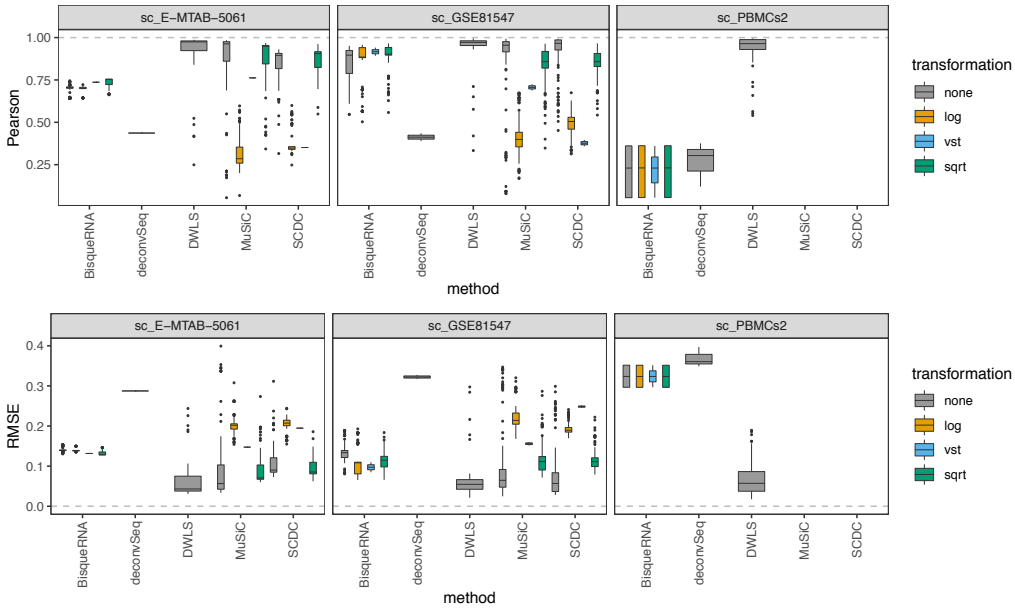
Supplementary Figure 1 – RAM memory (bytes) and time requirements (seconds) for the different transformations across datasets. "none" represents the data un-transformed, in linear scale; log = logarithmic; sqrt = square-root; vst = variance stabilization transformation.



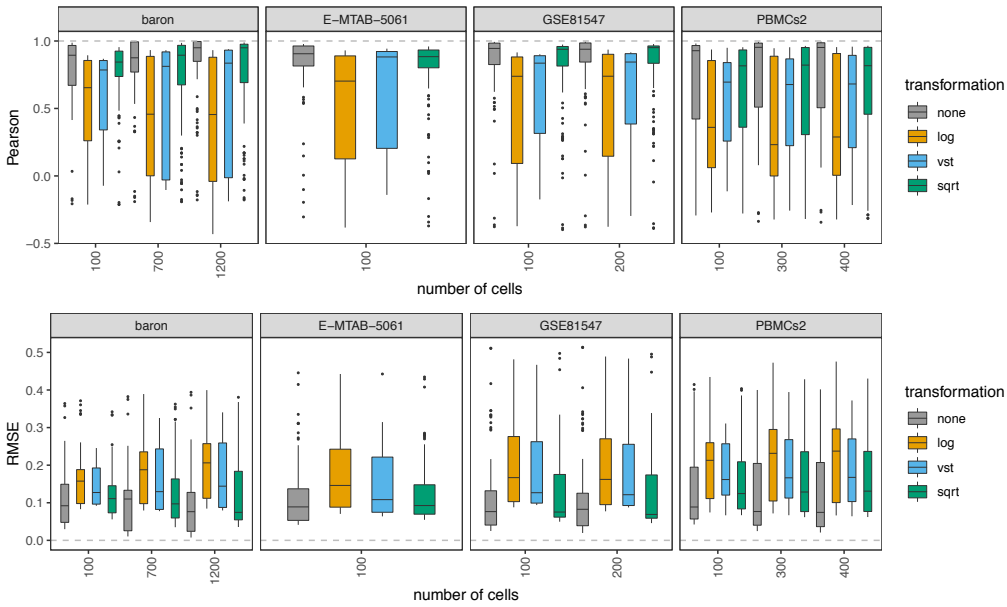
Supplemental Figure 2 – RAM memory (bytes) and time requirements (seconds) for the different scaling/normalization strategies across different single-cell datasets.



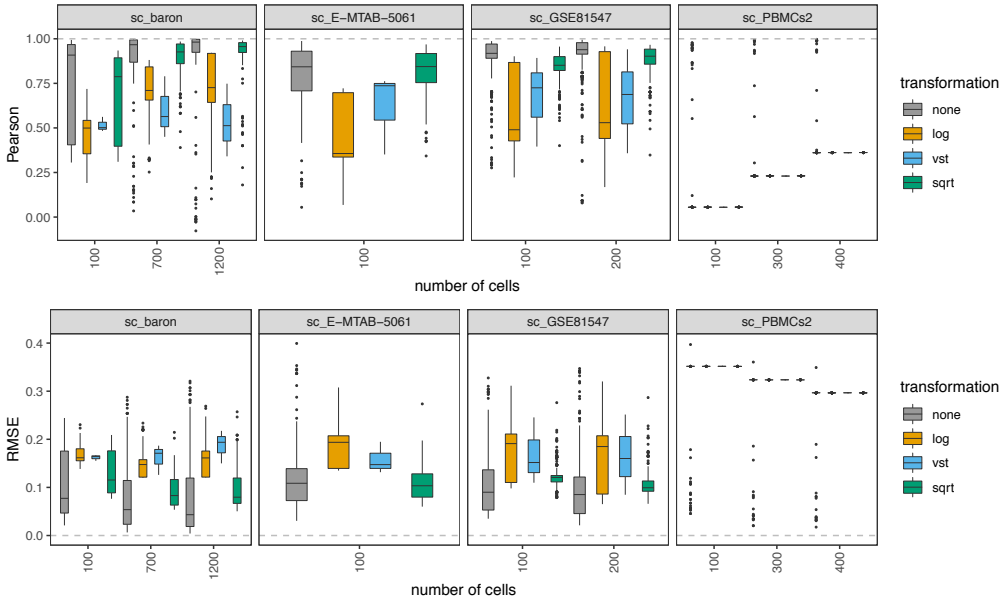
Supplementary Figure 3 – Pearson correlation (top panel) and RMSE values (bottom panel) between the known proportions in 1000 pseudo-bulk tissue mixtures from the E-MTAB-5061, GSE81547 and PBMCs datasets (pool size = 100 cells per mixture) and the predicted proportions from the different bulk deconvolution methods.



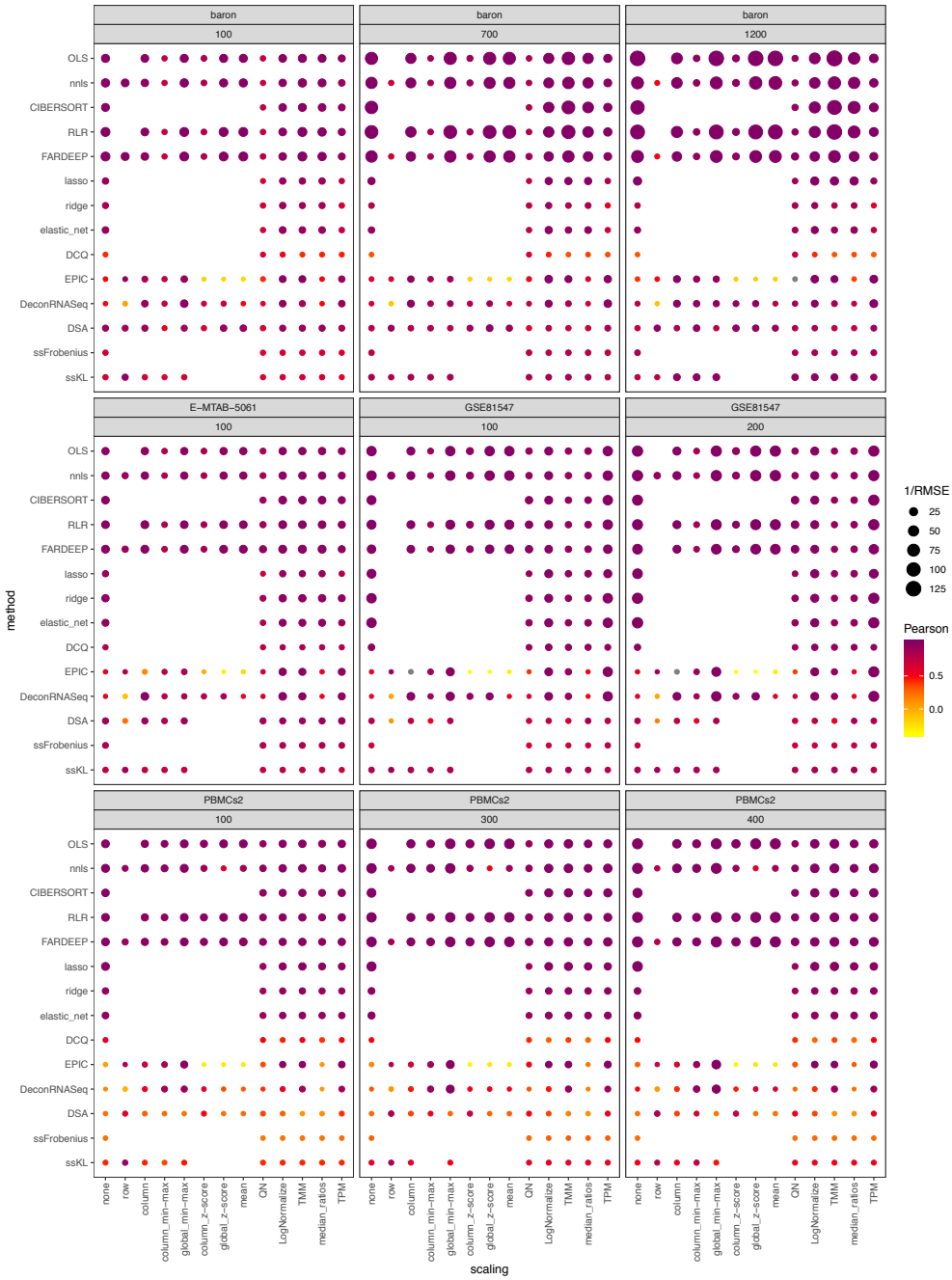
Supplementary Figure 4 – Pearson correlation (top panel) and RMSE values (bottom panel) between the known proportions in 1000 pseudo-bulk tissue mixtures from the E-MTAB-5061, GSE81547 and PBMCs datasets (pool size = 100 cells per mixture) and the predicted proportions from the different single-cell deconvolution methods. MuSiC and SCDC were not applicable to the PBMC dataset because it requires the number of samples to be greater than one. Each boxplot contains all normalization strategies that were tested in combination with a given method.



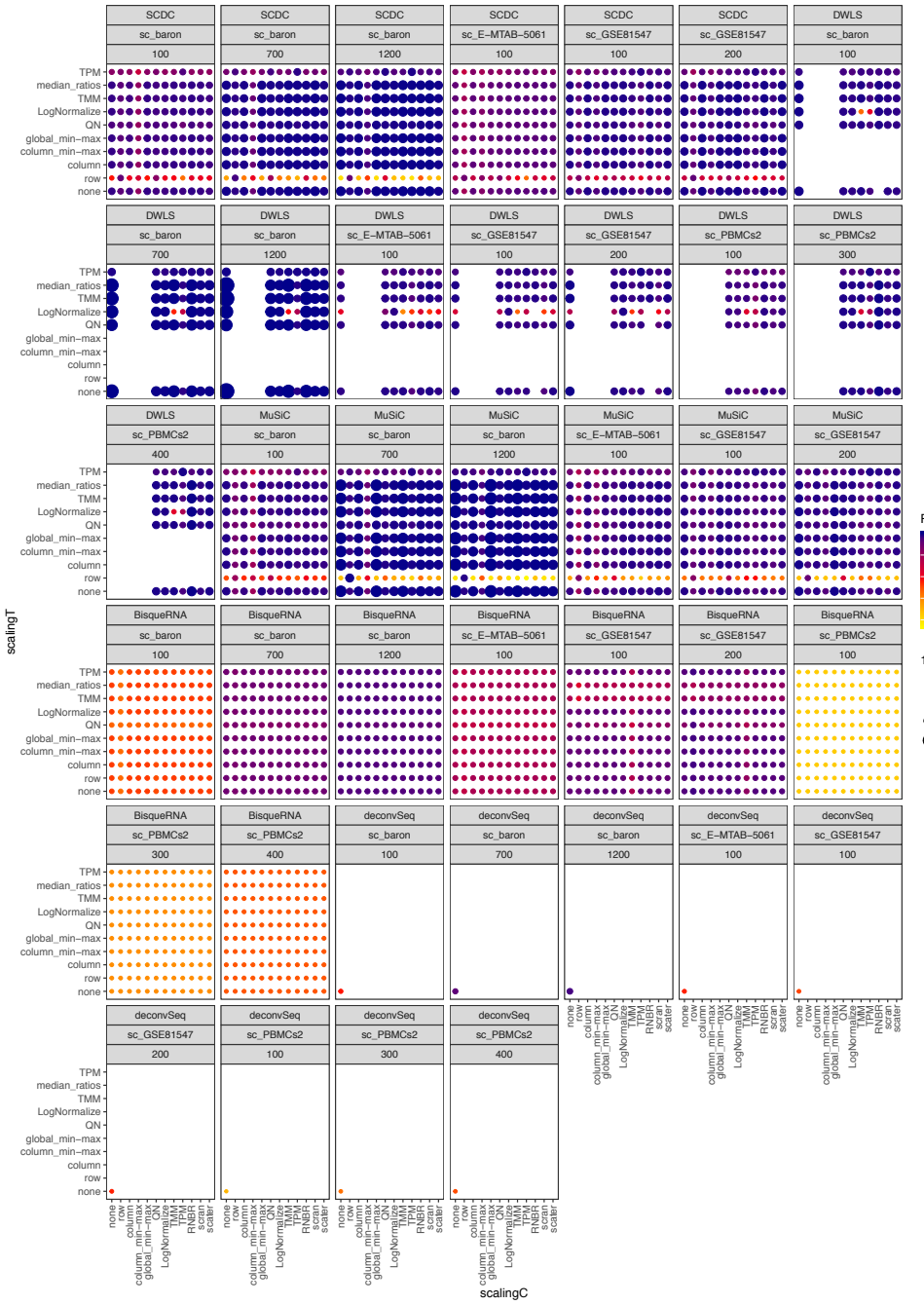
Supplementary Figure 5 – Pearson correlation (top panel) and RMSE values (bottom panel) between the known proportions in 1000 pseudo-bulk tissue mixtures and the predicted proportions from the different bulk deconvolution methods. Each boxplot contains all combinations of method and normalization strategies that were tested with a given cell pool size.



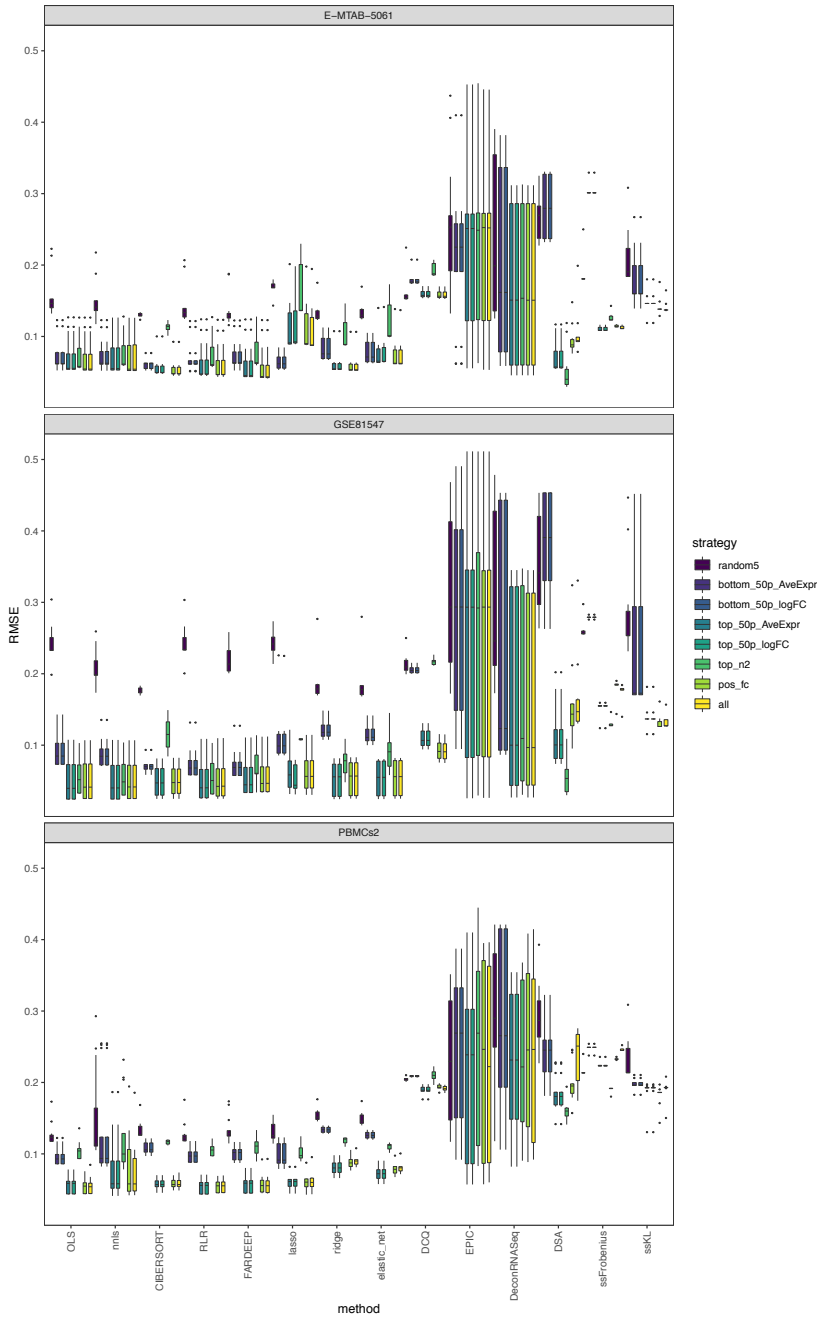
Supplementary Figure 6 – Pearson correlation (top panel) and RMSE values (bottom panel) between the known proportions in 1000 pseudo-bulk tissue mixtures and the predicted proportions from the different single-cell deconvolution methods. Each boxplot contains all combinations of method and normalization strategies that were tested with a given cell pool size.



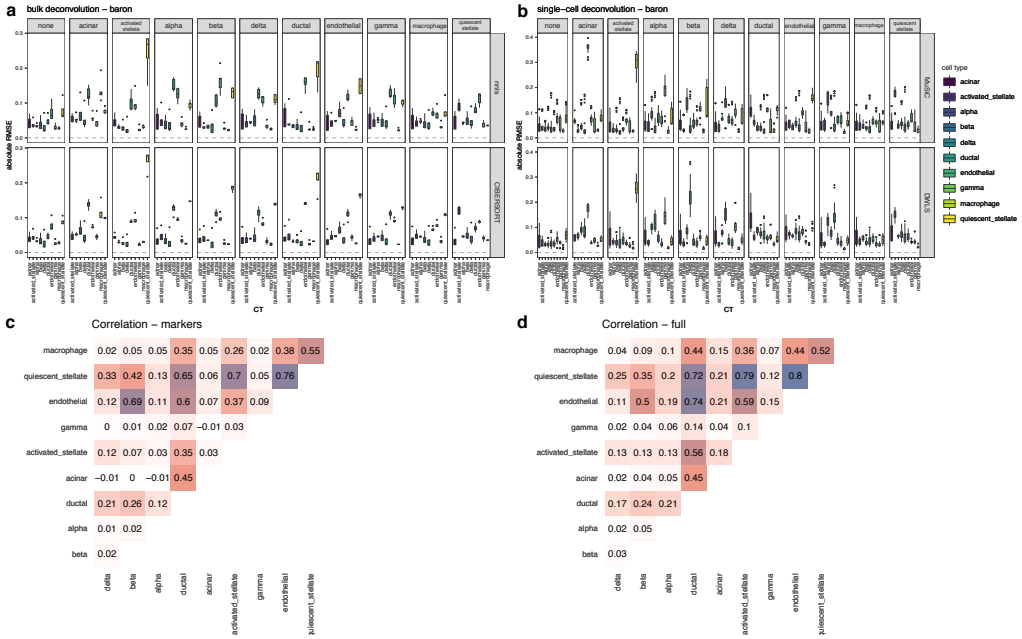
Supplementary Figure 7 – Pearson correlation values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures in linear scale (several pool sizes and datasets, as depicted in the grey labels) and the output proportions from the different bulk deconvolution methods. The darker the blue and the higher the area of the circle represents higher Pearson and lower RMSE values, respectively.



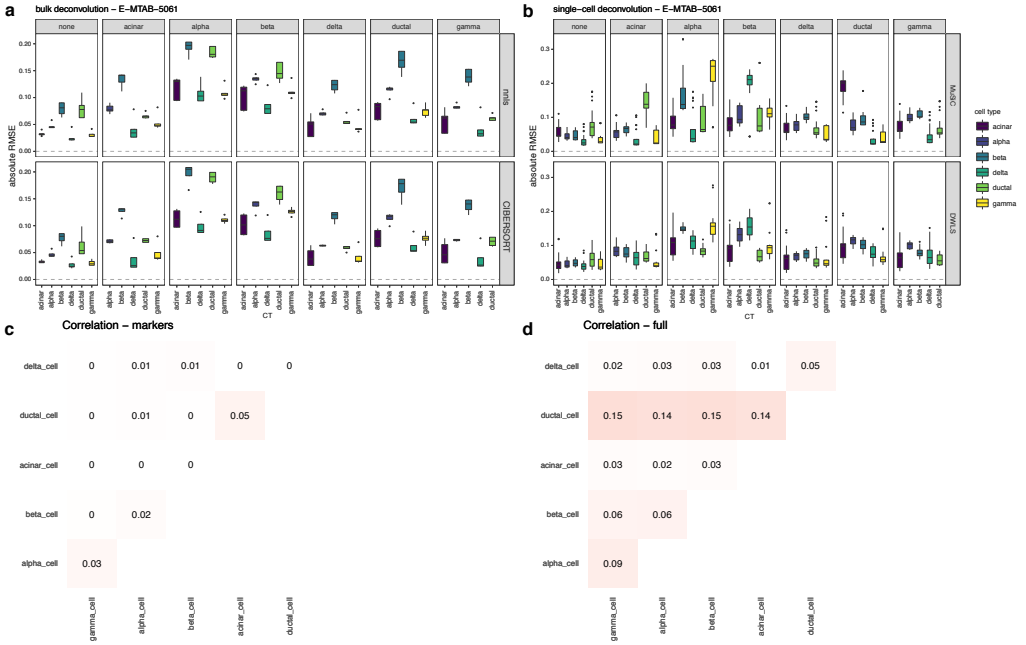
Supplementary Figure 8 – Pearson correlation values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures in linear scale (several pool sizes and datasets, as depicted in the grey labels) and the output proportions from the different single-cell deconvolution methods. The darker the blue and the higher the area of the circle represents higher Pearson and lower RMSE values, respectively.



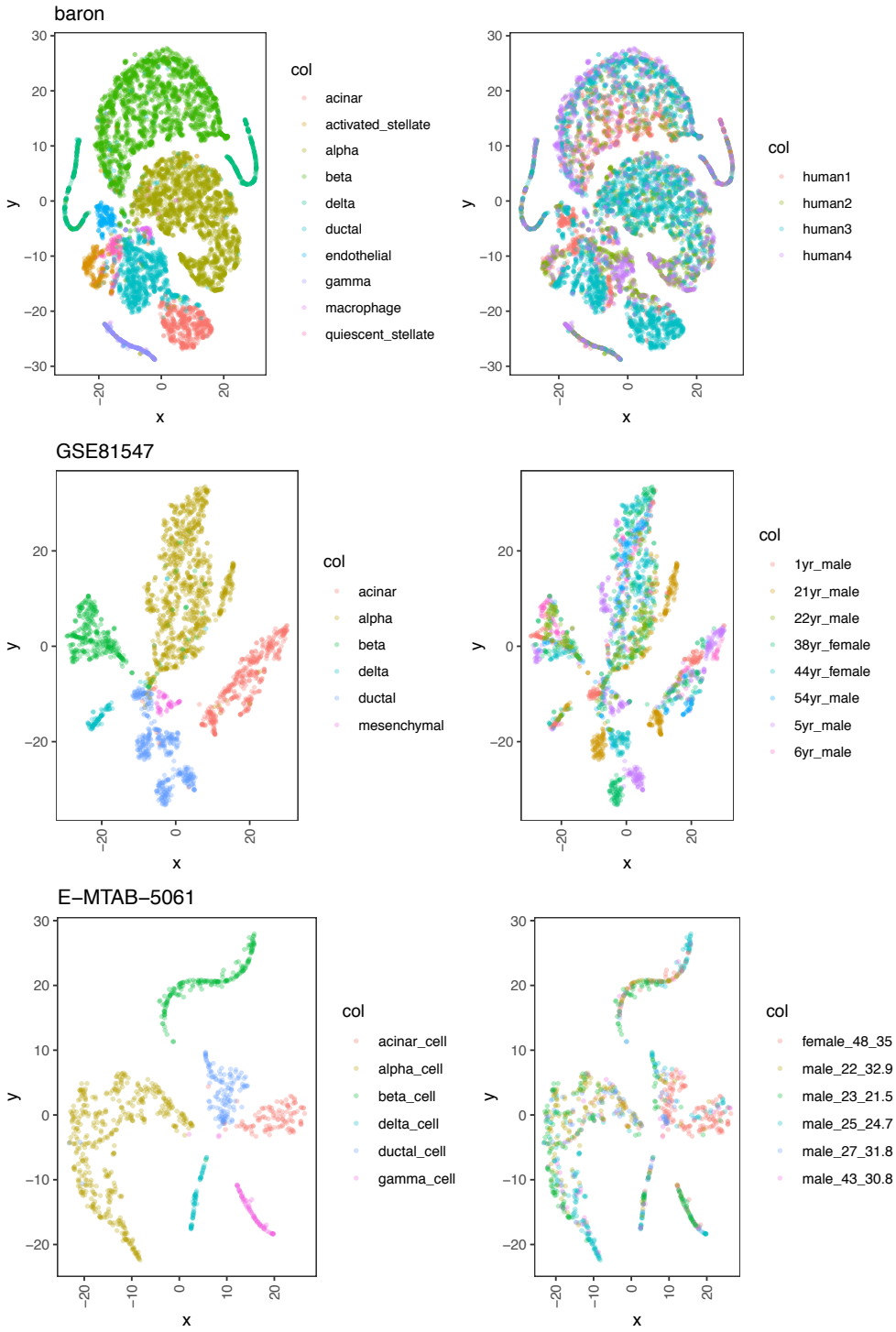
Supplementary Figure 9 – RMSE values between the expected (known) proportions in 1000 pseudo-bulk tissue mixtures (linear scale; pool size = 100 cells per mixture) and the output proportions from the E-MTAB-5061, GSE81547 and PBMCs datasets, using eight different marker selection strategies. Each boxplot contains all normalization strategies that were tested in combination with a given marker strategy across the different bulk deconvolution methods.



Supplementary Figure 10 – Effect of cell type removal on the deconvolution results using the baron dataset [100-cell pseudo-bulk mixtures in linear scale]. a) results using bulk deconvolution methods (nnls and CIBERSORT); b) results using single-cell deconvolution methods (MuSiC and DWLS); c) pairwise Pearson correlation values between expression profiles for the different cell types, using a subset of the reference matrix containing only the markers used in the bulk deconvolution; d) pairwise Pearson correlation values between complete expression profiles for the different cell types. In a) and b), each grey column represents a specific cell type removed. Each data point conforming a boxplot represents a different scaling/normalization strategy used.



Supplementary Figure 11 – Effect of cell type removal on the deconvolution results using the E-MTAB-5061 dataset [100-cell pseudo-bulk mixtures in linear scale]. a) results using bulk deconvolution methods (nnls and CIBERSORT); b) results using single-cell deconvolution methods (MuSiC and DWLS); c) pairwise Pearson correlation values between expression profiles for the different cell types, using a subset of the reference matrix containing only the markers used in the bulk deconvolution; d) pairwise Pearson correlation values between complete expression profiles for the different cell types. In a) and b), each grey column represents a specific cell type removed. Each data point conforming a boxplot represents a different scaling/normalization strategy used.



Supplementary Figure 12 – Dimensionality reduction plots (tSNE) by cell type (left) and donor (right) across all datasets after quality control.

REFERENCES

1. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

Part IV - Discussion and Future Perspectives

Current improvements and remaining challenges in transcript annotation

As mentioned in the first part of this thesis, one of the goals of RNA-sequencing is to provide an accurate quantification of the transcriptome, meaning a complete set of all transcripts present and their abundance. The development of the Zipper plot, using a combination of diverse omics datasets (=multi-omics approach) allows distinguishing truthful, relevant and potentially functional transcripts from noise or DNA contamination.

Nevertheless, it is worth mentioning that the majority of the sequencing data available today is actually short-read sequencing. The determination of accurate transcription start sites (TSSs), transcription end sites (TESs) or exon chaining events (=splicing) can sometimes be problematic with short-read sequencing, complicating transcript reconstruction and quantification tasks[1]. Single molecule RNA-sequencing is becoming more and more available, thanks to the refinement of long-read next-generation sequencing methods, such as those developed by Oxford Nanopore[2] or Pacific Biosciences (PacBio)[3]. Even though this type of sequencing also has other challenges to address (e.g. removing errors from long read sequence data[4]), its output enables a more accurate prediction of transcript models[5] and alternative isoforms[6].

Furthermore, increasing number of lncRNAs were found to contain small open reading frames (sORFs) encoding small peptides or "micropeptides" (shorter than 100 amino acids)[7], suggesting that the coding potential of lncRNAs might have been under-estimated and lncRNAs should be re-defined as RNA transcripts longer than 200 nucleotides not able to generate peptides longer than 100 amino acids. This is an additional challenge to be accounted for while performing transcriptome annotation.

Apart from its use in refining the RNA Atlas transcriptome, researchers have used the Zipper plot tool to refine the neuroblastoma lncRNome[8] and a neuronal enhancer network upstream of MEF2C in human neuronal cell types and brain tissues[9]. In its current form, our tool gathers information coming from the FANTOM5 and Roadmap Epigenomics Project. However, after the release of the Zipper plot in 2017, other valuable datasets have been published and can further help researchers with the transcript annotation task. For example, refTSS[10] contains refined TSS annotations (with their correspondent gene annotations for both human and mouse) based on several other databases. Furthermore, technological advances such as SLIC-CAGE[11] can achieve higher resolution of TSS mapping than the traditional CAGE while only requiring few nanograms of total RNA as input.

The applicability of our tool will be further expanded in the future by integrating publicly available sequencing datasets detecting open chromatin (ATAC-sequencing), nascent RNAs (GRO- and PRO-sequencing), chromatin states coming from the Roadmap Epigenomics project (as depicted on Figure 1 from Paper 2) and refTSS human peaks other than those coming from the FANTOM5 project: EPDnew[12] (~25,500 regions), RAMPAGE[13] (3,935 regions), ENCODE CAGE[14] (8,232), DBTSS[15] (4,753) and Stem cell CAGE[16] (11,082). Moreover, if researchers generate new data on their own and are willing to try out the Zipper plot with it, cloning the Zipper plot repository from Github (https://github.com/favilaco/Zipper_plot) and ensuring their data is formatted similarly as the database I built (see "Step 3" on Github) will suffice.

Imada *et al.*[17] recently showed the use of an improved version of the human transcriptome (being filtered by using FANTOM CAGE data) to re-process more than 70,000 RNA-seq samples present in the reCount2 database, leading to the identification of novel lncRNAs that showed differential expression across different cancer types. These are postulated as new candidates involved in tumor pathogenesis, emphasizing the importance of using a high-quality transcriptome reference and highlighting the need of re-analysing other publicly available RNA-seq datasets.

Comparison of the Zipper plot with other similar approaches

A comparison with tools available at the time of publishing the Zipper plot was difficult because our tool helps refining 5' transcript boundaries (TSSs) for any type of RNA transcript (coding or not) using marks indicative of active transcription, whereas other tools focus on lncRNA identification/annotation by computing a coding potential score base on the sequence of nucleotides composing the transcript and its conservation across multiple species (e.g. PhyloCSF[18] being one of the first appearing in 2011, slncky[19] or PLAR[20] (mentioned in our manuscript) or tools such as FEELnc[21], the coding potential calculator 2 (CPC2)[22] or LncRNAnet[23], all published on the same month or later than the Zipper plot). Moreover, as already pointed out in Paper 1, lncRNAs are generally poorly conserved across species, potentially damaging the applicability of these methods.

In parallel to the revision of our manuscript, Xu et al.[24] performed a combination of literature search (finding 21 lncRNA annotation resources) and mining of three databases (GENCODE, Lncipedia and NONCODE) to retrieve a collection of more than 200,000 lncRNA genes. These were closely investigated by using 4 histone marks (H3K4me3, H3K4me1, H3K27me3 and H3K27ac) in eight human cell lines (=a sub-set of what the

Zipper plot peak database contains) to discover enrichment of histone modifications in the promoter regions (-2.5 kb to +0.5 kb from the TSS) of lncRNA genes, followed by a conservation analysis for exons and the same promoter regions (-2.5 to +0.5kb from the TSS) for both lncRNAs and protein-coding genes.

An easy or direct comparison with Chen *et al.* (BMC Genomics 2016)[25] is not feasible because there was no tool developed, the input data is from *Drosophila* (not human) and would require re-processing of all raw data. Chen *et al.* used 30 RNA-seq datasets on *Drosophila* to perform *de novo* lncRNA annotation (splitting transcripts longer than 200 nt into coding and non-coding using the Coding Potential Calculator (CPC) tool) and discovered 462 novel lncRNA transcripts. Next, using 32 ChIP-seq datasets (H3K4me3, H3K36me3 and PolII) they discovered than half of those transcripts did not have chromatin signatures related to active transcription. For this reason, they decided to use RT-qPCR to investigate 42 lncRNAs and found out that 40 out 42 (95.24%) were detected as transcribed and independently of being associated with active chromatin signatures or not. Nevertheless, these conclusions may be somewhat biased because Chen *et al.* included only three ChIP-seq datasets and did not take CAGE-seq data into account whereas the Zipper plot includes both ChIP-seq on nine different histone marks (including H3K4me3 and H3K36me3) and CAGE-seq.

Finally, the ZENBU browser is a tool released by the FANTOM5 team to visualize CAGE-seq data (for human and mouse) on a genomic context. However, it only allows visualization for one genomic region at a time and needs to reload the CAGE track every time the user wants to zoom into a different genomic region, making the process very slow and inefficient, and it will not be a practical tool to use to investigate the complete transcriptome.

Reflection about the choice of different parameters in the deconvolution benchmark

There were two types of choices for the deconvolution methods included in the article: 1) choice of parameter values for a specific function; 2) selection a number of permutations/iterations.

Ridge regression (also known as Tikhonov regularization), lasso regression, elastic net and Digital Cell Quantifier (DCQ) are four different penalized regression approaches implemented using the glmnet function that vary depending on the values for the alpha

and lambda parameters. Specifically: a) ridge: alpha = 0, “grid search” for lambda; b) lasso: alpha = 1, “grid search” for lambda; c) elastic net: alpha = 0.2, “grid search” for lambda; d) DCQ: alpha = 0.05, lambda = 0.2.

The grid search for lambda consisted in scanning the data (in this case, one particular pseudo-bulk mixture at a time) to configure its optimal value through the “one standard error rule" ("lambda.1se” in R). Next, FARDEEP included a “perm” parameter (which was set to 10) and both ssKL and ssFrobenius included “maxIter”, which was set to 500. Furthermore, the choice of seed included ensures reproducibility of the results (given that the use of permutations in any method implies a stochastic nature).

All other methods did not include parameters that needed to be specified by the user prior to the deconvolution.

Shortcomings in the deconvolution benchmark

Even though different library preparation protocols were used to generate the four single-cell RNA-sequencing datasets used in this thesis (both full-length and tag-based), the inclusion of additional datasets across other biological fractions other than blood and human pancreas would definitely improve the robustness of our conclusions.

Furthermore, only partial (supervised) deconvolution methods generating cell type proportions as output were tested. In theory, supervised methods are likely to produce more accurate proportion estimates than the total deconvolution (= completely unsupervised) counterparts. The performance of non-guided methods strongly depends on the ability to recover meaningful gene signatures or expression profiles for the different cell types. Therefore, even though this was proven to be the case for the semi-supervised methods we tested (ssKL and ssFrobenius), methods such as Linseed, CDSeq or deconICA will need to be tested in the future. Complete deconvolution strategies will be able to work in scenarios where there is no a priori information of the cell type composition in a mixture or an incorrect cell type has been included as part of the reference matrix that is used by a partial deconvolution method. However, the remaining challenge for complete deconvolution strategies is selecting the correct number of components (“cell types”) present in the mixture and assigning a label (namely the cell type) to each of the components the algorithm has found (which is far from trivial).

Regarding the PBMC dataset, the multi-collinearity problem appeared as six different (but highly correlated) subsets of T cells. The original marker selection strategy that was used

for the other three datasets (TMM normalization followed by differential gene expression with limma-voom; see Paper 3 for more details) produced no markers able to distinguish these subpopulations. This issue was addressed by labelling all subsets as T cells. In contrast to this conservative approach, I could have allowed marker genes to be highly expressed not only in one but, two, three or all the six T-cell subpopulations with respect to the other cell types present in the reference matrix. This problem could be also alleviated by a hierarchical deconvolution approach similar to what is described in the original MuSiC article[51]: "*MuSiC constructs a hierarchical clustering tree reflecting the similarity between cell types. Based on this tree, the user can determine the stages of recursive estimation and which cell types to group together at each stage*".

Furthermore, the co-linearity at gene level also arises when there are multiple genes (rows) being highly correlated. This aspect was not included in the benchmark article (Paper 3) but, since we evaluated several marker selection strategies in the manuscript (therefore leading to different versions of the reference matrix used in the deconvolution), the condition number (CN) of each matrix could have also been computed. This could potentially be used to evaluate the robustness of the reference matrices (= low CN values) or lack thereof (=high CN) and, for those cases with high CN values, changes in the CN could be evaluated when iteratively removing highly correlated markers from the matrix.

The most important yet unsolved issue is the presence of unknown cell types in a mixture. Failure to include cell types in the reference matrix that are actually present in a mixture always led to substantially worse results (higher RMSE values). EPIC[52] shows a first attempt in alleviating this problem by considering an unknown cell type present in the mixture. Nevertheless, this is done using markers of non-malignant cells that are not expressed in cancer cells, not being a general solution yet and leaving room for future improvement.

Statistical mixture modeling and principal component analysis (PCA) can be used to determine the number of components (proxy for the number of cell types) present in a mixture. This can already allow researchers to investigate whether the reference matrix to be used in the deconvolution is complete or not. Secondly, I am currently investigating the usefulness of start with applying the deconvolution with the initial reference matrix that is suspected to lack a cell type followed by using those proportion estimates to fit the new regression model genome-wide and investigate the good goodness of fit for each gene. Those with a poor fit are, hypothetically, markers for the unknown cell type in the mixture. In

any case, the expression profile for the unknown cell type would still need to be "artificially created", so there research in this direction is still needed.

Finally, the role of factors affecting the deconvolution performance which could not be attributed to the presence or absence of cell types (such as cell cycle phases) was not discussed in the benchmark paper. When working in the laboratory, it may be feasible to synchronize the cell cycle of cells growing in a petri dish by nutrient starvation or adding diverse pharmacological agents. However, cells composing a tissue or organ are all in different stages of the cell cycle and each cell has different external conditions such as nutrient availability, hypoxia, etcetera.

Current deconvolution frameworks assume cell-type specific markers to be insensitive or invariant to those factors. However, if we had to analyse different spatial samples from the same tumor, the latter will be an important factor to take into account, since the tumor microenvironment is known to have a gradient of oxygen concentration[53].

As I stated in the introduction, Lu *et al.*[54] proposed the use of phase-specific markers (such as cyclin *CLN2* for phase G1 or *CLB4* for phase G2) to establish different time points of the cell cycle. The expression of these and other relevant markers could help to distinguish a set of "cell-cycle invariant" markers and the cell cycle stage could potentially be included as a covariate in the deconvolution when computing the cell type proportions.

Computational deconvolution: a bright future ahead

Bayesian and regression-based methodologies have been proven effective in the framework of the deconvolution problem. However, currently there is no tool addressing all the challenges we highlighted in the introduction and result sections of this thesis, leaving some room for improvement. The ideal tool should: 1) include alternatives to solve all formulations of the deconvolution problem described in the introduction, meaning supervised and completely unsupervised scenarios; 2) allow to study the changes in cell type proportions across multiple time points (such as DCQ[26]); 3) account for different phases of the cell cycle using markers such as CLN2 for phase G1; 4) account for small perturbations between reference expression profiles of pure cell types and those constituting the heterogeneous samples (such as PERT[27] or ISOpure(R)[28,29]); 5) be computationally efficient, with fast running time and rate of convergence; 6) be able to account for the presence of multiple correlated cell types in the mixture (such as CIBERSORT[30]).

For supervised deconvolution scenarios (=partial deconvolution), we argue against the use of non-informative (=random) initial estimates and recommend the use of one or more approaches described in "*Selection of cell type-specific markers or expression profiles*" (Part I: Introduction). Regarding the marker selection, the unsupervised geometric identification of markers proposed by UNDO[31] and CAM[32] (identification of vertices and resident genes of a K-dimensional polytope where K is the number of cell types present in a mixture), seems like a sensible and unbiased approach compared to the usage of external reference datasets (that might come from several technology platforms) or arbitrary log fold change and p-value thresholds.

Several authors stated that their deconvolution methods should specifically be applied to samples belonging to a common tumor (sub-)type[33,34] or to a common tissue[35]. Importantly, non-guided approaches such as non-negative matrix factorization (NMF) successfully identified different pancreatic ductal adenocarcinoma subtypes[36]. However, this only addresses the inter-tumor heterogeneity. In order to study intra-tumor heterogeneity, either single-cell profiling data or sequencing multiple locations from the same tumor would be needed.

The amount of gene expression data from single cells is growing exponentially, revealing information that is hidden in tissue-averaged expression measurements from heterogeneous samples. However, the expression levels are often smaller than the detection limits of current state-of-the-art single-cell technologies. To overcome the detection issue, an approach called "stochastic profiling" has been proposed[37–39]. Stochastic profiling consists of measuring the expression of random pools of cells (e.g. 10 cells) followed by modelling the expression of each gene as a binomial choice from a mixture of two different regulatory states: "ON" for cells expressing the gene and "OFF" for those that do not. Since the amount of input mRNA from a pool of cells is bigger than the mRNA from a single cell, this method offers more robust detection. The idea of stochastic profiling has been further pursued by Lun *et al.* with the development of scran[40] (one of the scRNA-seq specific normalization methods included in the benchmark paper), reducing the incidence of problematic zeroes by summing across cells.

Methods such as SCDC[41] smartly allow the use of multiple single-cell RNA-seq datasets to increase the robustness of the cell type proportions generated as output in a deconvolution framework. Furthermore, as it has been shown in this thesis, other RNA fractions (other than mRNAs) can be potentially used in the deconvolution.

I also foresee the inclusion of multi-omics factor analysis (MOFA[42]) frameworks in the context of computational deconvolution to identify hidden technical and biological sources of variability that need to be controlled for and to obtain more accurate results by integrating multiple omics datasets coming from the same cell: G&T-seq[43] (joint genome and transcriptome); CITE-seq[44] (joint proteome and transcriptome); SNARE-seq[45] (transcriptome and chromatin accessibility); sc-GEM[46] (joint transcriptome, methylome and genotype information) or scTrio-seq[47] (transcriptome, methylome and genome).

A deconvolution approach able to use several omics datasets at a time already exists. ("DC3")[48] is able to simultaneously use HiChIP, RNA-seq and ATAC-seq from a common cell population and is able to accurately identify different subpopulations and deconvolve bulk profiles into sub-population specific profiles. Having several omics datasets coming from the same sample will help answering whether a specific gene can only be (or is) useful in one layer or whether the really useful markers are those found to be relevant across multiple layers.

Another clever idea worth mentioning consists in borrowing information across different species. Butler et al. (2018)[49] showed that similar cell types in mice and humans share gene expression signatures, suggesting that the integration of scRNA-seq between these two species was possible. Donovan et al. (2020)[50] showed that signature genes from mice scRNA-seq can be used for deconvolution of human liver and skin samples from GTEx.

Moreover, Donovan *et al.*[50] applied computational deconvolution to the tissues present in GTEx, leading to new knowledge obtained from database that was already released several years ago. Therefore, since there are many other publicly available resources where the deconvolution can be applied, I predict this is only the first one of many more to come.

While single-cell and stochastic profiling are postulated as firm candidates to revolutionize the transcriptomics field with continuous improvements in terms of sensitivity and affordability, we foresee a rapid inclusion of deconvolution methodologies to existing pipelines for the analysis of omics data in the meantime, increasing the accuracy and reliability of downstream cell type-specific differential gene expression analysis without incurring in additional costs.

Computational deconvolution of circulating cell-free RNA in liquid biopsies: the next frontier

Novel insights into the genetic events driving cancer initiation and progression are currently fueling development of more effective and less toxic molecular therapies specifically targeting a molecular Achilles' heel and killing the cancer cells more efficiently (= precision medicine). This approach requires access to tumor material for genetic testing but metastasized tumors and even some primary tumors are inaccessible for surgical resection. Furthermore, surgical resection or tumor biopsy imply risks for the patient and may not capture the heterogeneity of the tumor. Biopsy is, however, crucial for diagnosis, prognostication and therapy stratification. Moreover, many patients acquire secondary mutations during the course of the treatment, endowing the tumor cells with drug resistance[55]. Therefore, identification of changes in the disease over time from the moment of diagnosis is essential to improve patient outcome and would require more than one biopsy.

For these reasons, a lot of efforts are currently being directed towards the ability to "biopsy" solid tumor diseases through non-invasive (or minimally invasive) sampling of blood and other human fluids (also known as 'liquid biopsy'). It is underexplored how well the tumor transcriptome is recapitulated in different fractions of blood samples (=circulating transcriptome) and whether the same conclusions concerning activated and druggable pathways in a tumor sample can be drawn from this data.

Circulating cell-free DNA (cfDNA) and RNA (cfRNA) are detectable in blood samples or other body fluids. However, these nucleic acids are released into the bloodstream from both normal and tumorous cells in a passive (through cell death) or active (by cell secretion) manner, complicating the retrieval of tumor-specific signal. Since these nucleic acids are not within cells but in circulation, single-cell technologies cannot be used. Therefore, the only mathematical approach that can be used under these scenarios (e.g. on blood or plasma samples from cancer patients) is computational deconvolution. There is evidence that both tissue-of-origin prediction (= deconvolution) and non-invasive cancer diagnosis is possible from cfDNA[56] and cfRNA[57] present in blood samples. To be able to account for the non-tumorous signal, gene expression data from large compendia of healthy tissues, like those reported by GTEx[58], FANTOM[59,60], the RNA Atlas[61] and, in the near future, the still ongoing Human Cell Atlas (HCA; whose ultimate goal is to generate a comprehensive reference of all human cells; https://www.humancellatlas.org/), can be used to select tissue and cell-type specific markers. By generating a comprehensive reference

matrix with as many cell types as possible, the likelihood of missing out relevant cell types present in a mixture will decrease, and thus a better performance could be achieved in the deconvolution.

On the other hand, the Human Biofluid RNA Atlas[62] provides a new and unprecedented set of (healthy) heterogeneous samples in which different deconvolution frameworks could be tested. Preliminary analyses show that relevant "biofluid – tissue of origin" pairs can be detected in such biofluids (e.g. seminal fluid – testicle; bronchoalveolar lavage – oesophagus; saliva – oesophagus).

The final goal would be to establish a bioinformatics pipeline for deconvolution of the circulating tumor transcriptome (e.g. from blood or other biofluid) for diagnosis, follow-up and potential drug target identification for cancer patients.

Conclusion

We developed the Zipper plot, a tool that can be used to assess the reliability of the annotation of thousands of transcripts using features that are indicative of independent transcription and that has been used to refine the human transcriptome generated using the RNA Atlas dataset. Furthermore, we reviewed more than fifty different computational deconvolution methods developed during the last two decades, evaluated the use of different RNA fractions (other than mRNAs) in the computational deconvolution of transcriptomics data and performed a comprehensive assessment of different key factors affecting the deconvolution results, including data transformation, scaling/normalization and marker selection strategies, the cell type composition of the reference matrix and the choice of method.

REFERENCES

1. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
2. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
3. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
4. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long read error correction methods. *bioRxiv* 519330 (2019) doi:10.1101/519330.
5. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
6. Kuosmanen, A., Norri, T. & Mäkinen, V. Evaluating approaches to find exon chains based on long reads. *Brief. Bioinform.* **19**, 404–414 (2018).
7. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
8. Rombaut, D. *et al.* Integrative analysis identifies lincRNAs up- and downstream of neuroblastoma driver genes. *Sci. Rep.* **9**, 1–13 (2019).
9. D'haene, E. *et al.* A neuronal enhancer network upstream of MEF2C is compromised in patients with Rett-like characteristics. *Hum. Mol. Genet.* **28**, 818–827 (2019).
10. Abugessaisa, I. *et al.* refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J. Mol. Biol.* **431**, 2407–2422 (2019).
11. Cvetesic, N. *et al.* SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Res.* gr.235937.118 (2018) doi:10.1101/gr.235937.118.
12. Dreos, R., Ambrosini, G., Cavin Périer, R. & Bucher, P. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* **41**, D157–D164 (2013).
13. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
14. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
15. Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y. & Nakai, K. DBTSS provides a tissue specific dynamic view of Transcription Start Sites. *Nucleic Acids Res.* **38**, D98–D104 (2010).
16. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566 (2014).
17. Imada, E. L. *et al.* Recounting the FANTOM CAGE–Associated Transcriptome. *Genome Res.* gr.254656.119 (2020) doi:10.1101/gr.254656.119.
18. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
19. Chen, J. *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* **17**, 19 (2016).
20. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
21. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57–e57 (2017).
22. Kang, Y.-J. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).

23. Baek, J., Lee, B., Kwon, S. & Yoon, S. LncRNAnet: long non-coding RNA identification using deep learning. *Bioinformatics* **34**, 3889–3897 (2018).
24. Xu, J. *et al.* A comprehensive overview of lncRNA annotation resources. *Brief. Bioinform.* **18**, 236–249 (2017).
25. Chen, M.-J. M. *et al.* Integrating RNA-seq and ChIP-seq data to characterize long non-coding RNAs in Drosophila melanogaster. *BMC Genomics* **17**, 220 (2016).
26. Altboum, Z. *et al.* Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10**, 720 (2014).
27. Qiao, W. *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**, e1002838 (2012).
28. Quon, G. *et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).
29. Anghel, C. V. *et al.* ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* **16**, 156 (2015).
30. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
31. Wang, N. *et al.* UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinforma. Oxf. Engl.* **31**, 137–139 (2015).
32. Wang, N. *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6**, 18909 (2016).
33. Quon, G. & Morris, Q. ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinforma. Oxf. Engl.* **25**, 2882–2889 (2009).
34. Ahn, J. *et al.* DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinforma. Oxf. Engl.* **29**, 1865–1871 (2013).
35. Frishberg, A., Steuerman, Y. & Gat-Viks, I. CoD: inferring immune-cell quantities related to disease states. *Bioinforma. Oxf. Engl.* **31**, 3961–3969 (2015).
36. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).
37. Janes, K. A., Wang, C.-C., Holmberg, K. J., Cabral, K. & Brugge, J. S. Identifying single-cell molecular programs by stochastic profiling. *Nat. Methods* **7**, 311–317 (2010).
38. Bajikar, S. S., Fuchs, C., Roller, A., Theis, F. J. & Janes, K. A. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc. Natl. Acad. Sci.* **111**, E626–E635 (2014).
39. Narayanan, M., Martins, A. J. & Tsang, J. S. Robust Inference of Cell-to-Cell Expression Variations from Single- and K-Cell Profiling. *PLOS Comput. Biol.* **12**, e1005016 (2016).
40. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
41. Dong, M. *et al.* SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References. *bioRxiv* 743591 (2019) doi:10.1101/743591.
42. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, (2018).
43. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
44. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
45. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

46. Cheow, L. F. *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* **13**, 833–836 (2016).
47. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).
48. Zeng, W. *et al.* DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat. Commun.* **10**, 1–11 (2019).
49. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
50. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 1–14 (2020).
51. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 1–9 (2019).
52. Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).
53. Byrne, M. B., Leslie, M. T., Gaskins, H. R. & Kenis, P. J. A. Methods to study the tumor microenvironment under controlled oxygen conditions. *Trends Biotechnol.* **32**, 556–563 (2014).
54. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci.* **100**, 10370–10375 (2003).
55. Longley, D. B. & Johnston, P. G. Molecular mechanisms of drug resistance. *J. Pathol.* **205**, 275–292 (2005).
56. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017).
57. Koh, W. *et al.* Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc. Natl. Acad. Sci.* **111**, 7361–7366 (2014).
58. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
59. The FANTOM Consortium and the RIKEN PMI and Clst (dgt) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
60. Rie, D. de *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).
61. Lorenzi, L. *et al.* The RNA Atlas, a single nucleotide resolution map of the human transcriptome. *bioRxiv* 807529 (2019) doi:10.1101/807529.
62. Hulstaert, E. *et al.* Charting extracellular transcriptomes in The Human Biofluid RNA Atlas. *bioRxiv* 823369 (2019) doi:10.1101/823369.

ACKNOWLEDGEMENTS

This is IT! I cannot believe my PhD has come to an end!

First and foremost: Katleen and Pieter, you are two of the most talented scientists I know and truly were an inspiration. I am and will always be thankful for giving me the opportunity to do research in such a great environment. Jo: even though you were not officially one of my supervisors, I also feel really lucky to have received such amazing insights and feedback from your side. I will always remember the following piece of advice you once gave me:

As a scientist, you can argue or disagree, or be more precise.

Try to be fair and balanced in your critiques.

Prof. Jo Vandesompele

To the members of the examination committee: thank you very much for being part of the jury, the thorough reading of the thesis and the valuable feedback provided during the past weeks.

Lucía and Annelien: I am very happy you accepted to be my paranimfs! Thanks for all your help (not only these final weeks) and, more importantly, for your friendship! Rotterdam, lemon pies, Basel walks and dinners, movies … ☺

Jasper, Kathleen, Kimberly, Nurten and Charlotte: thanks a million for all your time and effort helping me inside and outside the lab during my PhD. This thesis won't be written today if it wasn't for your help!

Sofía, Jacopo and Phoebe: not only you are my friends but also helped me and my family during some difficult times… I will always be grateful and in doubt with you!

Celine: we started and finished almost at the same time! For more scientific discussions, Gent jazz festivals and International dinners with Gentse stoverij to come!

Anneleen, Annelien, Carolina, Celine, Eva, Hetty, Alan, Dries, Fien, Jasper, Jilke, Jill, Marieke, Pauline, Pieter-Jan, Ruben and Sofía: with some in the same office for years, others for months… but in any case: thanks for making our workplace such a great place to go to every day!

Eva, Karen, Annelynn, Bieke, Kaat and Frank: Thanks for making New Mexico such a great experience! It was the first time I hitch-hiked! :P

Pieter-Jan and Fjoralba: thanks for being my "peter" and "meter" when I started! Gloria, Farzaneh, Frank, Jeroen and Carina: thanks for the guided tour in Antwerpen! Fien, Shanna, Eric, Christophe and Lucía: for many more tennis matches to come!

To the rest of my amazing CMGG colleagues (and ex-colleagues!): thanks for all the lunches in the restaurant, farewell dinners, movies, boardgame nights, Gent film festivals, wine tastings, laser-shootings and a never-ending list of activities we did all together. Another big thank you for everyone that came to visit me after my surgery: yet another proof of the amazing group of people I am lucky to work with!

Joseph: you also became an extra supervisor during my six months in Australia. Thanks for welcoming me in your lab and for everything you did to make my stay as memorable as it was, including the short trip to Lorne and the conference. It was phenomenal!

José: gracias por ser tan buen compañero, ¡ya sabes donde estoy si decides venirte a Europa!

Vikki, Marco, Seyhan, Drew, Rika, Brian, Angela, Vanessa and Rachael…thanks for all those lunches together and for being such wonderful colleagues. I hope to see you all again in some part of the world (or conference)!

Marce, Dani y Nati: nunca podré expresar suficientemente mi agradecimiento por vuestro recibimiento australiano… hicisteis que me sintiera como en casa aún estando a más de 17000km de distancia. Hikings en Manly, viajes a Canberra, yoga, unas Navidades en la playa, cenas con tortilla y un año nuevo con uvas y el Sydney Harbour Bridge de fondo. ¡Millones de gracias por todo!

Elise, Griet, Evy, Tania, Ken, Karim, prof. Brusselle, prof. Joos and everyone else from the respiratory medicine department in UZ Gent (and Hataitip in Groningen!): even though our common project on lncRNAs in COPD did not end as expected, it was a pleasure to collaborate across many other projects during the past few years! I learnt a lot from each of you and I am happy we got to share more than science.

I also want to thank many people from my stay in Ireland: Marcus Claesson, Anca Mustata, Damian Conway, Michael Cronin and specially Joseph Manning (who regretfully passed away while this thesis was being written): your passion and amazing lectures led me to the bioinformatics path I am in today.

A varios de mis profesores de la Universidad León (Rafael Santamaría Sánchez, María Teresa Trobajo de las Matas, Roberto Fraile Laíz, Luis Fernando Calvo Prieto, Luis Enrique Sáenz de Miera Carnicer, Juan José Arranz Santos, José Ignacio Rodríguez Barbosa, Xiomar Arleth Gómez Barrios, Gustavo Adolfo González Fernández, Miguel Ángel Chinchetru Manero, Héctor Díez Machío, María Margarita Marqués Martínez, Yolanda Bayón González, María del Carmen Pérez Díez, Beatriz Aguado Otero, Paulino de Paz Cabello y José Carlos Pena Álvarez), del I.E.S. Río Duero en Valladolid (Guadalupe Lozano Gutierrez, Begoña Lapeña Barrio, Carmen Alonso Alonso, Mª Pilar Cabezudo Cabezudo, Ángela Cabello, Milagros López Romero y Sagrario), Avelina y Mª Ángeles: ¡Gracias, gracias y más gracias! ¡Miro atrás y no tengo suficientes palabras para agradecer vuestras enseñanzas, consejos y recomendaciones durante este largo camino!

Laura, Cristina, Verito. ¡Qué os voy a decir! Pocos amig@s de verdad se encuentran en la vida, y yo tuve la grandísima suerte de encontraros hace más de una década... ¡Y por muchos años más!

Jime, Javi, Lu (de vuelta), Pablito, Teo & Marlena: lovely Gent and our Dutch lessons brought us together and now you are way more than friends… voor altijd! I love you guys!

Por último… mis pilares fundamentales: mamá, Du, Adolfo & Dzobi. Aunque estemos a decenas/cientos de kilómetros de distancia, os siento a mi lado cada día. Gracias a vosotros hoy estoy aquí y soy la persona que soy. Haciéndome más fuerte cada día y ayudándome a afrontar cada reto con toda mi ilusión y esfuerzo. ¡Os amo con todo mi corazón!

CURRICULUM VITAE

PERSONAL DETAILS

Francisco Avila Cobos
Center for Medical Genetics Ghent (CMGG), Ghent University
Blok B, room 100.011 (entrance 36)
C. Heymanslaan 10 — 9000 Gent, Belgium
+32 489 132 138
Francisco.AvilaCobos@UGent.be
favil90@gmail.com

EDUCATION

- PhD student (Feb 2015 – today): "*Addressing challenges in transcriptome annotation and cell-type heterogeneity through integration of omics datasets and computational deconvolution*".
  - Center for Medical Genetics Ghent (CMGG). Ghent University. (Belgium)
  - Special Research Fund (BOF) scholarship of Ghent University. Grant: BOF.DOC.2017.0026.01
  - Scholarship for a long stay abroad (V440318N) from the Fund for Scientific Research Flanders (FWO)
  - Thesis advisor: Prof. Katleen De Preter / Prof. Pieter Mestdagh

- MSc Bioinformatics and Computational Biology (Sept 2013 - Feb 2015): "*Qualitative insights into Meta-RNA-Seq analysis*".
  - University College Cork (Ireland)
  - Thesis advisor: Dr. Marcus Claesson

- BSc Biotechnology (Sept 2008 - June 2013): University of Leon (Spain)

SKILLS & RESEARCH INTERESTS

- Programming languages: R, Python, bash scripting, high-performance computing, MySQL.
- Bioinformatics: Data mining, statistical learning and data analysis, machine learning, multifactor dimensionality reduction, networks.

TRAINING

- HPC Unix command line, shell and Python scripting. (Ghent, May - June 2015)
- GATK best practices for variant discovery. (Ghent, 8 December 2015).Trainer: Geraldine Van Der Auwera. VIB Bioinformatics Training & Service Facility (BITS)
- The Non-Coding Genome. (Leuven, 12 May 2016). Including practical hands-on by Dr. Leonard Lipovich. VIB Bioinformatics Training & Service Facility (BITS)
- C003083 - Bioinformatics Algorithms (from MSc Bioinformatics). Ghent University. Academic course 2015-2016.
- Single Cell Analysis. (Leuven, 21 September 2016). VIB Bioinformatics Training & Service Facility (BITS)
- Specialist Workshop in Scientific Computing (SWSC2016): "Scaling your data analysis in Python with Pandas and Dask". (Ghent, 21 November 2016).
- 10x Genomics User Group Meeting in Sydney (collaboration with Decode Science). Garvan-Weizmann Center for Cellular Genomics. (December 11, 2018).

AWARDS

- GlaxoSmithKline Clinical Science Award at the Belgian Society for Pneumology Meeting (2015). Brussels, Belgium: Tania Maes, Francisco Avila Cobos et al. "*Identification of differentially expressed microRNAs in neutrophilic asthma*".
- Best poster award at the 14th ERS Lung Science Conference (LSC) "System Approaches in Lung Disease" in Estoril, Portugal (10-13 March, 2016): Griet Conickx, Pieter Mestdagh, Francisco Avila Cobos et al. "*miRNA profiling reveals a role for miRNA-218-5p in the pathogenesis of chronic obstructive pulmonary disease*".
- Best poster and short-pitch presentation at f-TALES: Cancer, an old dog with new tricks in Leuven, Belgium (6-7 March, 2017): Francisco Avila Cobos et al. "*Zipper plot: visualizing transcriptional activity of genomic regions*".
- 10X poster prize from Decode Science at Lorne Genome Conference in Lorne, Australia (17 - 19 February 2019): Francisco Avila Cobos. "*Assessing the biological signal of different RNA fractions for computational deconvolution of healthy tissues*".

PEER-REVIEWED JOURNAL ARTICLES

- Wallaert A, Durinck K, Van Loocke W, Van de Walle I, Matthijssens F, Volders PJ, **Avila Cobos F** et al. "*Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia*". Leukemia , (2016).
- Maes T, **Avila Cobos F**, Schleich F, Sorbello V, Henket M, De Preter K, Bracke K et al. "*Asthma inflammatory phenotypes show differential microRNA expression in sputum*". Journal of Allergy and Clinical Immunology , (2016).

- Conickx G, Mestdagh P, **Avila Cobos F**, Verhamme F, Maes T, Vanaudenaerde BM, Seys L et al. "*MicroRNA profiling reveals a role for microRNA-218-5p in the pathogenesis of chronic obstructive pulmonary disease*". American Journal of Respiratory and Critical Care Medicine, (2017).
- Conickx G*, **Avila Cobos F***, van den Berge M, Faiz A, Timens W, Hiemstra P, Joos G, Brusselle G, Mestdagh P and Bracke K. "*microRNA profiling in lung tissue and bronchoalveolar lavage of cigarette smoke-exposed mice and in COPD patients: a translational approach*". Scientific Reports , (2017).
- **Avila Cobos F**, Anckaert J, Volders PJ, Everaert C, Rombaut D, Vandesompele J, De Preter K and Mestdagh P. "*Zipper plot : visualizing transcriptional activity of genomic regions*". BMC Bioinformatics , (2017).
- **Avila Cobos F**, Vandesompele J, Mestdagh P and De Preter K. "*Computational deconvolution of transcriptomics data from mixed cell populations*". Bioinformatics , (2018).
- Verboom K, Van Loocke W, Volders PJ, Decaesteker B, **Avila Cobos, F**, Bornschein S, de Bock CE, Atak ZK, Clappier E, Aerts S, Cools J, Soulier J, Taghon T, Van Vlierberghe P, Vandesompele J, Speleman F and Durinck K. "*A comprehensive inventory of TLX1 controlled long non-coding RNAs in T-cell acute lymphoblastic leukemia through polyA+ and total RNA sequencing*". Haematologica , (2018).
- D'haene E, Bar-Yaacov R, Bariah I, Vantomme L, Van Loo S, **Avila Cobos F**, Verboom K, Eshel R, Alatawna R, Menten B, Birnbaum RY and Vergult S. "*A neuronal enhancer network upstream of MEF2C is compromised in patients with Rett-like characteristics*". Human Molecular Genetics , (2018).
- Lorenzi L, **Avila Cobos F**, Decock A, Everaert C, Helsmoortel H, Lefever S, Verboom K, Volders PJ, Speleman F, Vandesompele J and Mestdagh P. "*Long noncoding RNA expression profiling in cancer: Challenges and opportunities*". Genes, Chromosomes and Cancer , (2019).
- De Smet EG , Van Eeckhoutte HP, **Avila Cobos F**, Blomme E, Verhammea FM, Provoost S, Verleden SE, Venken K, Maes T, Joos GF, Mestdagh P, Brusselle GG, Bracke KR. "*The role of miR-155 in cigarette smoke-induced pulmonary inflammation and COPD*". Mucosal Immunology , (2019).
- Lucia Lorenzi, Hua-Sheng Chiu, **Francisco Avila Cobos**, Stephen Gross, Pieter-Jan Volders, Robrecht Cannoodt, Justine Nuytens, Katrien Vanderheyden, Jasper Anckaert, Steve Lefever, Tine Goovaerts, Thomas Birkballe Hansen, Scott Kuersten, Nele Nijs, Tom Taghon, Karim Vermaelen, Ken R. Bracke, Yvan Saeys, Tim De Meyer, Nandan Deshpande, Govardhan Anande, Ting-Wen Chen, Marc R. Wilkins, Ashwin Unnikrishnan, Katleen De Preter, Jorgen Kjems, Jan Koster, Gary P. Schroth, Jo Vandesompele, Pavel Sumazin, Pieter Mestdagh. "*The RNA Atlas, a single nucleotide resolution map of the human transcriptome*". bioRxiv , (2019).

- Eva Hulstaert, Annelien Morlion, **Francisco Avila Cobos**, Kimberly Verniers, Justine Nuytens, Eveline Vanden Eynde, Nurten Yigit, Jasper Anckaert, Anja Geerts, Pieter Hindryckx, Peggy Jacques, Guy Brusselle, Ken R. Bracke, Tania Maes, Thomas Malfait, Thierry Derveaux, Virginie Ninclaus, Caroline Van Cauwenbergh, Kristien Roelens, Ellen Roets, Dimitri Hemelsoet, Kelly Tilleman, Lieve Brochez, Scott Kuersten, Lukas Simon, Sebastian Karg, Alexandra Kautzky-Willers, Michael Leutner, Christa Nöhammer, Ondrej Slaby, Gary P. Schroth, Jo Vandesompele, Pieter Mestdagh. "*Charting extracellular transcriptomes in The Human Biofluid RNA Atlas*". bioRxiv , (2019).
- **Francisco Avila Cobos**, José Alquicira-Hernandez, Joseph Powell*, Pieter Mestdagh* and Katleen De Preter*. "*Comprehensive benchmarking of computational deconvolution of transcriptomics data*". bioRxiv , (2020).

CONFERENCES, WORKSHOPS & MEETINGS

- Keystone symposia on Molecular and Cellular Biology, Abstracts. Keystone symposia on Molecular and Cellular Biology: Noncoding RNAs in health and disease (Q5). Santa Fe, New Mexico (USA), 21-24 February 2016.
- Applied Bioinformatics in Life Sciences. Leuven, 17-18 March 2016. BIG-N2N Multidisciplinary Seminar Series on Bioinformatics. Academic course 2015-2016. f-TALES workshop: "Light on the dark side of the genome". Ghent, 15-16 September 2016.
- f-TALES workshop: "Cancer: an old dog with new tricks". Leuven, 6-7 March 2017.
- LKI Symposium: "Liquid Biopsies & Cancer". Leuven, 28-30 August 2017.
- 18th annual Belgian Society for Human Genetics: "The epigenome in development and disease". 16 February 2018. Ghent, Belgium.
- International Society for Computational Biology 2018 (ISMB). 6 - 10 July 2018. Chicago, Illinois (USA).
- Lorne Genome Conference. 17 - 19 February 2019. Lorne, Victoria (Australia).
- International Society for Computational Biology 2019 (ISMB). 21 - 25 July 2019. Basel, Switzerland.

ORAL PRESENTATIONS

- f-TALES: Cancer, an old dog with new tricks (6-7 March 2017. Leuven, Belgium). "*Zipper Plot: Visualizing transcriptional activity of genomic regions*".
- 18th annual Belgian Society for Human Genetics: the epigenome in development and disease. Session IV: Bio-informatics and Functional Genetics (16 February 2018. Ghent, Belgium). "*Zipper Plot: Visualizing transcriptional activity of genomic regions*".

- Statistical Bioinformatics - School of Mathematics and Statistics and the Integrative Systems and Modelling Theme at the Charles Perkins Centre (13 May 2019. University of Sydney, Australia). "*Impact of data transformation, pre-processing and choice of method in the computational deconvolution of transcriptomics data*".
- "*Comprehensive benchmarking of computational deconvolution of transcriptomics data*" - 2nd Health Data Challenge: Matrix factorization and deconvolution methods to quantify tumor heterogeneity in cancer research. (25 - 29 November 2019. Centre Paul Langevin. Aussois (France)).

POSTERS

- Francisco Avila Cobos, Ken Bracke, Jo Vandesompele, Kimberly Verniers, Guy F. Joos, Guy G. Brusselle, Pieter Mestdagh and Katleen de Preter. "*LncRNAs in the pathogenesis of Chronic Obstructive Pulmonary Disease (COPD) are enriched for genes involved in metabolism of xenobiotics and diverse immune responses".* Keystone Symposia Conference. Q5: Noncoding RNAs in Health and Disease. 21 - 24 February 2016. Santa Fe, New Mexico (USA).
- Francisco Avila Cobos, Jasper Anckaert, Pieter-Jan Volders, Jo Vandesompele, Katleen De Preter and Pieter Mestdagh. "*Zipper Plot: Visualizing transcriptional activity of genomic regions*". BIG N2N symposium. 19 May 2016.
- Francisco Avila Cobos, Lucia Lorenzi, Jo Vandesompele, Gary Schroth, Katleen De Preter,* and Pieter Mestdagh*. "*Assessing the biological signal of different RNA fractions for computational deconvolution of healthy tissues*". International Society for Computational Biology 2018 (ISMB). 6 - 10 July 2018. Chicago, Illinois (USA).
- Francisco Avila Cobos, Lucia Lorenzi, Jo Vandesompele, Gary Schroth, Katleen De Preter,* and Pieter Mestdagh*. "*Assessing the biological signal of different RNA fractions for computational deconvolution of healthy tissues*". Lorne Genome Conference. 17 - 19 February 2019. Lorne, Victoria (Australia).
- Francisco Avila Cobos, José Alquicira-Hernandez, Jo Vandesompele, Joseph Powell, Pieter Mestdagh and Katleen De Preter: "*Benchmarking the impact of data transformation, preprocessing and choice of method in the computational deconvolution of transcriptomics data*". International Society for Computational Biology 2019 (ISMB). 21 - 25 July 2019. Basel, Switzerland.

THESES SUPERVISED

- Fien Verhamme - Howest Brugge (Dec 2017 - June 2018). Bachelor after Bachelor in Bioinformatics: "*Exploratory data analysis on Total RNA Sequencing from COPD patients*".

- Elias Vermeiren - Ghent University (Feb - June 2017). Master in Biotechnology and Biochemistry (specialization: Systems Biology): "*Differential tRNA expression analysis in healthy versus prostate cancer samples*".
- Hataitip Tasena - Universitair Medisch Centrum Groningen (Nov - Dec 2015): "*MicroRNA profiling of COPD patients with and without mucus hypersecretion*". (Data analysis).