KU Leuven
Biomedical Sciences
Group Faculty of Medicine
Department of Oncology

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

BIOMEX: A SOFTWARE TO ALLEVIATE THE BOTTLENECKS IN THE ANALYSIS-TO-INTERPRETATION PIPELINE OF OMICS DATASETS

Federico TAVERNA

Supervisor: Prof. Peter Carmeliet
Co-supervisor: Prof. Tobias Karakach
Co-supervisor: Dr. Abhishek Subramanian
Chair examining committee: Prof. Ghislain Opdenakker
Chair public defense: Prof. Baki Topal

Jury members:
Prof. Massimiliano Mazzone
Prof. Bernard Thienpont
Prof. Ton Rabelink
Prof. Vincenzo Lagani

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Biomedical Sciences

January 2021

Table of Contents

List of abbreviations **1**
Summary **2**
Samenvatting **4**
Chapter I General introduction **6**
Data science in the omics field 6
Shifting the paradigm 6
The bottlenecks in omics data analysis and interpretation 6
Analysis of omics data 7
Raw data quantification 8
Data pretreatment 9
Transcriptomics by micro-array 9
Transcriptomics by RNA-sequencing 10
Transcriptomics by single cell RNA-sequencing 11
Metabolomics and Proteomics 12
Mass cytometry 13
Unsupervised downstream analyses 13
Dimensionality reduction 13
Clustering 14
Heatmaps 15
Supervised downstream analyses 15
Differential analysis 15
Set enrichment analysis 17
Marker set analysis 18
Correction for false discovery rate (FDR) 19
Visualization of trends 19
Single cell specific algorithms 19
Trajectory inference 19
Cluster projection 20
Feature engineering 20
Set variation analysis 21
Survival analysis 21
Machine learning 22
Meta-analysis 23
Bulk omics data 23
Single cell omics data 24
Accessible analysis software in bioinformatics 25
Chapter II Research objectives **26**
Chapter III BIOMEX **27**
Abstract 28
Introduction 28
Overview 29
Functional requirements and design rationale 29
Manual and video tutorials 31
Data import: data and metadata matrix 31
Data analysis modules 32

Module 1: Data pretreatment....32
Module 2: Feature engineering....32
Module 3: Visualization of trends....33
Module 4: Unsupervised analyses....33
Module 5: Supervised analyses....33
Module 6: Single cell specific algorithms....34
Module 7: Survival analysis....34
Module 8: Machine learning....34
Module 9: Meta-analysis of bulk omics data....35
Module 10: Single cell meta-analysis....35
Data export....35
Case studies....35
Bulk data: exploration of the TCGA cholangiocarcinoma dataset....36
Single cell data: re-analysis of the endothelial cell atlas dataset....38
Conclusion....38
Supplementary tables....40
Supplementary data....42
Author contributions....42
Conflict of interest declaration....42
Funding....42
Author information....42
References....42
Chapter IV EndoDB....47
Abstract....48
Introduction....48
Materials and Methods....49
Retrieval of EC transcriptomics data and database content....49
Pre-processing and quality control of bulk transcriptomics data....50
Pre-processing and quality control of single cell data....52
Metadata curation....52
Data analysis and visualization....52
Website implementation....53
Data availability and downloads....53
Documentation and manual....53
EndoDB functionality....55
Study-centered data exploration....55
Metabolism- and transcription factor centered data exploration....55
Gene- and pathway-centered data exploration....57
Single cell data exploration....57
Conclusion....59
Data availability....60
Supplementary data....60
Funding....60
Conflict of interest declaration....60
Author contributions....61
Acknowledgements....61
Author information....61

References....61
Chapter V EC heterogeneity at single cell level....65
Abstract....66
Significance....66
Introduction....66
Taxonomy results....67
Single cell atlas of EC phenotypes in human lung cancer....67
Possible clinical implications....72
Lung tumor EC phenotypes in other species and models....72
Effect of VEGF blockade on TEC phenotypes....75
Differential sensitivity of TEC phenotypes....75
Molecular signature of tumor vessel disorganization and effect of VEGF blockade....76
Identification of conserved TEC phenotypes and tip cell markers....76
Integrated analysis highlights collagen modifications as an angiogenic pathway....80
Conclusion....81
Supplementary data....84
Author contributions....85
Conflict of interest declaration....85
Funding....85
Author information....85
References....86
Chapter VI Discussion....88
Contribution of BIOMEX in research....88
BIOMEX showcase: the database of endothelial cell transcriptomics data (EndoDB)....89
BIOMEX showcase: the lung endothelial single cell taxonomy....91
Conclusion....93
Design philosophy....93
Future perspectives....93
Limitations....94
Software updates....95
Final words....96
Chapter VII References of Chapter I and VI....97
Chapter VIII Acknowledgements and personal contribution....103
Scientific acknowledgements....103
Personal acknowledgements....103
Personal contributions....103
Conflict of interest statement....104
Chapter IX Curriculum Vitae....105

List of abbreviations

List of abbreviations used in Chapter I and VI.

Abbreviation	Full name
AI	Artificial intelligence
ANN	Artificial neural networks
BIOMEX	Biological Interpretation Of Multi-omics EXperiments
CART	Classification and regression tree
DNA	Deoxyribonucleic acid
DT	Decision tree
FDR	False discovery rate
GSEA	Gene set enrichment analysis
GSVA	Gene set variation analysis
iid	Independent and identically distributed
IQR	Interquartile range
kNN	k-nearest neighbors
LLC	Lewis lung carcinoma
LOESS	Local regression
m/z	Mass-to-charge ratio
MACS	Magnetic-activated cell sorting
ML	Machine learning
MNAR	Missing not at random
mNN	Mutual nearest neighbors
mRNA	Messenger RNA
MS	Mass spectrometer
MSigDB	Molecular signatures database
NEC	Normal endothelial cells
NGS	Next generation sequencing
PCA	Principal component analysis
PCs	Principal components
PLS-DA	Partial least squares discriminant analysis
PP	Projection pursuit
QC	Quality control
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
SAM	Significance analysis of microarrays
scRNA-seq	Single cell RNA sequencing
SOM	Self-organizing maps
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
TEC	Tumor endothelial cells
TMM	Trimmed mean of M-values
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection
VEGF	Vascular endothelial growth factor
VSN	Variance stabilizing normalization

Summary

Biomedical research has witnessed a tsunami of data since the inception, more than two decades ago, of high throughput measurement technologies, including next-generation sequencing, mass spectrometry, fluorescence imaging and other assays. This growth was accelerated by the increased capabilities to acquire measurements at a single cell resolution, using technologies for single cell RNA and DNA sequencing as well as mass and flow cytometry. The combination of these technologies with innovative experimental designs and advanced data mining has provided unprecedented insights and yielded breakthroughs, except for one of the most pressing problems in biology: an unresolved challenge to move from descriptive analysis towards mechanistic insight and clinical translation. In contrast to the increasing commoditization of high-throughput technologies, there is a limited availability of software that allows experimental biologists, who are lacking bioinformatics expertise but are interested in solving biological questions, to discover novel patterns and targets, thereby creating a bottleneck in the data interpretation pipeline.

To alleviate these bottlenecks, I developed BIOMEX, a flexible and easy-to-use data mining software developed for the Biological Interpretation Of Multi-omics EXperiments to bridge the gap between experimental biologists and access to biological information embedded in omics data. BIOMEX is designed specifically to allow bench scientists with little to no bioinformatics expertise to analyze and visualize a large range of data derived from single cell and bulk omics experiments in a systematic manner. The software suite integrates the most current and validated data analysis methods and statistical tools into a unified, well-defined and guided stepwise workflow, including data pretreatment and normalization, dimensionality reduction, clustering, differential and enrichment analysis, pathway mapping, trajectory inference, survival analysis, predictive modeling with machine learning, and more. All results are displayed via interactive visualization using the R/Shiny framework together providing a flexible workflow that encourages analytical creativity and is at the same time sufficiently rigid to allow bench scientists to use advanced computational tools. Moreover, the software allows users to generate high quality publication-ready figures.

Previous software solutions were an important first step in unlocking the potential of omics data, but with the surge in complexity of both high-dimensional datasets and biological questions, a broadly accessible data mining software is needed to unravel their intricacies.

BIOMEX provides several advantages; it has a user-friendly, highly interactive web-

like interface that allows users who are not experts in bioinformatics, to analyze data generated to answer complex biological questions and fine-tune the analyses interactively. In addition, it allows different omics data, such as metabolomics, transcriptomics, proteomics, single cell RNA-sequencing and mass cytometry data to be simultaneously uploaded and analyzed into one consolidated platform. BIOMEX can facilitate the reproducibility of analyses, where results can be shared between scientists and uploaded into the software with ease. Also, it can execute meta-analyses between different experiments and enable the exploration of publicly available datasets. To make BIOMEX accessible to non-bioinformatics experimental biologists, I provide a comprehensive manual and video tutorials, where each step of the analysis is explained, potentially making the software a good starting point to introduce researchers to the field of biological data science.

BIOMEX is multi-platform, free, and has been field-tested in recent publications. Since BIOMEX can be easily expanded and customized with new algorithms, I anticipate that the software has the potential to extract insight from complex datasets and help alleviate the key bottlenecks in the interpretation of biological omics data.

In this doctoral thesis, three manuscripts are presented: one main manuscript focuses on how I developed BIOMEX and describes its functionalities and capabilities, and two supporting manuscripts, in which I showcase how BIOMEX can be used to analyze different omics datasets.

Samenvatting

Het biomedische onderzoek is de laatste 2 decennia getuige geweest van een data-tsunami door ‘high throughput’ meettechnologieën, waaronder next-generation sequenering, massaspectrometrie, fluorescentiemicroscopie en andere assays. Deze groei werd verder versneld door de toenemende mogelijkheid om metingen te doen met een resolutie van één enkele cel, daarbij gebruik makend van verschillende technologieën zoals single cell RNA- en DNA-sequenering, evenals massa- en flowcytometrie. De combinatie van deze technologieën tesamen met innovatieve experimentele ontwerpen en geavanceerde datamining heeft tot ongekende inzichten en doorbraken geleid. Op één uitzondering na, en dat is net één van de meest acute problemen binnen het biomedisch onderzoek: de onopgeloste uitdaging om van beschrijvende analyse te komen tot mechanistische inzichten en klinische toepassingen. In tegenstelling tot de toenemende inburgering van high-througput technologieën, is er maar een beperkte beschikbaarheid aan software waarmee experimentele biologen zonder bio-informatica expertise overweg kunnen. Hierdoor ontstaat er een knelpunt in de gegevensverwerking en interpretatie, wat de ontrafeling van biologische vraagstukken en de ontdekking van nieuwe patronen en ‘targets’ vertraagt.

Om deze kloof tussen experimentele biologen en toegang tot biologische informatie die in omics-data verscholen zit, te overbruggen, heb ik BIOMEX ontwikkeld. BIOMEX is een flexibele en gebruiksvriendelijke datamining-software die helpt met de biologische interpretatie van multi-omics-data en is specifiek ontworpen voor wetenschappers met weinig tot geen bio-informatica expertise. Met behulp van BIOMEX kunnen zij op een systematische manier een breed gamma aan data afkomstig van single cell en bulk omics-data analyseren en visualiseren. De software integreert de meest actuele en gevalideerde data-analyse methoden en statistische tests in een uniforme, goed gedefinieerde en stapsgewijze workflow. Hierin zit vervat: de datavoorbehandeling en normalisatie, dimensionele reductie, clustering, differentiële- en verrijkingsanalyse, pathway mapping, pseudotemporale ordening, overlevingsanalyse, voorspellende modellering met machine learning en nog veel meer. Alle resultaten worden weergegeven via een interactieve visualisatie waarbij gebruik gemaakt wordt van het R/Shiny-framework. Dat, samen met de flexibele workflow van BIOMEX, stimuleert de analytische creativiteit van wetenschappers om toch geavanceerde computertools te gebruiken terwijl de data-analyse voldoende rigide blijft. Bovendien stelt de software gebruikers in staat om hoogwaardige publicatieklare figuren te genereren.

Bestaande softwareoplossingen waren al een belangrijke eerste stap in het verwerken

van het hoge potentieel aan omics-data. Maar met de toenemende complexiteit van zowel hoog-dimensionele datasets en biologische vragen, is er echter een breed toegankelijke datamining-software nodig om de kleinste details te ontrafelen.

BIOMEX biedt verschillende voordelen; het is een zeer gebruiksvriendelijke, interactieve website-gelijkende interface waarmee gebruikers zonder al te veel bio-informatica kennis de door hen gegenereerde data kunnen analyseren. Zo kunnen ze complexe biologische vraagstukken eenvoudiger beantwoorden en interactieve analyses uitvoeren om hun inzicht in de data te vergroten. Bovendien kunnen verschillende omics-data, zoals metabolomics, transcriptomics, proteomics, single cell RNA-sequencing en massa-cytometrie data gelijktijdig opgeladen en geanalyseerd worden en dat op één en hetzelfde platform. BIOMEX kan de reproduceerbaarheid van analyses vergemakkelijken door wetenschappers hun resultaten gemakkelijk te laten delen en in te laden in de software. Daarbovenop kunnen meta-analyses tussen verschillende experimenten worden uitgevoerd wat ook verder onderzoek van beschikbare datasets mogelijk maakt. Om BIOMEX toegankelijk te maken voor experimentele biologen zonder bio-informatica kennis is er een uitgebreide handleiding en video tutorials voorzien waarin de analyses stap voor stap worden uitgelegd. Zo krijgen alle onderzoekers de mogelijkhied om geïntroduceerd te worden in het gebied van de bio-informatica.

BIOMEX is een gratis en multi-platform software waarvan het nut al werd aangetoond in recente publicaties. BIOMEX kan makkelijk uitgebreid en aangepast worden met nieuwe algoritmen. Ik verwacht dan ook dat de software het potentieel in zich heeft om verder inzicht te verwerven in complexe datasets en kan bijdragen om de belangrijkste struikelblokken van biologische omics-data interpretatie weg te nemen.

In deze doctoraatsthesis worden drie manuscripten gepresenteerd. Eén hoofdmanuscript focust op hoe ik BIOMEX heb ontwikkeld en beschrijft de functionaliteiten en mogelijkheden ervan. Daarnaast worden er nog twee ondersteunende manuscripten gepresenteerd, waarin ik laat zien hoe BIOMEX kan gebruikt worden om verschillende omics-datasets te analyseren.

Chapter I General introduction

This chapter is based on the supplementary material of the publication shown in Chapter III.
https://academic.oup.com/nar/article/48/W1/W385/5835814#supplementary-data

Data science in the omics field

Shifting the paradigm

Before the advent of omics technologies, there was no straightforward method to study or measure the abundance of molecules at a global level in a biological system. For example, traditional molecular biology techniques employed to measure the levels of mRNA such as Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR) (1), while highly accurate in measuring the expression of single genes, cannot be used effectively to measure the expression of the global mRNA present in the transcriptome. Approximately, there are around 20,000-30,000 protein encoding genes in the human genome (2), which would be impractical to analyze with traditional techniques.

With the introduction of unbiased high-throughput profiling technologies, the paradigm of how data is generated and analyzed changed remarkably (3), enabling scientists to measure the abundance of many molecules simultaneously (4). With the current mass spectrometry and next generation sequencing technologies (NGS), the simultaneous measurement of thousands of proteins, metabolites and genes (mRNA) is now possible; the full human genome can even be sequenced in under one hour (5). Thanks to recent advancements, molecular measurements can be carried out specifically at a single cell level (6). This revolutionary approach has shifted the paradigm even more, since it allows the phenotypic characteristic of each single cell to be studied in detail (7).

The bottlenecks in omics data analysis and interpretation

The commoditization of high-throughput (single cell) omics technologies has spurred the exponential growth of valuable biomedical data, but the capacity for non-bioinformatics experts to analyse these increasingly complex datasets did not scale accordingly (8). Indeed, across fields, an unresolved challenge is how to move from descriptive analysis of biomarkers and cellular taxonomies towards mechanistic insight, therapeutic targets and clinical translation (9-12).

Availability of large and complex omics datasets introduces both computational and biological challenges: the data has to be processed and analyzed, which requires

computational expertise, but the data needs to be biologically interpreted, which requires in-depth biological expertise. It is seldom that an individual will have both the detailed computational and biological knowledge required to handle emergent omics data, creating a bottleneck in the data-analysis-to-interpretation pipeline.

The bottleneck in the analysis and interpretation pipeline of omics data is becoming one of the more prominent and recognized issues in biomedical research: “the bottleneck in omics approaches is becoming less and less about data generation and more and more about data management, integration, analysis, and interpretation” quoted from (13).

Making advanced state-of-the-art computational tools to analyse high-dimensional data available to bench scientists via easy-to-use data mining software could effectively address these bottlenecks. This can be achieved by providing a workflow for automated data parsing, cleaning and script customization to accelerate turn-around-time (14), while capitalizing on the domain-knowledge of bench scientists to increase the depth and biological/clinical relevance of each analysis (8,15).

Analysis of omics data

Many different omics data types exist, and even though each omics data type is generated via different techniques, a common high-level pipeline can be used to analyze such data. Analyzing omics data is a multi-step process, which can be divided in four main steps:

1. Raw data quantification: the raw data is numerically quantified, which means the abundance of each molecule (*e.g.* mRNAs, metabolites, proteins) is assessed (16-18).
2. Pretreatment: unwanted technical variation and systematic biases are removed, as the underlying assumption of the next analysis step is that the most significant source of variation in the data should be of the biological process of interest (19).
3. Downstream analysis: biological insights into the processes under study are provided based on statistical and methodological inferences (20). This includes unsupervised and supervised exploratory analyses, machine learning and many others, including custom-made tailored methods.
4. Meta-analysis: these approaches can be used to analyze multiple datasets simultaneously, combining the results of independent experiments in one singular, robust result (*e.g.* combining downstream analysis results together) (21).

An extensive amount of tools was created to meet the challenges related to the aforementioned analysis steps, each with their advantages and disadvantages. In the

following sections, the baseline principles which these tools are based on will be explained.

Raw data quantification

Biological samples are processed and quantified with different instruments, selecting the right technology based on the type of molecules that needs to be measured. The computational structure of the data files obtained from these instruments is entirely dependent on the employed technology: these files are called raw data, and they need to be computationally processed using dedicated tools, to enable extraction of biological insight.

There are different ways to obtain genomic sequences. One common approach is to label the DNA using fluorophores, marking the different nucleotides with different colors. The light signal is then acquired via a photo-sensor that discriminates between the different colors, enabling the reconstruction of the DNA sequence. The DNA sequence is saved in *.FASTQ* files (22). A similar approach can be used to sequence mRNA fragments (transcriptomics). The sequence of these fragments is also saved in *.FASTQ* files, which are subsequently mapped to a predefined genome (*e.g.* the human genome) to quantify the amount of mRNA present in each gene (6).

On the other hand, metabolites abundances and protein sequences can be obtained by using mass spectrometers. In mass spectrometry-based metabolomics, charged molecules can be separated based on their mass-to-charge (m/z) ratio and their intensities are recorded in *.RAW* files to produce a chromatogram, in which the intensity of each peak is proportional to the intensity of the molecule represented by that peak (23). In mass-spectrometry based proteomics, protein fragments called peptides are measured. These peptides are fragmented, and in general the difference between the m/z peaks of the peptide fragment ions can be used to reconstruct the peptide sequence (24).

The abundance of mRNA sequences is quantified by calculating the number of reads mapped to a reference genome. The abundance of metabolites/proteins is calculated by integrating their chromatogram peaks (*i.e.* calculating the area of the peaks).

Regardless of the type of molecules being analyzed, the output of this step is a data matrix, in which every molecule in each sample is numerically quantified. The rows of the matrix represent the biological molecules taken into consideration, which are referred to as features. The columns of the matrix represent the sample/cell identifiers, which are referred to as observations. Each element in the matrix represents the abundance of any given biological molecule quantified with respect to a given sample or cell.

Data pretreatment

The data pretreatment step is completely dependent on the type of omics data under analysis. For each common omics data type, a brief description is provided. The output of this step is a clean data matrix that can be used during downstream analysis.

Transcriptomics by micro-array

One of the most widely used platforms for transcriptomics by microarrays is the single channel Affymetrix Arrays. This technology differs from the two-color arrays that employ the principle of competitive hybridization where observations for each feature can only be treated as a ratio. In the Affymetrix Arrays, each feature measurement represents an absolute signal intensity corresponding to the transcript abundance hybridized on a given probe set (25). Because of the relatively large dynamic range in the absolute signal measured by these arrays, the data are typically transformed into a log scale. This transformation allows the data distribution to approximate the Gaussian probability distribution with well-defined characteristics that allow the data to be analyzed using linear methods. The choice of the base 2 log transformation is historical rather than mathematical; in the most common experimental design paradigms in biology (case-control), a base 2 log transformation provides symmetry.

The data has to be corrected for background intensity, which is a vital step in all microarray data analyses. This need arises because the probe intensities measured correspond to two portions of the signal. One portion can be attributed to non-specific hybridizations and other factors related to the optical detection systems as covered in detail in the previously mentioned reference (25). This portion of the signal is referred to as the “background” and must be removed from the observed signal to give the accurate measurement of the signal due to probe-specific hybridization.

Given that Affymetrix does not utilize the principle of competitive hybridization, each sample is therefore measured on an independent array. In order to compare measurements across these different arrays, therefore, it is important to equalize the signal intensities across these arrays and minimize systematic variation resulting from many sources. This process of normalization can also be referred to as “batch correction” and in many applications is predicated on the simple hypothesis that in a typical microarray experiment, any robust measure of central tendency (*e.g*. median, inter-quantile range (IQR), etc.) should remain constant across all measurements. The different arrays are therefore corrected to equalize this measure of central tendency (26).

Mathematically, data preprocessing in microarrays can be easily represented by two

equations. Supposing that a measured signal for transcript *i*, on array *j* is represented as $t_{i,j}$, the observed signal is however a function of other factors represented as:

$$y_{i,j} = B + \alpha_j t_{i,j} \quad (1)$$

where *B* is the background and α_j is the array specific systematic coefficient that must be corrected via normalization. After background correction and normalization, the true signal can now be represented as:

$$\log(t_{i,j}) = \mu + \theta + \epsilon \quad (2)$$

where μ represents the logarithm of the true abundance while the second and third parameters on the right-hand side represent probe specific effects and random errors respectively. Motakis et al. (27) describe the distributional characteristics of this signal, which is the basis for the variance stabilizing transformation (VSN) approach used to pre-process microarray data.

Transcriptomics by RNA-sequencing

In contrast to microarray technologies (where fluorophores are used as reporters and quantified to represent molecular intensities as surrogates for expression levels), gene expression by RNA sequencing (RNA-seq) involves directly counting the number of sequenced RNA fragments that map to a reference genome or transcriptome. Inherently, this produces count data that, in principle, can be modelled using Poisson statistics. These data have, nonetheless, been shown to follow a negative binomial distribution (which is closely related to the Poisson distribution) and are modelled as such by using approaches such as DESeq and DESeq2 (28). However, following log transformation, the distribution of these data closely approach the Gaussian distribution and can then be modelled using classical linear regression based methods such as limma (29).

The standard workflow for analysis of RNA-seq data is similar to that used in microarray data. The first step is the filtering of the data. Filtering simply involves removing features that have no expression values across a number of measurements. Although the goal of many transcriptomics experiments (following the case-control design) is to examine genes that are differentially expressed between the conditions, some features are undetected throughout all samples. In fact, the number of such features can amount to up to 20% in some studies.

Similar to microarray experiments, non-biological factors that arise from sample preparation or the sequencing process can affect the overall level of gene expression in individual samples. During differential expression analysis, it is assumed that the majority of genes in all samples exhibit a similar linear dynamic range and distribution of expressed values.

There are several strategies for normalizing the data in order to ensure that the measurements are compared on a scale that follows this basic assumption so that the distribution of gene expression values for all samples is comparable across the entire experiment. These methods have been covered extensively in different reviews (30,31) and include Trimmed Mean of M-Values (TMM) and was developed by Robinson and Oshlack (32). This method is based on the core principle hypothesis in transcriptomics experiments that most genes are not differentially expressed. Therefore, a trimmed weighted mean of log expression values is calculated (after removing the upper and lower percentage of the data), and is used as the normalization factor.

Transcriptomics by single cell RNA-sequencing

The basic principle of single cell RNA-sequencing (scRNA-seq) is the same as that employed for bulk RNA-seq with the caveat that gene transcripts are extracted from a single cell rather than averaged from a sample homogenate. The distributional characteristics have a distinct difference from bulk RNA-seq data because of the increased number of missing values. A big challenge for scRNA-seq data is that it is difficult to determine if missing values are due to technical issues (as with bulk RNA-seq) or if these arise because of the biological phenomenon referred to as transcriptional bursting (33). Nonetheless, a pragmatic approach is taken in most cases where an arbitrary threshold (such as 0.01) is selected to represent the minimum average expression allowed for a given transcript. In addition, because of empirical protocols for single cell isolation, these data often contain cells that are considered “dead” and will therefore have a significant number of expressed mitochondrial RNA. Such cells are excluded from further analyses (based on the percentage of expression of the mitochondrial genes).

As one can imagine, data normalization in scRNA-seq (34) presents a different challenge than for their bulk counterparts. Issues such as variation in sequencing depth (number of genes or molecules detected per cell) and the linear dynamic range spanning orders of magnitude are common between the bulk RNA-seq and scRNA-seq. They are however, accentuated in scRNA-seq data due to the extreme sparsity of the data matrices. A popular normalization method is the one found in the Seurat package (35) and is loosely

referred to as the Seurat normalization (library size normalization followed by a natural log transformation). This can be followed by batch correction to remove unwanted technical variation. There are several methods for batch correction, among them principal component analysis (PCA) and many other latent variable regression models in general. The basic goal of batch correction is to find a regression coefficient that maps data set *A* to data set *B* such that the residuals of that regression are independent, identically distributed and follow a normal distribution (*iid* normal, $E \sim N(0,1)$). This effectively removes the "batch effect" which is the regression coefficient. Among methods for this kind of regression is the mutual nearest neighbor method (mNN), which is a variant to the more familiar k-nearest neighbor (kNN) method for non-parametric regression.

Metabolomics and Proteomics

Metabolomics and proteomics data measured by mass spectrometry exhibit the same general data characteristics. Analysis of the data follows the same framework as those of transcriptomics in the sense that we employ the workflow of data filtering and normalization.

In most cases, this kind of data contains features that are not quantified in all replicates, resulting in data with missing values, which can be filled in using imputation methods. Data imputation can only be done for features that meet a cut-off threshold of missing values for all samples. For example, we can discard proteins/metabolites that are not detected in at least 10% of the samples (different thresholds can be selected). Missing values that survive the original filtering criterion can then be imputed. Typically, missing values can be randomly distributed throughout the dataset (technical missingness) or they may also be missing because they are not detected in one condition relative to another (biological missingness). The presence of missing not at random (MNAR) can indicate that features are below the detection limit in specific samples, which could be very well the case in proteomics experiments. Missing value can be imputed by using the mean, median and minimum abundances values for each feature in the data.

In omics measurements, systematic fluctuations in the signal intensity levels between different experimental runs/timeblocks can often dominate the biological information. In addition to quality control (QC) samples that are interspersed within an experimental timeblock, normalization methods are often used to remove these biases that are, in most case, observable among these QC samples. One of the most prominent approaches for this is the locally weight scatter smoothing (LOESS/LOWESS) (36) approach, which is a flexible method for linear or non-linear regression. LOWESS is a flexible method that constructs local regressions on a collection of points in a similar way to a moving average. During the

regression, each point within (the span of local points) is weighted according to its distance around the predictors in order to estimate its ordinal value. LOESS has been used extensively in microarray normalization (37) and also found widespread use in LC-MS metabolomics and/or proteomics (38).

Signal realization for MS data exhibit similar characteristics as those of micro-array data and are therefore treated the same way regarding log transformation. This allows the data to be on a comparable scale while homogenizing the variance.

Mass cytometry

Mass cytometry is a counterpart to scRNA-seq and fuses flow cytometry and elemental time-of-flight mass spectrometry to increase the number of cellular parameters that can be quantified simultaneously (39). The data characteristics are therefore a cross between those of proteomics by MS but also exhibit the sparsity of scRNA-seq data.

CyTOF data typically exhibit skewed distributions with a large linear dynamic range as is common among these data types. Unlike RNA-seq and MS data in which the data are log transformed, the arcsinh (40) has been shown to be the most appropriate transformation to symmetrize the distributions and allow the data to be on a comparable scale.

Unsupervised downstream analyses

The premise for most omics studies is that under different pathophysiological conditions, it is possible to classify samples based on the expression levels of key identifying features (*e.g.* proteins, metabolites or transcripts). This classification is achieved using the so-called exploratory data analysis techniques that allow visualization of sample differences and determine the molecules that are responsible for such differences.

Unsupervised methods of exploratory data analysis take no input from the user about sample identities and instead rely on inherent degree of (dis)similarities among the measured variables. There are many choices for what the measure of (dis)similarity is and prominent ones include variance, as implemented via Principal Component Analysis (PCA), Gaussianity, skewness or kurtosis as implemented in projection pursuit (PP) among other methods.

Dimensionality reduction

Methods such as PCA are referred to as bilinear latent variable methods because the original data are projected onto a lower dimensional coordinate system. The original data can be summarized by a set of two linear matrices of lower dimension (but equal rank) to the original

raw data matrix, and their outer product should in principle reproduce the original raw data, without the spurious noise signal. When plotted against each other, columns or rows of these data reveal underlying (dis)similarities among the samples and the features that are important for those (dis)similarities.

These methods are, however, applied under the assumption that the measured data exhibit a multivariate normal distribution and that the measurement uncertainties are independent, identical and follow a normal distribution (*iid* normal). To achieve this, the data are often scaled using different approaches including auto-scaling (z-score normalization), range scaling, Pareto scaling among other approaches discussed in van den Berg et al. (41).

Although PCA has been successful in analyzing bulk RNA-seq data that have more variables than samples, their applicability in scRNA-seq and CyToF data has been met with mixed success. Instead, t-Distributed Neighbor Embedding (t-SNE) (42) has become almost the *de facto* approach for analyzing these data. This is because, unlike PCA which is a linear projection, t-SNE use local projections that allow non-linearity in the data to be modelled. Other methods that have become common for scRNA-seq data analysis include Uniform Manifold Approximation and Projection (UMAP) (43), which claims to better conserve the local structure of the data compared to the t-SNE approach.

Sun et al. (44) has reviewed a collection of dimensionality reduction methods, and assessed how they behave in different scenarios. PCA, t-SNE and UMAP were deemed robust methods, as testing them with different data normalization approaches (log CPM and z-scores) did not appear to substantially influence the performance of these methods: the normalized mutual information score was similar across the different normalization methods.

Clustering

Clustering methods allow biological entities with similar properties to be assigned to the same cluster by using various (dis)similarity measures. Clustering can reveal meaningful patterns or unknown associations in data. Clustering methods include Hierarchical clustering, K-means clustering, graph-based method such as the ones included in the Seurat package (35) and self-organizing maps (SOM) included in the FlowSOM package (45) (which is used for CyTOF data). While classic methods like Hierarchical clustering and K-means are widely used, methods like Seurat and FlowSOM are better suited to specifically analyze single cell data.

An extensive review by Duò et al. (46) shows the performance of classic methods and a variety of single cell specific methods (including Seurat and FlowSOM). In the review, clustering algorithms were compared on real and synthetic datasets, calculating the Hubert-

Arabie Adjusted Rand Index (47) score to assess how the inferred clusters approximated the true populations. Seurat was one of the top clustering methods, together with SC3 (48). It was found that FlowSOM performs well as long as the number of putative clusters selected by the user is increased beyond the true number of clusters.

Heatmaps

Heatmaps are used, in general, to represent data in order to visualize the differential expression of features across the samples. They may also be combined with a variety of clustering methods to group features or samples together based on some similarity measure. The most common approach in omics data analysis is to combine heatmaps with hierarchical clustering which, from a technical point of view, entails iteratively calculating the (dis)similarity between samples and joining them (into small clusters) to form increasingly larger ones. This continues until all samples have been accounted for at the top of the hierarchy. This is often represented using dendrograms that are appended on either side of a heatmap to represent observation or feature clustering. The confidence of each branch of the dendrogram can be estimated by using a bootstrap resampling approach (49).

Supervised downstream analyses

Differential analysis

Differential expression analysis is a concept that is uniquely rooted in molecular biology and has been in use even before high throughput measurements became pervasive. Under these circumstances, the goal was simply to determine differences in the absolute expression levels of one or two features at a time. In the advent of omics measurements however, the meaning of differential analysis took a different definition because the measurement characteristics became more complex, given that the data were acquired simultaneously. Nonetheless, differential analysis remains an attractive way of examining important differences in levels of key features in response to a given condition. Several statistical methods have been developed to facilitate this in view of the complexity in data structures that are commensurate with the complexity in measurement technologies (50-53).

These methods can be classified into two categories: parametric and non-parametric. Parametric methods assign a statistical distribution to the data and assume that the data follow this distribution. They then model the parameters of this distribution to capture the information in these data. In differential gene expression analysis, for example, it is assumed that, after normalization, the measured expression value for each gene can be mapped onto a normal distribution (DNA microarrays) or Poisson/Negative Binomial (RNA-seq). This is in contrast to

non-parametric methods that do not impose a rigid parametric model about the data distribution.

The most prominent parametric methods for differential gene expression analysis are edgeR (54), DESeq (28) and limma (29). Given that RNA-seq experiments arise from counting the number of reads that map to a given location on the genome/transcriptome, it follows then that they would naturally follow a Poisson distribution represented as:

$$P(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{K!} \quad (3)$$

where *k* is the number of read counts and λ is the expected number of counts. In fact, both the Poisson and Negative Binomial (NB) distributions can be represented by Eq. 3 but an important characteristic of the Poisson distribution is that the expected number of counts (also referred to as the expectation/mean) is equal to the variance of that distribution and can be written as:

$$Var(k) = \lambda \quad (4)$$

while for the NB distribution, this variance is represented as:

$$Var(k) = \lambda + \frac{\rho^2}{\alpha} \quad (5)$$

where the last term on the right-hand side of Eq. 5 is referred to as the over dispersion parameter. The mean-variance relationship in RNA-seq data has been shown to follow Eq. 5 rather than Eq. 4 and is the basis of the differences in methods for differential analysis, edgeR and DESeq/DESeq2 (the methods differ by approach used to estimate this dispersion parameter).

Another important property of the NB distributions is that when the data are log transformed, the distribution approximates a Gaussian distribution. Limma takes advantage of this property to transform the data and model them using classical linear regression methods with some modifications that include empirical Bayesian statistics (29). The coefficients of the linear regression are then compared to determine significant differences between treatment groups.

Single cell RNA-sequencing data are important in characterizing the heterogeneity of cellular responses to pathophysiological changes, including their transitions between different

states. Methods developed for differential expression analysis of microarrays or bulk RNA-seq data have been applied to the analysis of scRNA-seq data despite the differences in their general characteristics from those of bulk measurements. Encouragingly, some methods have recently been developed to analyse these data given their unique characteristics including high drop-out rates, multi-modal distributions and very high heterogeneity. These methods include Model-based Analysis of Single-cell Transcriptomics (MAST) (55), Single-Cell Differential Expression (SCDE) (56) and DEsingle (57) among others that have been expertly reviewed (51).

There are several reviews on the subject of differential expression analysis in omics data. Costa-Silva et al. (53) shows that in RNA-seq datasets, limma (combined with VOOM) is one of the best performing methods. Together with DESeq, limma shows high specificity and true positive rate when used on real datasets obtained from the Microarray Quality Control project (58). Limma, contrary to DESeq, can also be used to model not only transcriptomics data, but all the varieties of abundance-based omics technologies (*e.g.* metabolomics, proteomics).

Set enrichment analysis

Set enrichment analysis is one of the final objectives of functional genomics studies: to understand the role of a set of features (feature sets, *e.g.* gene sets) and the pathways that influence key phenotypes. It is used to detect a concordant pattern of differentially expressed features for groups of features that are known to be annotated to some biological function, processes, and so on. With this approach, the analysis of the data is no longer based on the empirically measured expression levels of individual features, it is centered around sets of related features. This then allows more biology-driven conclusions to be drawn from the analyses.

Statistically, determining whether a feature set is associated with a given functional role has been formalized via hypothesis testing using several methods. There are two general approaches used in set enrichment analysis to formulate the comparisons against which the null hypotheses are tested: competitive or self-contained set analysis. In the former, a set of features is compared against its standard complement set within a database. In the latter, the features are compared to a fixed set (the features outside that set are not taken into consideration) (59-61).

Naturally, these methods assign a p-value to each set, which are tested for significance using some permutation tests. Unfortunately, permutation tests have known limitations

including the need for moderate to large sample sizes and their unsuitability to complex designs that may exhibit batch effects. Alternative methods have therefore been developed to improve this, including ROAST (62), which introduced rotation testing for self-contained enrichment analysis with an ability to mitigate permutation tests and handle complex experimental designs.

For the competitive set enrichment analysis, Gene Set Enrichment Analysis (GSEA) (63) is perhaps the most common method due in part to its longevity and conceptual simplicity. Its principle is to search for sets of features whose functional annotation is significantly over-represented in a given list of features, compared to a background set of features. This functional annotation can be found in databases such as the Molecular Signatures Database (MSigDB) (64).

While many other enrichment methods exist (*e.g.* sigPathway (61) and SAFE (65)), GSEA is one of the most widely used, also thanks to the MSigDB repository, where a variety of pre-made feature sets can be explored and used in the enrichment analysis. A review by Mathur et al. (66) compares different enrichment methods on synthetic datasets using false positive rates and (statistical) power as metrics. The review shows that GSEA is one the best performing methods, with the caveat that the q-values should be used when assessing the significance of the feature sets.

Marker set analysis

In the process of exploring the data, one of the crucial steps is to characterize different biological processes or functions. When two groups need to be compared, a differential analysis is sufficient to understand which features are differentially expressed. When multiple groups need to be compared, we use an intra-dataset meta-analysis approach to obtain an easily interpretable ranked set of marker features for each group. This can be done, for example, via a two-step approach. First, we perform a differential analysis for each group against all the other groups separately. Subsequently, we rank the marker features (based on the effect size, the log fold change) for each group using a product-based meta-analysis (67) although other statistical measures of central tendency such as the median or sum can be used to rank the marker features. This ranking can be preceded by a step that filters out features that are not consistently differentially expressed (*e.g.* we can take into consideration only the features that are upregulated). The output is a list of features that are exclusively differentially expressed in each group (compared to all the other groups).

Seurat, and many others tools, offer the possibility to obtain a set of markers, given a

number of clusters/groups. While no comparisons exist between all these methods, using a meta-analysis approach might, in theory, provide results that are less influenced by outliers, due to the intrinsically robust nature of meta-analysis methods (68).

Correction for false discovery rate (FDR)

As stated earlier, differential expression analysis derives from legacy experimental designs in molecular biology where a conventional null hypothesis of a nonzero difference in average expression was tested one feature at a time. However, due the large number of such hypotheses tested in a typical microarray or RNA-seq data set, correction for multiple comparisons is extremely important.

Benjamini and Hochberg (69) provide an attractive measure of control for multiple testing in differential expression analysis where the FDR can be estimated from accurate p-values. Other methods for estimating the FDR include the permutation-based approaches such as Significance Analysis of Microarrays (SAM), which have been in the literature since the early 2000s (70).

Visualization of trends

How the data can be explored and visualized is perhaps the most important aspect of data analysis, especially in the context of multi-dimensional data. In general statistical frameworks, data are presented in box plots, bar plots, pie charts, density plots, etc., since they can easily be interpreted by the human eye. These plots often summarize underlying trends in the data including, but not limited to, means, medians, range and spread of the data. These are standard plots that are extensively used in the field (71-73).

Single cell specific algorithms

Trajectory inference

One of the defining characteristics of scRNA-seq data is that, although the experiments are conducted on a case-control design paradigm, the data inherently exhibit temporal characteristics. The experiments can be viewed as a pulse experiment in which the cells are sampled at a given time but, at the time of sampling, those same cells may be undergoing different transcriptional programs. As a simple example, consider cells going through the developmental cell cycle. At the time of sampling, cells are collected at different stages of the cycle and this is represented by their transcriptional profiles.

In order to deconvolve this temporal structure, there are methods developed to infer their status. The caveat is that although the transcriptional profiles of cells differ at different

stages, it is not possible to define the finite time scale under which they are. A pseudo time scale is therefore defined for this and there are over fifty methods developed to make this inference. These methods include Monocle (74), SCORPIUS (75), slingshot (76), among others. It is therefore understandably difficult to choose a method for trajectory inference. A comprehensive comparison of twenty-nine of these methods was carried out by Saelens et al. (77) using both simulated and real data sets.

Cluster projection

Cluster projection is an approach that allows phenotypic similarities between different cell types to be inferred from the expression of characteristic signatures, which are assumed to be conserved. For example, the marker set for a certain cluster of cells could be expected to remain largely the same for the same organ even when experiments are conducted on different instruments, days, animals, or labs.

For scRNA-seq, cluster projection therefore allows clusters from one experiment to be projected onto the clusters of another in order to validate those findings. This is an important procedure because it allows spurious clusters such as those arising due to batch or other experimental factors to be eliminated while at the same time providing incontrovertible biological validation of cluster identities.

One of the classic methods for cluster projection is Principal Component Analysis (PCA) where principal component loadings derived from one data set can be used to determine the principal components (PCs) of another set. When these PCs are plotted, the clusters from one set will, in principle, recapitulate similar clusters in the second set. In scRNA-seq, scmap (78) has emerged as one of the most commonly used approaches for this projection which, unlike PC loadings, uses other similarity measures between clusters and a reference set to perform the projection. These similarity measures include the Pearson/Spearman correlations and cosine similarity.

Feature engineering

The concept of “the curse of dimensionality” is well known in multivariate data analysis (79) and occurs when analyzing high-dimensional data with a goal to relating them to few empirically meaningful factors. For example, in a case-control study examining differential expression between a tumor and normal tissue, it is often only a small set of genes that are differentially expressed yet over 20,000 gene expression values are measured for these two samples. If we then increase this number of samples to include 6 replicates of each, the number of deregulated features remains the same. It is difficult to represent this type of

measurement in 2 or 3 dimensions since our experiment has yielded a matrix of dimensions 12 x 20,000. To summarize these relationships, methods such as Principal Component Analysis (PCA) define a different co-ordinate system that captures this information in reduced (latent) variable space allowing only the important features to be retained. Feature engineering allows high dimensional data to be represented in lower dimensional space by selecting only those features that carry important information.

Set variation analysis

The ultimate goal of functional genomics analyses is to understand the role of a set of features and the pathways that influence key phenotypes. In many bioinformatics analyses however, a large list of features is generated and designated as differentially expressed features. Unfortunately, such list per se does not provide much information to the molecular biologists. Instead, it is the function of those features that matters and interpreting this list of features is often done using many approaches.

Gene Set Variation Analysis (GSVA) (80) is an approach that is similar in principle to Gene Set Enrichment Analysis (GSEA) (63) with major statistical and functional improvements (*e.g.* it uses nonparametric assumptions and it is unsupervised unlike GSEA). GSVA estimates the relative enrichment of a given feature set over a population of samples in contrast to GSEA. It therefore permits one to observe variations in the functional activity of a feature set (*e.g.* pathway) with respect to a biological condition. GSVA calculates sample-wise enrichment scores (ES) which can be examined for associations with clinical features of interest. By using GSVA, the features-by-observations data matrix can be converted into an engineered sets-by-observations data matrix. The newly created engineered data can then be used to perform other downstream analyses, such as differential analysis.

Other unsupervised methods have been developed to compute set scores on single samples. Single sample GSEA (ssGSEA) (81), the combined z-score method (82) and Pathway Level Analysis of Gene Expression (PLAGE) (83) are between the most known methods. Still, GSVA has proved to attain higher statistical power than the other three methods, with significant improvements in performances (80).

Survival analysis

In translational molecular biology, the goal of omics studies is to identify a feature from which biological inference can be made in relation to a disease. For example, in longitudinal cancer studies such as those deposited in The Cancer Genome Atlas (TCGA) repository (84), it is possible to extract features that have potential diagnostic or prognostic capacity. The most

common application is to evaluate the degree to which changes in the expression of a certain feature (*e.g.* a gene) affects a patient's survival time. This type of analysis is valuable because it can be used to provide some clinical inference regarding a feature's relationship with key properties of a disease, such as tumor grade and molecular subtype, or the expected length of survival time of a patient expressing that feature.

To assign such predictive capacity on an identified feature, there are many statistical methods that have been adapted from canonical survival analysis techniques including the Kaplan-Meier approach (85) combined with the log-rank test (86). In general, these methods are employed to model binary inputs like the presence or absence of a drug treatment, and so on. In omics analyses where the measurements are a continuous output of expression/abundance values, standard workflows are adopted to discretize the data and transform the inputs into a binary one with a defined threshold. The data is divided into quantiles based on the expression of a feature, which is then split based on a threshold (*e.g.* the median). The survival model can then be created using the Kaplan-Meier approach. Significance between the survival curves can be assessed by using the log-rank test.

Advanced modelling techniques, such as Cox regression (87) (and its variants *e.g.* lasso Cox (88), elastic-net Cox (89)) and machine learning-based survival trees (90) are gaining traction. Due to the historical use and simplicity of the Kaplan-Meier approach, this is still the most widely used survival model. A review by Lee et al. (91) delves into both the classic and new survival analysis methodologies, which shows that adapting classical statistical methods with machine learning techniques can improve survival predictability.

Machine learning

Supervised methods depend on input from the user about sample identity and are useful in cases where these measures of central tendency are not strong enough to associate (dis)similarity among the measurements. These methods include the so-called machine learning approaches such as random forest developed by Leo Breiman (92) and partial least squares with discriminant analysis (PLS-DA) (93). Although supervised classification methods are an essential part of exploratory data analysis, their applications must be treated with caution because they require test data to validate the trained models.

In recent years, supervised classification methods have been increasingly referred to as Machine Learning (ML) methods and they fall under a branch of Artificial Intelligence (AI), where the goal is to learn from data samples the underlying information that can facilitate behavioral inference. ML approaches consist of two processes. First, unknown dependencies

(relationship between sample and features) in the data are estimated given a data set. Second, these dependencies are then used to predict the characteristics of future samples given this dependency. This is similar to a classical regression problem where the regression coefficients are learnt and used to predict the response of future observations given predictor variables.

In a supervised classification approach, a labeled set of training data are used to estimate the parameters (features) that can categorize the input data into the discretized, finite class labels. These parameters are then used to decipher the class labels for future samples given measurements of those parameters. For example, given a set of gene expression data corresponding to 20,000 genes in 100 samples of which half are from a cancer patient and the other half from normal samples, ML approaches can be used to determine which genes (features) discriminate cancer from normal samples. Later, with gene expression measurement in a sample of unknown origin, it is possible to determine whether or not the sample is from a cancer patient.

The most widely used ML techniques include Artificial Neural Networks (ANN), Random Forests, Support Vector Machines (SVM), Self Organizing Maps (SOM), Decision Trees (DTs) among others. Perhaps the most important caveat with ML methods is that they are prone to over-fitting training data. For this reason, important metrics are used to decide the quality of a ML model before it is used in prediction. These include the receiver operating characteristic curves (ROCs) that measures specificity and sensitivity, and bootstrap aggregating in which several models are developed by bootstrapping samples. In bootstrap aggregation, the samples that are left out of the bootstrap are then predicted using the model and an error rate is calculated. This error is referred to as the out-of-bag error and can give an indicator of the suitability of a model. Other metrics for model accuracy are common in the statistical and machine learning literature (94).

Meta-analysis

Bulk omics data

Many RNA-seq, proteomics or metabolomics studies are characterized by low sample sizes. In the recent years however, the requirement that published data are stored in public repositories has alleviated this drawback because these data offer an invaluable resource for reuse, augmenting laboratory-acquired data to generate new knowledge. Meta-analysis methods can facilitate the ability to extract biologically relevant information from data integrated from multiple sources, allowing extraneous variation to be removed from that data while preserving biologically relevant information.

There are three main approaches for meta-analysis, and they are based on: (*a*) the effect size; (*b*) the test of significance and (*c*) the rank.

In differential expression analysis, the effect size is the magnitude of change between conditions and explains the difference between the strength of a given phenomenon (*e.g.* a disease) on the expression of a feature. This is often calculated using Hedge's *g* estimator given as:

$$T_i = k\frac{R_e - G_c}{\sigma} \quad (6)$$

where T_i is the effect size for one feature in the ith study, *k* is a factor that corrects for bias while R_e and G_c are the mean expression levels for that feature in the two conditions in that study and σ is the standard deviation. Typically, T_i can also be calculated as the log ratio of the experimental and reference condition (log fold change). Values of T_i and their uncertainties are then calculated for all conditions and combined to evaluate the effect of each feature across all studies.

Other approaches for meta-analysis that use significance testing, for example, utilize hypergeometric tests to evaluate the strength of differentially expressed features across samples. Rank-based methods are also employed: they use various nonparametric tests such as rank product or rank sums in the same framework as discussed for effect size approaches (67).

Single cell omics data

One of the most documented advantages of scRNA-seq analysis is that it can help characterize cellular heterogeneity within an organ. Such information is useful because different cell types have been shown to respond differently to treatment (95,96). Between multiple organs, it is useful to show that the (dis)similarities between the cell types is conserved especially as they relate to treatment. The most common approach for showing this type of gene expression (dis)similarities between organs and even species is via meta-analyses (97).

In a typical scRNA-seq workflow, cell types are identified via cluster analysis and each cluster is then given a unique set of marker genes that differentiate it from all other clusters and therefore assign it an identity. This is, essentially, accomplished by comparing every cluster to all other clusters by treating the cells within each putative cluster as replicates and performing differential analysis. Features that are uniquely expressed in a given cluster are then said to be markers of that cluster and can be used to identify the cell types in other

organs, species or the same organ/species after treatment.

To quantify the differences between marker sets of clusters that have been derived from different experiments, the Jaccard similarity index (98) can be employed, which is a similarity metrics used to compare sets (set of features, marker sets):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| + |A \cap B|} \quad (7)$$

While we developed this methodology specifically for single cell data, this approach is also applicable to bulk omics data. To quantify the differences between marker sets of groups (treatments, conditions, etc., not necessarily single cell clusters), Formula 7 can still be applied without any change.

Since single cell technologies are relatively new, classic methods are still widely used (described above in the “Bulk omics data” section) to analyze single cell omics datasets (99,100). As more single cell data is generated, the need to combine the results of multiple studies at a single cell level is going to increase quickly, accelerating the development of ad-hoc single cell meta-analysis methods.

Accessible analysis software in bioinformatics

In this chapter, I have presented an overview on how omics data can be analyzed and which methods could be employed to do so.

While the number of algorithms and methods to analyze omics data is steadily increasing, the number of easily accessible software made specifically for non-bioinformatician experimental biologists is struggling to catch up. This, combined with the exponential increase of the amount of generated biomedical data and the lack of easy-to-use software for non-bioinformatics experts is one of the main bottlenecks in the data interpretation pipeline.

Providing the non-bioinformatician experimental biologists with bioinformatics methods to analyse large-scale biological data is becoming a prominent issue in biomedical research.

Chapter II Research objectives

Alleviating the data analysis and interpretation bottlenecks can expedite the process of analyzing omics datasets and improve the quality of the biological interpretation of the data. To alleviate the bottleneck, the aim is to provide an easy-to-use software to allow non-bioinformatician experimental scientists to perform interactive data mining of bulk and single cell omics datasets. I defined the following aims, in which the software should:

1. Integrate algorithms into one unified workflow.
2. Suggest a “best-practice” data analysis workflow but also be flexible enough for users to choose their own strategy.
3. Be able to analyze different omics experiments, such as metabolomics, transcriptomics, proteomics, and single cell data.
4. Handle complex experimental designs, perform analysis using covariates (*e.g.* clinical or experimental data), and individual experiments can be contextualized by comparing to similar previously performed analyses (meta-analysis).
5. Be easily extendible to analyze data derived from novel techniques using the latest algorithms and state-of-the art methods.
6. Be highly interactive, allowing users to ask increasingly complex questions and performing increasingly fine-tuned analyses.
7. Provide high-quality publication ready figures.

Based on our analysis of the bottlenecks in omics data analysis and interpretation, I aimed to develop the software as a tool that could assist users in rapidly, but accurately, analyzing a large variety of complex datasets to derive new biological knowledge. The software should not simply execute what researchers ask but systematically guide them through the analysis so that maximum insight is captured. Moreover, the software should educate researchers on the possibilities of advanced data analysis techniques and algorithms so that they can design increasingly complex and powerful experiments.

The ultimate goal is not to deliver a “silver bullet” software that fully automates the extraction of meaningful data from complex datasets without any human intervention. This model is unlikely to be successful in the biomedical sciences where algorithms seldom provide sufficiently accurate results and where extensive domain knowledge is required to biologically interpret the results. The software should rather be a guide that aims to efficiently show researchers all aspects of their data to optimize the biological interpretation of omics data.

Chapter III BIOMEX

This chapter is based on the first author research article published in Nucleic Acid Research.
https://academic.oup.com/nar/article/48/W1/W385/5835814

This is the main manuscript of the thesis and is meant to showcase the functionalities of the BIOMEX software.

BIOMEX: AN INTERACTIVE WORKFLOW FOR (SINGLE CELL) OMICS DATA INTERPRETATION AND VISUALIZATION

Federico Taverna, Jermaine Goveia*, Tobias K. Karakach, Shawez Khan, Katerina Rohlenova, Lucas Treps, Abhishek Subramanian, Luc Schoonjans, Mieke Dewerchin, Guy Eelen & Peter Carmeliet#*

*: equal co-authorship; #: corresponding author

Editorial correspondence:

Peter Carmeliet; peter.carmeliet@kuleuven.vib.be

Personal contribution: I managed, designed, developed, tested and deployed the BIOMEX software (and manual). The design was performed together with Jermaine Goveia. The testing was performed together with members of the laboratory of Peter Carmeliet.

Abstract

The amount of biological data, generated with (single cell) omics technologies, is rapidly increasing, thereby exacerbating bottlenecks in the data analysis and interpretation of omics experiments. Data mining platforms that facilitate non-bioinformatician experimental scientists to analyze a wide range of experimental designs and data types can alleviate such bottlenecks, aiding in the exploration of (newly generated or publicly available) omics datasets. Here, we present BIOMEX, a browser-based software, designed to facilitate the Biological Interpretation Of Multi-omics EXperiments by bench scientists. BIOMEX integrates state-of-the-art statistical tools and field-tested algorithms into a flexible but well-defined workflow that accommodates metabolomics, transcriptomics, proteomics, mass cytometry and single cell data from different platforms and organisms. The BIOMEX workflow is accompanied by a manual and video tutorials that provide the necessary background to navigate the interface and get acquainted with the employed methods. BIOMEX guides the user through omics-tailored analyses, such as data pretreatment and normalization, dimensionality reduction, differential and enrichment analysis, pathway mapping, clustering, marker analysis, trajectory inference, meta-analysis and others. BIOMEX is fully interactive, allowing users to easily change parameters and generate customized plots exportable as high-quality publication-ready figures. BIOMEX is open source and freely available at https://www.vibcancer.be/software-tools/biomex.

Introduction

The recent growth of unbiased high-throughput sequencing and profiling technologies has revolutionized the generation and analysis of biological data (1). The commoditization, exponential growth and increased throughput of these technologies (2) has helped the community to develop breakthroughs in bioanalytical research. For example, using next generation sequencing technologies, it is possible to analyze whole genome and transcriptome sequences within an hour (3). In addition, using mass spectrometry, thousands of proteins and metabolites can be measured simultaneously (4,5). Until recently, traditional profiling methods could be applied only to ‘bulk’ samples homogenized from whole tissue or organ extracts. With the advent of single cell genomics, transcriptomics and proteomics profiling technologies, individual cell contents can now be measured (6) allowing the characteristics of individual cells to be studied (7,8), further increasing the volume of data to analyze.

Availability of such large and complex datasets introduces multiple challenges. Computational challenges relate to the handling, processing and analysis of the data; new bioinformatics tools are continuously developed in tandem with technological advances.

Biological challenges stem from the need to understand the biological significance of the information in the data and require in-depth knowledge of the biological question. Consequently, detailed biological interpretation of omics data requires a synthesis of domain-knowledge and computational skills, which continues to inspire interdisciplinary projects between researchers with complementary expertise. Various tools exist to meet the challenges related to analyzing omics data, and can be loosely defined as workflow tools (*e.g.* Galaxy (9), Taverna (10)), pre-processing tools (*e.g.* XCMS Online (11), MaxQuant (12)), specialized tools and more broad data analysis platforms. Easy-to-use data analysis platforms that allow experimental scientists to autonomously analyze omics data have been highly successful as solutions to bridge the gap between data generation and interpretation (*e.g.* Perseus (13), EXPANDER (14), InstantClue (15), MetaboAnalyst (16)). However, currently available data analysis platforms mostly focus on analyses of bulk omics data types, and in many cases they are not tailored to support the pretreatment and analysis of a wide range of omics experiments within the same interactive framework (*e.g.* RNA-sequencing, gene expression microarrays, metabolomics, proteomics), making the unified analysis of all these varieties of omics data challenging. Moreover, these tools do not scale, nor provide a structured data mining workflow to explore single cell datasets (*e.g.* single cell RNA-sequencing and mass cytometry).

Here we present BIOMEX, a data mining software developed for the Biological Interpretation Of Multi-omics EXperiments. BIOMEX integrates a range of publicly available algorithms and field-tested data analysis approaches into a well-defined and guided stepwise workflow that accommodates a wide variety of experimental designs and multi-omics data, including metabolomics, transcriptomics, proteomics, single cell RNA-sequencing and mass cytometry experiments (Figure 1). The software is capable of handling large-scale data such as single cell omics experiments, scaling from tens to hundreds of thousands of cells.

BIOMEX aims to alleviate the bottlenecks in the biological data-analysis-to-interpretation pipeline of omics experiments (13,17): these bottlenecks have become prominent in light of the increased need to understand the underlying biological phenomena that are now measureable at a much higher resolution.

Overview

Functional requirements and design rationale

BIOMEX is designed to allow non-bioinformatician experimental scientists to perform interactive data mining of bulk and single cell omics datasets.

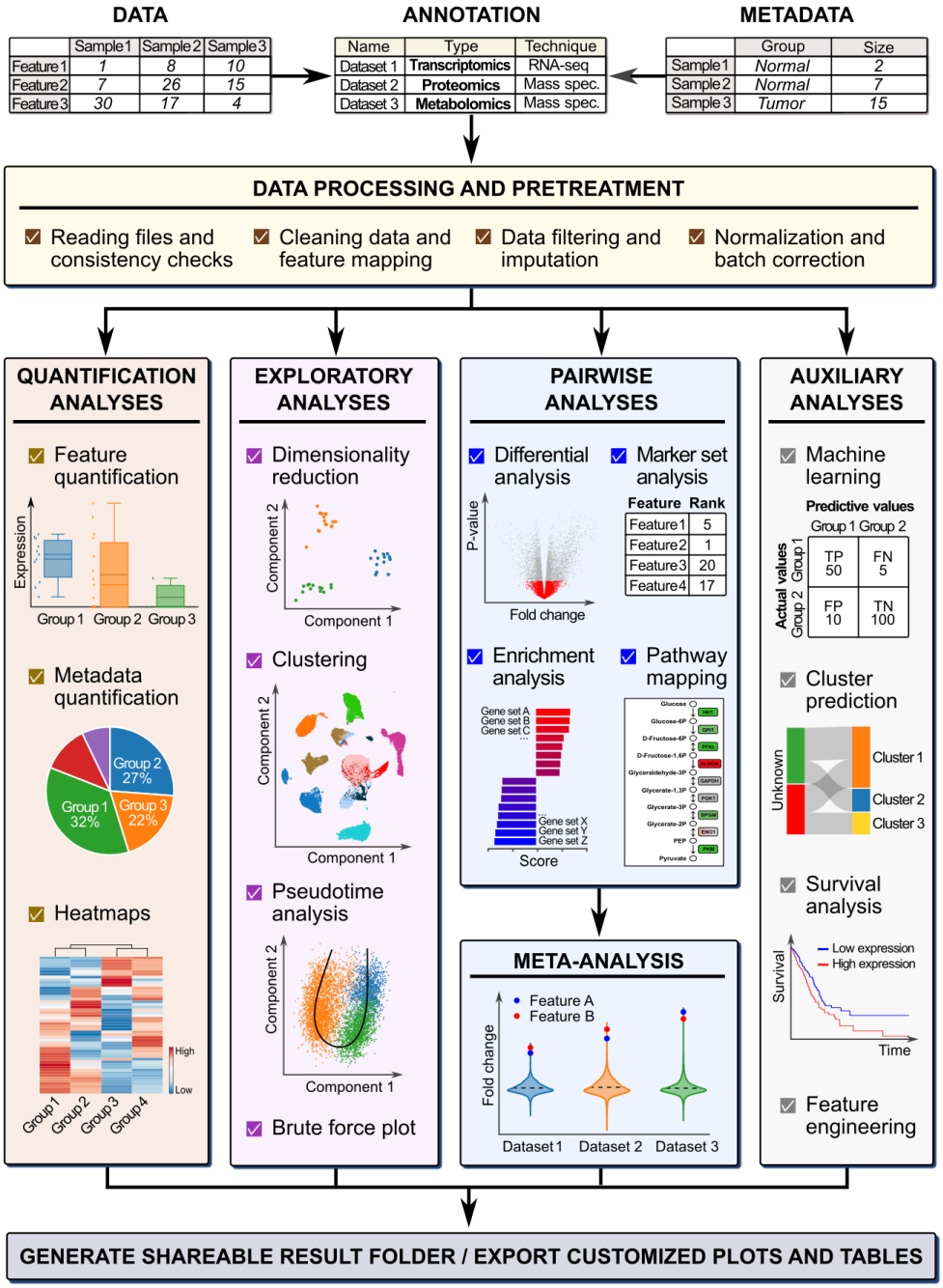

Figure 1: The BIOMEX workflow.

The workflow guides the user through distinct analysis steps. The data (and metadata) need to be uploaded, and each uploaded dataset must be annotated with the relevant information (*i.e.* the omics data type, technology, feature identifier, etc.). In the processing step, the data is cleaned and consistency checks are performed to verify that the data was uploaded in the correct format. After the feature identifiers are mapped to feature names, the data is filtered and normalized in the pre-treatment step. Depending on the omics data type, the data can also be imputed or batch corrected. Once the data is pre-treated, it is ready for downstream analysis. The different analyses available in BIOMEX can be divided into five categories: i) quantification analyses quantify the abundance level of the features in the data (and metadata); ii) exploratory analyses assist in understanding the underlying structure of the data; iii) pairwise analyses reveal the functional differences between groups; iv) meta-analysis combines results from different studies in a singular, unique and robust result; and v) auxiliary analyses (*e.g.* machine learning and survival analysis). Ultimately, all the analyses can be saved in a self-contained folder that can be shared between scientists and results can be customized and exported either as tables or high quality publication-ready figures. Abbreviations: TP, true positive, FN, false negative, TN, true negative, FP, false positive. Note: The term 'features' is used to indicate genes, metabolites and proteins.

We therefore defined the following functional requirements for the software:

1. To accommodate multi-omics data across select biological species and computational platforms.
2. To allow users to interactively analyze complex experimental designs using state-of-the-art and field-tested algorithms.
3. To facilitate the re-use of publicly available data.
4. To provide a flexible, well-defined data analysis workflow.
5. To provide self-contained, non-technical background information to aid the meaningful use of each analysis module.
6. To enable the generation of highly-customizable publication-ready plots and figures.

BIOMEX is implemented in the open source R programming language (https://cran.r-project.org/); the majority of algorithms required for biological data mining are available through open source R packages. BIOMEX integrates these algorithms and packages into a workflow, in which parameters can be interactively tuned using the Shiny web framework (https://shiny.rstudio.com/, the full list of packages used in BIOMEX is available in Supplementary Table 1) (18). Together, BIOMEX creates a workflow that allows users to iteratively fine-tune complex analyses in order to facilitate detailed biological interpretation through interactive visualizations.

Manual and video tutorials

A comprehensive web manual that describes all functionalities, data formats, parameters and analyses related to the workflow is provided within the BIOMEX software. The manual guides the user through the step-by-step procedure required to execute the workflow and is complemented with video tutorials that introduce users to all interface elements and software functionalities.

Data import: data and metadata matrix

BIOMEX requires two files for each experiment, the data and metadata matrix.

Data matrix: the data matrix contains typical omics (*i.e.* transcriptomics, metabolomics, proteomics, mass cytometry) output in a text (.txt) or comma separated values (.csv) format. The data file is organized such that the first column contains feature identifiers (*i.e.*, genes, metabolites, proteins), while the first row contains descriptors (*i.e.* sample or cell IDs). The data matrix can be uploaded as unprocessed gene expression values (raw read counts, unique molecular identifier counts for single cell RNA-sequencing, non-log transformed

intensities for microarrays) or absolute abundances for metabolomics and proteomics data. Alternatively, BIOMEX accepts preprocessed data (*e.g.* filtered and normalized, batch corrected, etc.).

BIOMEX automatically checks the uniformity and compatibility of the data, while also dealing with irrelevant (empty) observations and features. For transcriptomics and proteomics, the feature (gene or protein) identifiers are mapped to feature names to allow downstream interpretation of the results and further analyses.

Metadata matrix: the second required file is the metadata file (.txt or .csv format) that contains all the auxiliary information about the experimental design (variables). The metadata file is organized such that the first column contains descriptors matching with the data file, while the first row contains the variables (*e.g.* factors, numeric, etc.). This file can be modified interactively within the software.

Example datasets: we provide example datasets (the data and metadata matrices) that are readily available to be uploaded into the software. These example datasets encompass all the omics data types supported by BIOMEX, and they include case studies (described below) available directly from the software via the “Case studies” section.

Data analysis modules

The BIOMEX workflow contains 10 data analysis modules, which are briefly described below.

Module 1: Data pretreatment

After data upload, the data is filtered to remove low quality features, normalized, and, if necessary, corrected for unwanted technical variation (*e.g.* batch effects) (19-22). Algorithms for quality filtering, normalization and regression are often omics-type specific, BIOMEX automatically suggests applicable algorithms and parameter settings depending on the type of data being analyzed. For example, single cell RNA-sequencing data can be corrected for batch effects by using the mutual nearest neighbor (MNN) method (22).

The output of this module is a clean/corrected data matrix that can be used for downstream analysis.

Module 2: Feature engineering

Complementing unbiased and automated methods, domain knowledge can be used to craft new features from the existing features in the data to estimate biological variation (*e.g.* from gene expression to pathway activity). During this (optional) step, BIOMEX employs gene set

variation analysis (23) (GSVA) to convert the features-by-observations data matrix (output of module 1) into an engineered sets-by-observations data matrix. The newly created engineered data can then be used to perform downstream analysis, including differential analysis. BIOMEX includes the KEGG sets by default, but users can also upload custom sets. Alternatively, feature engineered data, created with independent methods, can be directly uploaded to BIOMEX and subsequently used in downstream analysis.

Module 3: Visualization of trends

Feature magnitudes and trends are visualized in bar plots, box plots, violin plots and density kernel estimation plots. These plots are grouped based on the information present in the metadata. The uploaded metadata can also be explored through pie charts and horizontal bar plots.

Module 4: Unsupervised analyses

Unsupervised analysis aims to unbiasedly detect patterns in the data. For dimensionality reduction and visualization, BIOMEX includes Principal Component Analysis (24) (PCA, flashPCA package (25)), t-distributed Stochastic Neighbor Embedding (26) (t-SNE, Rtsne package) and Uniform Manifold Approximation and Projection (27) (UMAP, umap package). Also, BIOMEX supports K-means, hierarchical and graph-based clustering (Seurat (21) and FlowSOM (28) packages). The output of hierarchical clustering can be visualized via dendrograms, and the associated uncertainty can be assessed using multi-scale bootstrap resampling (pvclust package (29)). BIOMEX provides interactive heatmaps (heatmaply package (30)) to visualize inherent associations between groups or clusters.

Module 5: Supervised analyses

Supervised pairwise analyses are used to explore quantitative differences in expression or abundance levels between groups (differential analysis). BIOMEX uses linear models (limma and MAST packages (31,32)) to describe the relationship between expression levels of features between two groups. This enables handling of complex experimental designs, and allows including covariates in the modeling process. The magnitude of differential expression (log fold change) and the p-values are provided for each feature, together with the false discovery rate adjusted p-values calculated with the Benjamini-Hochberg method (33). Volcano plots are used to visually represent the differential analysis results.

As an extension of pair-wise differential analysis, BIOMEX includes marker analysis that can be used to detect key discriminating features between multiple groups (or clusters in single cell data). This analysis consists of a two-step intra-dataset meta-analysis approach.

First, BIOMEX performs a differential analysis for each group against all the other groups separately and filters out features that are not consistently differentially expressed (34). Subsequently, marker features are ranked using a product-based meta-analysis (median-, sum- or p-value-based meta-analysis can be used to rank features) (35).

Functional analysis of omics data can be performed in BIOMEX using several tools. These include Gene Set Enrichment Analysis (36) (GSEA, clusterProfiler (37) package) used to perform the competitive set enrichment analysis, and rotation gene set tests (ROAST) (38) to perform self-contained set enrichment analysis. Results of such analyses can be either displayed as a waterfall plot or a horizontal bar plot. These analyses provide functional information regarding, for example, pathways or biological processes that may be deregulated in a select set of conditions. The default setting in the software is to use the (metabolic) KEGG pathway sets (39), but this can be changed by the user to include other pathways or biological processes. BIOMEX uses the KEGG pathways to map features in the data (genes, proteins, metabolites) to well defined and constructed pathways using the pathview (40) package. The pathway visualizations are interactive and can be customized by the user to incorporate *a priori* biological insight (*e.g.* irrelevant isoforms can be manually excluded).

Module 6: Single cell specific algorithms

Single cell data can be used to infer differentiation trajectories using computational methods. BIOMEX includes Monocle (41) and SCORPIUS (42) to infer branched and linear cell trajectories, respectively. BIOMEX also uses locally estimated scatterplot smoothing (LOESS) regression to subsequently model the dynamic behavior of features in pseudotime. As a second single cell-specific approach, BIOMEX includes scmap (43) to project cluster identities from a reference dataset to another non-clustered dataset by calculating the similarities between cells of the non-clustered dataset and the cluster centroids in the reference dataset.

Module 7: Survival analysis

Survival analysis is implemented in BIOMEX in order to link omics data to a disease outcome. For example, using The Cancer Genome Atlas (TCGA) (44) and other resources, it is possible to infer the effect of deregulation of a given gene to a treatment outcome or patient survival. BIOMEX uses the Kaplan-Meier (45) test to generate the survival functions, and the logrank test (46) to assess the significance of those survival functions (survival package).

Module 8: Machine learning

Machine learning is a set of approaches that can model the relationship between a set of variables (features) and instances (observations) based on a given training dataset. BIOMEX

includes the ranger (47) implementation of the random forest model to perform classification and regression tasks (48). Recursive feature elimination (RFE) (49) is used as the feature selection method of choice to select the most predictive features. The machine learning pipeline is based on the caret package and includes cross-validation strategies to assess the predictive performance of the model (50).

Module 9: Meta-analysis of bulk omics data

Integrative data analysis approaches have been successfully used to analyze multiple datasets simultaneously to compare the results of independent experiments (51,52). With the availability of added-value databases, publicly available preprocessed data can be easily accessed by scientists and used to perform meta-analyses. In a meta-analysis, (i) a pair-wise differential analysis is performed for each dataset independently; (ii) we rank the features in each dataset by a metric (*e.g.* fold change), and (iii) we combine the rank numbers for all features using a product-based (or median-, sum- and p-value-based) meta-analysis approach. As a result, we obtain a ranked list of features, which are consistently differentially expressed across all the selected comparisons (*i.e.* differential analyses) in different datasets. The results can be visually explored through violin plots.

Module 10: Single cell meta-analysis

Meta-analysis can also be performed by measuring the similarity between clusters (53). This analysis, developed specifically for single cell omics data, assesses the conservation of cell phenotypes between different tissues, organs, studies, conditions, etc. BIOMEX performs the cluster similarity analysis by combining the results obtained during the marker set analysis. Similarity between the clusters present in the marker set results are calculated using the pairwise Jaccard similarity coefficients (54) for all clusters against all other clusters. The output of this analysis is a similarity score matrix, which describes quantitatively how each cluster is similar to other clusters. PCA is applied to the pairwise Jaccard similarity coefficient matrix to visually represent the similarity between clusters.

Data export

All the plots and tables (plotly, ggplot2 (55), DT packages) can be fully customized and exported in a variety of high quality formats (*e.g.* vectorized image format). BIOMEX saves all parameters and results in a self-contained folder, which can be shared between users and loaded into BIOMEX, improving the reproducibility of analyses.

Case studies

To showcase the analysis modules implemented in BIOMEX, we provide two case studies. A

step-by-step tutorial on how to reproduce the results obtained in both case studies is available in the manual, which includes all the parameters used to perform the analyses and to generate the plots.

Bulk data: exploration of the TCGA cholangiocarcinoma dataset

To provide an illustrative example on how bulk data can be analyzed, we explored a publicly available TCGA dataset (TCGA-CHOL) on cholangiocarcinoma (CCA), a cancer from the bile duct that represents the second most commonly diagnosed primary liver tumor (56). Even when diagnosed at an early stage, CCA is a very aggressive malignancy with poor patient outcome and limited treatment opportunities (56). According to their anatomical location, CCAs are classified as intrahepatic, hilar-perihilar and distal, which represent respectively 88.2%, 5.9% and 5.9% of the patients in this analysis (Figure 2A). Dimensionality reduction (PCA) and correlation heatmap analyses showed that biopsies from normal tissue have a clearly distinct transcriptomic signature compared to intrahepatic CCA resections (Figure 2B-C). Although there is a strong inter tumor sample heterogeneity between patients (Figure 2C), we aimed at determining transcriptomic similarities that could be involved in overall CCA pathogenesis. Enrichment analysis of normal *versus* intrahepatic CCA samples indicated that cell cycle and extracellular matrix (ECM)-receptor interaction gene sets were the most upregulated (Figure 2D). Consistently, differential gene expression analysis showed that several key mitotic checkpoints (*e.g. CDK1*, *E2F1*, *CDC45*, *SFN*) and mitotic spindle assembly/control genes (*e.g. CDC20*, *CDC25*, *TUBB3*), as well as numerous genes encoding laminin, integrin, collagen and ECM-secreted proteins (*e.g. SPP1*, *COMP*, *TNC*) were upregulated in the tumor samples (Figure 2E-G and not shown). To explore whether this signature was conserved across the other classes of CCA, we performed a meta-analysis of normal *versus* tumor samples from intrahepatic, hilar-perihilar and distal CCAs. We identified two genes, namely *CEACAM5* and *AFAP1-AS1*, ranking in the top 2 after a product-rank meta-analysis (Figure 2H). Interestingly, *CEACAM5* (carcinoembryonic antigen, CEA) is a well-established prognostic marker in CCA (57), while the long non-coding RNA *AFAP1-AS1* has been linked to metastasis (a process requiring complex ECM remodeling) and cancer cell proliferation in CCA. Hence, the meta-analysis results further supported the importance of cell proliferation and ECM-cell adhesion in CCA.

A crucial step during invasion and metastasis is the remodeling of the ECM by proteolytic degradation, involving matrix metalloproteinases (MMPs) as pivotal actors. Interestingly, *MMP11* has been correlated to poor survival in several cancer types including CCA (58), breast and pancreatic cancers. Consistently, we found that high *MMP11* expression was correlated to poor survival in patients with intrahepatic CCA (p-value=0.0015) (Figure 2I).

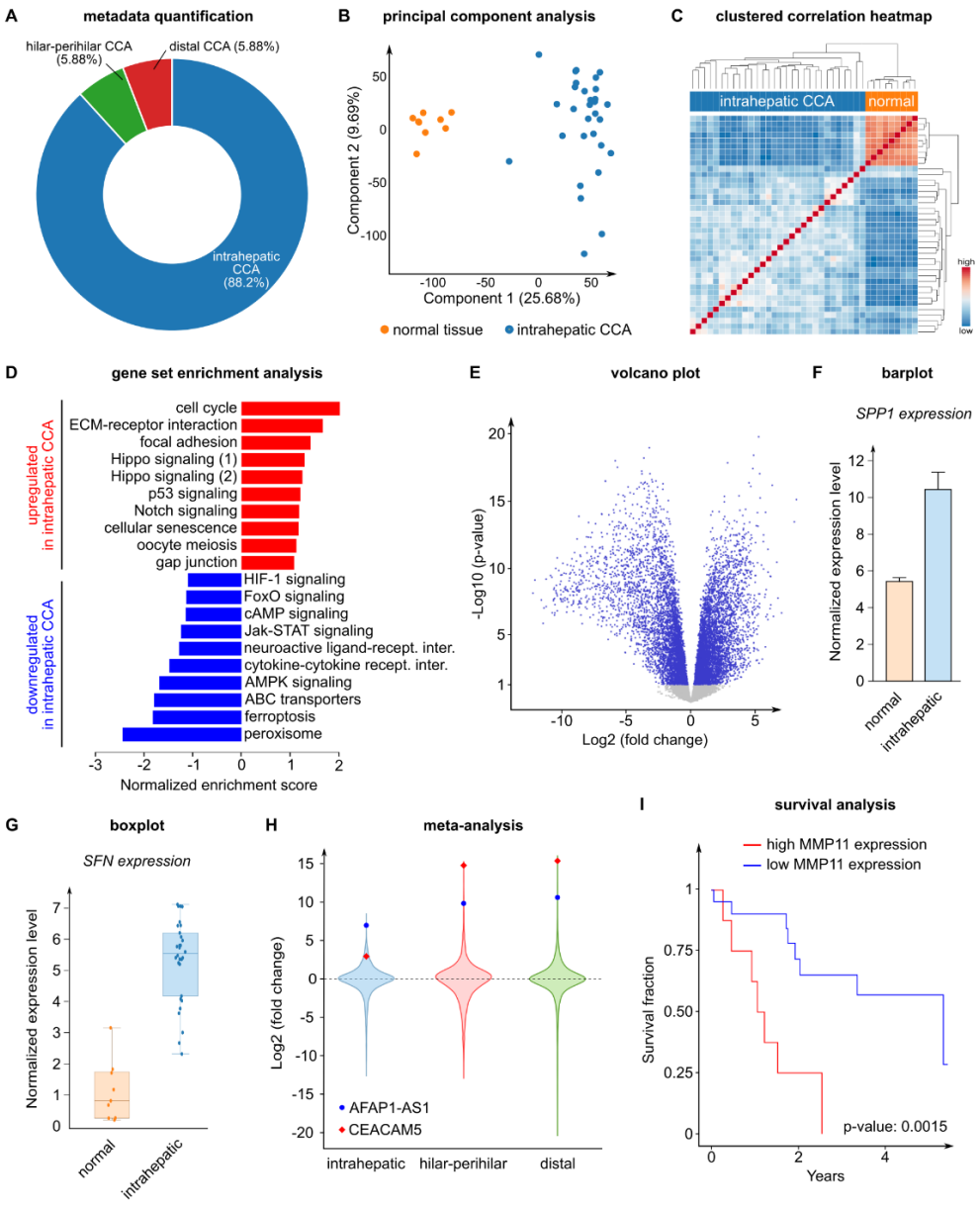

Figure 2: Cholangiocarcinoma TCGA data analysis results.

(**A**) Overall histological type percentage of the TCGA cholangiocarcinoma dataset. (**B**) PCA of normal and intrahepatic tumor samples. (**C**) Clustered heatmap based on the correlation of normal and intrahepatic tumor samples. (**D**) Competitive enrichment analysis of normal *versus* intrahepatic tumor samples using the gene sets related to the "Environmental Information Processing" and "Cellular Process" KEGG pathway maps. The upregulated gene sets are shown in red, the downregulated gene sets are shown in blue. Note: There are two enriched KEGG gene sets related to Hippo signaling, indicated separately in the figure. Hippo signaling (1): KEGG Hippo signaling pathway; Hippo signaling (2): KEGG Hippo signaling pathway - multiple species. (**E**) Differential analysis of normal *versus* intrahepatic tumor samples shown in a volcano plot. The significantly different genes ($p < 0.05$) are shown in blue, the non-significant genes are shown in grey. (**F**) Barplot visualization of SPP1 expression in normal and intrahepatic tumor samples. The error bar represents the standard error. (**G**) Boxplot visualization of SFN expression in normal and intrahepatic tumor samples. The box represents the range between the first quartile (Q1) and the third quartile (Q3), the horizontal line represents the median, the whiskers represent the interquartile ranges (IQR, 1.5 × IQR below Q1 and 1.5 × IQR above Q3). (**H**) Meta-analysis of intrahepatic, hilar-perihilar and distal tumor types. Each violin plot represents the differential analysis of normal *versus* the corresponding tumor type. The top 2 most consistently upregulated genes (*CEACAM5* and *AFAP1-AS1*) are highlighted. (**I**) Survival analysis based on *MMP11* gene expression of intrahepatic tumor samples. All the results shown in the figure can be directly explored in the BIOMEX "Case studies" section. The parameters used to generate these plots can be found in the manual.

Together, these findings show that cell proliferation and ECM adhesion/remodeling are conserved features across the CCA classes. Hence, (novel) insights can be derived by exploring publicly available data, showcasing the potential and value of BIOMEX.

Single cell data: re-analysis of the endothelial cell atlas dataset

To provide an illustrative example on how single cell data can be analyzed, we selected heart, liver and lung endothelial cells (ECs) from the recently published murine EC atlas scRNA-seq dataset (59). We intended to showcase a logical sequence of analyses that a BIOMEX user can employ to (re-)analyze single cell data.

ECs line the lumen of blood vessels and are known to be heterogeneous along the vascular tree. Consistently, dimensionality reduction and visualization using PCA and t-SNE indicated that ECs from the lung, heart and liver vascular beds have a distinct transcriptional profile (Figure 3A). To explore heterogeneity of ECs within a single vascular bed, we performed dimensionality reduction using UMAP to visualize the subclusters as they were detected in the EC atlas (59) (Figure 3B). Next, we performed rank-product based marker set analysis, and visualized the top 5 marker genes for each cluster using a heatmap (Figure 3C). Marker genes were consistent with previously described markers of (sublineages of) arterial, capillary, venous, lymphatic and proliferating ECs (59). Quantification of the number of cells per cluster showed that capillary ECs constitute the majority of the liver single cell population (Figure 3D). Further, unbiased linear trajectory inference reconstructed a phenotypic continuum of arterial, capillary and venous phenotypes, consistent with the known anatomical topography of liver ECs (Figure 3E). To explore whether the cluster signatures are conserved across tissues, we performed a similar analysis for the lung and heart ECs (not shown), and subsequently performed a Jaccard similarity analysis. This analysis revealed that marker genes of arterial, capillary and venous phenotypes are conserved across vascular beds (Figure 3F). Finally, we used scmap to project liver ECs from an independent reference dataset (Tabula Muris dataset (8)), onto the cluster identified in the EC atlas liver ECs (Figure 3G).

Together, this sequence of relatively simple analysis steps shows the power of the BIOMEX workflow to easily explore single cell datasets in detail.

Conclusion

To facilitate bench scientists in solving the computational problems arising from omics experiments, we designed and developed BIOMEX, a data mining software for the Biological Interpretation Of Multi-omics EXperiments.

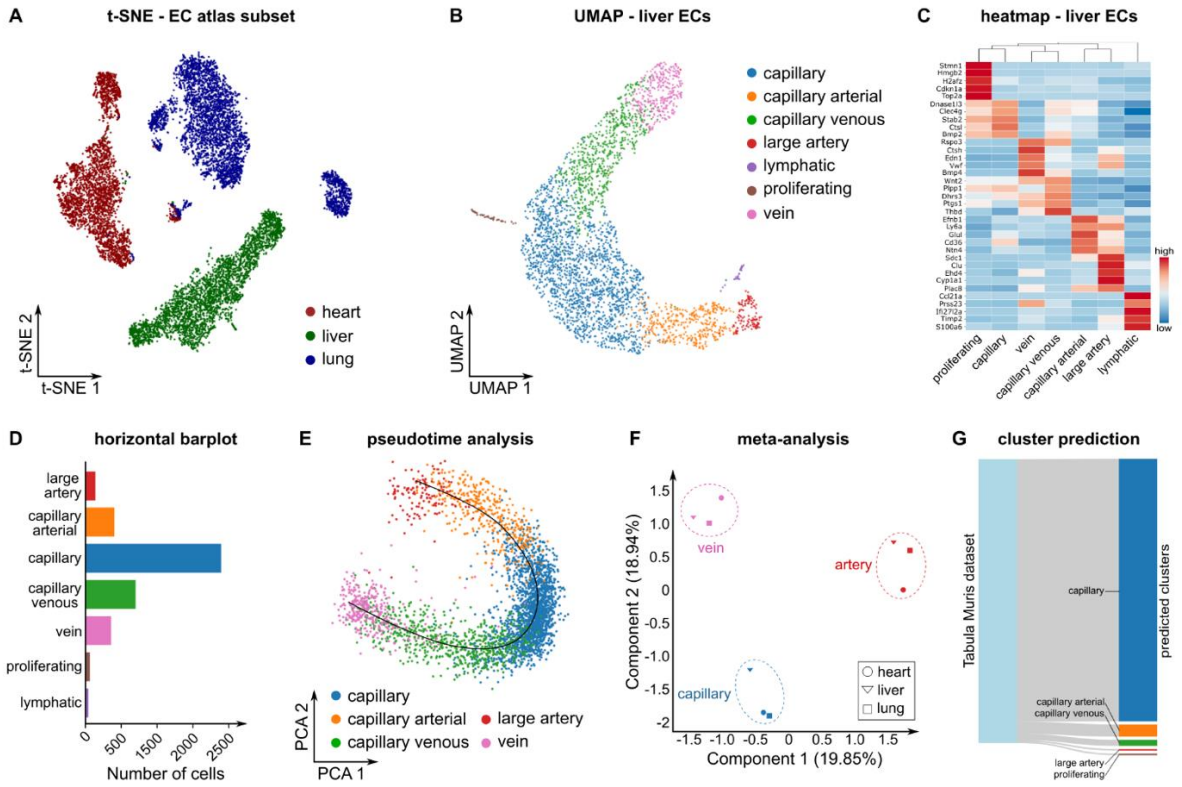

Figure 3: Endothelial cell atlas data analysis results.

(**A**) t-SNE plot of ECs from three murine tissues (heart, liver, lung). (**B**) UMAP of liver tissue showing the endothelial cell clusters as described in the EC atlas. (**C**) Clustered heatmap showing the top 5 marker genes for each cluster. Colors represent row-wise scaled gene expression with a mean of 0 and a standard deviation of 1 (Z scores). (**D**) Number of cells for each cluster in liver ECs. (**E**) Differentiation trajectory of the classic EC phenotypes (arteries, capillaries, veins) in liver. (**F**) PCA on the pairwise Jaccard similarity coefficients between the top 50 marker genes of the classic EC phenotypes (arteries, capillaries and veins) in heart, lung and liver. (**G**) Sankey diagram showing the scmap cluster projection of the EC atlas liver data on the Tabula Muris EC liver data. All the results shown in the figure can be directly explored in the BIOMEX “Case studies” section. The parameters used to generate these plots can be found in the manual.

BIOMEX aims to alleviate the data-analysis-to-interpretation bottlenecks, lowering the barriers needed to extract the biological information embedded in omics measurements. With its user-friendly, highly interactive web-like interface, users can address complex biological questions by using advanced computational tools and fine-tune the analyses in real time. In addition, its design is unconstrained and allows multi-omics data to be simultaneously uploaded and analyzed into one unified framework, providing a well-defined workflow to analyze, interpret and visualize large-scale data such as single cell measurements. BIOMEX also aids the exploration of datasets generated from publicly available profiling efforts (*e.g.* Tabula Muris (8), Human Cell Atlas (60), The Cancer Genome Atlas (44)), repositories (*e.g.* ArrayExpress (61), Gene Expression Omnibus (62)) and added-value databases (*e.g.* EndoDB (63)). Furthermore, it facilitates the shareability of results, reproducibility of analyses and execution of meta-analyses between different experiments. Due to its convenient user interface and comprehensive manual, BIOMEX could also be used as a didactical tool to introduce

researchers to the field of biological data science.

To further promote detailed data mining of (single cell) omics datasets accessible to non-bioinformatician experimental scientists, we made BIOMEX freely available for Windows and Linux at https://www.vibcancer.be/software-tools/biomex. The source code is deposited at https://bitbucket.org/ftaverna/biomex.

Supplementary tables

Name	Version	Source
R distribution and repositories		
base	3.4.4	https://www.R-project.org/
stats	3.4.4	https://www.R-project.org/
bioconductor	3.6	https://www.bioconductor.org/
CRAN	-	https://cran.r-project.org/
Framework		
shiny	1.1.0	https://CRAN.R-project.org/package=shiny
shinydashboard	0.7.1	https://CRAN.R-project.org/package=shinydashboard
shinyjs	1.0	https://CRAN.R-project.org/package=shinyjs
shinyBS	0.62	https://CRAN.R-project.org/package=shinyBS
shinyjqui	0.3.2	https://CRAN.R-project.org/package=shinyjqui
shinycssloaders	0.2.0	https://CRAN.R-project.org/package=shinycssloaders
Visualization		
plotly	4.9.0	https://CRAN.R-project.org/package=plotly
ggplot2	3.1.1	https://CRAN.R-project.org/package=ggplot2
DT	0.6	https://CRAN.R-project.org/package=DT
rhandsontable	0.3.7	https://CRAN.R-project.org/package=rhandsontable
VennDiagram	1.6.20	https://CRAN.R-project.org/package=VennDiagram
Data handling		
data.table	1.12.2	https://CRAN.R-project.org/package=data.table
SingleCellExperiment	1.0.0	https://bioconductor.org/packages/SingleCellExperiment
dplyr	0.8.0.1	https://CRAN.R-project.org/package=dplyr
Annotation		
AnnotationDbi	1.40.0	https://bioconductor.org/packages/AnnotationDbi
org.Hs.eg.db	3.5.0	https://bioconductor.org/packages/org.Hs.eg.db
org.Mm.eg.db	3.5.0	https://bioconductor.org/packages/org.Mm.eg.db
org.Rn.eg.db	3.5.0	https://bioconductor.org/packages/org.Rn.eg.db
org.Bt.eg.db	3.5.0	https://bioconductor.org/packages/org.Bt.eg.db
org.Dr.eg.db	3.5.0	https://bioconductor.org/packages/org.Dr.eg.db
org.Ss.eg.db	3.5.0	https://bioconductor.org/packages/org.Ss.eg.db
Analyses		
edgeR	3.20.9	https://bioconductor.org/packages/edgeR
scran	1.6.9	https://bioconductor.org/packages/scran
GSVA	1.26.0	https://bioconductor.org/packages/GSVA
flashpcaR	2.0	https://github.com/gabraham/flashpca

Name	Version	Source
Rtsne	0.15	https://CRAN.R-project.org/package=Rtsne
umap	0.2.0.0	https://CRAN.R-project.org/package=umap
fastcluster	1.1.25	https://CRAN.R-project.org/package=fastcluster
Seurat	2.3.4	https://CRAN.R-project.org/package=Seurat
FlowSOM	1.10.0	https://bioconductor.org/packages/FlowSOM
heatmaply	0.16.0	https://CRAN.R-project.org/package=heatmaply
dendextend	1.12.0	https://CRAN.R-project.org/package=dendextend
pvclust	2.0.0	https://CRAN.R-project.org/package=pvclust
limma	3.34.9	https://bioconductor.org/packages/limma
MAST	1.4.1	https://bioconductor.org/packages/MAST/
clusterProfiler	3.6.0	https://bioconductor.org/packages/clusterProfiler
pathview	1.18.2	http://bioconductor.org/packages/pathview/
monocle	2.6.4	http://bioconductor.org/packages/monocle/
SCORPIUS	1.0.2	https://CRAN.R-project.org/package=SCORPIUS
scmap	1.1.5	http://bioconductor.org/packages/scmap/
survival	2.44.1.1	https://CRAN.R-project.org/package=survival
caret	6.0.81	https://CRAN.R-project.org/package=caret
ranger	0.11.2	https://CRAN.R-project.org/package=ranger
Parallel computing		
parallel	3.4.4	https://www.R-project.org/
BiocParallel	1.12.0	http://bioconductor.org/packages/BiocParallel/
doParallel	1.0.14	https://CRAN.R-project.org/package=doParallel
Utilities		
colourpicker	1.0	https://CRAN.R-project.org/package=colourpicker
R.devices	2.16.0	https://CRAN.R-project.org/package=R.devices
zip	2.0.1	https://CRAN.R-project.org/package=zip
rChoiceDialogs	1.0.6	https://CRAN.R-project.org/package=rChoiceDialogs
varhandle	2.0.3	https://CRAN.R-project.org/package=varhandle
RCurl	1.95.4.12	https://CRAN.R-project.org/package=RCurl
viridis	0.5.1	https://CRAN.R-project.org/package=viridis
viridisLite	0.3.0	https://CRAN.R-project.org/package=viridisLite
RColorBrewer	1.1.2	https://CRAN.R-project.org/package=RColorBrewer
digest	0.6.18	https://CRAN.R-project.org/package=digest
png	0.1.7	https://CRAN.R-project.org/package=png
rJava	0.9.11	https://CRAN.R-project.org/package=rJava
base64enc	0.1.3	https://CRAN.R-project.org/package=base64enc
Interactive manual generation		
bookdown	0.9	https://CRAN.R-project.org/package=bookdown
htmltools	0.3.6	https://CRAN.R-project.org/package=htmltools

Supplementary Table 1: List of R packages used by BIOMEX.

Complete list of R packages included in BIOMEX. The name, version and source of the packages are shown. The full table is available in the “Supplementary data” section.

Supplementary data

Supplementary data are available at NAR Online.

https://academic.oup.com/nar/article/48/W1/W385/5835814#supplementary-data

The Supplementary data contains the following files:

- Supplementary Table 1: list of R packages used by BIOMEX.
- Additional information - methods overview: overview of the analysis and statistical methods used in BIOMEX. This file was used and incorporated in Chapter I of the thesis.

Author contributions

P.C. & J.G. conceptualized the BIOMEX design; F.T. developed the software; F.T., J.G., S.K., K.R., L.T., A.S. evaluated the software; F.T., T.K., J.G., L.T, P.C. wrote the manuscript; L.T., F.T. made the figures; K.R., A.S., S.K., G.E, M.D., L.S. provided advice and helped improve the manuscript. All authors commented on the manuscript.

Conflict of interest declaration

None of the authors have competing financial interests to declare.

Funding

This work was supported by the Fonds voor Wetenschappelijk Onderzoek (FWO-Vlaanderen) (to J.G., K.R., L.T., P.C.); Kom op Tegen Kanker (Stand up to Cancer, Flemish Cancer Society) (to S.K., P.C.); Devotion (Developmental Origins of Chronic Diseases in Children) Network (to T.K.); Methusalem funding (to P.C.); Foundation against Cancer [2016-078 to P.C.]; European Research Council Proof of Concept [ERC-713758 to P.C.] and Advanced European Research Council Grant [EU-ERC743074 to P.C.].

Author information

Correspondence and requests for materials should be addressed to PC (peter.carmeliet@kuleuven.vib.be).

References

1. Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A. and Ferrari, R. (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform, 19, 286-302.

2. Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S. and Robinson, G.E. (2015) Big Data: Astronomical or Genomical? PLoS Biol,

13, e1002195.

3. Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E. and Blayney, J.K. (2019) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform, 20, 1795-1811.

4. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. Nature, 422, 198-207.

5. Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007) Mass spectrometry-based metabolomics. Mass spectrometry reviews, 26, 51-78.

6. Hwang, B., Lee, J.H. and Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med, 50, 96.

7. Bhaduri, A., Nowakowski, T.J., Pollen, A.A. and Kriegstein, A.R. (2018) Identification of cell types in a mouse brain single-cell atlas using low sampling coverage. BMC Biol, 16, 113.

8. Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M.B. et al. (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature, 562, 367-372.

9. Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol, Chapter 19, Unit 19.10.11-21.

10. Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P. et al. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res, 41, W557-561.

11. Tautenhahn, R., Patti, G.J., Rinehart, D. and Siuzdak, G. (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem, 84, 5035-5039.

12. Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 26, 1367-1372.

13. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M. and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature methods, 13, 731-740.

14. Hait, T.A., Maron-Katz, A., Sagir, D., Amar, D., Ulitsky, I., Linhart, C., Tanay, A., Sharan, R., Shiloh, Y., Elkon, R. et al. (2019) The EXPANDER Integrated Platform for Transcriptome Analysis. Journal of Molecular Biology, 431, 2398-2406.

15. Nolte, H., MacVicar, T.D., Tellkamp, F. and Kruger, M. (2018) Instant Clue: A Software Suite for Interactive Data Visualization and Analysis. Scientific reports, 8, 12648.

16. Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic acids research, 37, W652-W660.

17. Alyass, A., Turcotte, M. and Meyre, D. (2015) From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics, 8, 33.

18. Mattmann, C.A. (2013) Computing: A vision for data science. Nature, 493, 473-475.

19. Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X. and Zhu, F. (2017) NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res, 45, W162-w170.

20. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford,

England), 26, 139-140.

21. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol, 33, 495-502.

22. Haghverdi, L., Lun, A.T.L., Morgan, M.D. and Marioni, J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol, 36, 421-427.

23. Hanzelmann, S., Castelo, R. and Guinney, J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. BMC bioinformatics, 14, 7.

24. Jolliffe, I.T. and Cadima, J. (2016) Principal component analysis: a review and recent developments. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374, 20150202.

25. Abraham, G. and Inouye, M. (2014) Fast principal component analysis of large-scale genome-wide data. PloS one, 9, e93766.

26. van der Maaten, L.J.P. and Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.

27. McInnes, L., Healy, J. and Melville, J. (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

28. Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T. and Saeys, Y. (2015) FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry. Part A : the journal of the International Society for Analytical Cytology, 87, 636-645.

29. Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics (Oxford, England), 22, 1540-1542.

30. Galili, T., O'Callaghan, A., Sidi, J. and Sievert, C. (2018) heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics (Oxford, England), 34, 1600-1602.

31. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 43, e47.

32. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M. et al. (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol, 16, 278.

33. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. and Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. Behavioural brain research, 125, 279-284.

34. Bruning, U., Morales-Rodriguez, F., Kalucka, J., Goveia, J., Taverna, F., Queiroz, K.C.S., Dubois, C., Cantelmo, A.R., Chen, R., Loroch, S. et al. (2018) Impairment of Angiogenesis by Fatty Acid Synthase Inhibition Involves mTOR Malonylation. Cell metabolism, 28, 866-880.e815.

35. Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L. and Chory, J. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics (Oxford, England), 22, 2825-2827.

36. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

37. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology, 16, 284-287.

38. Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.L., Visvader, J.E. and Smyth, G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics (Oxford, England), 26, 2176-2182.

39. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28, 27-30.

40. Luo, W. and Brouwer, C. (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics (Oxford, England), 29, 1830-1831.

41. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol, 32, 381-386.

42. Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D. and Saeys, Y. (2016) SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv, 079509.

43. Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018) scmap: projection of single-cell RNA-seq data across data sets. Nature methods, 15, 359-362.

44. Tomczak, K., Czerwinska, P. and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology (Poznan, Poland), 19, A68-77.

45. Goel, M.K., Khanna, P. and Kishore, J. (2010) Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research, 1, 274-278.

46. Bland, J.M. and Altman, D.G. (2004) The logrank test. BMJ (Clinical research ed.), 328, 1073.

47. Wright, M.N. and Ziegler, A. (2017) ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software; Vol 1, Issue 1 (2017).

48. Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

49. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46, 389-422.

50. Refaeilzadeh, P., Tang, L. and Liu, H. (2009) In Liu, L. and ÖZsu, M. T. (eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 532-538.

51. Chen, J., Qian, Z., Li, F., Li, J. and Lu, Y. (2017) Integrative Analysis of Microarray Data to Reveal Regulation Patterns in the Pathogenesis of Hepatocellular Carcinoma. Gut and liver, 11, 112-120.

52. Cahan, P., Rovegno, F., Mooney, D., Newman, J.C., St Laurent, G., 3rd and McCaffrey, T.A. (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene, 401, 12-18.

53. Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.C., Pircher, A., Geldhof, V., de Rooij, L., Kalucka, J., Sokol, L. et al. (2020) An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. Cancer Cell, 37, 21-36.e13.

54. Levandowsky, M. and Winter, D. (1971) Distance between Sets. Nature, 234, 34-35.

55. Wickham, H. (2009) ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company, Incorporated.

56. Rizvi, S., Khan, S.A., Hallemeier, C.L., Kelley, R.K. and Gores, G.J. (2018) Cholangiocarcinoma - evolving concepts and therapeutic strategies. Nature reviews. Clinical oncology, 15, 95-111.

57. Loosen, S.H., Roderburg, C., Kauertz, K.L., Koch, A., Vucur, M., Schneider, A.T., Binnebösel, M., Ulmer, T.F., Lurje, G., Schoening, W. et al. (2017) CEA but not CA19-9 is an independent prognostic factor in patients undergoing resection of cholangiocarcinoma. Scientific reports, 7, 16975.

58. Zhong, W., Dai, L., Liu, J. and Zhou, S. (2018) Cholangiocarcinomaassociated genes identified by integrative analysis of gene expression data. Molecular medicine reports, 17, 5744-5753.

59. Kalucka, J., de Rooij, L.P.M.H., Goveia, J., Rohlenova, K., Dumas, S.J., Meta, E., Conchinha, N.V., Taverna, F., Teuwen, L.-A., Veys, K. et al. (2020) Single-Cell Transcriptome Atlas of Murine Endothelial Cells. Cell.

60. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. et al. (2017) The Human Cell Atlas. eLife, 6.

61. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T. et al. (2015) ArrayExpress update--simplifying data submissions. Nucleic Acids Res, 43, D1113-1116.

62. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res, 35, D760-765.

63. Khan, S., Taverna, F., Rohlenova, K., Treps, L., Geldhof, V., de Rooij, L., Sokol, L., Pircher, A., Conradi, L.C., Kalucka, J. et al. (2019) EndoDB: a database of endothelial cell transcriptomics data. Nucleic Acids Res, 47, D736-d744.

Chapter IV EndoDB

This chapter is based on the first author research article published in Nucleic Acid Research.
https://academic.oup.com/nar/article/47/D1/D736/5144156

This is a supporting manuscript of the thesis and is meant to showcase the application of BIOMEX in analyzing publicly available data.

ENDODB: A DATABASE OF ENDOTHELIAL CELL TRANSCRIPTOMICS DATA

Shawez Khan, Federico Taverna*, Katerina Rohlenova*, Lucas Treps, Vincent Geldhof, Laura de Rooij, Liliana Sokol, Andreas Pircher, Lena-Christin Conradi, Joanna Kalucka, Luc Schoonjans, Guy Eelen, Mieke Dewerchin, Tobias Karakach, Xuri Li#, Jermaine Goveia# and Peter Carmeliet#*

*: equal co-authorship; #: co-corresponding author

Editorial correspondence

Peter Carmeliet: peter.carmeliet@kuleuven.vib.be

Jermaine Goveia: jermaine.goveia@kuleuven.vib.be

Xuri Li: lixr6@mail.sysu.edu.cn

Personal contribution: I created the code used to perform the downstream analysis on all the collected transcriptomics data for the database and developed the back-end and front-end of the EndoDB database.

The supplemental tables can be found in the online version of the publication.
https://academic.oup.com/nar/article/47/D1/D736/5144156#supplementary-data

Abstract

Endothelial cells (ECs) line blood vessels, regulate homeostatic processes (blood flow, immune cell trafficking), but are also involved in many prevalent diseases. The increasing use of high-throughput technologies such as gene expression microarrays and (single cell) RNA sequencing generated a wealth of data on the molecular basis of EC (dys-)function. Extracting biological insight from these datasets is challenging for scientists who are not proficient in bioinformatics. To facilitate the re-use of publicly available EC transcriptomics data, we developed the endothelial database EndoDB, a web-accessible collection of expert curated, quality assured and pre-analyzed data collected from 360 datasets comprising a total of 4,741 bulk and 5,847 single cell endothelial transcriptomes from 6 species. Unlike other added-value databases, EndoDB allows to easily retrieve and explore data of specific studies, determine under which conditions genes and pathways of interest are deregulated, and assess reprogramming of metabolism via principal component analysis, differential gene expression analysis, gene set enrichment analysis, heatmaps and metabolic and transcription factor analysis, while single cell data is visualized as gene expression color-coded t-SNE plots. Plots and tables in EndoDB are customizable, downloadable and interactive. EndoDB is freely available at https://vibcancer.be/software-tools/endodb, and will be updated to include new studies.

Introduction

Endothelial cells (ECs) line the lumen of blood vessels, are metabolically active and orchestrate important processes such as vasomotor tone, coagulation, permeability, tissue vascularization and immune response (1). The inability of ECs to fulfill their physiological functions is a key feature of multiple prevalent diseases including hypertension, atherosclerosis, diabetes, and cancer (2). In the past three decades, hypothesis-driven studies have started to unravel the molecular mechanisms that underlie EC function in health and disease, which resulted in clinically-approved EC-targeting drugs such as bevacizumab and ranibizumab for cancer and age-related wet macular degeneration, respectively (3). Recent technological advances have made it possible to perform global profiling of transcript (transcriptomics), protein (proteomics) and metabolite levels (metabolomics), even in single cells. These technologies have provided unprecedented insight in EC biology and generated many novel hypotheses (4).

To facilitate the re-use of omics datasets, it is now a standard requirement of most scientific journals to make published data publicly available via repositories such as ArrayExpress (5) and Gene Expression Omnibus (GEO) (6). However, the main purpose of

these primary archives is to store raw data in a format that is accessible by bioinformaticians, not to enable bench scientists and experimentalists to directly explore and re-use data to answer biological questions (7). In the cancer field, added-value databases (databases that are specifically designed to allow bench scientists to explore pre-selected, highly-curated and pre-analyzed data) have been transformative and facilitated several breakthrough discoveries (8-13). A few independent research groups have made the results of their EC profiling efforts web-accessible (14,15), but currently no comprehensive EC-specific added-value databases exist to facilitate the reuse of gene expression data by the vascular biology community. Here we describe the development of the endothelial cell database (EndoDB), the first freely available added-value database of EC (single cell) transcriptomics studies, designed to allow vascular biologists and other bench scientists to unlock the untapped potential of publicly available data via an easy-to-use interactive web interface (https://vibcancer.be/software-tools/endodb). The user-interface and functionality of the EndoDB were developed in close collaboration between software developers, bioinformaticians and vascular biologists. The EndoDB is easy-to-use, allows to interactively explore data using powerful statistical and bioinformatics approaches, and has been field-tested and used in several recent publications (16,17).

Materials and Methods

Retrieval of EC transcriptomics data and database content

We first aimed to collect and organize EC transcriptomics datasets available in the public domain. To do this, we used a two-step approach (Figure 1 and Supplementary Table 1). First, we used a broad and sensitive filter ("endothelial OR endothelium") to search ArrayExpress and GEO for EC transcriptomics datasets, which returned 1,121 studies (as of June 1st 2018). We then screened the title, abstract and if necessary sample information to determine which of these studies performed gene expression profiling in ECs specifically. We excluded 663 studies based on the title abstract screen and 113 studies for which data was not made available in the public domain, resulting in a total of 345 studies comprising 357 bulk transcriptomics datasets (a single study can contain multiple independent datasets).

Second, we complemented our search with a pubmed screen to identify recently published single cell transcriptomics studies of ECs, which retrieved 10 single cell RNA sequencing (scRNA-seq) studies (as of June 1st 2018). To ensure high quality scRNA-seq data in the EndoDB, we only considered studies that aimed to perform scRNA-seq in ECs specifically, used isolation protocols optimized for this purpose, and sequenced >1,000 ECs because: i) ECs are a minority cell type in all tissues, ranging from ~1-5% in tumor tissue to

~20% in highly vascularized organs such as the lung (18,19); ii) generic dissociation protocols fail to adequately capture ECs, resulting in a further depletion of the EC fraction (20); iii) ECs are sensitive to dissociation induced artifacts (20), which biases data interpretation (21), together indicating that dissociations protocols have to be specifically optimized to isolate a high number of high quality ECs; iv) the reliability of single cell analyses depends on the number of cells analyzed (22,23); and v) the number of ECs that can be sequenced with more advanced technology (e.g. droplet-based sequencing) will increase by orders of magnitude. Based on these criteria, we excluded eight studies (Supplementary table 2), and finally included two studies comprising three datasets.

The final database contains 4,744 bulk and 5,874 single cell EC transcriptomes from 6 different model organisms (*Homo sapiens* [human], *Mus musculus* [mouse], *Rattus norvegicus* [rat], *Bos taurus* [cow*], Danio rerio* [zebrafish], [pig]), generated using a variety of technologies (micro-arrays, (single cell) RNA sequencing) and platforms (Affymetrix, Agilent, Illumina).

Pre-processing and quality control of bulk transcriptomics data

To maximize the number of datasets included in the EndoDB, we aimed to include data from all micro-array and RNA-sequencing platforms. For micro-arrays, we retrieved raw data files when possible, but included log2-transformed pre-processed data when raw data was not available. Raw data from the affymetrix platform was downloaded as binary or text CEL files, normalized using the robust multi-array average (RMA) using the R-packages affy and oligo, followed by quantile normalization, log2 transformation and probe-set summarization (24,25). Raw data files from Illumina BeadChips were normalized using the normexp algorithm using control probes (neqc) from the limma package (26). Raw data from Agilent platforms was normalized using the normalizeWithinArrays and normalizebetweenArrays functions available from limma package (26). Probe set identifiers were mapped to Entrez identifiers, official gene symbols and gene names using annotation packages for the corresponding platform. We applied standard quality controls using the arrayQualityMetrics R package to identify low-quality arrays (27). In this approach, samples were scored on the basis of three metrics, the distances between the arrays, Bland-Altman plot (MA plots) and boxplots, samples that were flagged as outliers based on more than one of these metrics were removed from the analysis in the EndoDB, but the raw data of all samples can be downloaded for down-stream analyses in independent software pipelines. Finally, we randomly selected 25% of studies and cross-checked our results for congruency with the original publication.

For RNA-sequencing, data was retrieved in fastq format, and alignment to the

reference genome and quantification of transcript abundances was performed using the Kallisto software (28). We performed trimmed mean of M-values (TMM)-normalization using the R-package edgeR (29) and VOOM-normalization using limma before down-stream differential analysis.

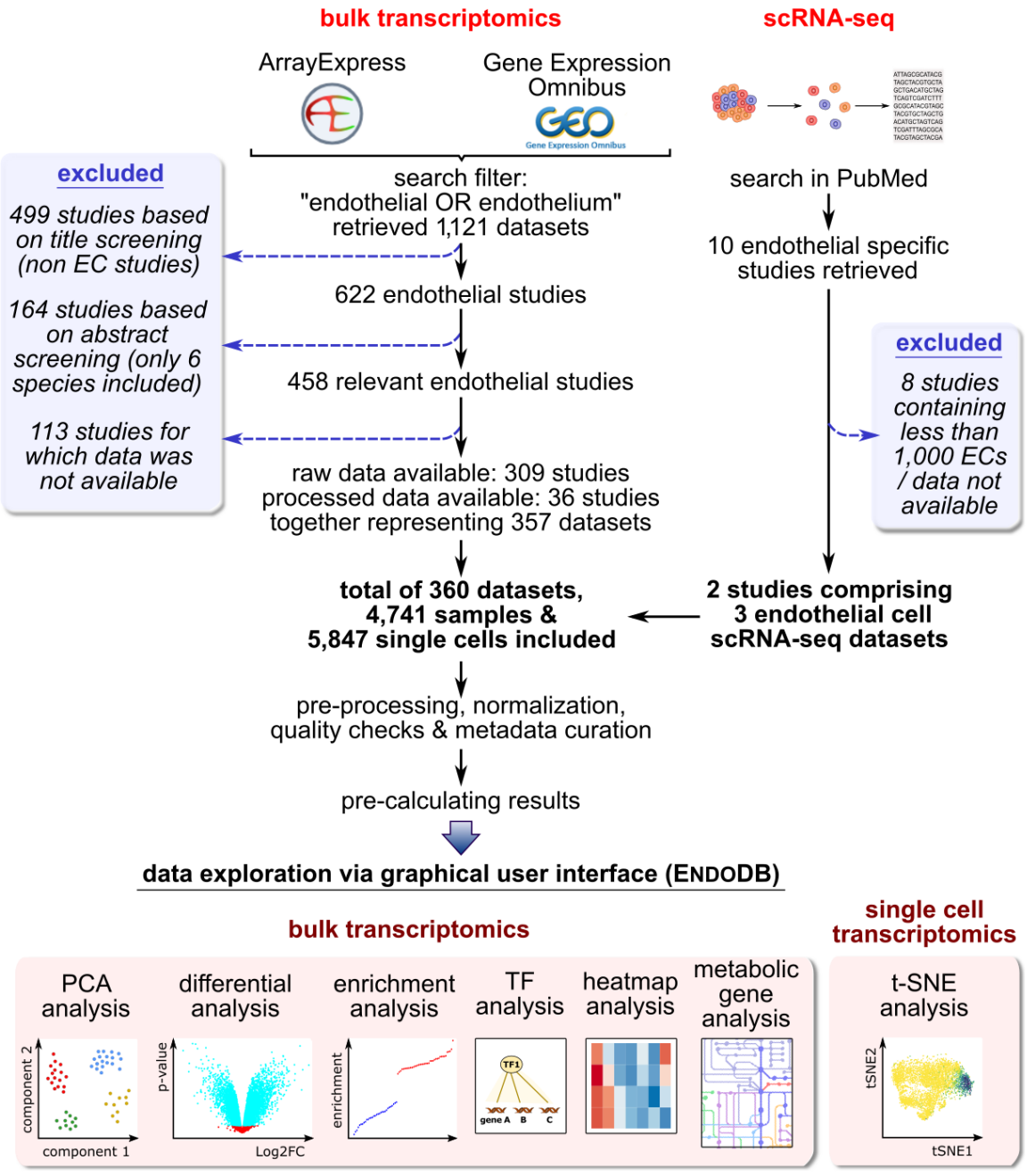

Figure 1: Reconstruction and contents of the EndoDB.

Schematic overview of EndoDB construction and functionalities.

Pre-processing and quality control of single cell data

Recent technological breakthroughs have made it possible to RNA sequence transcriptomes of single cells (scRNA-seq) (30). These studies have great resource value and are beginning to elucidate the EC heterogeneity at the single cell level. However, so far, no general searchable database of EC-specific scRNA-seq data exists, which limits the general use of these large-scale sequencing efforts. To increase the resource value of scRNA-seq EC datasets, we have developed a dedicated scRNA-seq EndoDB module. For each dataset, we removed low quality and apoptotic cells based on the number of expressed genes (<200) and the percentage of reads assigned to mitochondrial genes (>5%) (31), and performed library size normalization followed by natural log transformation using log1p available via the Seurat package (32). After initial graph-based clustering using Seurat, we *in silico* selected ECs based on the expression of canonical marker genes (*PECAM1*, *CDH5*) and the absence of leukocytes (*PTPRC*), fibroblasts (*COL1A1*) and pericytes (*PDGFRB*).

Briefly, to *in silico* select the ECs in a dataset, we used an unbiased clustering approach, in which we clustered the cells together based on their expression profile, and then assessed which clusters express canonical EC marker genes (*PECAM1*, *CDH5*). We further excluded clusters that expressed markers for leukocytes (*PTPRC*), pericytes (*PDGFRB*) and fibroblasts (COL1A1). Using a clustering approach rather than marker gene gating on a cell per cell basis is more robust, since it allows to overcome misclassification of cells as non-EC due to gene drop out.

Metadata curation

Metadata was downloaded from GEO or ArrayExpress and manually curated by a team of expert vascular biologists. Repetitive and redundant information was removed, technical information such as the chip design and library preparation was standardized, and each array was manually annotated with biological information such as the model organism, vascular bed, endothelial cell type, and the experimental conditions (Figure 2, Table 1-3 and Supplementary Table 4-5). In a second step, we used the OpenRefine software (33) to standardize spelling across all arrays and datasets. In addition, a description of each study, the GEO or ArrayExpress identifiers, and references to the original publication were included in the metadata.

Data analysis and visualization

For all bulk transcriptomics datasets independently, we performed: i) principal component analysis (PCA) and visualized the results as two-dimensional plots where individual samples are color-coded according to the experimental design; ii) differential gene expression analysis

(DGEA) (34) and visualized the results as volcano plots, bar plots and browsable tables; iii) gene set enrichment analysis (GSEA) (35,36) and visualized the results as waterfall plots; and iv) heatmap analysis to show gene expression changes of a panel of user-selected genes. Together, these analyses cover some of the most commonly used analyses and allow for detailed data interpretation. EC single cell data are provided as interactive t-SNE plots using the Rtsne and plotly packages (see Supplementary Table 3 for scRNA-seq pre-processing and visualization parameters). All analytical results are pre-calculated and stored in the EndoDB to reduce computation time and to allow fast data retrieval.

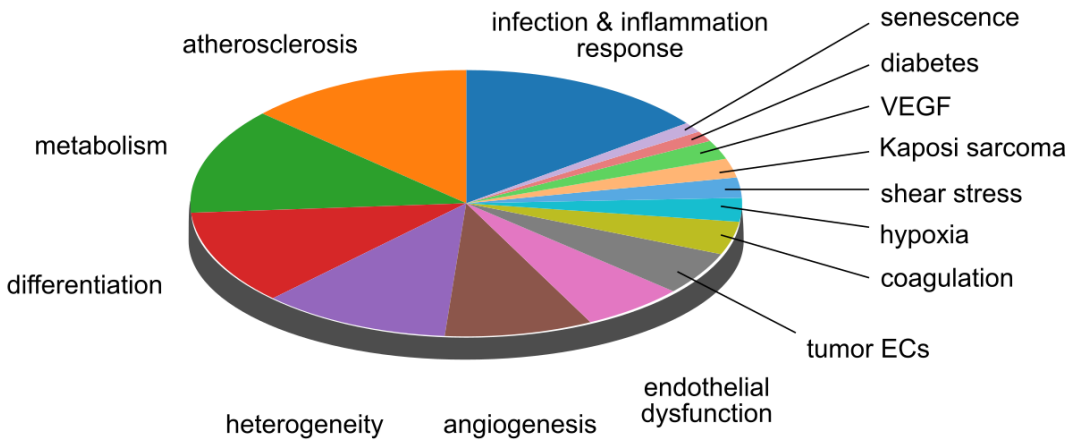

Figure 2: Most common experimental conditions in the EndoDB.

Relative representation of the 15 most common sample types in the EndoDB grouped by experimental conditions. Abbreviations - EC: endothelial cell; *VEGF*: vascular endothelial growth factor.

Website implementation

The EndoDB is implemented as an interactive web application using the R/Shiny web framework (37,38). We used the plotly package for data visualization and the data.table and DT packages to provide searchable tables (39).

Data availability and downloads

All pre-processed data and curated metadata are available for download in comma-separated flat file format. Visualizations can be downloaded as PNG or HTML files that preserve the same interactivity as available from the web application, tables can be downloaded in comma-separated value format.

Documentation and manual

A detailed user manual and video tutorials are available via the web application.

Species	Colloquial species name	Number of datasets in EndoDB
Homo sapiens	Human	270
Mus musculus	Mouse	76
Rattus norvegicus	Rat	5
Bos taurus	Cow	3
Danio rerio	Zebrafish	3
Sus scrofa	Pig	3

Table 1: List of species in the EndoDB.

Complete list of species included in the EndoDB, the number of datasets per species is indicated.

Organ	Number of datasets in EndoDB
Umbilical cord	149
Skin	42
Aorta	28
Lung	28
Heart	25
Brain	22
Liver	16
Eye	15
Blood	13
Lymphatic system	13

Table 2: Top 10 most common tissue types in the EndoDB.

Top 10 most common tissue types in the EndoDB, the number of datasets per tissue type is indicated.

Cell type	Number of datasets in EndoDB
Umbilical vein ECs	142
Aorta ECs	29
Dermal ECs	23
Lymphatic ECs	21
Coronary artery ECs	17
Pulmonary ECs	16
Brain ECs	11
Retinal ECs	11
Dermal lymphatic ECs	10
Blood outgrowth ECs	9

Table 3: Top 10 most common EC types in the EndoDB.

Top 10 most common EC types in the EndoDB, the number of datasets per cell type is indicated.

EndoDB functionality

The functionality of the EndoDB is centered on four analytical approaches to extract biologically meaningful information from (single cell) transcriptomics data. Below, we briefly discuss each approach and provide illustrative examples focused on tumor ECs (TECs). TECs are targets of clinically approved anti-angiogenic therapy but remain poorly characterized. Isolating TECs from human or murine tumors is notoriously challenging and laborious, re-use of publicly available data to investigate gene expression reprogramming in TECs is therefore a potentially favorable and cost effective complementary approach. The examples below are for illustrative purposes only, detailed transcriptome analysis and biological interpretation is beyond the scope of this manuscript.

Study-centered data exploration

An important purpose of the EndoDB is to facilitate the re-use of publicly available datasets to address novel and unresolved questions, and to allow comparative analyses of similar studies to resolve study-specific biases. The study-centered data exploration options allow users to retrieve studies based on the study title, pubmed identifier, GEO and ArrayExpress identifiers, and manually annotated keywords. After selecting a study, users can download all data and metadata or explore the data using the results from PCA, DGEA, GSEA, heatmap analysis and bar plots.

Example: We used the study-centered data exploration functionality to re-analyze a previously published dataset on ECs isolated from normal murine hindbrain, Sonic hedgehog (Shh)-driven or Wnt-driven medulloblastoma (40). PCA showed that normal brain ECs (NECs) have a clearly distinct transcriptional profile compared to Shh- and Wnt-driven medulloblastoma tumor ECs (TECs) (Figure 3A). To determine which pathways are most upregulated in TECs, we assessed differences in gene expression signatures in NECs *versus* Shh-medulloblastoma TECs (Shh-TECs). This analysis revealed upregulation of cell cycle and DNA replication (Figure 3B). Consistently, DGEA showed that several proliferation markers (e.g. *Mki67*, *Top2a*) ranked in the top 25 most upregulated genes (Figure 3C-E). Together, these data suggest that Shh-TECs undergo an angiogenic switch and adopt a proliferating phenotype.

Metabolism- and transcription factor centered data exploration

Emerging evidence indicates that EC metabolism can overrule fundamental signaling cascades, and several metabolic enzymes and transporters have been identified as therapeutic targets to inhibit pathological angiogenesis in ocular disease, inflammation and cancer (41-43).

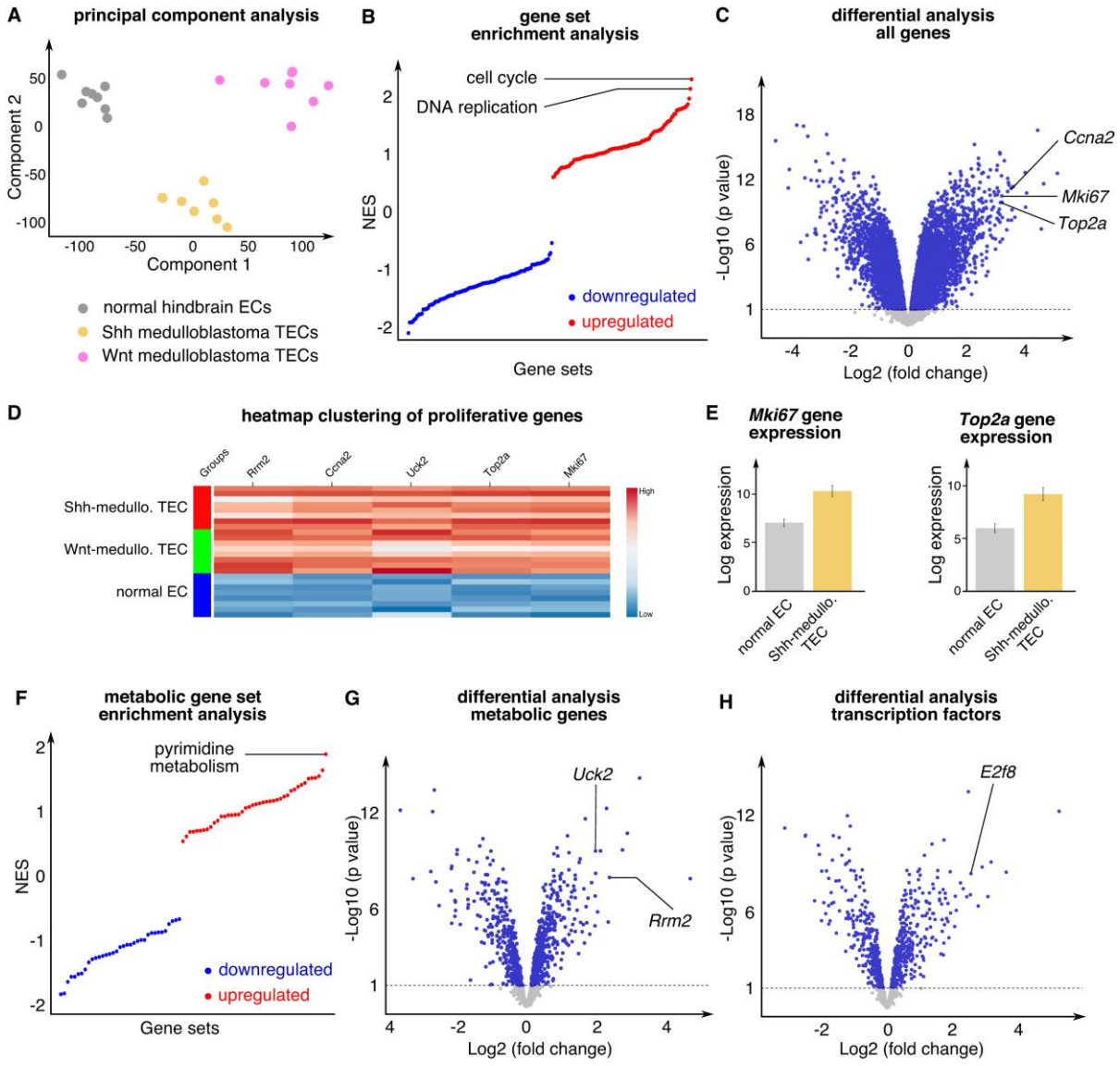

Figure 3: Study-centered data exploration.

(**A**) PCA of normal hindbrain, Shh-medulloblastoma and Wnt-medulloblastoma ECs (study E-GEOD-73753). (**B**) Gene set enrichment analysis of normal hindbrain ECs *versus* Shh-medulloblastoma ECs. The upregulated gene sets are shown in red, the downregulated gene sets are shown in blue. (**C**) Differential analysis of normal hindbrain ECs *versus* Shh-medulloblastoma ECs shown in volcano plot; some highly deregulated proliferation-associated genes are indicated. (**D**) Gene expression heatmap of normal hindbrain ECs *versus* Shh-medulloblastoma and Wnt-medulloblastoma. The high-gene expression levels are shown in red, the low-expression levels in blue. (**E**) Expression of the indicated genes in normal hindbrain ECs and Shh-medulloblastoma ECs. (**F**) Metabolic gene set enrichment analysis of normal hindbrain ECs *versus* Shh-medulloblastoma ECs. The upregulated gene sets are shown in red, the downregulated gene sets are shown in blue. (G and H) Differential analysis of normal hindbrain ECs *versus* Shh-medulloblastoma ECs for the subset of metabolic genes (**G**) and transcription factors (**H**) shown as a volcano plot. Abbreviations - *Ccna2*: cyclin A2; *Mki67*: marker of proliferation ki67; NES: normalized enrichment score; *Rrm2*: ribonucleotide reductase regulatory subunit M2; Shh: Sonic Hedgehog; TEC: tumor endothelial cell; *Top2a*: topoisomerase 2 alpha; *Uck2*: uridine-cytidine kinase 2.

While the importance of EC metabolism is becoming increasingly clear, still little is known about how metabolism supports EC functions in health and disease (44). Analyses of the subset of metabolic genes and gene sets can provide important insight in metabolic reprograming (45,46). We therefore performed metabolic gene expression analysis and metabolic GSEA for all studies in the EndoDB. Similarly, to determine which transcription factors are deregulated in response to experimental manipulation, we performed analysis in the subset of transcription factors.

Example: Performing metabolic GSEA in Shh-TECs *versus* NECs revealed that pyrimidine metabolism is the most upregulated metabolic pathway. Consistently, metabolic GSEA and DGEA revealed upregulation of several pathways (pyrimidine metabolism) and genes (*Rrm2*, *Uck2*) involved in nucleotide biosynthesis (Figure 3F-G). Together, these results may indicate that proliferating Shh-TECs upregulate metabolic pathways to sustain nucleotide production required for DNA synthesis. Consistently, the cell cycle related transcription factor *E2f8* was among the most upregulated transcription factors (Figure 3H).

Gene- and pathway-centered data exploration

The EndoDB contains the results of pair-wise DGEA and GSEA results between all experimental conditions within all datasets. The gene- and pathway-centered data exploration functionality allows to unbiasedly determine under which (patho)-physiological and experimental conditions the expression of a particular gene or pathway of interest is most deregulated.

Example: Having identified *Mki67* as a gene of interest in Shh-TECs, we interrogated the EndoDB to determine under which other conditions this gene is upregulated. To do this, we performed a gene-centered search, which showed that *Mki67* is highly upregulated in the brain endothelioma cell line (b.End5) compared with primary cultures of normal brain ECs (Figure 4A-B). DGEA and GSEA analysis further showed that *Mki67*, *Uck2* and DNA replication and pyrimidine metabolism are among the most upregulated genes and pathways in b.End5 endothelioma cells (Figure 4C-E). Together, these findings show that proliferation is a general phenotype of both TECs and malignant endothelial cells.

Single cell data exploration

An increasing number of studies investigates EC heterogeneity at the single cell level (14,47). scRNA-seq studies are expensive large-scale sequencing efforts that generate high-dimensional datasets with high resource value.

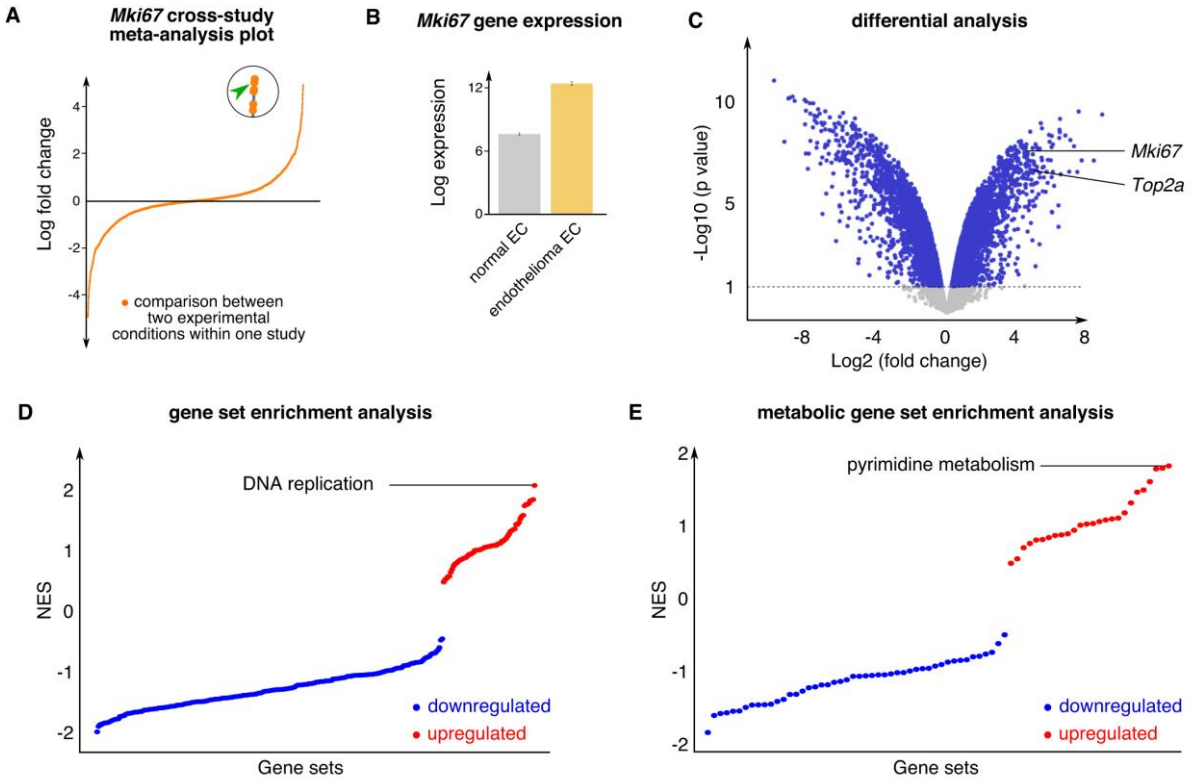

Figure 4: Gene- and pathway-centered data exploration.

(**A**) Dot plot showing log fold change of *Mki67* gene expression in all sample comparisons in EndoDB datasets. The dot indicated by the green arrow head in the enlarged circle represents the pair-wise comparison selected for further analysis (data from study E-GEOD-14375). (**B**) Expression of the *Mki67* gene in normal hindbrain ECs and endothelioma ECs. (**C**) Differential analysis of normal brain ECs *versus* endothelioma ECs for all genes shown in volcano plot. *Mki67* and *Top2a* genes are indicated. (**D** and **E**) Gene set enrichment analysis of normal brain ECs *versus* endothelioma ECs for all genes (**D**) and for the subset of detected metabolic genes (**E**). The upregulated gene sets are shown in red, the downregulated gene sets are shown in blue. Abbreviations - *Mki67*: marker of proliferation ki67; NES: normalized enrichment score; *Top2a*: topoisomerase 2 alpha.

We included a specific module in the EndoDB to enable exploration of transcriptional heterogeneity in scRNA-seq datasets via t-SNE plots color-coded for the expression of all detected genes.

Example: Based on the study- and gene-centered analyses described above, we hypothesized that ECs in tumors adopt a proliferative phenotype. From bulk transcriptomics analysis alone it is not possible to determine whether all TECs, or only a subset, are actively proliferating. To determine whether the TEC-associated genes *Mki67*, *Top2a*, and *Rrm2* are upregulated in a specific subset of TECs, we interrogated scRNA-seq data of TECs isolated from a colorectal cancer (COLO205) xenograft model (47). Interestingly, all three genes marked a specific subset of TECs suggesting that the observed upregulation of cell cycle-related genes in bulk transcriptomics data results from increased expression in a specific subset of TECs (Figure 5).

Together, these illustrative examples demonstrate the benefit and power of the

EndoDB to capitalize on publicly available data to generate novel testable hypotheses in clinically relevant settings.

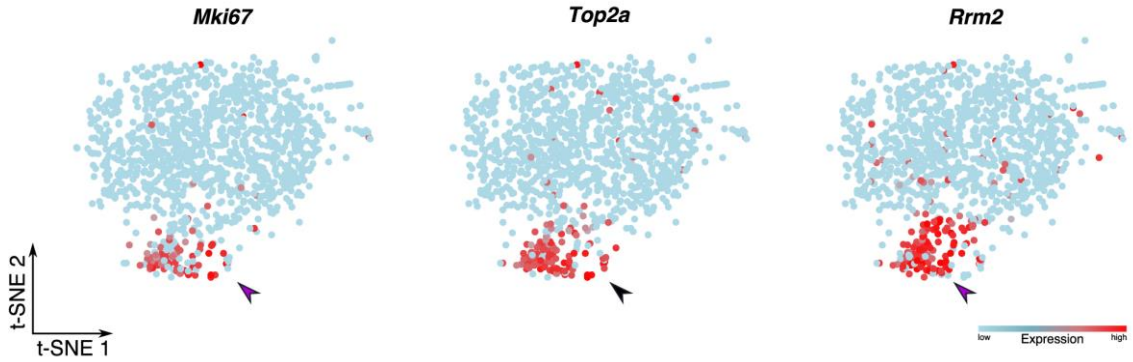

Figure 5: Single cell data exploration.

t-SNE plots of COLO205 TECs color-coded for the expression of the indicated proliferation-related genes (GSE110501). Abbreviations - *Mki67*: marker of proliferation ki67; *Rrm2*: ribonucleotide reductase regulatory subunit M2; *Top2a*: topoisomerase 2 alpha. The arrow head indicates the cluster of cells with highest expression.

Conclusion

EndoDB facilitates bench scientists to unlock the untapped potential of publicly available transcriptomics data, but can also be used by bioinformaticians as a resource of expert curated data. EndoDB has been constructed using field-tested methods, is tailored to extract biologically meaningful insight from gene expression profiling efforts, covers emerging fields such as EC metabolism, and is a first-in-class resource of EC scRNA-seq data. EndoDB distinguishes itself from other databases by its flexible user interface, providing an exceptionally large variety of pre-calculated data and analyses in customizable plots downloadable tables, allowing direct generation and implementation of working hypotheses.

The EndoDB was developed anticipating that an increasing number of EC profiling datasets will become available in the near future. The EndoDB can be further expanded by including datasets derived from other vascular cells, such as vascular smooth muscle cells and pericytes. The EndoDB can also be expanded to include other omics data such as metabolomics and proteomics.

Since ECs have key roles in health and disease, and (in)directly influence all cells in the body, we expect that EndoDB will be a useful tool, for both vascular biologists and the broader biomedical research community, to derive unprecedented insight in EC heterogeneity, identify new (metabolic) therapeutic targets and generate novel hypotheses. EndoDB is freely available at https://vibcancer.be/software-tools/endodb, and will be regularly updated.

Data availability

All pre-processed data and curated metadata are available for download in comma-separated flat file format.

Supplementary data

Supplementary data are available at NAR Online.

https://academic.oup.com/nar/article/47/D1/D736/5144156#supplementary-data

The Supplementary data contains the following files:

- Supplementary Table 1: data mining steps for inclusion of datasets in EndoDB.
- Supplementary Table 2: evaluated single cell RNA-seq datasets.
- Supplementary Table 3: complete list of datasets included in the EndoDB grouped by cell type.
- Supplementary Table 4: complete list of datasets included in the EndoDB grouped by organ type.
- Supplementary Table 5: parameters for single cell RNA-seq re-analysis.

Funding

S.K. is funded by a grant of Kom op Tegen Kanker (Stand up to Cancer, Flemish Cancer Society). K.R., J.K and J.G. are supported by the 'Fonds voor Wetenschappelijk Onderzoek' (FWO-Vlaanderen). L.T. is funded by a Marie Curie-IEF Fellowship and the Bettencourt Schueller Foundation. V.G. is funded by Strategisch Basisonderzoek Fonds Wetenschappelijk Onderzoek-Vlaanderen (SB-FWO). A.P. is funded by the Austrian Science Fund (FWF) (J3730-B26). L.-C.C. is funded by a fellowship of the Else Kröner-Fresenius-Stiftung and the Fritz Thyssen Stiftung (10.16.2.017MN). X.L. is supported by the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center at the Sun Yat-SenUniversity and by the National Natural Science Foundation of China (81330021, 81670855). The work of P.C. is supported by the VIB TechWatch program, long-term structural Methusalem funding by the Flemish Government, grants from the 'Fonds voor Wetenschappelijk Onderzoek' (FWO-Vlaanderen), Foundation against Cancer (2012-175 and 2016-078) and ERC Advanced Research Grant (EU-ERC743074).

Conflict of interest declaration

None of the authors have competing financial interests to declare.

Author contributions

J.G., K.R., S.K. searched primary archives and PubMed to identify EC transcriptomics studies; F.T., J.G., S.K. developed the R scripts for data retrieval and analysis; F.T. implemented the database; A.P., J.G., K.R., L.-C.C, L.R., L. Sokol, L.T., V.G. curated metadata; J.G., K.R., S.K. clustered single cell data; K.R., L.T., S.K. made the figures; J.G., P.C. conceptualized the EndoDB and supervised the study; J.G., K.R., S.K, P.C wrote the manuscript; J.G., P.C., X.L. contributed to the execution, support and analysis of experiments, data interpretation and/or advice; G.E., J.K., M.D., L. Schoonjans, T.K. provided advice and discussed results. All authors commented on the manuscript.

Acknowledgements

We acknowledge the help of Mathias Declercq and Valentina Lykhopiy for their feedback on the database design and manuscript preparation.

Author information

Correspondence and requests for materials should be addressed to PC (peter.carmeliet@kuleuven.vib.be)

References

1. Aird, W.C. (2012) Endothelial cell heterogeneity. Cold Spring Harbor perspectives in medicine, 2, a006429.

2. Goveia, J., Stapor, P. and Carmeliet, P. (2014) Principles of targeting endothelial cell metabolism to treat angiogenesis and endothelial cell dysfunction in disease. EMBO molecular medicine, 6, 1105-1120.

3. Carmeliet, P. and Jain, R.K. (2011) Molecular mechanisms and clinical applications of angiogenesis. Nature, 473, 298-307.

4. Heath, J.R., Ribas, A. and Mischel, P.S. (2016) Single-cell analysis tools for drug discovery and development. Nature reviews. Drug discovery, 15, 204-216.

5. Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M. et al. (2005) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic acids research, 33, D553-555.

6. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic acids research, 35, D760-765.

7. Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. Nature reviews. Genetics, 14, 89-99.

8. Rhodes, D.R., Yu, J.J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004) ONCOMINE: A cancer microarray database and integrated data-mining platform. Neoplasia (New York, N.Y.), 6, 1-6.

9. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R. et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, N.Y.), 310, 644-648.

10. Bos, P.D., Zhang, X.H., Nadal, C., Shu, W., Gomis, R.R., Nguyen, D.X., Minn, A.J., van de Vijver, M.J., Gerald, W.L., Foekens, J.A. et al. (2009) Genes that mediate breast cancer metastasis to the brain. Nature, 459, 1005-1009.

11. Qian, B.Z., Li, J., Zhang, H., Kitamura, T., Zhang, J., Campion, L.R., Kaiser, E.A., Snyder, L.A. and Pollard, J.W. (2011) CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. Nature, 475, 222-225.

12. Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J. et al. (2007) Integrative molecular concept modeling of prostate cancer progression. Nature genetics, 39, 41-51.

13. Erez, N., Truitt, M., Olson, P., Arron, S.T. and Hanahan, D. (2010) Cancer-Associated Fibroblasts Are Activated in Incipient Neoplasia to Orchestrate Tumor-Promoting Inflammation in an NF-kappaB-Dependent Manner. Cancer cell, 17, 135-147.

14. Vanlandewijck, M., He, L., Mae, M.A., Andrae, J., Ando, K., Del Gaudio, F., Nahar, K., Lebouvier, T., Lavina, B., Gouveia, L. et al. (2018) A molecular atlas of cell types and zonation in the brain vasculature. Nature, 554, 475-480.

15. Salvucci, O., Ohnuki, H., Maric, D., Hou, X., Li, X., Yoon, S.O., Segarra, M., Eberhart, C.G., Acker-Palmer, A. and Tosato, G. (2015) EphrinB2 controls vessel pruning through STAT1-JNK3 signalling. Nature communications, 6, 6576.

16. Vandekeere, S., Dubois, C., Kalucka, J., Sullivan, M.R., Garcia-Caballero, M., Goveia, J., Chen, R., Diehl, F.F., Bar-Lev, L., Souffreau, J. et al. (2018) Serine Synthesis via PHGDH Is Essential for Heme Production in Endothelial Cells. Cell metabolism.

17. Bruning, U., Morales-Rodriguez, F., Kalucka, J., Goveia, J., Taverna, F., Queiroz, K.C.S., Dubois, C., Cantelmo, A.R., Chen, R., Loroch, S. et al. (2018) Impairment of Angiogenesis by Fatty Acid Synthase Inhibition Involves mTOR Malonylation. Cell metabolism.

18. Singer, B.D., Mock, J.R., D'Alessio, F.R., Aggarwal, N.R., Mandke, P., Johnston, L. and Damarla, M. (2016) Flow-cytometric method for simultaneous analysis of mouse lung epithelial, endothelial, and hematopoietic lineage cells. American journal of physiology. Lung cellular and molecular physiology, 310, L796-801.

19. Sedlacek, H.H. (2001) Pharmacological aspects of targeting cancer gene therapy to endothelial cells. Critical reviews in oncology/hematology, 37, 169-215.

20. Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwe, H., Pircher, A., Van den Eynde, K. et al. (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. Nature medicine, 24, 1277-1289.

21. van den Brink, S.C., Sage, F., Vertesy, A., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C. and van Oudenaarden, A. (2017) Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nature methods, 14, 935-936.

22. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. and Enard, W. (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. Molecular cell, 65, 631-643 e634.

23. Bacher, R. and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. Genome biology, 17, 63.

24. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.

Bioinformatics (Oxford, England), 19, 185-193.

25. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England), 4, 249-264.

26. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43, e47.

27. Kauffmann, A., Gentleman, R. and Huber, W. (2009) arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics (Oxford, England), 25, 415-416.

28. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34, 525-527.

29. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26, 139-140.

30. Linnarsson, S. and Teichmann, S.A. (2016) Single-cell genomics: coming of age. Genome biology, 17, 97.

31. Lun, A.T., McCarthy, D.J. and Marioni, J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research, 5, 2122.

32. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. Nature biotechnology, 33, 495-502.

33. Ham, K. (2013) OpenRefine (version 2.5). Free, open-source tool for cleaning and transforming data. Journal of the Medical Library Association : JMLA, 101, 233-234.

34. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 3, Article3.

35. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

36. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology, 16, 284-287.

37. Team, R.C. (2018). R Foundation for Statistical Computing, Vienna, Austria.

38. Chang, W.C., J., Allaire, J.J.; Xie, Y.; McPherson, J. (2018) shiny: Web Application Framework for R.

39. Inc., P.T. (2015). Plotly Technologies Inc. , Montréal.

40. Phoenix, T.N., Patmore, D.M., Boop, S., Boulos, N., Jacus, M.O., Patel, Y.T., Roussel, M.F., Finkelstein, D., Goumnerova, L., Perreault, S. et al. (2016) Medulloblastoma Genotype Dictates Blood Brain Barrier Phenotype. Cancer cell, 29, 508-522.

41. Cantelmo, A.R., Conradi, L.C., Brajic, A., Goveia, J., Kalucka, J., Pircher, A., Chaturvedi, P., Hol, J., Thienpont, B., Teuwen, L.A. et al. (2016) Inhibition of the Glycolytic Activator PFKFB3 in Endothelium Induces Tumor Vessel Normalization, Impairs Metastasis, and Improves Chemotherapy. Cancer cell, 30, 968-985.

42. Schoors, S., Bruning, U., Missiaen, R., Queiroz, K.C.S., Borgers, G., Elia, I., Zecchin, A.,

Cantelmo, A.R., Christen, S., Goveia, J. et al. (2015) Fatty acid carbon is essential for dNTP synthesis in endothelial cells. Nature, 520, 192-197.

43. Schoors, S., Cantelmo, A.R., Georgiadou, M., Stapor, P., Wang, X., Quaegebeur, A., Cauwenberghs, S., Wong, B.W., Bifari, F., Decimo, I. et al. (2014) Incomplete and transitory decrease of glycolysis: a new paradigm for anti-angiogenic therapy? Cell cycle (Georgetown, Tex.), 13, 16-22.

44. Rohlenova, K., Veys, K., Miranda-Santos, I., De Bock, K. and Carmeliet, P. (2018) Endothelial Cell Metabolism in Health and Disease. Trends in cell biology, 28, 224-236.

45. Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N.G., Strittmatter, L., Kampf, C., Huang, J., Asplund, A. and Mootha, V.K. (2014) Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. Nature communications, 5, 3128.

46. Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., Vander Heiden, M.G. and Vitkup, D. (2013) Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature biotechnology, 31, 522-529.

47. Zhao, Q., Eichten, A., Parveen, A., Adler, C., Huang, Y., Wang, W., Ding, Y., Adler, A., Nevins, T., Ni, M. et al. (2018) Single-Cell Transcriptome Analyses Reveal Endothelial Cell Heterogeneity in Tumors and Changes following Antiangiogenic Treatment. Cancer research, 78, 2370-2382.

Chapter V EC heterogeneity at single cell level

This chapter is based on a first author research article published in Cancer Cell.
https://www.cell.com/cancer-cell/fulltext/S1535-6108(19)30571-9

This is a supporting manuscript of the thesis and is meant to showcase the application of BIOMEX in analyzing in-house generated high-dimensional single cell data.

AN INTEGRATED APPROACH TO PROFILE LUNG TUMOR ENDOTHELIAL CELL HETEROGENEITY ACROSS SPECIES AND MODELS AND TO IDENTIFY ANGIOGENIC CANDIDATES

Jermaine Goveia[*]*, Katerina Rohlenova*[*]*, Federico Taverna*[*]*, Lucas Treps*[*]*, Lena-Christin Conradi, Andreas Pircher, Vincent Geldhof, Laura de Rooij, Joanna Kalucka, Liliana Sokol, Junbin Qian, Laure-Anne Teuwen, Shawez Khan, Bram Boeckx, Els Wauters, Herbert Decaluwé, Paul De Leyn, Johan Vansteenkiste, Birgit Weynand, Xavier Sagaert, Erik Verbeken, Albert Wolthuis, Baki Topal, Wouter Everaerts, Hanibal Bohnenberger, Alexander Emmert, Dena Panovska, Frederik De Smet, Frank J.T. Staal, Rene J. Mclaughlin, Francis Impens, Stefan Vinckier, Massimiliano Mazzone, Luc Schoonjans, Mieke Dewerchin, Guy Eelen, Tobias K. Karakach, Huanming Yang, Jian Wang, Lars Bolund, Lin Lin, Bernard Thienpont, Xuri Li*[#]*,Diether Lambrechts, Yonglun Luo*[#] *& Peter Carmeliet*[#]

*: equal co-authorship; #: co-corresponding author

Editorial correspondence

Peter Carmeliet: peter.carmeliet@kuleuven.vib.be

Yonglun Luo: alun@biomed.au.dk

Xuri Li: lixr6@mail.sysu.edu.cn

Personal contribution: I implemented the pipeline used to analyze all the single cell (transcriptomics, mass cytometry) and bulk (transcriptomics, proteomics) omics data and developed the back-end and front-end of the Lung ECTax repository resource.

The supplemental figures, tables and STAR methods can be found in the online version of the publication.
https://www.cell.com/cancer-cell/fulltext/S1535-6108(19)30571-9#supplementaryMaterial

Abstract

Heterogeneity of lung tumor endothelial cell (TEC) phenotypes across patients, species (human/mouse) and models (*in vivo*/*vitro*) remains poorly inventoried at the single-cell-level. We single-cell RNA-sequenced 56,771 endothelial cells from human/mouse (peri)-tumoral lung and cultured human lung TECs, detected 17 known and discovered 16 novel phenotypes, including TECs putatively regulating immune surveillance. We resolved the canonical tip TECs into a known migratory tip and a putative basement-membrane remodeling *breach* phenotype. Tip-TEC signatures correlated with patient-survival, and tip/breach TECs were most sensitive to VEGF-blockade. Only tip-TECs were congruent across species/models and shared conserved markers. Integrated analysis of the scRNA-seq data with orthogonal multi-omics and meta-analysis data across different human tumors, validated by functional analysis, identified collagen-modification as an angiogenic candidate pathway.

Significance

Therapeutic targeting of endothelial cells (ECs) to block tumor angiogenesis is clinically approved but suffers resistance. Here we surveyed EC phenotypes in human and murine lung cancer using single cell RNA sequencing in combination with orthogonal bulk multi-omics approaches to provide a molecular atlas of tumor EC (TEC) phenotypes. Classical angiogenic tip and proliferating ECs comprised only a minority of TECs, and we identified distinct subpopulations expressing gene signatures related to basement membrane breaching, immune cell recruitment, and semi-professional antigen presentation. Integrated analysis revealed previously overlooked tip cell markers including transcription factors, matricellular proteins and collagen cross linking enzymes. Functional analysis confirmed novel angiogenic candidates and highlights the potential of our resource to revisit anti-angiogenic strategies.

Introduction

During vessel sprouting in a retinal vascularization model (1), ECs exhibit heterogeneity: a navigating tip EC leads, while proliferating stalk ECs elongate the sprout. Tip and stalk ECs are not genetically predetermined fixed states, but dynamically interchangeable phenotypes (1). In tumors and eye diseases, the vasculature is structurally disorganized and functionally abnormal (2), but the transcriptome heterogeneity of tumor ECs (TECs) at the single cell level across patients, species (human versus mouse) and models (freshly isolated versus cultured), analyzed in a single study, has not been inventoried.

In the tumor angiogenesis field, some scRNA-seq studies reported a descriptive list of previously known TEC phenotypes (3,4), but did not identify previously unknown functionally

validated angiogenic candidates. It can even be questioned if a single scRNA-seq study of one tumor type is sufficiently powerful at all to prioritize angiogenic candidates. Indeed, tip-like TECs in two human colon carcinoma xenograft models in a single study already expressed different markers (4). Here, we aimed to overcome species- and model-dependent variability and used an integrated approach, comprising in addition to scRNA-seq, validation by orthogonal multi-omics approaches, to identify conserved phenotypes and markers across patients, tumor types, species and models, which are more likely to be of functional relevance.

Taxonomy results

Single cell atlas of EC phenotypes in human lung cancer

Focusing on human non-small cell lung cancer (NSCLC), we profiled freshly isolated human tumor ECs (hTECs) and (paired from the same patient) human (peri-tumoral) non-tumor pulmonary ECs (hpNECs) from 1 large cell carcinoma, 4 squamous cell carcinoma and 3 adenocarcinoma treatment-naïve patients (Table S1). Single-cell suspensions, MACS-depleted for $CD45^+$ leukocytes and enriched for $CD31^+$ ECs (Figure 1A), were subjected to scRNA-seq using a 10x Genomics-based single-tube protocol. After quality filtering for the number of detected genes and mitochondrial read counts, duplicates were assessed and unique transcripts were normalized for total read depth (Star Methods; Table S2).

ECs were *in silico* selected (Figure S1A; S1B), and up to 12,323 hTEC and 8,929 hpNEC transcriptomes from 8 patients were pooled, batch-corrected, clustered and visualized using t-SNE plots (Figure 1B,C; Figure S1C; Table S3). To ensure that batch correction did not remove relevant biological features, we also analyzed each sample separately. Clusters detected in batch corrected data were largely similar to those obtained without batch correction (not shown), any relevant differences are indicated below. To assess cluster reproducibility (5), we performed hierarchical clustering and bootstrap analysis (Figure S1D). For biologically relevant subclusters that were not resolved by bootstrapping (e.g. lymphatic ECs from tumor and peritumor tissue), we ascertained that they were statistically separable using pair-wise differential analysis (6) (Figure S1D, Table S4). Each of the 13 phenotypes is numbered as in Figure 1C (H1, H2, etc).

Clusters were biologically annotated based on the relative abundance of top-ranking marker genes in hpNECs and hTECs, and differences in marker gene expression levels shown in display items were statistically quantified to support the qualitative biological annotation (Figure 1D-G; S1E; Table S4, S5).

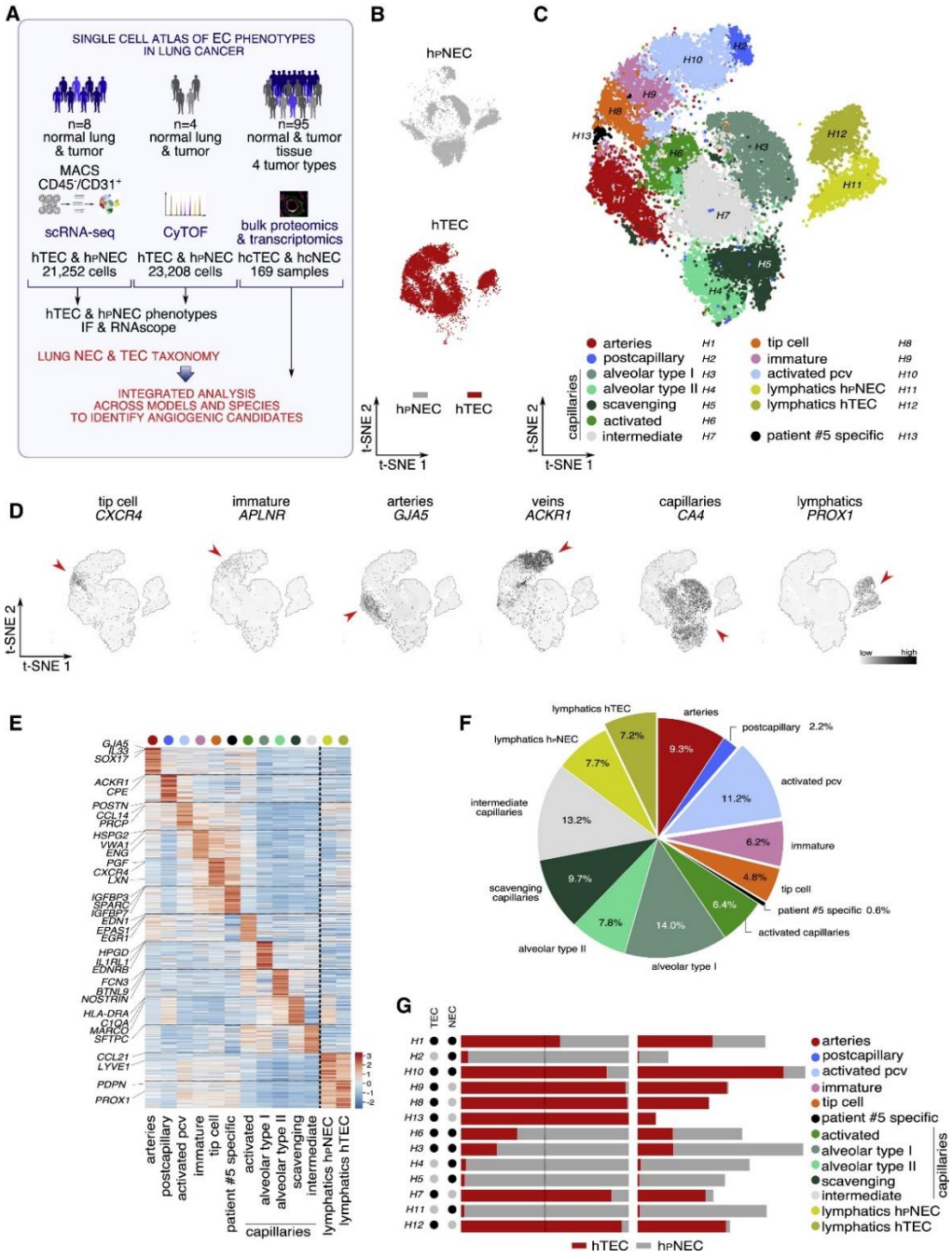

Figure 1: Construction of the hTEC and hpNEC taxonomy.

(**A**) Study design. CyTOF, mass cytometry of time-of-flight; EC, endothelial cell; IF, immunofluorescence; MACS, magnetic-activated cell sorting; NEC, normal EC; TEC, tumor EC. (**B**) t-SNE plot color-coded for ECs from peritumoral non-malignant lung (hpNEC; grey) and tumor tissue (hTEC; red). (**C**) t-SNE plot of hTEC and hpNEC transcriptomes, color- and number-coded for the 13 phenotypes identified by graph-based clustering. (**D**) t-SNE plots, color-coded for expression of indicated marker genes (red arrowheads). (**E**) Gene expression levels of top-ranking marker genes in different EC phenotypes. In this and all further heat maps depicting marker genes, colors represent row-wise scaled gene expression with a mean of 0 and a standard deviation of 1 (z-scores). (**F**) Relative abundance of each phenotype in hTECs and hpNECs. (**G**) Relative contribution of hTECs (red) and hpNEC (grey) to each phenotype. Left: To correct for differences in absolute numbers of hTECs *versus* hpNECs, numbers in each phenotype were divided by the total number of hTECs and hpNECs, respectively. Black vertical line: equal relative contribution (50%). Right: contribution of hTECs and hpNECs in absolute numbers. Note the underrepresentation of capillary hTECs. Dots on the left indicate the relative absence (grey) or presence (black) of the phenotype in hTECs and hpNECs, where presence is defined as a relative contribution of > 1% to the total pool of hTECs and hpNECs, respectively. See also Figure S1 and Figure S2 and Tables S1-S5.

To obtain a more robust interpretation, we focused on clusters defined by gene set signatures rather than single genes (Table S4), and based cluster annotation on the expression of canonical gene signatures (e.g. arterial, capillary, venous, lymphatic, tip and immature stalk phenotypes (listed in Tables S5)). To improve the generalizability of our analysis, we discuss all clusters at least once but focused mainly on clusters that were reproducible in multiple patients (Table S3). A detailed description of previously documented roles of the top ranking markers, used to infer the putative biological role of each cluster, is provided in Table S4 and S5.

Arterial ECs (cluster H1) expressed genes involved in vascular integrity, homeostasis and vasotonus (Figure 2A), while postcapillary venous ECs (H2) expressed genes implicated in leukocyte recruitment, tissue perfusion and pulmonary blood pressure (Figure 2B). We identified a type I (H3) and type II (H4) alveolar capillary EC phenotype, characterized by differential expression of von Willebrand Factor (*vWF*, upregulated in type II compared to type I [fold change = 3.63, adjusted p value = $2.36x10^{-77}$]) and endomucin (*EMCN*, upregulated in type I compared to type II [fold change = 2.89, adjusted p value = $1.22x10^{-68}$]), which might be involved in respectively vasoregulation and anti-microbial defense (Figure 2C-E). Compared to all other clusters, capillary ECs expressed a signature of genes involved in MHC-II mediated antigen presentation and processing at higher levels than other phenotypes, but the co-stimulatory molecules *CD80* and *CD86* were not detectable (Figure 2F), suggesting a function as semi-professional antigen presentation cells (APCs). We also identified 2 novel capillary phenotypes that might be induced by tumor-derived cytokines: scavenging capillaries (H5) were by bootstrap analysis similar to type II capillary ECs, but upregulated scavenging receptors and genes associated with macrophages and antigen processing (Figure 2G,H; S1D), while activated capillaries (H6) expressed EC activation markers (Figure 2I). An intermediate capillary EC phenotype (H7) resembled activated capillaries, but consisted mostly of cells derived from a single patient (Figure S1D; Table S3).

Traditional angiogenic phenotypes such as tip and proliferating ECs (presumed targets of anti-angiogenic therapy (AAT)) were detectable only in hTECs (Figure 3A, S1E, S2A; Table S3). Tip cells (H8), expressing gene signatures associated with EC migration, matrix remodeling and VEGF signaling (Figure 3B), comprised only a minority (< 10%) of hTECs (Figure S1E; Table S3). Notably, tip hTECs showed highly restricted expression of the disease-specific molecule *PGF* (placental growth factor; PlGF [adjusted p value = $2.41x10^{-14}$]) (Table S4). We also identified an immature hTEC phenotype (H9), which was similar to tip cells but upregulated genes involved in the maturation of newly-formed vessels, vessel barrier integrity and Notch-signaling, possibly resembling stalk-like ECs (Figure 3C; S1D).

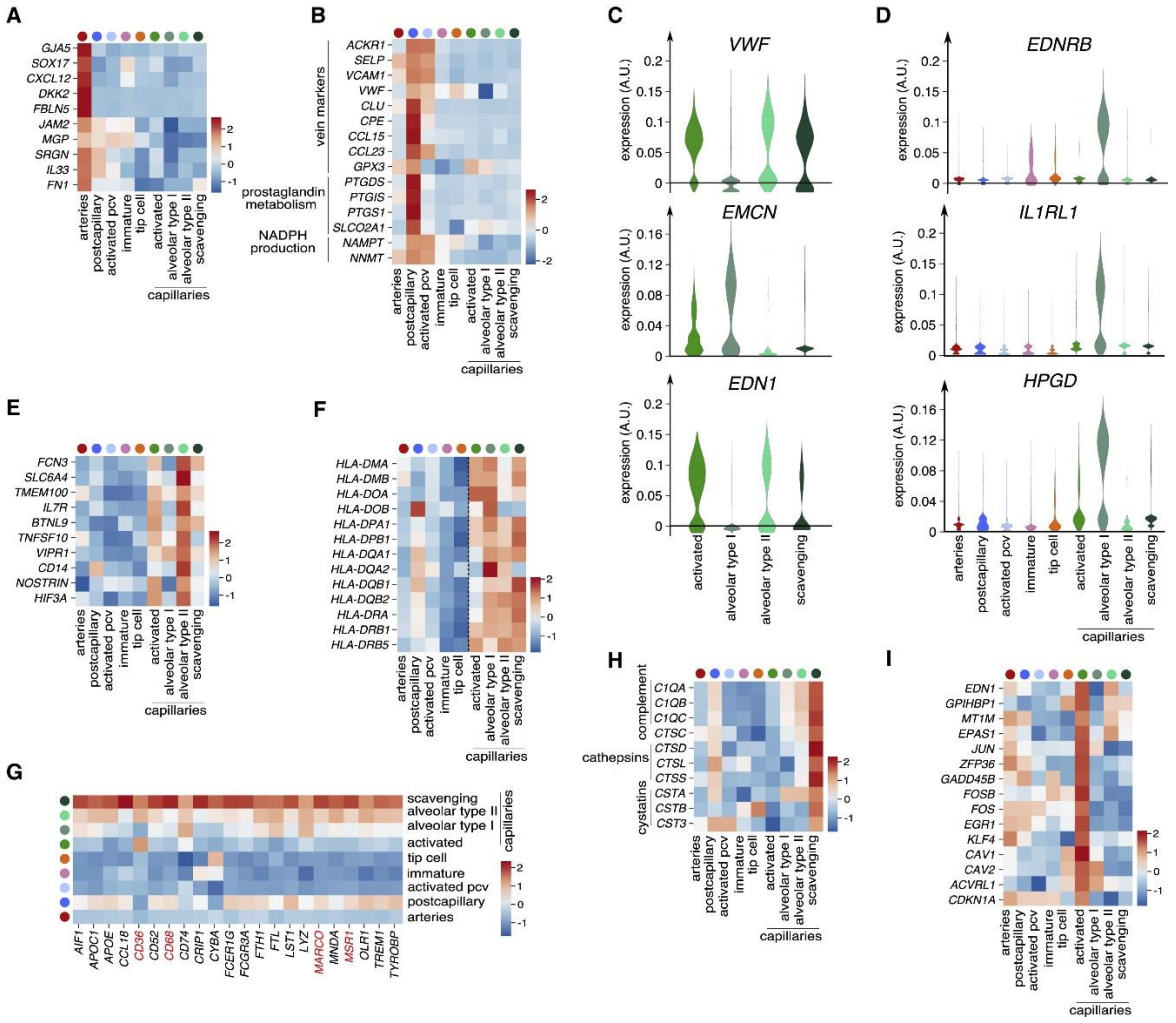

Figure 2: Quiescent and activated normal pulmonary EC phenotypes.

(**A,B**) Expression levels of arterial (A) and postcapillary venous (B) EC markers. (**C,D**) Expression levels of the indicated genes of capillary EC phenotypes (C) and the type I alveolar capillary EC phenotype (D). A.U., arbitrary units. (**E-I**) Expression levels of markers of type II alveolar capillary ECs (E), of genes involved in MHC II antigen presentation (F), of scavenging EC markers (G, with scavenging receptors in red), of defense marker genes (complement, cathepsins and cystatins) (H), and of genes associated with activated alveolar capillary ECs (I). See also Figure S1-S3 and Tables S4-S5.

Proliferating hTECs were only detected in a subset of patients, representing <1% of hTECs, and were no longer detected as separate cluster after batch correction (Figure S2A). In hTECs, an activated postcapillary vein phenotype (H10) upregulated immunomodulatory factors and ribosomal proteins (Figure 3D), features of high endothelial venules (HEVs) in inflamed tissues (7). In contrast to the extensive phenotypic heterogeneity of blood vascular ECs, the gene expression signatures of tumor and peritumoral lymphatic ECs (LEC; H11, H12) were highly similar and LEC subpopulations were not detected (Figure 3E; S1D; Table S4). An ambiguous phenotype, expressing markers of tip and arterial ECs (Figure 1C; S1D), predominantly consisted of cells from a single patient (patient 5, H13) and is not discussed.

To explore the generalizability of our NSCLC EC taxonomy, we used top-ranking

marker genes to train a machine learning algorithm (8) to automatically annotate 574 hpNECs and 638 hTECs, *in silico* selected from a 52,698-cell catalogue from tumors of 5 treatment-naïve NSCLC patients (3). The majority of ECs were annotated with high confidence (similarity threshold > 0.5) (Figure 3F), indicating that our taxonomy is an externally valid resource that comprises all major EC phenotypes detectable in 13 lung cancer patients.

We validated the taxonomy using orthogonal techniques. Immunostaining combined with quantitative RNAscope to count transcript numbers confirmed that *PGF* levels were upregulated in $CD31^{+}$/$CXCR4^{high}$ tip hTECs, while triple immunostaining showed co-localization of CXCR4 and PlGF in TECs (Figure 3G,H; S2B). We also confirmed by immunostaining of: (i) SELP the signature of activated postcapillary vein ECs (Figure S2C,D); (ii) VWF and EFNB2 respectively the venous and arterial phenotype (Figures S2E-H); and (iii) the HEV-specific MECA-79 antigen the signature of the HEV phenotype (Figures S2I). RNAscope, combined with staining for the EC marker CD31, revealed that arterial (*EFNB2*) and venous (*ACKR1*) marker transcripts did not colocalize in the same hTECs (Figure S3A,B), while activated postcapillary vein hTECs expressed the marker CCL14 (Figure S3C). Further, transcripts of the scavenging capillary markers CD52 and CD68 , and of the type I capillary markers EDNRB and IL1RL1 co-localized in the same hTECs (Figure S3D,E).

We used CyTOF to quantify protein levels of marker genes in single cancer, EC and stromal cells, freshly isolated from NSCLC samples, using metal-conjugated antibodies against 26 preselected markers of these cells. Unbiased clustering and t-SNE visualization of the CyTOF single cell data revealed separate clusters of ECs, stromal and cancer cells (Figure 3I; S4A-D). Consistent with scRNA-seq, we detected type I (vWF^{low}) and type II (vWF^{high}) alveolar capillary ECs, postcapillary vein ECs ($ACKR1^{high}$/vWF^{high}) and lymphatic hpNEC and hTEC phenotypes ($Prox1^{high}$) (Figure S4E). A phenotype expressing the highest CXCR4 levels (tip cell marker) but downregulating the expression of capillary markers (CD36, CA4, HLA II proteins) likely represented angiogenic hTECs. We also detected a poorly resolved cluster, putative activated capillaries that, compared to type I and type II capillaries, expressed reduced capillary markers (CA4, CD36, HLA-II), but increased levels of the vein TEC marker VWF and the tip hTEC marker CXCR4 (Figure 3J; S4E). CyTOF confirmed the increased levels of HLA and CD36 in capillary ECs and their downregulation in hTECs (Figure 3J; S4F). To validate the observation that the populations identified using scRNA-seq and CyTOF co-incided, we performed hierarchical clustering using key surface markers (Methods) (Figure 3K). LECs and capillary ECs clustered together, while vein ECs clustered with angiogenic ECs. Notably, multiscale bootstrap analysis revealed that scRNA-seq resolved type I and type II capillary ECs as most similar to their CyTOF counterparts, cross-validating the presence of

two distinct capillary phenotypes at the mRNA and protein level.

Possible clinical implications

Despite all patients having early stage treatment-naïve disease, we observed inter-patient heterogeneity in the relative fraction of the different EC phenotypes (Figure 4A; Table S3). In hpNECs, arterial, capillary, venous and lymphatic ECs were detected at variable proportions, especially for the capillary hpNEC phenotypes. In hTECs, the number of capillary ECs was reduced (Table S3), a finding confirmed by immunohistochemistry (Figure 4B), though not to the same extent in each patient. Statistically, immature hTECs were more abundant in adenocarcinoma than squamous cell carcinoma patients (Table S3), though larger cohorts are required for meaningful interpretation (see below).

To correlate expression of EC phenotype-specific gene signatures with NSCLC patient survival, we took advantage of a large bulk RNA-sequencing dataset of 1,024 NSCLC patients of the Cancer Genome Atlas (TCGA). After identifying EC-specific marker gene sets for each phenotype utilizing a publicly available resource (3) (Methods), we used gene set variation analysis (GSVA) to score the enrichment of these signatures in each of the 1,024 NSCLC patients. Notably, squamous (but not adenocarcinoma (not shown)) NSCLC patients, who expressed high levels of gene set signatures of angiogenic tip, immature, activated postcapillary or lymphatic TECs, had shorter overall survival, presumably because these signatures reflect active angiogenesis and lymphatic spread (Figure 4C).

Lung tumor EC phenotypes in other species and models

To compare lung TEC taxonomies across species and models for identification of common angiogenic TEC phenotypes and targets in an integrated congruency analysis, we used a similar strategy to construct two additional lung TEC taxonomies, one of 29,007 murine mNECs and mTECs microdissected from Lewis lung carcinoma (LLC) (another species), treated or not with anti-VEGF, and another of 6,512 human cultured hcTECs from human lung tumor (another model) (Table S2-4). We used the LLC model, since it relies on angiogenic vessel sprouting (9) (the target of AAT), unlike other preclinical mouse NSCLC tumor models (9).

As in human lung, we identified mNEC-enriched phenotypes, expressing typical markers of arterial (M1), alveolar type I and II capillary (M2, M3), venous (M4) and lymphatic (M5) ECs (Figure 5A-D; S4G-I; S5A-D; Table S4; S6). Type II alveolar capillary ECs expressed the highest levels of genes involved in MHC-II antigen presentation, processing and loading, but did not express the co-stimulatory genes *Cd80* and *Cd86* (Figure S4H; S5E).

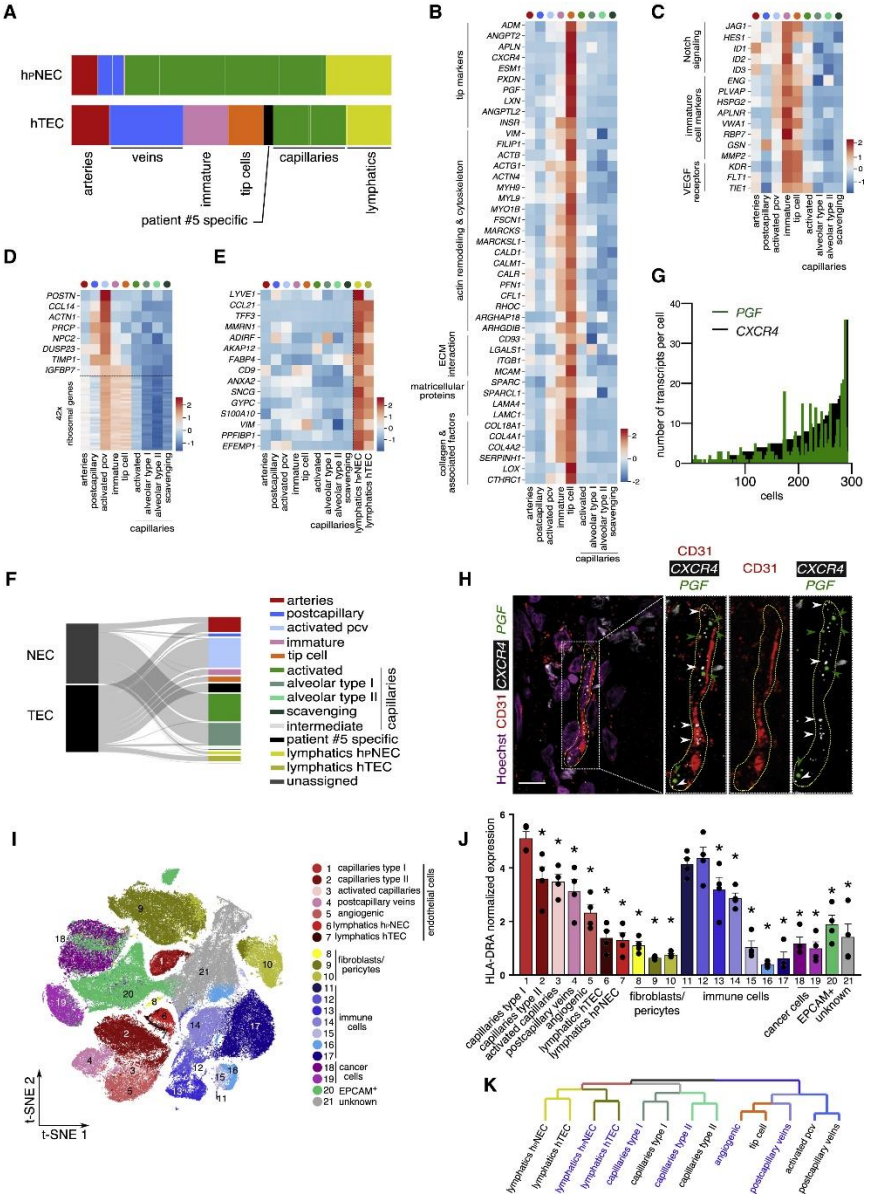

Figure 3: Pathological tumor EC phenotypes and validation by CyTOF.

(**A**) Relative abundance of EC phenotypes in hTECs and hpNECs, weighted by the number of cells per patient. (**B-E**) Expression levels of marker genes of tip (B), immature (C), activated postcapillary vein (D), and lymphatic ECs (E). (**F**) Assignment of hTEC and hpNEC phenotypes (identified in the taxonomy) to hTECs and hpNECs freshly isolated from an independent cohort of 5 NSCLC patients (publicly available data (3)). (**G**) Number of *CXCR4* and *PGF* mRNA transcripts per cell in NSCLC tumor vessels as determined by RNAscope (n=3 patients). *CXCR4* and *PGF* positively correlate (Spearman r = 0.584; two-sided p value < 0.0001). (**H**) RNAscope images of a NSCLC tumor vessel probed for *CXCR4* (white) and *PGF* (green) mRNA. ECs are immunostained for CD31 (red). Nuclei are stained with Hoechst (magenta). Dotted lines denote (part of) the vessel; boxed area is magnified on the right. Green and white arrowheads indicate *PGF* and *CXCR4* mRNA transcripts, respectively. Scale bar, 20 μm. (**I**) t-SNE plot of CyTOF data, color-coded for stromal and cancer cell phenotypes. For visualization, all non-EC types have been randomly down-sampled to no more than the total number of ECs. (**J**) Quantification of HLA II expression by CyTOF in stromal cell subpopulations. Asterisks indicate clusters in which HLA II expression is significantly downregulated compared to type I capillary ECs (mean ± SEM; n=4, p value < 0.05 by one-way ANOVA with Dunnett's method). (**K**) Dendrogram visualization of hierarchical clustering analysis of EC phenotypes identified by scRNA-seq and CyTOF analysis. Color differences in dendrogram indicate clusters that were resolved by multiscale bootstrap analysis. Clusters identified by CyTOF are indicated in blue. See also Figures S1-S4 and Table S4-S5.

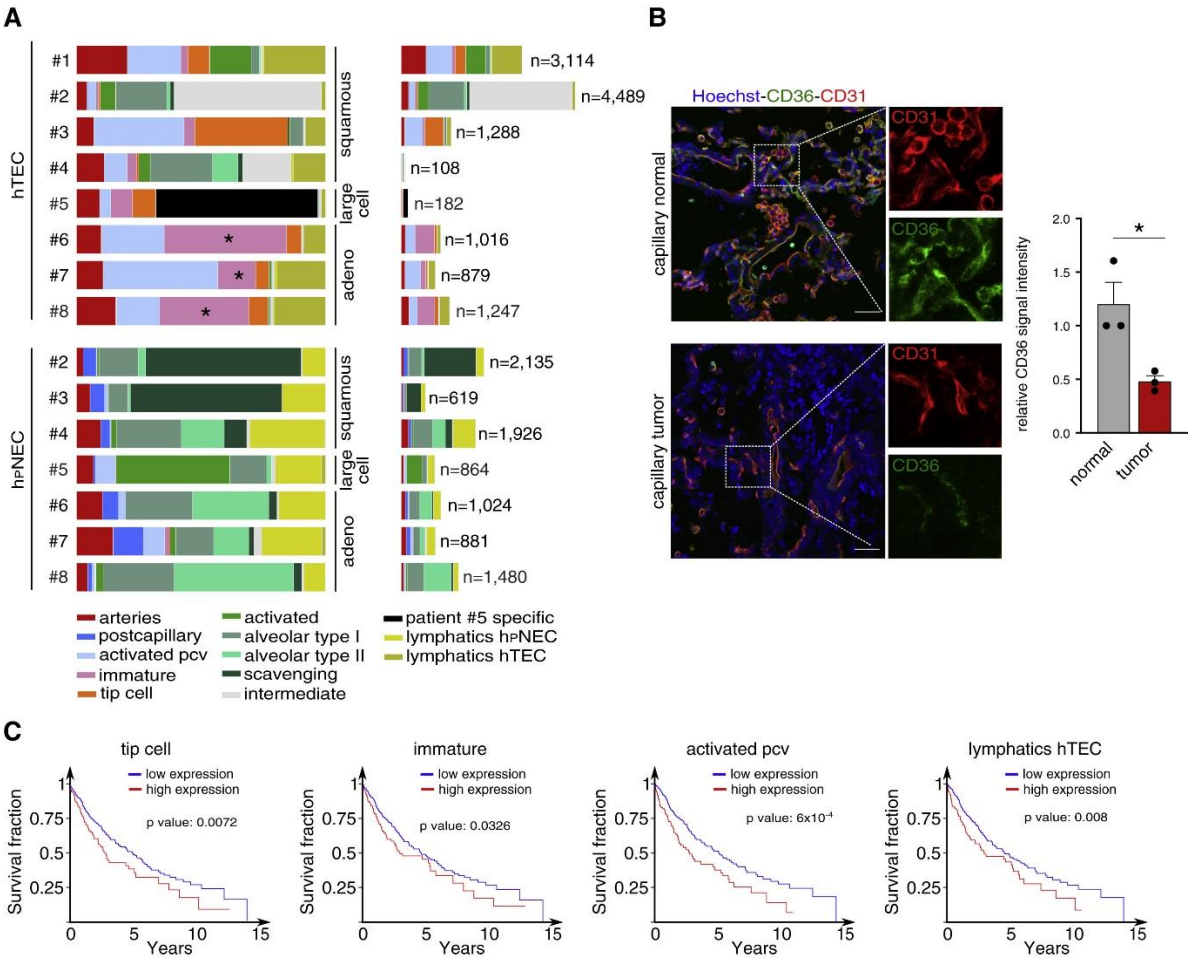

Figure 4: Translational and possible therapeutic implications of the NSCLC EC taxonomy.

(**A**) EC phenotype composition in tumor and normal tissue from individual patients. Left: Relative contribution of each phenotype scaled to 100%. Right: Contribution of each phenotype in absolute numbers; total number of analyzed ECs on the right. Asterisks indicate the significantly increased fraction of immature TECs in adenocarcinoma compared to squamous cell carcinoma patients (n=4 squamous and n=3 adenocarcinoma patients, p value < 0.05 by two-tailed unpaired *t*-test). (**B**) Representative micrograph of human peritumoral tissue (upper) and NSCLC tumor (lower) section immunostained for CD31 and CD36. Nuclei are stained with Hoechst. Images to the right are magnifications of the respective boxed areas. Scale bar, 50 μm. Right: Quantification of the CD36 signal intensity (mean ± SEM; n=3 patient sections, p value < 0.05 by two-tailed unpaired *t*-test). (**C**) Overall survival of 502 lung squamous cell carcinoma patients selected from the TCGA dataset and stratified by the indicated gene set expression signature scores. See also Table S1 and Table S3.

As in NSCLC, the fraction of capillary ECs was underrepresented in tumors (Figure S4G; Table S3), and mTECs downregulated the typical capillary gene signature, including *MHC-II* expression (Figure S5E). mTECs also downregulated *Cd36* (a gene involved in fatty acid uptake), except in one capillary TEC phenotype ([adjusted p value = 7.81×10^{-39}]; M6) (Figure 5D). Unlike in human lung, the vein EC phenotype in murine lung expressed resident endothelial stem cell (ESC) markers (*Cd200* [adjusted p value = 4.61×10^{-26}], *Bst1* [adjusted p value = 2.80×10^{-20}]), previously identified in large pulmonary vessels (10) (Figure S5F).

Proliferating (M7) and tip (M8) mTECs were also detected (Figure 5C,E); proliferating mTECs were more abundant than in human lung tumors, consistent with the faster growth of

murine lung tumors and a possible different type of tumor vascularization in NSCLC (Figure 5D; S5B). An immature mTEC phenotype (M9), lacking a strong transcriptome signature, was also recognized (Figure 5C). Consistent with human ECs, we observed high venous plasticity and detected TEC-enriched phenotypes including large (M10) and postcapillary (M11) vein ECs (Figure 5C). The postcapillary vein mTECs upregulated high endothelial venule (HEV) markers (Figure S5G,H) confirmed by immunostaining (Figure S6A,B).

The remaining murine phenotypes were predominantly detected in mTECs (Figure 5C; S5C; Table S3). A neo-phalanx mTEC phenotype (M12) expressed capillary and arteriole markers, while activated artery mTECs (M13) upregulated neo-arteriogenesis markers. Notably, only mTECs, not mNECs, exhibited a phenotype characterized by the expression of an interferon (IFN) response gene signature and chemokines, involved in immune cell recruitment and angiostasis (M14) (Figure S4I; Table S3, S4). We also identified a previously unrecognized EC phenotype, which we termed breach cells (M15), and their putative precursors, pre-breach cells (M16) (Figure 5E-G). Breach cells upregulated expression of not only tip TEC markers, but also of genes previously involved in VEGF-induced podosome rosette-mediated basement membrane (BM) and collagen remodeling (Table S4; S6) (11). Immunostaining confirmed the expression of signature markers of arterial, capillary, venous and tip mNEC and mTEC phenotypes at the protein level (Figure S6).

In cultured hcTECs, the in vivo artery, capillary and vein EC phenotypes were no longer detectable (Figure S7A,B; Table S2-S4). It was therefore surprising that the typical tip EC phenotype (C1), which is a plastic transient phenotype (1), was detectable in hcTECs. As expected for propagating ECs in culture, proliferating hcTECs (C2) were detected. Perhaps related to the culture conditions (presence of TGFß, an inducer of endothelial-to-fibroblast transition in serum), a phenotype (C3), exhibiting a signature of endothelial-to-mesenchymal transition (*TAGLN, SERPINE1, FN1, CD44* [adjusted p value < $5x10^{-4}$ for all genes]) (12) was identified. We also detected an intermediate transitioning phenotype (C4) that upregulated ribosomal genes, suggesting EC phenotypes on their way to adopting another phenotype (Figure S7B).

Effect of VEGF blockade on TEC phenotypes

Differential sensitivity of TEC phenotypes

Using the mTEC taxonomy we explored if specific EC phenotypes were differentially sensitive to anti-VEGF AAT (VEGFR2 antibody (DC101); VEGFR tyrosine kinase inhibitor (PTK787)) (Figure 5H). These compounds were tested at doses that inhibit pathological angiogenesis and reduce EC numbers in tumors (13), and inhibited tumor growth (Figure S7C). Notably,

control- and AAT-treated ECs were comprised of the same clusters (Figure 5H; Table S3), suggesting that AAT did not alter the global transcriptome signatures of the different EC phenotypes. However, quantification of individual EC phenotypes revealed that tip and breach TECs were most sensitive, consistent with human colon carcinoma xenografts showing tip cell sensitivity to aflibercept (4). Postcapillary vein and proliferating TECs were less sensitive to VEGF blockade, while capillary TECs were less sensitive to PTK787 treatment (Figure 5H). Whether this is due to the tumor switching from vessel sprouting to vessel cooption is unknown.

Molecular signature of tumor vessel disorganization and effect of VEGF blockade

Tumor vessels are structurally disorganized and functionally abnormal (14-16), but a detailed unbiased molecular footprint is lacking. Traditionally, AAT is considered to *prune* angiogenic ECs, though VEGF-blockade can also normalize the abnormal tumor vasculature ("tumor vessel normalization") (15). We explored if VEGF blockade tuned ECs by inducing more subtle gene expression changes. We therefore first constructed "normal" and "tumor vessel disorganization" gene signatures by comparing angiogenic versus non-angiogenic phenotypes in a pair-wise differential analysis to identify genes that were significantly (adjusted p value < 0.01) differentially expressed from an effective fold change threshold of ± 1.5. This revealed 18 and 28 genes that were significantly higher expressed in disorganized (among which tip, breach and immature EC marker genes) and normal (including capillary marker genes) phenotypes, respectively (Figure 5I,J; S7D and Table S4).

GSVA using the 18 disorganization gene signature revealed that tumor vessel disorganization was partially reversed by VEGF blockade (Figure 5I,J). Gene set enrichment analysis further showed that both VEGFR inhibitors reduced the expression of tip EC gene signatures (glycolysis (17), BM remodeling), while increasing the expression of signatures associated with mature homeostatic functions (antigen presentation, barrier and blood vessel development) (Figure S7E,F). Thus, VEGF blockade not only pruned but also tuned TECs to promote a more quiescent and mature tumor vasculature with homeostatic functions, suggesting partial molecular tumor vessel normalization.

Identification of conserved TEC phenotypes and tip cell markers

Anti-VEGF therapy efficacy is limited by resistance due to upregulation of alternative pro-angiogenic signals (18). To identify alternative angiogenic candidates, we hypothesized that TEC phenotypes, conserved across species and models, and congruently overexpressed genes might be stronger and more robust angiogenic candidates. We thus performed an integrated combined single cell transcriptomics, bulk multi-omics and transcriptomics meta-

analysis to identify such candidates (Figure S7G).

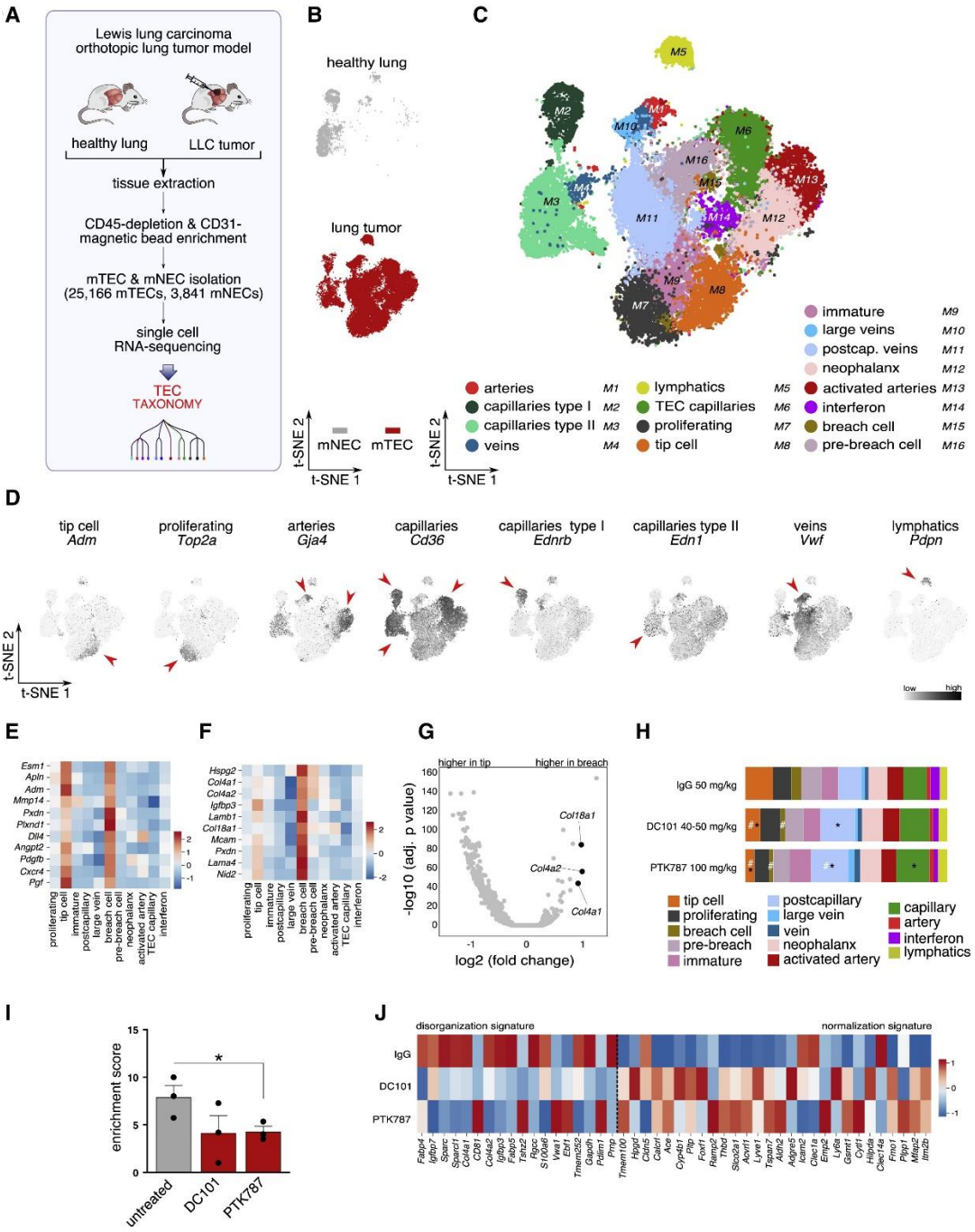

Figure 5: Construction of the mTEC and mNEC taxonomy.

(**A**) Experimental design. (**B**) t-SNE plot, color-coded for mNECs from healthy lung (grey) and mTECs from tumor-bearing mice (red). (**C**) t-SNE plot, color-coded for the pulmonary mNEC and mTEC phenotypes. (**D**) t-SNE plots of mNECs and mTECs, color-coded for the expression of indicated marker genes (red arrowheads). (**E,F**) Expression levels of tip (E) and breach mTEC marker genes (F). (**G**) Differentially expressed genes in breach cells *vs* tip cells. (**H**) Relative composition of mTEC phenotypes upon treatment with control IgG, DC101 or PTK787. *p value < 0.05 by Dunnett's method; #p value < 0.05 by two-tailed unpaired *t*-test. (**I**) Expression level of the 18-gene disorganization signature (mean ± SEM; n=3, *p value < 0.05 by one-tailed unpaired *t*-test) (I). (**J**) Expression levels of the genes in the disorganization signature (18 genes on the left of the black dotted line in the control (IgG) condition (top row)) and more normal signature (28 genes on the right of the black dotted line in the control (IgG) condition (top row)), and effects of DC101- and PTK787-treatments on these signatures (middle and bottom rows). See also Figures S4-S7 and Tables S2-S4 and S6.

We used the scRNA-seq data to assess similarity of angiogenic EC phenotypes across species and models using pairwise Jaccard similarity coefficients of marker gene sets and applied PCA for visualization (Figure 6A; S7G). The in vivo artery, capillary and vein phenotypes were lost in hcTECs in vitro; in contrast, the tip and breach EC transcriptome signature of in vivo hTECs and mTECs was conserved in cultured hcTECs. ECs with a proliferation signature were only detected in cultured and murine TECs, and in hTECs from some NSCLC patients (see above). Freshly-isolated human vein and murine immature TECs, and intermediate hcTECs highly expressed ribosomal genes, suggesting plastic phenotypes, transitioning to other angiogenic EC phenotypes. All other phenotypes were model- or species-specific. Since tip TECs expressed a similar marker gene signature across all species and models tested, we focused on identifying congruent tip TEC markers.

To construct a ranked list of tip TEC markers, we first selected 296 conserved tip cell-enriched genes that were consistently highest expressed in tip TECs across species and models. Second, for each conserved gene, we defined a rank score by calculating the product of the marker gene rank in each of the 3 datasets (in this list, genes that are consistently among the most upregulated genes in tip cells rank the highest) (Figure 6B). For subsequent analyses, we focused on the top-ranking 50 enriched markers (adjusted p value < $5x10^{-3}$). Validating our approach, we confirmed the inclusion of tip EC markers (*ANGPT2*, *APLN*, *FSCN1*, *PGF*, *PLXND1*, *ADM*, *PDGFB*, *CXCR4*, others), previously detected in tip ECs but not necessarily further characterized at the expression pattern and functional level (4,19,20). Half of the 50 top-ranking genes were not previously described as tip TEC markers (not detected in transcriptomics studies of ECs in tumor or physiological angiogenesis). Congruent tip TEC markers were associated with the migratory tip EC phenotype, including laminins (*LAMA4*, *LAMC1*, *LAMB1*), matricellular proteins (*SPARC*, *LXN*), cytoskeleton-associated genes (*VIM*, *MARCKS*, *MYH9*, *MYO1B*) and cell adhesion molecules (*CD93*, *MCAM*, *ITGA5*). We also identified novel tip TEC-enriched transcription factors (*TCF4*, *SOX4*, *SMAD1*).

The disorganized vascular architecture in tumors impedes/precludes topographical identification of tip TECs: Figure S6J shows heterogeneous expression of a tip cell marker, but its position at the forefront of the tumor vessel sprout and its expression in morphological tip cells cannot be unambiguously identified. We therefore used the postnatal retinal angiogenesis model to validate the expression of LXN (Latexin) and FSCN1 (Fascin) in tip ECs at the vascular front (Figure S8A). To document their functional role, we examined if silencing these tip cell markers affected tip cell competitivity (17). We silenced LXN or FSCN1 expression in human umbilical vein ECs (HUVECs; >70-80% silencing efficiency; Figure S8B) and generated mosaic spheroids containing a 1:1 mixture of control wild-type HUVECs (red)

and HUVECs silenced for LXN or FSCN1 (green). Silenced cells were less often at the tip position, confirming the tip cell role of these markers (Figure 6C,D).

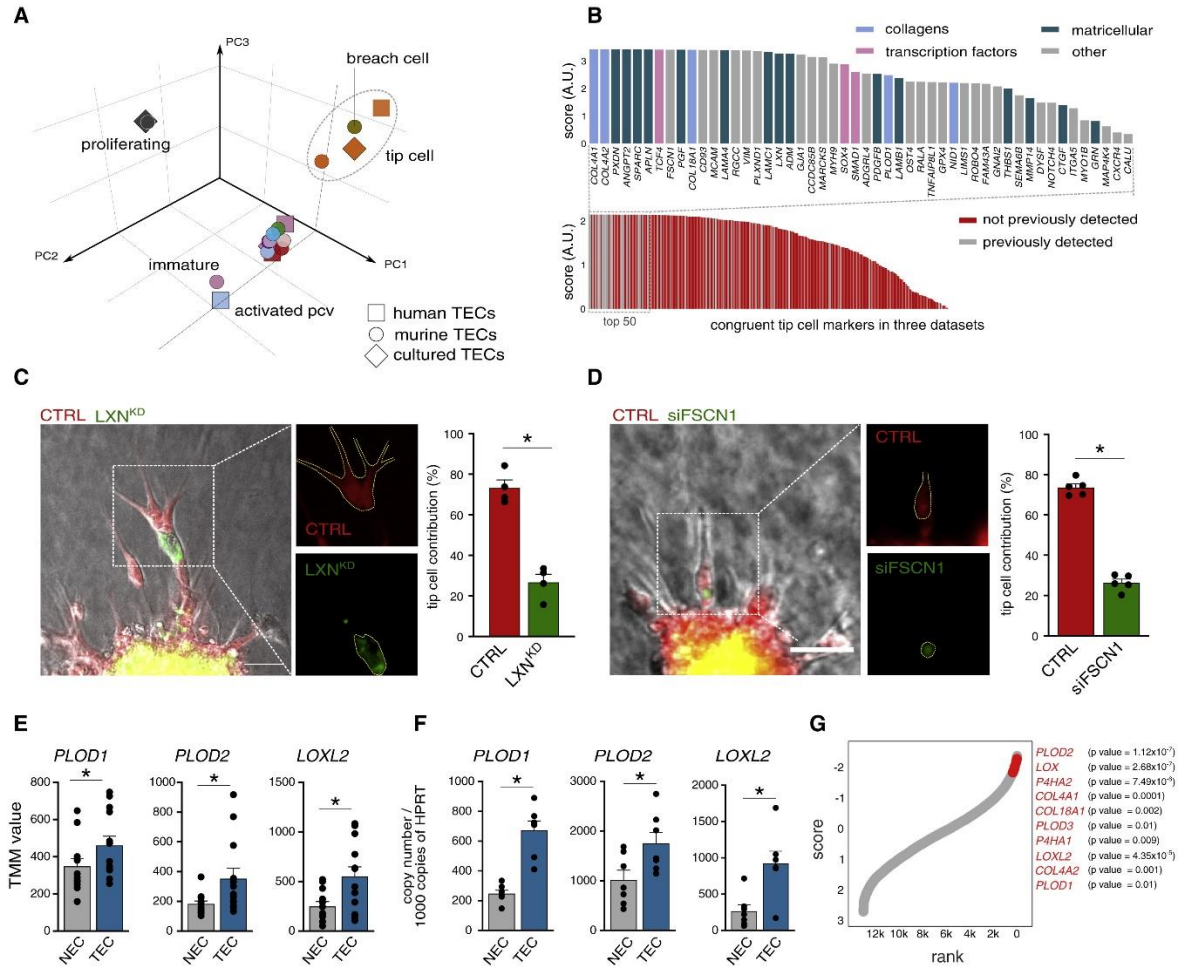

Figure 6: Congruency analysis and transcriptomics meta-analysis.

(**A**) Principal component analysis (PCA) on the pair-wise Jaccard similarity coefficients between the top 50 marker genes of freshly isolated human and murine, and cultured human TEC phenotypes. (**B**) Bar plot visualization depicting the congruency score of conserved tip cell marker genes and their intersection with benchmark data sets (bottom panel) and the 50 top-ranking marker genes categorized by the indicated biological processes (top panel). A.U., arbitrary units. (**C,D**) Mosaic EC spheroid competition of CTR^{red} ECs and LXN$^{KD/green}$ (C; KD denotes shRNA knockdown) or siFSCN1^{green} ECs (D; si denotes siRNA silencing). Left: Representative micrographs of mosaic EC spheroid. Scale bar, 50 μm. Right: Quantification of the fraction of tip cells with the indicated genotype. Data are mean ± SEM; n=4 for LXN , n=5 for FSCN1; *p value < 0.05 by unpaired two-tailed *t*-test. (**E,F**) Expression levels of genes encoding collagen hydroxylation and cross-linking enzymes measured by bulk RNA-sequencing (E) or qRT-PCR (F). Data are mean ± SEM; n=13 (D) and n=7 (E); *p value < 0.05 by paired two-tailed *t*-test. TMM, trimmed mean of *M*-values. (**G**) Gene expression meta-analysis of 6 publicly available NEC *versus* TEC datasets from 5 different tumor types. The S-curve has 12,000 dots, representing genes that were detected in all 6 datasets. X-axis: rank numbers from 1 to 12,000 (consistently overexpressed genes in TECs have a low rank number, consistently downregulated genes have a high rank number); Y-axis: the scaled meta-analysis score (consistently overexpressed genes in TECs have a low meta-analysis score, consistently downregulated genes have a high meta-analysis score). Genes involved in collagen modification or encoding collagens are shown as red dots and listed on the right (p value < 0.01 by rank product analysis). See also Figure S7 and Figure S8 and Table S3 and Table S4.

Integrated analysis highlights collagen modifications as an angiogenic pathway

Genes encoding collagens (*COL4A1*, *COL4A2*, *COL18A1*) ranked in the top 10 most congruent tip cell markers, while other collagen modifying (cross-linking) enzymes (*PXDN*, *PLOD1*) ranked in the top 50 (Figure 6B). Bulk RNA-seq analysis of 13 independent NSCLC patients (Figure S7G; Table S3) and confirmatory RT-PCR analysis showed that the expression of genes involved in collagen cross-linking (*LOXL2*, fold change = 2.5; p value = $5.15x10^{-3}$) and hydroxylation (*PLOD1-3*, fold change 1.24-1.74; p value < 0.05 for all genes) was higher in TECs than NECs (Figure 6E,F). Procollagen-lysine, 2-oxoglutarate 5-dioxygenase (PLOD) isoenzymes (known as lysyl hydroxylases) intracellularly hydroxylate lysine residues in collagen, while lysyl oxidase (LOX) or its homologue LOXL2 extracellularly catalyze the first step in collagen crosslink formation (21). To explore upregulation of collagen modifying enzymes in TECs from other tumor types, we performed a meta-analysis of 6 publicly available NEC *versus* TEC bulk transcriptomics datasets, freshly isolated from patients with 5 different tumor types (Table S4). Also in this analysis, transcripts encoding collagen modifying enzymes were enriched (p value < 0.05 for all genes) and ranked among the top 1% to 5% most consistently upregulated genes in TECs (Figure 6G).

To explore upregulation of collagen modifying enzymes in TECs at the protein level, we performed proteomics analysis on an in-house generated cohort of 144 prospectively collected TEC and NEC samples from lung, kidney and colorectal (CRC) tumors and colorectal liver metastasis (CRCLM) (Figure S7G; Table S3). In TECs from NSCLC patients (n=27 patients), LOXL2 was the highest upregulated protein (>4-fold; p value = $6.89x10^{-7}$), while PLOD1 and PLOD2 were upregulated 1.7-(p value = $5.52x10^{-5}$) and 1.9-fold (p value = $4.39x10^{-4}$), respectively (Figure 7A). Meta-analysis across all four tumor types identified 288 consistently higher expressed proteins, including upregulated LOX (p value = $7.88x10^{-3}$), LOXL2 (p value = $1.21x10^{-6}$), PLOD1 (p value = $7.19x10^{-3}$) and PLOD2 (p value = $4.52x10^{-5}$) (Figure 7B; Table S4). Gene ontology (GO) analysis on the subset of consistently higher expressed proteins revealed an enrichment of glycolytic enzymes (p value = $4.98x10^{-4}$), growth (p value = $4.80x10^{-4}$) and actin cytoskeleton organization (p value = $1.16x10^{-4}$), as well as protein hydroxylation and cross-linking (p value = $7.55x10^{-4}$) (Figure 7C; S8C; Table S7). We consistently identified PLOD1, PLOD2 and LOXL2 and protein hydroxylation as highly ranked targets, and thus assessed if these genes functionally regulate vessel sprouting.

In agreement with the above GO analysis, migration, proliferation and vessel sprouting were higher in TECs than NECs (Figure 8A-C).

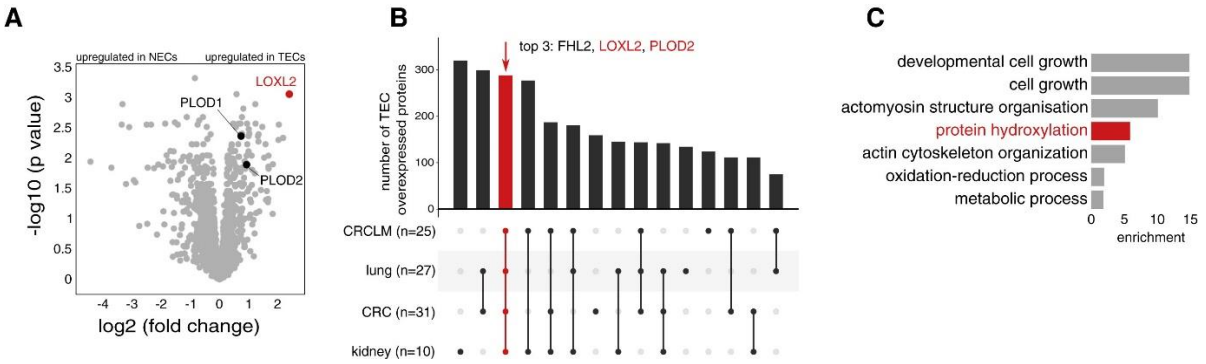

Figure 7: Proteomics and Gene Ontology analysis.

(**A**) Volcano plot of differentially expressed proteins in TECs compared to NECs. (**B**) Upset plot visualization of the results of a cultured hcNEC *versus* hcTEC differential protein expression meta-analysis of 144 bulk proteomics samples (n=93 patients), showing the number of proteins that were higher expressed in TECs than NECs isolated from the indicated tumor type. The bar graph represents the number of proteins, detected in the tumor type(s) indicated by the dot plot panel below. 288 proteins were consistently higher expressed in TECs than NECs in all 4 tumor types (red bar graph and intersection). LOXL2 and PLOD2 were ranked in the top 3. (**C**) Bar graph representation of significantly enriched gene ontology terms (p value < 10^{-3}). See also Figure S7 and S8; Table S3, S4 and S7.

Silencing of PLOD1, PLOD2 or LOXL2 in HUVECs (>70-80% silencing efficiency (Figure S8D)) reduced EC migration (Figure 8D) and impaired vessel sprouting in HUVEC spheroids (Figure 8E,F). Minoxidil, a pharmacological PLOD blocker (22), reduced EC migration in vitro (Figure 8G). Administration of minoxidil or the lysyl oxidase inhibitor beta-aminopropionitrile (BAPN) (23) inhibited corneal angiogenesis in vivo (Figure 8H). We cannot also not exclude the possibility that this blocker impairs corneal angiogenesis through indirect effects on other cell types. Overall, functional validation revealed that targeting genes identified by integrated analysis inhibits vessel sprouting.

Conclusion

Previous bulk analysis-based transcriptomics studies revealed only a limited number of EC phenotypes (24-27). Highly powered single cell transcriptomics studies now allow characterization of EC phenotypes in more detail and enable revision of the traditional outlook on the tumor endothelium. Our data suggests a possible role for peritumoral (scavenging) capillary NECs and TECs with a transcriptome signature of HEVs or semi-professional APCs in tumor immune surveillance. Notably, TECs represent the first line defense contact for immune cells in the tumor micro-environment (TME). We also identified ECs that expressed markers of tip cells and an EC type, implicated in the initiation of vessel sprouting by creating an opening in the basement membrane to assist invasive tip cells to sprout (11), which we coined breach ECs. We additionally identified vein ECs with a resident endothelial stem cell signature in mice, which might contribute to lung tumor vascularization, though this remains to be further studied.

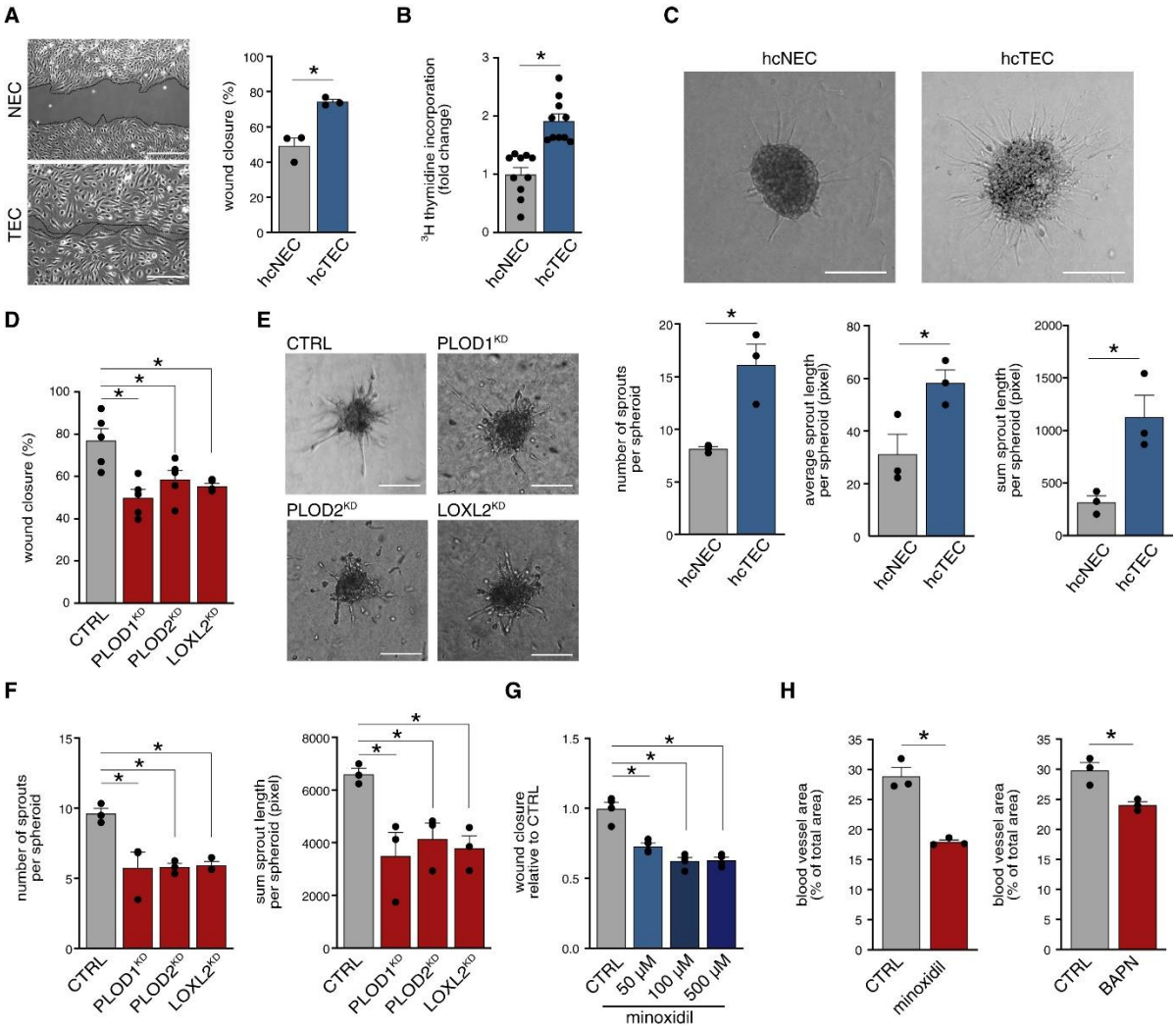

Figure 8: Inhibition of collagen hydroxylation/cross-linking impairs vessel sprouting.

(**A**) Micrographs (left) and quantification (right) of hcNEC and hcTEC migration in scratch wound assays (mean ± SEM; n=3; *p value < 0.05, unpaired two-tailed *t*-test). Scale bar, 300 µm. (**B**) ^{3}H-thymidine incorporation assay (mean ± SEM, n=10, *p value < 0.05, unpaired two-tailed *t*-test). (**C**) Bright field photographs (top) and morphometric quantification (bottom) of hcNEC and hcTEC spheroid sprouting (mean ± SEM; n=3; *p value < 0.05 versus NECs, unpaired two-tailed *t*-test). Scale bars, 100 µm. (**D**) scratch wound migration assay with control and PLOD1, PLOD2 or LOXL2 silenced (KD) ECs (mean ± SEM, n=5 in (E), *p value < 0.05, one-way ANOVA with Dunnett's test). (**E,F**) Bright field photographs (E) and morphometric quantification (F) of spheroid sprouting using control and PLOD1, PLOD2 or LOXL2 silenced (KD) ECs (F). Scale bars, 100 µm. Data are mean ± SEM; n=3; *p value < 0.05 versus control, one-way ANOVA with Dunnett's test. (**G**) Scratch wound migration assay of ECs without or with pharmacological inhibition of PLODs by minoxidil (mean ± SEM, n=3, *p value < 0.05,). (**H**) Corneal angiogenesis upon corneal cauterization-induced injury in mice treated with vehicle (CTRL), minoxidil (MXD) or beta-aminopropionitrile (BAPN). Data are mean ± SEM; n=3 independent experiments of six mice per group; *p value < 0.05; unpaired, two-tailed *t*-test. See also Figure S7 and Figure S8.

Extending previous findings (4), our data show that a VEGFR2-selective inhibitor not only affects more sensitively tip but also breach mTECs, while having smaller effects on vein and capillary mTECs. Additionally, this study provides an unbiased molecular characterization of disorganized TECs and shows that VEGF-blockade induces partial molecular tumor vessel normalization, including the induction of the activated postcapillary vein phenotype

(resembling HEVs) and other more quiescent NEC phenotypes. This raises the question if AAT, in addition to pruning the small number of angiogenic and proliferating TECs, might benefit from tuning the more abundant TEC phenotypes contributing to tumor vessel disorganization. A concern is that tip TECs in human lung tumors make up only <10% of all TECs, raising the question if targeting such a small TEC subpopulation suffices to inhibit tumor angiogenesis and if the paucity of angiogenic tip and proliferating TECs in human lung tumors contributes to the insufficient efficacy and resistance to VEGF-blockade AAT. Regardless, our study provides initial evidence for a correlation of angiogenic signatures with NSCLC patient survival.

Our integrated scRNA-seq and multi-omics approach showed that only the tip TEC phenotype was conserved across species and models. This was surprising, given that the tip EC phenotype is not a genetically predetermined state (1), but may be due to the presence of VEGF in the culture medium, an inducer of the tip EC phenotype (28). The identification of the tip TEC phenotype in culture raises the opportunity to study tip TEC targets in vitro, in contrast to arterial, venous, capillary and other in vivo TEC phenotypes that are lost in culture.

To overcome inter-patient heterogeneity and contextual differences in marker gene signatures due to differences in species (human versus mouse), tumor type (NSCLC versus LLC) and model (freshly isolated versus cultured TECs), we used scRNAseq in combination with orthogonal multi-omics approaches to identify multiple genes involved in posttranslational collagen modification, illustrating that the discovery of these genes as angiogenic candidates was not accidental. We confirmed LOXL2 as known angiogenic candidate (21,29-31), and identified and functionally validated procollagen crosslinking PLOD isoenzymes. Collagens have been implicated in tumor angiogenesis, albeit contextually (32,33), but PLOD isoenzymes have not been substantially and consistently involved in angiogenesis yet. Hence, an integrated approach, based on orthogonal multi-omics single cell and bulk methods, may offer improved chances to identify conserved TEC markers that are biologically and translationally relevant.

This study provides a public resource for data exploration, available at https://www.vibcancer.be/software-tools/lungTumor_ECTax and via the publicly available EndoDB database (59). While our study shows how this integrated approach can identify highly relevant angiogenic targets, we acknowledge limitations of our work. First, the inferred biological role for each EC phenotype is putative and requires functional validation to probe their biological role. Second, larger numbers of patients must be analyzed to probe inter-patient heterogeneity and identify all possible TEC phenotypes at deep resolution. Third, it should not be surprising that mouse and human TEC taxonomies do not completely overlap,

given that tumors in both species grow differently, and patients exhibit larger genetic and environmental heterogeneity than mice.

Supplementary data

Supplementary data are available at Cancer Cell Online.

https://www.cell.com/cancer-cell/fulltext/S1535-6108(19)30571-9#supplementaryMaterial

The Supplementary data contains the following files:

- Table S1: characteristics of patients included in the study (related to Figure 1).
- Table S2: key quality metrics of the scRNA-seq process (related to Figures 1 and 5).
- Table S3: the number of samples and cells per dataset (related to Figures 1 and 3–7).
- Table S4: marker genes of each cluster and differential gene expression analyses (related to Figures 1–3 and 5–7).
- Table S5: molecular taxonomy of phenotypes of freshly isolated human (peri-tumoral) non-tumor ECs in human NSCLC (related to Figure 1).
- Table S6: molecular taxonomy of EC phenotypes of normal ECs (NEC) and tumor ECs (TEC) in a mouse lung tumor model (related to Figure 5).
- Table S7: GoRilla gene ontology analysis results (related to Figure 7).
- Table S8: oligonucleotide sequences and parameter settings of scRNA-seq data preprocessing, visualization, and clustering (related to STAR Methods).
- Figure S1: in silico EC selection and subtype quantification of hpNECs and hTECs (related to Figure 1).
- Figure S2: immunostaining of hpNEC and hTEC phenotypes in human patient samples (related to Figure 3).
- Figure S3: RNAscope in situ hybridization of EC markers in human NSCLC samples (related to Figure 3).
- Figure S4: CyTOF analysis validation of EC phenotypes in human NSCLC samples (related to Figure 3).
- Figure S5: heterogeneity of murine lung NECs and TECs (related to Figure 5).
- Figure S6: immunostaining of mNEC and mTEC phenotypes (related to Figure 5).
- Figure S7: treatment effect, human cultured TEC phenotypes and silencing efficiency (related to Figure 5).
- Figure S8: validation of congruent tip cell markers and proteomics gene ontology analysis (related to Figure 6-8).

Author contributions

J.G., K.R., F.T., L.T. designed and analyzed experiments. J.G., L.T., L.-C.C., A.P., J.K., M.G-C. set up EC isolation protocols. L.-C.C., E.W., H.D, P.DL., J.V., B.W., X.S., E.V., A.W., B. Topal, W.E., H.B., A.E. provided human tumor samples. K.R., L.T., L.-C.C., V.G., J.K., L. Sokol, J.Q, L.-A.T. prepared scRNA-seq samples. L.L., B. Thienpont, D.L., Y.L. did 10x single-cell sequencing. S.K., B.B., T.K. processed scRNA-seq data. L.dR., J.K., D.P., F.D.S., F.J.T.S., R.J.M., Y.Z. did CyTOF. L.T., J.K., S.V. did immunohistochemistry. F.I. did proteomics analysis. V.L. provided advice on the statistical methodology. F.D.S., M.M., G.E., M.D., L. Schoonjans, H.Y., J.W., L.B., L.L., B. Thienpont, X.L., D.L., Y.L. provided advice and discussed results. J.G., K.R., P.C. wrote the manuscript. P.C. conceptualized the study. All authors discussed results and manuscript.

Conflict of interest declaration

None of the authors have competing financial interests to declare.

Funding

We acknowledge K. Peeters, T. Van Brussel and R. Schepers for assistance. J.G., K.R. and J.K. are supported by the 'Fonds voor Wetenschappelijk Onderzoek' (FWO); L.T. by a Marie Curie-IEF ; L.-C.C. by the Else Kröner-Fresenius and Fritz Thyssen Stiftung (10.16.2.017MN); A.P. by the Austrian Science Fund (J3730-B26); L.-A.T. by Antwerp University; V.G. by Strategisch Basisonderzoek FWO (SB-FWO); S.K. by Kom op Tegen Kanker (Flemish Cancer Society); M.M. by Methusalem funding (Flemish Government); L.B. & Y.L. by Sanming Project of Medicine in Shenzhen (SZSM201612074); L.L. by Lundbeck Foundation (R219–2016-1375) and DFF Sapere Aude Starting grant (8048-00072A); Y.L. by BGI-Research, Danish Research Council for Independent Research (DFF–1337–00128), Sapere Aude Young Research Talent Prize (DFF-1335–00763A) & Aarhus University Strategic Grant (AU-iCRISPR); X.L. by State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center (Sun Yat-Sen University) and National Natural Science Foundation of China (81670855), Key Program of Guangzhou Scientific Research Plan (3030901006074); P.C. by VIB TechWatch, Methusalem funding, FWO, Foundation against Cancer (2016-078), Kom op Tegen Kanker, ERC Proof of Concept (ERC-713758) & Advanced ERC Research Grant (EU-ERC743074).

Author information

Correspondence and requests for materials should be addressed to PC (peter.carmeliet@kuleuven.vib.be).

References

1. Blanco, R. and Gerhardt, H. (2013) VEGF and Notch in tip and stalk cell selection. *Cold Spring Harbor perspectives in medicine*, **3**, a006569.

2. Carmeliet, P. and Jain, R.K. (2011) Molecular mechanisms and clinical applications of angiogenesis. *Nature*, **473**, 298-307.

3. Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwe, H., Pircher, A., Van den Eynde, K. *et al.* (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature medicine*, **24**, 1277-1289.

4. Zhao, Q., Eichten, A., Parveen, A., Adler, C., Huang, Y., Wang, W., Ding, Y., Adler, A., Nevins, T., Ni, M. *et al.* (2018) Single-Cell Transcriptome Analyses Reveal Endothelial Cell Heterogeneity in Tumors and Changes following Antiangiogenic Treatment. *Cancer Res*, **78**, 2370-2382.

5. Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S. *et al.* (2018) Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, **563**, 72-78.

6. Innes, B.T. and Bader, G.D. (2018) scClustViz - Single-cell RNAseq cluster assessment and visualization. *F1000Research*, **7**.

7. Girard, J.P. and Springer, T.A. (1995) High endothelial venules (HEVs): specialized endothelium for lymphocyte migration. *Immunology today*, **16**, 449-457.

8. Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods*, **15**, 359-362.

9. Eldridge, L., Moldobaeva, A., Zhong, Q., Jenkins, J., Snyder, M., Brown, R.H., Mitzner, W. and Wagner, E.M. (2016) Bronchial Artery Angiogenesis Drives Lung Tumor Growth. *Cancer Res*, **76**, 5962-5969.

10. Wakabayashi, T., Naito, H., Suehiro, J.I., Lin, Y., Kawaji, H., Iba, T., Kouno, T., Ishikawa-Kato, S., Furuno, M., Takara, K. *et al.* (2018) CD157 Marks Tissue-Resident Endothelial Stem Cells with Homeostatic and Regenerative Properties. *Cell stem cell*, **22**, 384-397 e386.

11. Seano, G., Chiaverina, G., Gagliardi, P.A., di Blasio, L., Puliafito, A., Bouvard, C., Sessa, R., Tarone, G., Sorokin, L., Helley, D. *et al.* (2014) Endothelial podosome rosettes regulate vascular branching in tumour angiogenesis. *Nature cell biology*, **16**, 931-941, 931-938.

12. Dejana, E., Hirschi, K.K. and Simons, M. (2017) The molecular basis of endothelial cell plasticity. *Nat Commun*, **8**, 14361.

13. Liu, Y., Poon, R.T., Li, Q., Kok, T.W., Lau, C. and Fan, S.T. (2005) Both antiangiogenesis- and angiogenesis-independent effects are responsible for hepatocellular carcinoma growth arrest by tyrosine kinase inhibitor PTK787/ZK222584. *Cancer Res*, **65**, 3691-3699.

14. Jain, R.K. (2014) Antiangiogenesis strategies revisited: from starving tumors to alleviating hypoxia. *Cancer cell*, **26**, 605-622.

15. Jain, R.K. (2005) Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science (New York, N.Y.)*, **307**, 58-62.

16. Carmeliet, P. and Jain, R.K. (2011) Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases. *Nature reviews. Drug discovery*, **10**, 417-427.

17. De Bock, K., Georgiadou, M., Schoors, S., Kuchnio, A., Wong, B.W., Cantelmo, A.R., Quaegebeur, A., Ghesquiere, B., Cauwenberghs, S., Eelen, G. *et al.* (2013) Role of PFKFB3-driven glycolysis in vessel sprouting. *Cell*, **154**, 651-663.

18. Ebos, J.M. and Kerbel, R.S. (2011) Antiangiogenic therapy: impact on invasion, disease

progression, and metastasis. *Nature reviews. Clinical oncology*, **8**, 210-221.

19. Strasser, G.A., Kaminker, J.S. and Tessier-Lavigne, M. (2010) Microarray analysis of retinal endothelial tip cells identifies CXCR4 as a mediator of tip cell morphology and branching. *Blood*, **115**, 5102-5110.

20. del Toro, R., Prahst, C., Mathivet, T., Siegfried, G., Kaminker, J.S., Larrivee, B., Breant, C., Duarte, A., Takakura, N., Fukamizu, A. *et al.* (2010) Identification and functional analysis of endothelial tip cell-enriched genes. *Blood*, **116**, 4025-4033.

21. Gilkes, D.M., Semenza, G.L. and Wirtz, D. (2014) Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nature reviews. Cancer*, **14**, 430-439.

22. Shao, S., Zhang, X., Duan, L., Fang, H., Rao, S., Liu, W., Guo, B. and Zhang, X. (2018) Lysyl Hydroxylase Inhibition by Minoxidil Blocks Collagen Deposition and Prevents Pulmonary Fibrosis via TGF-beta(1)/Smad3 Signaling Pathway. *Medical science monitor : international medical journal of experimental and clinical research*, **24**, 8592-8601.

23. Rodriguez, H.M., Vaysberg, M., Mikels, A., McCauley, S., Velayo, A.C., Garcia, C. and Smith, V. (2010) Modulation of lysyl oxidase-like 2 enzymatic activity by an allosteric antibody inhibitor. *The Journal of biological chemistry*, **285**, 20964-20974.

24. Nolan, D.J., Ginsberg, M., Israely, E., Palikuqi, B., Poulos, M.G., James, D., Ding, B.S., Schachterle, W., Liu, Y., Rosenwaks, Z. *et al.* (2013) Molecular signatures of tissue-specific microvascular endothelial cell heterogeneity in organ maintenance and regeneration. *Developmental cell*, **26**, 204-219.

25. Marcu, R., Choi, Y.J., Xue, J., Fortin, C.L., Wang, Y., Nagao, R.J., Xu, J., MacDonald, J.W., Bammler, T.K., Murry, C.E. *et al.* (2018) Human Organ-Specific Endothelial Cell Heterogeneity. *iScience*, **4**, 20-35.

26. Sabbagh, M.F., Heng, J.S., Luo, C., Castanon, R.G., Nery, J.R., Rattner, A., Goff, L.A., Ecker, J.R. and Nathans, J. (2018) Transcriptional and epigenomic landscapes of CNS and non-CNS vascular endothelial cells. *eLife*, **7**.

27. Coppiello, G., Collantes, M., Sirerol-Piquer, M.S., Vandenwijngaert, S., Schoors, S., Swinnen, M., Vandersmissen, I., Herijgers, P., Topal, B., van Loon, J. *et al.* (2015) Meox2/Tcf15 heterodimers program the heart capillary endothelium for cardiac fatty acid uptake. *Circulation*, **131**, 815-826.

28. Siemerink, M.J., Klaassen, I., Vogels, I.M., Griffioen, A.W., Van Noorden, C.J. and Schlingemann, R.O. (2012) CD34 marks angiogenic tip cells in human vascular endothelial cell cultures. *Angiogenesis*, **15**, 151-163.

29. Zaffryar-Eilot, S., Marshall, D., Voloshin, T., Bar-Zion, A., Spangler, R., Kessler, O., Ghermazien, H., Brekhman, V., Suss-Toby, E., Adam, D. *et al.* (2013) Lysyl oxidase-like-2 promotes tumour angiogenesis and is a potential therapeutic target in angiogenic tumours. *Carcinogenesis*, **34**, 2370-2379.

30. Osawa, T., Ohga, N., Akiyama, K., Hida, Y., Kitayama, K., Kawamoto, T., Yamamoto, K., Maishi, N., Kondoh, M., Onodera, Y. *et al.* (2013) Lysyl oxidase secreted by tumour endothelial cells promotes angiogenesis and metastasis. *British journal of cancer*, **109**, 2237-2247.

31. Baker, A.M., Bird, D., Welti, J.C., Gourlaouen, M., Lang, G., Murray, G.I., Reynolds, A.R., Cox, T.R. and Erler, J.T. (2013) Lysyl oxidase plays a critical role in endothelial cell stimulation to drive tumor angiogenesis. *Cancer Res*, **73**, 583-594.

32. Sottile, J. (2004) Regulation of angiogenesis by extracellular matrix. *Biochimica et biophysica acta*, **1654**, 13-22.

33. Fang, M., Yuan, J., Peng, C. and Li, Y. (2014) Collagen as a double-edged sword in tumor progression. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, **35**, 2871-2882.

Chapter VI Discussion

After decades of computational research in the bioinformatics omics field, countless tools and scripts have been created to tackle important problems arising from the need to fully understand biological datasets. These tools are usually available in publicly available repositories (such as CRAN and Bioconductor (101)), which expert bioinformatics users can use and apply to their own datasets. As of June 2020, there are around 1,900 tools available on Bioconductor. These repositories of tools are a goldmine for people accustomed with the world of bioinformatics, but these tools would be even more useful in the hands of people who have the domain-knowledge to interpret biological data. Since these tools are not exposed in a user-friendly way, they do not manage to reach an important part of the audience, namely the non-bioinformatician experimental scientists. Developing new software that can bridge the bioinformatics tools to analyse large-scale biological data to the non-bioinformatician experimental biologists has become a key component for advancing modern biomedical research (102).

In this doctoral thesis, I aimed to provide a solution to the bottleneck problem by presenting an easy-to-use software for the visualization and interpretation of omics data, and showcase the application of such software in recently published works. BIOMEX (103) was developed specifically for non-bioinformatician experimental biologists, with the intent of making bioinformatics methods easily available to the scientists with the domain-knowledge to interpret biological data. Over the course of its development, BIOMEX has been fully tailored to the experimental biologists' needs: i) an extensive manual and video tutorials are provided, which in combination with the convenient user interface, make the software an excellent entry-point to introduce researchers to the field of biological data science; ii) results can be conveniently shared between scientists, improving reproducibility; and iii) all plots and tables are highly customizable and can be exported as high quality publication ready figures.

In summary, BIOMEX enables researches to analyze transcriptomics, metabolomics, proteomics and single cell datasets under a single platform via a friendly user interface.

Contribution of BIOMEX in research

In a software, the effectiveness, approachability and widespread application are the most fundamental aspects (104). BIOMEX has been successfully field-tested in recent peer-reviewed publications, including single cell studies (105-107) and bulk studies (108-110). The software was used not only in the publication of internal projects, but also in the publication of

external collaborative projects (110).

To exhibit the capabilities of BIOMEX, an explanation on how the software was used to analyze (single cell) omics data in two recently published projects from our lab is provided:

1. The re-analysis of publicly available data to create EndoDB, the database of endothelial cell transcriptomics data (111). This is related to the supporting manuscript presented in Chapter IV.
2. The analysis of lung endothelial cell heterogeneity to create the first high-resolution endothelial single cell lung taxonomy across species (105). This is related to the supporting manuscript presented in Chapter V.

Each section will contain a small introduction to the project and which BIOMEX modules were used to analyze the data.

BIOMEX showcase: the database of endothelial cell transcriptomics data (EndoDB)

EndoDB is a web-accessible EC (endothelial cell) specific database, developed in order to promote the re-use of publicly available EC transcriptomics data (111). EndoDB is a collection of pre-analyzed data which was manually curated and quality checked (360 datasets in total): 4,741 bulk and 5,847 single cell endothelial transcriptomes from six different organisms were obtained and analyzed. To create the database, we first searched for EC transcriptomics datasets available in the public domain in repositories like ArrayExpress and Gene Expression Omnibus. We collected the raw data and metadata for these EC transcriptomics datasets. The data was pre-processed to obtain the transcript abundances, and the metadata was manually curated by a team of vascular biology experts. Subsequently, the data and metadata were converted to a tabular format suitable to be used by BIOMEX.

Once the data and metadata preparation was finalized (360 datasets in total), to analyze the data we used BIOMEX, following these steps:

1. (BIOMEX Data Import Section) The data matrices were loaded into BIOMEX.
2. (BIOMEX Module 1) The data was pretreated using the software standard settings.
3. (BIOMEX Module 4) PCA and t-SNE were performed and the results were visualized as 2D plots where individual samples are color-coded according to the experimental design. Additionally, heatmaps were computed to show gene expression changes in a panel of genes.

4. (BIOMEX Module 3 and 5) Differential gene expression analysis and gene set enrichment analysis were performed and the results were visualized as volcano plots, bar plots, waterfall plots and browsable tables.

Together, these analyses cover some of the most commonly used methods and allow for detailed interpretation of biological data. Additionally, if desired, the data and metadata can be downloaded from EndoDB and directly loaded into BIOMEX, making it possible to easily analyze the data using the full set of functionalities/modules provided by BIOMEX. This makes BIOMEX and EndoDB perfectly interoperable, removing an additional barrier in the analysis of omics data. A schematic overview of the construction of the database and at which step the analyses were performed with BIOMEX can be seen in Figure 1. For more details on the BIOMEX modules, please refer to Chapter III. For more details about EndoDB, please refer to Chapter IV.

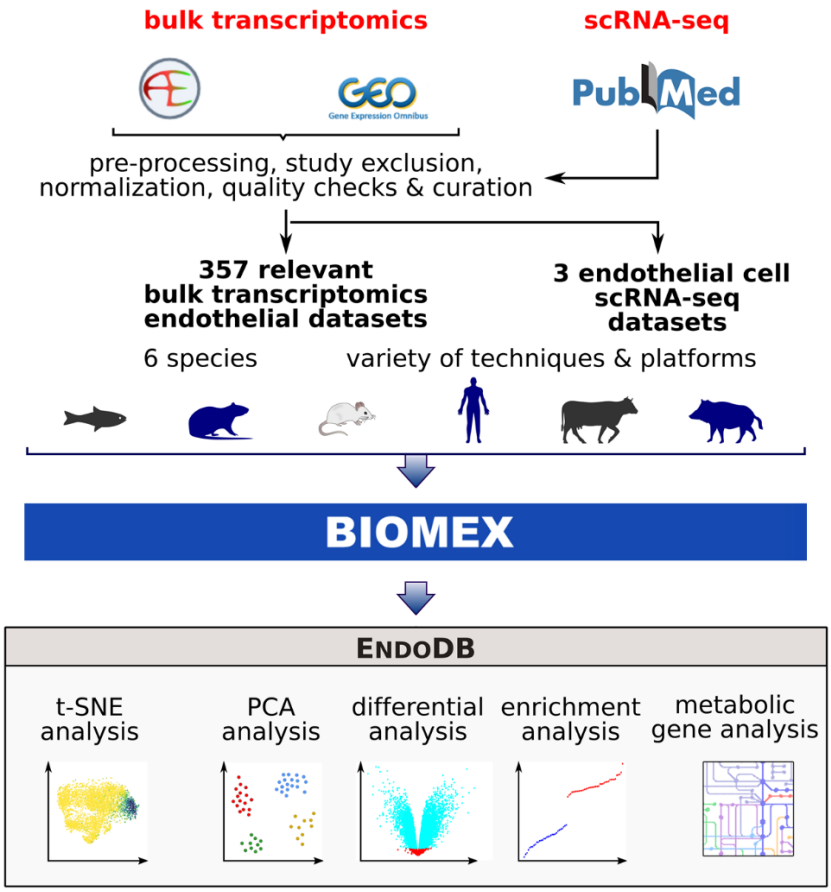

Figure 1: Schematic overview of the EndoDB construction.

This is a modified version of the Figure shown in Chapter IV. In this Figure, the role of BIOMEX is highlighted: all the 360 datasets were analyzed by using the BIOMEX software. Due to this, BIOMEX and EndoDB are completely interoperable, since the data matrices available on EndoDB can be downloaded and directly uploaded into BIOMEX without requiring any additional steps.

BIOMEX showcase: the lung endothelial single cell taxonomy

At a single cell level and across patients and species, the heterogeneity of lung tumor endothelial cell (TEC) phenotypes was poorly inventoried. For this reason, we profiled freshly isolated human tumor ECs (hTECs) and (paired) human (peri-tumoral) non-tumor pulmonary ECs (hPNECs) from patients with non-small cell lung cancer (NSCLC). Single cells were sorted by using magnetic-activated cell sorting (MACS), enriching for endothelial cells (CD31+ cells) and depleting for leukocytes (CD45+ cells). The cells were then sequenced by using the 10x Genomics platform (Figure 2A). A similar strategy was used to construct a murine lung taxonomy of freshly isolated normal ECs (mNECs) and tumor ECs (mTECs) from Lewis lung carcinoma (LLC). Single cells from murine treated with anti-VEGF were also obtained (Figure 2B). For both species, the data was pre-processed to obtain the transcript abundances and converted to a tabular format suitable to be used by BIOMEX.

To create the endothelial single cell lung taxonomy for both species (human and murine), we used BIOMEX, following these steps:

1. (BIOMEX Data Import Section) The transcript abundance data (unique molecular identifier counts) and corresponding metadata were loaded into BIOMEX.
2. (BIOMEX Module 1) The low-quality genes were removed (average expression count < 0.01), together with the cells with high mitochondrial gene expression (percentage expression > 5%). The data was normalized (Seurat normalization) to account for the different sequencing depths between cells. Additionally, the data was batch corrected (mutual nearest neighbor method) to account for unwanted technical variation between samples.
3. (BIOMEX Module 4) PCA was performed on the batch corrected data, and the first 15 PCA components were used to generate the t-SNE plot visualization. These 15 PCA components were also used to find the endothelial cell clusters, using the Seurat graph-based clustering method.
4. (BIOMEX Module 3 and 5) Endothelial cell clusters were annotated with a biological name by finding the gene marker set for each cluster. In the end, 17 known and 16 novel phenotypes were detected. For example, the canonical tip TEC phenotype was categorized into a putative basement-membrane remodeling breach phenotype and a known migratory tip phenotype.
5. (BIOMEX Module 2 and 7) Survival analysis was performed on 502 lung squamous cell carcinoma patients selected from The Cancer Genome Atlas, resulting in tip TEC signatures correlating with patient-survival.

6. (BIOMEX Module 9 and 10) A meta-analysis was performed using data spanning across different human tumors, showing that the tip TECs were the most congruent/similar phenotype across species, as they shared the most conserved set of markers.

This sequence of steps shows how BIOMEX can be used to explore high-dimensional single cell transcriptomics data, leading to potential novel findings. For more details on the BIOMEX modules, please refer to Chapter III. For more details about the EC taxonomies, please refer to Chapter V.

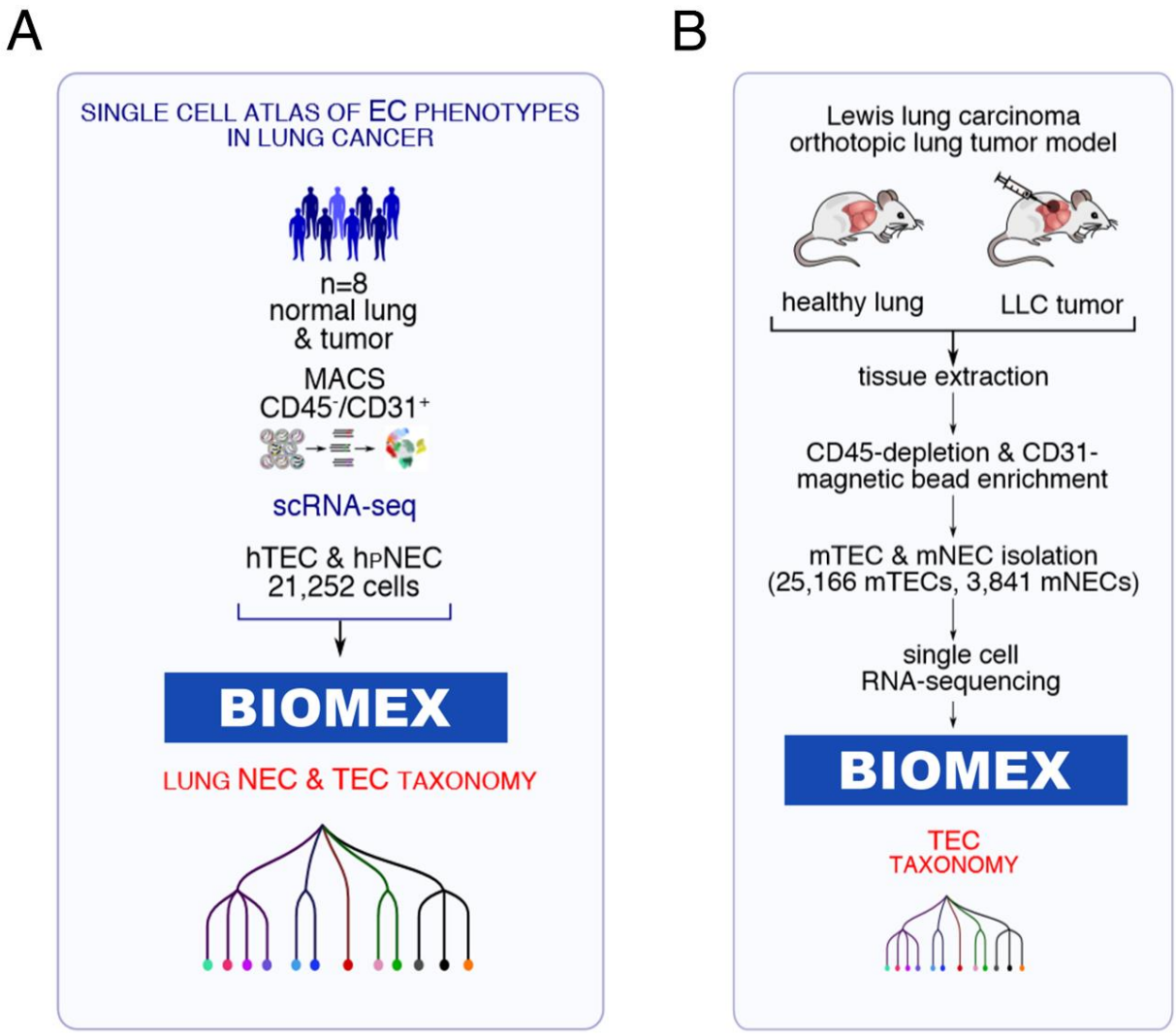

Figure 2: Experimental design of the endothelial cell lung taxonomy.

This is a modified version of the Figures shown in Chapter V. In this Figure, the role of BIOMEX is highlighted: the in-house generated single cell (human and mouse) transcriptomics datasets were analyzed by using the BIOMEX software.

Conclusion

Design philosophy

In biological sciences it is unlikely that a suite of integrated algorithms exist to model all data types without human intervention (112). As previously said in Chapter II, BIOMEX is not a “silver bullet” software but a tool that allows biomedical researchers to optimize the data analysis and interpretation processes, and to make data-driven decisions without extensive knowledge of the bio-computational background.

BIOMEX aims to solve the most common data analysis problems by making the typical data analysis methods available to the experimental biologists in a simple way, it does not aim to provide all the solutions to solve each particular data analysis problem. Hence, BIOMEX has to harmonize its usefulness with its user-friendliness. By including countless methods into the software and increasing the number of parameters the user has to control, we could improve the overall capacity of the software to solve data analysis problems, but at the cost of usability and user-friendliness. Conversely, focusing on pure usability might lead the software in a direction where it could not solve some of the basic problems in omics data analysis. Striking this balance was paramount during the development of BIOMEX.

BIOMEX is not the first and will not be the last software suite dedicated to solve the many data analysis problems present in the omics data science field. Each currently available software has its own design philosophy, making them useful for different reasons compared to BIOMEX. For example, Perseus (113) is one of the most used software for the analysis of proteomics datasets, and the core of the software revolves around the creation of custom workflows. This design philosophy makes Perseus extremely flexible in analyzing omics data, as the user can decide how to approach the analysis independently from the software itself. That flexibility can be exploited to the fullest only when associated with expertise: a bioinformatician might want to use Perseus to analyse proteomics data, as it provides all the capabilities necessary to perform custom and in-depth analyses. Vice versa, an experimental biologist with no bioinformatics expertise might find it difficult to use such software, and would prefer to use a software with a design philosophy closer to that of BIOMEX. As usual, using the right tool for the right job and expertise is still important even in the omics data science field.

Future perspectives

Due to the fast growing nature of the omics data analysis field, BIOMEX should be updated to reflect the improvements made to the analysis methods and algorithms. Additionally, it should

be updated to reflect the ever growing increase in the complexity of data being analyzed.

A decade ago, only few measurements would be taken into consideration during an analysis. Today, thanks to the advent of single cell technologies, hundreds of thousands of measurements are generated. In the future, several million measurements for each dataset could be the norm. Enabling BIOMEX to analyze such amount of data is a priority if we want the software to remain relevant. Using more efficient methods to analyse such data, even for the most basic analyses (*e.g.* dimensionality reduction, differential analysis, etc.) is going to be the focus of many laboratories developing methods for the analysis of omics data, and they should be included in BIOMEX as they become available to the scientific community.

Technological advances should also be taken into consideration in the future development of BIOMEX. New technologies are re-shaping the way we think about biological data analysis. Metagenomics (114) and epigenomics (115) are all new omics technologies which should be carefully considered for addition in BIOMEX. Adding a new analysis workflow for even one of these technologies would entail a major update to the software and a considerable time investment, but the insights they would be able to unlock if incorporated in a framework such as BIOMEX would make the effort worthwhile.

Another important update would be to develop BIOMEX for Mac. As of now, BIOMEX is available on Windows and Linux, and not available on Mac. Mac users are common in the scientific environment, so developing a native Mac version of BIOMEX has the potential to reach many more users than is now possible.

Limitations

As shown in Chapter III, BIOMEX uses R packages to employ the analysis methods into the software. This approach has both advantages and limitations. The advantage is that, since the vast majority of the analysis methods are released as R packages on the CRAN and Bioconductor repositories, the time necessary to implement them in BIOMEX is quite low, usually few days at maximum. This quick turn-around time for the implementation means the new analysis packages can be deployed and field-tested rapidly. The limitation of this approach is that, since a large number of R packages are used into the software, and many of them must communicate between each other, the interoperability between these packages must be checked competently and with care. Updating one R package could have an effect on many modules of the BIOMEX software, and careful testing should be performed to ensure that each module works correctly after the update, which is time consuming. Software that do not use pre-made packages, but implement the analysis methods from scratch can avoid such

circumstances, but in turn the implementation process is much longer (weeks to months), which means additions to the software must be thoughtfully planned.

Another limitation stems from the fact that BIOMEX, being a stand-alone application, has to be run on a local machine. Many machines cannot meet the requirements needed to run BIOMEX, limiting the number of users able to use the software. To mitigate this issue, a "light" version of BIOMEX able to analyze small bulk omics datasets could be developed and deployed as a full web-based service running on Google Cloud through the Google Kubernetes Engine. In that case, no installation would be required by the user, removing the barrier to use the software to analyze at least small datasets, possibly introducing more users to the software.

Software updates

To keep BIOMEX relevant, BIOMEX would need to be updated, by implementing additional relevant tools/techniques and by keeping the R packages up-to-date. The software should also be kept in working shape by providing bug fixes and minor maintenance updates when needed. Since BIOMEX is free-for-everyone, there are different ways in which the software will continue to be updated in the long term. These include:

- Internal updates: made by the original lab that conceptualized and published the software. This is by far the most common way a software can be updated as shown by tools like Perseus, MetaboAnalyst and XCMS Online. Through this route, BIOMEX will be updated either by appointing a dedicated software developer/bioinformatician who can improve the source code, update the R packages and deploy regular maintenance update. Alternatively, the software can be kept up-to-date by employing personnel via periodic consultancies.

- External updates: made by external users interested in improving the software to subsequently use it to analyze their own data. Since BIOMEX is publicly available on a BitBucket repository, users have two options to access the source code; they could create *pull requests*, which then have to be accepted by the repository manager, or *fork* the BIOMEX repository, and implement any change they want to the software. The fork repository can either remain independent of the main repository or, in the end, merged to the main repository as per the repository manager discretion.

- Re-use of code: while this is not specifically a way to update a software, I envision it will be one of the ways with which the BIOMEX software code base will be used and improved by external developers. For example, the BIOMEX code base is being used to develop a cloud based, completely independent version of the software (project under NDA).

Final words

Currently, BIOMEX is on version 1.0-4, and updates have already been made available since its release. The annotation package of the Crab-eating Macaque (*Macaca fascicularis*) was added, and stability fixes for bugs reported by external users (both via email and BitBucket) were implemented. Since BIOMEX is an open-source software and the source code is deposited online, anyone can propose modifications or extensions to the software, which is a fundamental factor for keeping BIOMEX relevant for the years to come.

I believe and anticipate that the software has the potential to unlock insight from complex omics datasets and will help to address key bottlenecks in the biological interpretation of omics data.

Chapter VII References of Chapter I and VI

1. Freeman, W.M., Walker, S.J. and Vrana, K.E. (1999) Quantitative RT-PCR: pitfalls and potential. *Biotechniques*, **26**, 112-122, 124-115.

2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

3. Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A. and Ferrari, R. (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform*, **19**, 286-302.

4. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, **6**, 377-382.

5. Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E. and Blayney, J.K. (2019) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief Bioinform*, **20**, 1795-1811.

6. Hwang, B., Lee, J.H. and Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*, **50**, 96.

7. Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D. *et al.* (2019) A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci*, **22**, 2087-2097.

8. Chang, J. (2015) Core services: Reward bioinformaticians. *Nature*, **520**, 151-152.

9. Tanay, A. and Regev, A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331-338.

10. Johnson, C.H., Ivanisevic, J. and Siuzdak, G. (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews. Molecular cell biology*, **17**, 451-459.

11. Altelaar, A.F., Munoz, J. and Heck, A.J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics*, **14**, 35-48.

12. Goveia, J., Pircher, A., Conradi, L.C., Kalucka, J., Lagani, V., Dewerchin, M., Eelen, G., DeBerardinis, R.J., Wilson, I.D. and Carmeliet, P. (2016) Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO molecular medicine*, **8**, 1134-1142.

13. Alyass, A., Turcotte, M. and Meyre, D. (2015) From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics*, **8**, 33.

14. Weber, R.J.M., Lawson, T.N., Salek, R.M., Ebbels, T.M.D., Glen, R.C., Goodacre, R., Griffin, J.L., Haug, K., Koulman, A., Moreno, P. *et al.* (2017) Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics : Official journal of the Metabolomic Society*, **13**, 12.

15. Brazas, M.D., Blackford, S. and Attwood, T.K. (2017) Training: Plug gap in essential bioinformatics skills. *Nature*, **544**, 161.

16. Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, **26**, 1367-1372.

17. Du, Y., Huang, Q., Arisdakessian, C. and Garmire, L.X. (2020) Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment. *G3 (Bethesda)*, **10**, 1775-1783.

18. Tautenhahn, R., Patti, G.J., Rinehart, D. and Siuzdak, G. (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal Chem*, **84**, 5035-5039.

19. Karaman, I. (2017) Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis. *Adv Exp Med Biol*, **965**, 145-161.

20. Dunkler, D., Sánchez-Cabo, F. and Heinze, G. (2011) Statistical analysis principles for Omics data. *Methods Mol Biol*, **719**, 113-131.

21. (2016) Meta-analysis in basic biology. *Nature Methods*, **13**, 959-959.

22. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017) Transcriptomics technologies. *PLoS Comput Biol*, **13**, e1005457.

23. Dettmer, K., Aronov, P.A. and Hammock, B.D. (2007) Mass spectrometry-based metabolomics. *Mass Spectrom Rev*, **26**, 51-78.

24. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.

25. Karakach, T.K., Flight, R.M., Douglas, S.E. and Wentzell, P.D. (2010) An introduction to DNA microarrays for gene expression analysis. *Chemometrics and Intelligent Laboratory Systems*, **104**, 28-52.

26. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

27. Motakis, E.S., Nason, G.P., Fryzlewicz, P. and Rutter, G.A. (2006) Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics*, **22**, 2547-2553.

28. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

29. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47.

30. Abrams, Z.B., Johnson, T.S., Huang, K., Payne, P.R.O. and Coombes, K. (2019) A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics*, **20**, 679.

31. Li, P., Piao, Y., Shon, H.S. and Ryu, K.H. (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, **16**, 347.

32. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.

33. Kaern, M., Elston, T.C., Blake, W.J. and Collins, J.J. (2005) Stochasticity in gene expression: from theories to phenotypes. *Nature reviews. Genetics*, **6**, 451-464.

34. Hafemeister, C. and Satija, R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*, **20**, 296.

35. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*, **33**, 495-502.

36. Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

37. Fujita, A., Sato, J.R., Rodrigues Lde, O., Ferreira, C.E. and Sogayar, M.C. (2006) Evaluating different methods of microarray data normalization. *BMC Bioinformatics*, **7**, 469.

38. Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X. and Zhu, F. (2017) NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res*, **45**, W162-w170.

39. Bandura, D.R., Baranov, V.I., Ornatsky, O.I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J.E. and Tanner, S.D. (2009) Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem*, **81**, 6813-6822.

40. Bendall, S.C., Simonds, E.F., Qiu, P., Amir el, A.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687-696.

41. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. and van der Werf, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

42. van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579--2605.

43. McInnes, L., Healy, J. and Melville, J. (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

44. Sun, S., Zhu, J., Ma, Y. and Zhou, X. (2019) Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, **20**, 269.

45. Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T. and Saeys, Y. (2015) FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A*, **87**, 636-645.

46. Duò, A., Robinson, M.D. and Soneson, C. (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res*, **7**, 1141.

47. Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, **2**, 193-218.

48. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, **14**, 483-486.

49. Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540-1542.

50. Huang, H.C., Niu, Y. and Qin, L.X. (2015) Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. *Cancer Inform*, **14**, 57-67.

51. Wang, T., Li, B., Nelson, C.E. and Nabavi, S. (2019) Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, **20**, 40.

52. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D. and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, **14**, R95.

53. Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*, **12**, e0190152.

54. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

55. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*, **16**, 278.

56. Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods*, **11**, 740-742.

57. Miao, Z., Deng, K., Wang, X. and Zhang, X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, **34**, 3223-3224.

58. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151-1161.

59. Rahmatallah, Y., Emmert-Streib, F. and Glazko, G. (2014) Comparative evaluation of gene set analysis approaches for RNA-Seq data. *BMC Bioinformatics*, **15**, 397.

60. Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-987.

61. Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, **102**, 13544-13549.

62. Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.L., Visvader, J.E. and Smyth, G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176-2182.

63. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545.

64. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.

65. Barry, W.T., Nobel, A.B. and Wright, F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943-1949.

66. Mathur, R., Rotroff, D., Ma, J., Shojaie, A. and Motsinger-Reif, A. (2018) Gene set analysis methods: a systematic comparison. *BioData Mining*, **11**, 8.

67. Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L. and Chory, J. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825-2827.

68. Rau, A., Marot, G. and Jaffrézic, F. (2014) Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, **15**, 91.

69. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.

70. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**, 5116-5121.

71. Allen, E.A., Erhardt, E.B. and Calhoun, V.D. (2012) Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron*, **74**, 603-608.

72. Slutsky, D.J. (2014) The effective use of graphs. *J Wrist Surg*, **3**, 67-68.

73. Chia, P.L., Gedye, C., Boutros, P.C., Wheatley-Price, P. and John, T. (2016) Current and Evolving Methods to Visualize Biological Data in Cancer Research. *J Natl Cancer Inst*, **108**.

74. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, **32**, 381-386.

75. Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D. and Saeys, Y. (2016) SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv*, 079509.

76. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E. and Dudoit, S. (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.

BMC Genomics, **19**, 477.

77. Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019) A comparison of single-cell trajectory inference methods. *Nat Biotechnol*, **37**, 547-554.

78. Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods*, **15**, 359-362.

79. Altman, N. and Krzywinski, M. (2018) The curse(s) of dimensionality. *Nat Methods*, **15**, 399-400.

80. Hanzelmann, S., Castelo, R. and Guinney, J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

81. Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108-112.

82. Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. and Lee, D. (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*, **4**, e1000217.

83. Tomfohr, J., Lu, J. and Kepler, T.B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

84. Tomczak, K., Czerwinska, P. and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, **19**, A68-77.

85. Goel, M.K., Khanna, P. and Kishore, J. (2010) Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, **1**, 274-278.

86. Bland, J.M. and Altman, D.G. (2004) The logrank test. *BMJ (Clinical research ed.)*, **328**, 1073.

87. Breslow, N.E. Analysis of Survival Data under the Proportional Hazards Model.

88. Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Stat Med*, **16**, 385-395.

89. Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*, **39**, 1-13.

90. Michael, L. and John, C. Survival Trees by Goodness of Split. *Journal of the American Statistical Association*, **88**, 457-467.

91. Lee, S. and Lim, H. (2019) Review of statistical methods for survival analysis using genomic data. *Genomics Inform*, **17**, e41.

92. Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

93. Szymanska, E., Saccenti, E., Smilde, A.K. and Westerhuis, J.A. (2012) Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics : Official journal of the Metabolomic Society*, **8**, 3-16.

94. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015) Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, **13**, 8-17.

95. Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.C., Pircher, A., Geldhof, V., de Rooij, L., Kalucka, J., Sokol, L. *et al.* (2020) An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell*, **37**, 21-36.e13.

96. Rohlenova, K., Goveia, J., García-Caballero, M., Subramanian, A., Kalucka, J., Treps, L., Falkenberg, K.D., de Rooij, L.P.M.H., Zheng, Y., Lin, L. *et al.* (2020) Single-Cell RNA Sequencing Maps Endothelial Metabolic Plasticity in Pathological Angiogenesis. *Cell Metabolism*, **31**, 862-877.e814.

97. Sudmant, P.H., Alexis, M.S. and Burge, C.B. (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol*, **16**, 287.

98. Levandowsky, M. and Winter, D. (1971) Distance between Sets. *Nature*, **234**, 34-35.

99. Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z. and Bar-Joseph, Z. (2018) A web server for comparative analysis of single-cell RNA-seq data. *Nature Communications*, **9**, 4768.

100. Hutchison, L.A.D., Berger, B. and Kohane, I.S. (2020) Meta-analysis of Caenorhabditis elegans single-cell developmental data reveals multi-frequency oscillation in gene activation. *Bioinformatics*, **36**, 4047-4057.

101. Reimers, M. and Carey, V.J. (2006) Bioconductor: an open source framework for bioinformatics and computational biology. *Methods in enzymology*, **411**, 119-134.

102. Mangul, S., Mosqueiro, T., Abdill, R.J., Duong, D., Mitchell, K., Sarwal, V., Hill, B., Brito, J., Littman, R.J., Statz, B. *et al.* (2019) Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS Biol*, **17**, e3000333.

103. Taverna, F., Goveia, J., Karakach, T.K., Khan, S., Rohlenova, K., Treps, L., Subramanian, A., Schoonjans, L., Dewerchin, M., Eelen, G. *et al.* (2020) BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. *Nucleic Acids Res*.

104. List, M., Ebert, P. and Albrecht, F. (2017) Ten Simple Rules for Developing Usable Software in Computational Biology. *PLoS Comput Biol*, **13**, e1005265.

105. Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.C., Pircher, A., Geldhof, V., de Rooij, L., Kalucka, J., Sokol, L. *et al.* (2020) An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell*, **37**, 21-36.e13.

106. Dumas, S.J., Meta, E., Borri, M., Goveia, J., Rohlenova, K., Conchinha, N.V., Falkenberg, K., Teuwen, L.A., de Rooij, L., Kalucka, J. *et al.* (2020) Single-Cell RNA Sequencing Reveals Renal Endothelium Heterogeneity and Metabolic Adaptation to Water Deprivation. *J Am Soc Nephrol*, **31**, 118-138.

107. Kalucka, J., de Rooij, L., Goveia, J., Rohlenova, K., Dumas, S.J., Meta, E., Conchinha, N.V., Taverna, F., Teuwen, L.A., Veys, K. *et al.* (2020) Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell*, **180**, 764-779.e720.

108. Kalucka, J., Bierhansl, L., Conchinha, N.V., Missiaen, R., Elia, I., Bruning, U., Scheinok, S., Treps, L., Cantelmo, A.R., Dubois, C. *et al.* (2018) Quiescent Endothelial Cells Upregulate Fatty Acid beta-Oxidation for Vasculoprotection via Redox Homeostasis. *Cell Metab*, **28**, 881-894.e813.

109. Bruning, U., Morales-Rodriguez, F., Kalucka, J., Goveia, J., Taverna, F., Queiroz, K.C.S., Dubois, C., Cantelmo, A.R., Chen, R., Loroch, S. *et al.* (2018) Impairment of Angiogenesis by Fatty Acid Synthase Inhibition Involves mTOR Malonylation. *Cell Metab*, **28**, 866-880.e815.

110. Bauer, R., Udonta, F., Wroblewski, M., Ben-Batalla, I., Santos, I.M., Taverna, F., Kuhlencord, M., Gensch, V., Pasler, S., Vinckier, S. *et al.* (2018) Blockade of Myeloid-Derived Suppressor Cell Expansion with All-Trans Retinoic Acid Increases the Efficacy of Antiangiogenic Therapy. *Cancer Res*, **78**, 3220-3232.

111. Khan, S., Taverna, F., Rohlenova, K., Treps, L., Geldhof, V., de Rooij, L., Sokol, L., Pircher, A., Conradi, L.C., Kalucka, J. *et al.* (2019) EndoDB: a database of endothelial cell transcriptomics data. *Nucleic Acids Res*, **47**, D736-d744.

112. Mattmann, C.A. (2013) Computing: A vision for data science. *Nature*, **493**, 473-475.

113. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M. and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, **13**, 731-740.

114. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, **35**, 833-844.

115. Stricker, S.H., Köferle, A. and Beck, S. (2017) From profiles to function in epigenomics. *Nature Reviews Genetics*, **18**, 51-66.

Chapter VIII Acknowledgements, personal contribution and conflict of interest statement

Scientific acknowledgements

I would like to express my deep gratitude to the TransBioInfo team (Shawez Khan, Tobias Karakach, Abhishek Subramanian, Jermaine Goveia and Katerina Rohlenova), which contributed to all the works presented in this thesis. I would also like to thank all the co-authors of the presented publications in this thesis, as they were essential in making these works come to fruition. The same applies to all senior/staff scientists who helped me achieve all PhD milestones. Additionally, I want to recognize the help many of the lab members have provided by using and testing the BIOMEX software and the EndoDB database, which aided me immensely in the development phase of these software tools.

Personal acknowledgements

I would like to say thanks to my promotor, Prof. Peter Carmeliet, for giving me the possibility to work in such a dynamic and interesting lab. It was a challenging but also fulfilling experience. I would like to thank again the TransBioInfo team, not only for the scientific contributions, but for the great moments we passed together, which made for a friendly environment to work in. The same applies to my office buddies, I will miss our chats and gossiping. Finally, I want to thank my family for always supporting and encouraging my endeavors for the full duration of my PhD.

Personal contributions

I, together with my promotor Prof. Peter Carmeliet, head of the Laboratory of Angiogenesis and Vascular Metabolism (VIB KU Leuven, Center for Cancer Biology, Campus Gasthuisberg, Leuven, Belgium), designed the structure of this doctoral thesis. I personally coordinated and performed this work with the aid of my co-promotors Tobias Karakach and Abhishek Subramanian, and with the support of Mieke Dewerchin and Guy Eelen.

My contributions related to the publications presented in this thesis are:

Chapter III: I managed, designed, developed, tested and deployed the BIOMEX software (and manual). The design was performed together with Jermaine Goveia. The testing was performed together with members of the laboratory of Peter Carmeliet.

Chapter IV: I created the code used to perform the downstream analysis on all the collected transcriptomics data for the database and developed the back-end and front-end of the EndoDB database.

Chapter V: I implemented the pipeline used to analyze all the single cell (transcriptomics, mass cytometry) and bulk (transcriptomics, proteomics) omics data and developed the back-end and front-end of the Lung ECTax repository resource.

Conflict of interest statement

I have no competing financial interests to declare.

Chapter IX Curriculum Vitae

FEDERICO TAVERNA

Personal information

Surname:	Taverna
Name:	Federico
Sex:	Male
Date of Birth:	19-12-1990
Nationality:	Italian
Languages:	Italian (mother tongue), English
Address:	Viale Europa 1, 26010 Izano, CR, Italy
E-mail:	federico.taverna@outlook.com

Work experience

Jun 2016 – Jun 2020	Software Engineer & Bioinformatician (PhD training), Center for Cancer Biology, KU Leuven, Belgium
Mar 2015 – Mar 2016	Graduate internship (mentor: Nicola Segata), Trento, Italy, University of Trento
Aug 2012 – Dec 2012	Undergraduate internship (mentor: Alberto Ceselli), Crema, Italy, University of Milano

Education

2016 – Present	PhD (Bioinformatics), Peter Carmeliet lab, Center for Cancer Biology, KU Leuven, Belgium
2012 – 2016	M.Sc. (Computer Science), University of Trento, Trento, Italy
2013 – 2014	Erasmus (Bioinformatics), VU University, Amsterdam, Netherlands
2009 – 2012	B.Sc. (Computer Science), University of Milano, Milano, Italy

Research fellowship

2009	University of Milano scholarship (B.Sc.)
2013	University of Trento scholarship (M.Sc. Erasmus)
2016 – 2020	PhD fellowship, research fund KU Leuven.

Publications (first authorship)

Taverna F*, Goveia J*, Khan S, Karakach T, Rohlenova K, Treps L, Subramanian A, Schoonjans L, Dewerchin M, Eelen G, Carmeliet P. *BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization.* Nucleic Acid Res. 2020 May 11.

Goveia J*, Rohlenova K*, **Taverna F***, Treps L*, Conradi LC, Pircher A, Geldhof V, de Rooij LPMH, Kalucka J, Sokol L, García-Caballero M, Zheng Y, Qian J, Teuwen LA, Khan S, Boeckx B, Wauters E, Decaluwé H, De Leyn P, Vansteenkiste J, Weynand B, Sagaert X, Verbeken E, Wolthuis A, Topal B, Everaerts W, Bohnenberger H, Emmert A, Panovska D, De Smet F, Staal FJT, Mclaughlin RJ, Impens F, Lagani V, Vinckier S, Mazzone M, Schoonjans L, Dewerchin M, Eelen G, Karakach TK, Yang H, Wang J, Bolund L, Lin L, Thienpont B, Li X, Lambrechts D, Luo Y, Carmeliet P. *An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates.* Cancer Cell. 2020 Jan 13.

Khan S*, **Taverna F***, Rohlenova K*, Treps L, Geldhof V, de Rooij L, Sokol L, Pircher A, Conradi LC, Kalucka J, Schoonjans L, Eelen G, Dewerchin M, Karakach T, Li X, Goveia J, Carmeliet P. *EndoDB: a database of endothelial cell transcriptomics data*. Nucleic Acids Res. 2019 Jan 8.

Publications (co-authorships)

Rohlenova K, Goveia J, García-Caballero M, Subramanian A, Kalucka J, Treps L, Falkenberg KD, de Rooij LPMH, Zheng Y, Lin L, Sokol L, Teuwen LA, Geldhof V, **Taverna F**, Pircher A, Conradi LC, Khan S, Stegen S, Panovska D, De Smet F, Staal FJT, Mclaughlin RJ, Vinckier S, Van Bergen T, Ectors N, De Haes P, Wang J, Bolund L, Schoonjans L, Karakach TK, Yang H, Carmeliet G, Liu Y, Thienpont B, Dewerchin M, Eelen G, Li X, Luo Y, Carmeliet P. *Single-Cell RNA Sequencing Maps Endothelial Metabolic Plasticity in Pathological Angiogenesis.* Cell Metab. 2020 Apr 7.

Kalucka J, de Rooij LPMH, Goveia J, Rohlenova K, Dumas SJ, Meta E, Conchinha NV, **Taverna F**, Teuwen LA, Veys K, García-Caballero M, Khan S, Geldhof V, Sokol L, Chen R, Treps L, Borri M, de Zeeuw P, Dubois C, Karakach TK, Falkenberg KD, Parys M, Yin X, Vinckier S, Du Y, Fenton RA, Schoonjans L, Dewerchin M, Eelen G, Thienpont B, Lin L, Bolund L, Li X, Luo Y, Carmeliet P. *Single-Cell Transcriptome Atlas of Murine Endothelial Cells.* Cell. 2020 Feb 20.

Dumas SJ, Meta E, Borri M, Goveia J, Rohlenova K, Conchinha NV, Falkenberg K, Teuwen LA, de Rooij L, Kalucka J, Chen R, Khan S, **Taverna F**, Lu W, Parys M, De Legher C, Vinckier S, Karakach TK,, Schoonjans L, Lin L,, Bolund L, Dewerchin M, Eelen G, Rabelink TJ, Li X, Luo Y, Carmeliet P. *Single-Cell RNA Sequencing Reveals Renal Endothelium Heterogeneity and Metabolic Adaptation to Water Deprivation.* J Am Soc Nephrol. 2020 Jan.

Bruning U, Morales-Rodriguez F, Kalucka J, Goveia J, **Taverna F**, Queiroz KCS, Dubois C, Cantelmo AR, Chen R, Loroch S, Timmerman E, Caixeta V, Bloch K, Conradi LC, Treps L, Staes A, Gevaert K, Tee A, Dewerchin M, Semenkovich CF, Impens F, Schilling B, Verdin E, Swinnen JV, Meier JL, Kulkarni RA, Sickmann A, Ghesquière B, Schoonjans L, Li X, Mazzone M, Carmeliet P. *Impairment of Angiogenesis by Fatty Acid Synthase Inhibition Involves mTOR Malonylation.* Cell Metab. 2018 Dec 4.

Kalucka J, Bierhansl L, Conchinha NV, Missiaen R, Elia I, Brüning U, Scheinok S, Treps L, Cantelmo AR, Dubois C, de Zeeuw P, Goveia J, Zecchin A, **Taverna F**, Morales-Rodriguez F, Brajic A, Conradi LC, Schoors S, Harjes U, Vriens K, Pilz GA, Chen R, Cubbon R, Thienpont B, Cruys B, Wong BW, Ghesquière B, Dewerchin M, De Bock K, Sagaert X, Jessberger S, Jones EAV, Gallez B, Lambrechts D, Mazzone M, Eelen G, Li X, Fendt SM, Carmeliet P. *Quiescent Endothelial Cells Upregulate Fatty Acid β-Oxidation for Vasculoprotection via Redox Homeostasis.* Cell Metab. 2018 Aug 20.

Bauer R, Udonta F, Wroblewski M, Ben-Batalla I, Santos IM, **Taverna F**, Kuhlencord M, Gensch V, Päsler S, Vinckier S, Brandner JM, Pantel K, Bokemeyer C, Vogl T, Roth J, Carmeliet P, Loges S. *Blockade of Myeloid-Derived Suppressor Cell Expansion with All-Trans Retinoic Acid Increases the Efficacy of Antiangiogenic Therapy.* Cancer Res. 2018 Jun 15.

Pfisterer L, Derlet A, Jae N, Hartung T, Michalik L, Goveia J, **Taverna F**, Chen W, Carmeliet P, Zeiher A, Boon R, Dimmeler S. *Endothelial Cell Metabolism is Controlled by the Flow-induced Long Non-coding Rna Lncflow2.* Circulation. 2018 Jun 9.

Thompson AA, Dickinson RS, Murphy F, Thomson JP, Marriott HM, Tavares A, Willson J, Williams L, Lewis A, Mirchandani A, Dos Santos Coelho P, Doherty C, Ryan E, Watts E, Morton NM, Forbes S, Stimson RH, Hameed AG, Arnold N, Preston JA, Lawrie A, Finisguerra V, Mazzone M, Sadiku P, Goveia J, **Taverna F**, Carmeliet P, Foster SJ, Chilvers ER, Cowburn AS, Dockrell DH, Johnson RS, Meehan RR, Whyte MK, Walmsley SR. *Hypoxia determines survival outcomes of bacterial infection through HIF-1alpha dependent re-programming of leukocyte metabolism.* Sci Immunol. 2017 Feb 10.