Aix*Marseille université
Socialement engagée

NNT : 000AIXM0000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 29 Octobre 2024 par

Margaux HAERING

Compréhension de la contribution des types cellulaires aux ataxies cérébelleuses grâce à l'exploration approfondie des données de séquençage Single Cell à l'aide des outils mitoXplorer et ataxiaXplorer

Discipline
Biologie Santé

Spécialité
Bioinformatique et Génomique

ED 62 - Sciences de la vie et de la Santé

Laboratoire
Institut de Biologie du développement de Marseille
IBDM - CNRS UMR 7288

IBDM

Composition du jury :

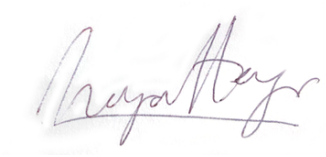
Marie SÉMON **ENS LYON**	Rapporteur
Carles CANTO **EPFL**	Rapporteur
Josselin NOIREL **GBCM - CNAM**	Examinateur
Cédric MAURANGE **IBDM - CNRS UMR 7288**	Examinateur
Andrew SAURIN **IBDM - CNRS UMR 7288**	Président du jury
Bianca HABERMANN **IBDM - CNRS UMR 7288**	Directrice de thèse

Affidavit

I, undersigned, **Margaux Haering**, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of **Bianca H. Habermann**, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, October 29th, 2024



Liste des publications et participations aux conférences

Publications

Haering M, Habermann BH. RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. F1000Res. 2021 Jul 26;10:654. doi: 10.12688/f1000research.54533.2. PMID: 35186266; PMCID: PMC8825645.

Meiler A, Marchiano F, **Haering M**, Weitkunat M, Schnorrer F, Habermann BH. AnnoMiner is a new web-tool to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators. Sci Rep. 2021 Jul 29;11(1):15463. doi: 10.1038/s41598-021-94805-1. PMID: 34326396; PMCID: PMC8322331.

Marchiano F, **Haering M**, Habermann BH. The mitoXplorer 2.0 update: integrating and interpreting mitochondrial expression dynamics within a cellular context. Nucleic Acids Res. 2022 Jul 5;50(W1):W490-W499. doi: 10.1093/nar/gkac306. PMID: 35524562; PMCID: PMC9252804.

Conférences

2022 :

COLLOQUE EDSVS 2022 - campus CNRS Joseph Aiguier à Marseille, France
Treat-ARCA/PROSPAX Berlin meeting - Berlin, Germany

2023 :

COLLOQUE EDSVS 2023 - campus CNRS Joseph Aiguier à Marseille, France

2024 :

JOBIM 2024 - Toulouse, France
REBICA 2024 - Nice, France

Résumé

Les données omiques, telles que le séquençage de l'ARN ou la protéomique, sont des techniques puissantes qui permettent d'étudier les profils d'expression cellulaire dans de nombreuses conditions différentes, telles que le développement, l'homéostasie, la progression d'une maladie ou la réponse à un traitement. Lorsqu'elle est réalisée au niveau des tissus, cette technique peut fournir des informations très précieuses sur la dynamique d'expression spécifique du tissu dans un ensemble varié de conditions. Toutefois, les tissus sont composés de nombreux types de cellules différents, dont les profils d'expression peuvent également varier. Pour vraiment comprendre la dynamique d'expression des types de cellules ou des cellules individuelles, nous devons réaliser de telles expériences au niveau de la cellule unique. Le single-cell RNA sequencing (sc-RNA-seq) est une technique de pointe qui permet d'établir le profil de l'expression des gènes au niveau de la cellule unique et peut donc donner un aperçu des processus biologiques à un niveau de détail et de résolution beaucoup plus élevé.

Les mitochondries sont des organites hérités de la mère qui jouent un rôle central dans plusieurs processus cellulaires, notamment le métabolisme, la production d'énergie, la signalisation et la régulation de l'apoptose. La compréhension de la fonction mitochondriale et de son dérèglement dans les maladies nécessite des outils bioinformatiques spécifiques. mitoXplorer et mitoXplorer 2.0 sont de tels outils qui aident à analyser les données -omiques d'un point de vue mitochondrial. Toutefois, ces outils sont actuellement limités à l'analyse des données de bulk RNA-seq. Lorsqu'on étudie les troubles neurodégénératifs, tels que les ataxies spinocérébelleuses (SCA), qui peuvent souvent être liés à la perte de types de cellules spécifiques dans le cerveau, l'analyse en masse sera limitée par la résolution des données et la conclusion qui peut en être tirée.

Dans cette thèse, je présente le développement d'outils conviviaux et basés sur le web pour plusieurs types de données transcriptomiques. Tout d'abord, j'ai développé une application R-shiny pour analyser les données bulk RNA-seq, RNfuzzyApp. Elle offre plusieurs outils pour l'analyse classique de l'expression différentielle. De plus, RNfuzzyApp est jusqu'à présent la seule application qui aide les utilisateurs à effectuer des analyses de séries temporelles avec l'algorithme de fuzzy clustering Mfuzz, en automatisant partiellement l'analyse de ces données. Deuxièmement, j'ai développé une mise à jour de la plateforme mitoXplorer 2.0 pour l'analyse de bulk centrées sur les mitochondries. Avec cette mise à jour vers mitoXplorer 3.0, la plateforme est désormais capable d'intégrer les données sc-RNA-seq.

Pour le prétraitement des données, j'ai développé un script automatisé qui génère d'abord des données pseudo-bulk et effectue une analyse d'expression différentielle basée sur le pseudo-bulk en fonction des types de cellules annotés. Deuxièmement, le script extrait les mitogènes de la matrice de comptage des cellules uniques, créant un fichier sc-RNA-seq spécifique aux mitogènes prêt à l'importation dans la nouvelle plateforme mitoXplorer 3.0, avec de nouvelles interfaces interactives que j'ai développées. Enfin, j'ai participé au développement de la première plateforme sœur de mitoXplorer, ataxiaXplorer, afin d'explorer et d'analyser les données transcriptomiques bulk et single cell dans le contexte des ataxies cérébelleuses.

Abstract

Omics data, such as RNA-sequencing or proteomics, are powerful techniques to interrogate cellular expression profiles in many different conditions, such as development, homoeostasis, disease progression or treatment response. When done on tissue-level, it can give highly valuable insight into the specific expression dynamics of the tissue in a diverse set of conditions. Tissues, however, are composed of many different cell types, which can also vary in their expression patterns. To truly understand the expression dynamics of cell types or individual cells, we need to perform such experiments on the single-cell level. Single-cell RNA sequencing (sc-RNA-seq) is a cutting-edge technique that enables the profiling of gene expression at the single-cell level and thus can give insights into biological processes at a much greater detail and resolution.

Mitochondria are maternally inherited organelles that play a pivotal role in several cellular processes including metabolism, energy production, signaling and regulation of apoptosis. Understanding mitochondrial function and its dysregulation in diseases requires specific bioinformatics tools. mitoXplorer and mitoXplorer 2.0 are such tools that aid in analyzing -omics data from a mitochondrial perspective. However, these tools are currently limited to analyzing bulk RNA-sequencing data. When looking at neurodegenerative disorders, such as Spinocerebellar Ataxias (SCAs), which often can be linked to the loss of specific cell types in the brain, bulk analysis will be limited by the resolution of the data and the conclusion that can be drawn from it.

In this thesis, I present the development of user-friendly and web-based tools for several types of transcriptomics data. First, I developed a R-shiny app to analyze bulk RNA-seq data, RNfuzzyApp. It offers several tools for classical bulk differential expression analysis. Moreover, RNfuzzyApp is the so far only app that helps users perform time-series analysis with the fuzzy clustering algorithm Mfuzz, by partially automating the process for fuzzy clustering analysis. Secondly, I developed an update of the mitoXplorer 2.0 platform for mitochondria-centric data analysis of RNA sequencing data. With this update to mitoXplorer 3.0, the platform is now able to integrate sc-RNA-sequencing data. For data pre-processing, I developed an automated script that first generates pseudo-bulk data and performs pseudo-bulk based differential expression analysis based on annotated cell types. Secondly, the script extracts the mito-genes from the single-cell count matrix, creating a mito-gene specific scRNA-seq dataset for direct upload and data mining in the mitoXplorer 3.0 platform, with novel interactive interfaces that I have developed. Finally, I have been involved in the development of the first sister platform of mitoXplorer, ataxiaXplorer, to explore and analyze bulk- and single-cell transcriptomics data in the context of cerebellar ataxias.

Acknowledgment

I wish to express my gratitude to Bianca for graciously welcoming me into the team during my Master 2 internship and affording me the opportunity to engage in this research and complete this thesis. I extend my appreciation to Theo for providing invaluable assistance with programming, to Ismahene for our insightful discussions, and to the remaining team members for our shared experiences over the years.

My heartfelt thanks go to my close friends Milla and Ju for the countless afternoons spent unraveling mean stuff and equations over bubble teas, for the shared laughter, adventures, and for being exceptional companions.

I am thankful to Maxence, my partner, for unwavering support, particularly throughout this past year, enabling me to find clarity, exploring the countryside on our motorcycles, and engaging in skydiving adventures. Your presence has consistently brought joy, love and levity to every moment.

I'd like to express my gratitude to my parents for their support and for patiently listening to my rehearsals and all the great moments.

I'm also deeply appreciative of Claire and Taher for their encouragement and invaluable feedback, which has offered me fresh perspectives on my work.

Lastly, I can't forget to mention my beloved cats - Osiris, le Zouz, and Calypso - whose comforting presence has been a constant source of support.

And finally…



Abbreviations

Bioinformatics terms

DE Differentially Expressed
DEA Differential Expression Analysis
DEG Differentially Expressed Gene
GO Gene Ontology
NGS Next Generation Sequencing
SQL Structured Query Language
TC Time Course

Biological terms

ATP Adenosine Triphosphate
cDNA Complementary DNA
TCA Tricarboxylic acid cycle
FACS Fluorescence-Activated Cell Sorting
DNA Deoxyribonucleic Acid
mRNA messenger RNA
mtDNA mitochondrial DNA
OXPHOS Oxidative phosphorylation
RNA Ribonucleic Acid
ROS Reactive oxygen species
rRNA ribosomal RNA
SCA SpinoCerebellar Ataxia
CA Cerebellar Ataxia

Molecular techniques

RNA-seq Bulk ribonucleic Acid sequencing
sc-RNA-seq Single Cell Ribonucleic Acid sequencing
sn-RNA-seq Single Nuclei Ribonucleic Acid sequencing
sc/n RNA-seq Single Cell/Nuclei Ribonucleic Acid sequencing

Statistics

NB Negative Binomial
FDR False Discovery Rate
Log2FC Log2-fold change

Table of contents

Affidavit ***1***

Liste des publications et participations aux conférences ***3***

Résumé ***5***

Abstract **6**

Acknowledgment ***7***

Abbreviations ***8***

Table of contents ***9***

List of figures **10**

Introduction **11**

Bulk RNA-sequencing 11
Single cell/nuclei RNA-sequencing 13
Mitochondria 17
The cerebellum 21
Ataxia and SCA 25
Methods in web tool development 27

Thesis aims **30**

Results **32**

Chapter 1 - RNfuzzyApp **32**

Chapter 2 - mitoXplorer 3.0 **47**

Chapter 3 - ataxiaXplorer **74**

Chapter 4 - mitoXplorer 2.0 and AnnoMiner **77**

Discussion **110**

References ***119***

List of figures

Figure 1 : Bulk RNA-seq workflow 10

Figure 2 : Bulk RNAseq vs single cell/nuclei workflow 12

Figure 3 : Single cell RNA-seq workflow 14

Figure 4 : Mitochondrial functions relying on cellular interactions 16

Figure 5 : Electron transport chain 18

Figure 6 : Lobes of the cerebellum 19

Figure 7 : The cerebellar architecture 21

Figure 8 : The cerebellum communication through the brainstem 22

Figure 9 : Repeat expansions causing ataxia 24

Figure 10 : Web tool structure 25

Figure 11 : ataxiaXplorer interactome building 75

Figure 12a : SCA1 condition : all genes vs mitogenes in PC subclustering 113

Figure 12b : WT condition : all genes vs mitogenes in PC subclustering 114

Figure 13 : Subpopulations' gene expression of mitogenes vs all data 115

Introduction

Bulk RNA-sequencing

RNA sequencing (RNA-seq) is a genomic method used to identify and quantify mRNA molecules in biological samples. It is a valuable tool for investigating cellular responses as it provides a view of gene expression within a whole organ or tissue. The method is predominantly employed for differential expression analysis (DEA) in between conditions enabling the identification of genes that exhibit significant expression variations among conditions, shedding light on specific biological processes and providing insights into the underlying biological features of an organism or tissue.

After collecting the biological sample, RNA is extracted, often accompanied by mRNA enrichment, mRNAs undergo fragmentation, the attachment of adaptor sequences and the construction of a cDNA library, which is crucial for the sequencing procedure. Subsequent to this conversion, each cDNA molecule undergoes Next Generation Sequencing (NGS) to generate short sequences from either one end (single-end sequencing) or both ends (paired-end sequencing) (Figure 1a).

The preprocessing pipeline for bulk RNA-seq begins with a quality control step, typically applied to the raw FASTQ files using FastQC [1]. If issues such as low-quality reads or adapter contamination are detected, trimming can be executed using tools like Cutadapt [2] or Trimmomatic [3]. Subsequently, the reads must be aligned to the genome, a task often accomplished using algorithms such as STAR [4].
The alignment step is essential as it assigns each sequenced read to a specific DNA region associated to a gene. Following this, the assigned fragments need to be quantified to estimate gene abundance. Tools like featureCounts [5] or HTseq [6] are commonly employed for this purpose. The outcome of this step is a raw read count matrix, which, however, requires further processing before it can yield meaningful biological insights.

From the preprocessed data (raw count matrix), bulk RNA-seq data undergo several steps to complete the analysis (Figure 1b). Normalization of read counts accounts for differences in sequencing depth and other biases, often using methods like TPM (Transcripts Per Million), RPKM (Reads Per Kilobase Million), or DESeq2's [7] normalization. Next, differential expression analysis (DEA) is performed to identify dysregulated genes across the compared conditions using statistical tests.

RNA-seq data typically exhibit overdispersion, where the variance exceeds the mean. The negative binomial (NB) distribution is a generalization of the Poisson distribution that can model this overdispersion by incorporating an extra parameter, the dispersion, to account for the variability. While the Poisson distribution assumes that the mean and variance are equal, the NB distribution can model situations where the variance is greater than the mean.

For each gene, the read counts across samples are modeled using an NB distribution, where the probability of observing a specific count depends on the mean expression level and the dispersion. The mean expression level and dispersion are estimated from the data. Read counts are normalized to account for differences in sequencing depth and RNA composition. This normalization can involve scaling factors or size factors to ensure fair comparisons across samples.

To identify differentially expressed genes, statistical tests are conducted comparing the read counts between conditions. The NB model provides a framework for these tests, typically using methods like the Wald test. By using the NB distribution, RNA-seq analysis can more accurately reflect the underlying biological and technical variability in the data, leading to more reliable and interpretable results.

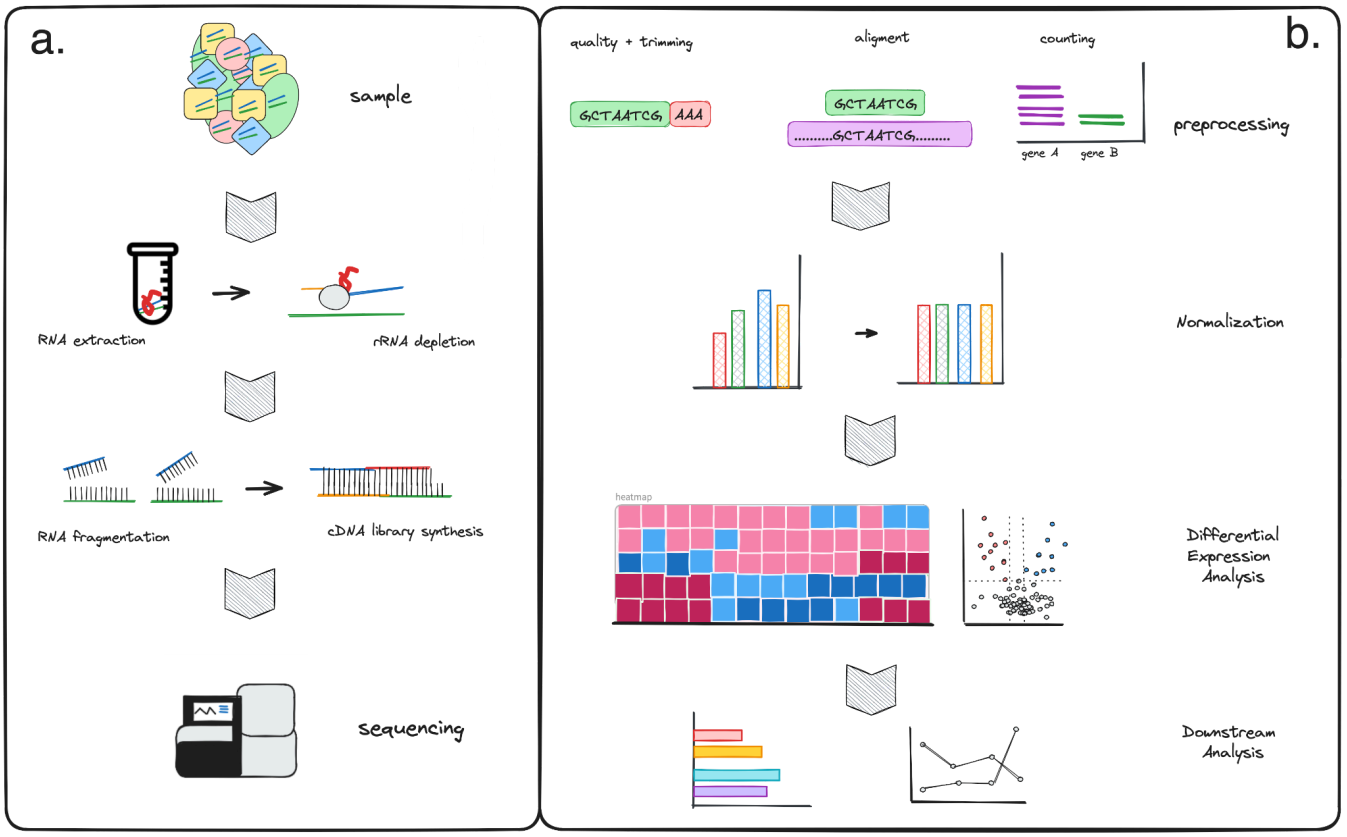


Figure 1 : Bulk RNA-seq analysis workflow. ***a.*** *Data collection to sequencing.* ***b.*** *From preprocessed data, they underdo normalization, followed by the identification of differentially*

expressed genes. Finally, some downstream analysis can be further done like functional enrichment or time series analysis.

Using bulk RNA sequencing, one can also look at temporal processes, for instance during development, disease progression or to observe the temporal response to a treatment. These objectives can be effectively pursued by analyzing and studying data that captures temporal dynamics or time-varying patterns. Using time-series data, the dynamic nature of such biological phenomena can be captured, gaining valuable insights and helping to understand the underlying mechanisms involved.

Time course data can be periodical [8], or not [9] and their respective analysis methods differ [10] as one emphasizes recognizing and modeling regular cycles such as the cell cycle or the circadian clock; or on detecting trends, changes, and irregular patterns.
As in bulk RNA-seq data, regular and irregular patterns can co-exist in a dataset. To analyze time-series data, clustering algorithms are very useful and popular. Data clustering is an essential technique in data mining that involves grouping data elements into distinct clusters or groups. The process aims to ensure that elements within the same cluster share a high degree of similarity, while elements in different clusters exhibit significant dissimilarities [11]. This approach facilitates the identification of underlying patterns, structures, or relationships within the data, enabling effective data exploration and analysis.
One of the clustering methods is fuzzy clustering, also known as soft clustering. Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster. This is in contrast to hard clustering, where each data point belongs to exactly one cluster [12]. Fuzzy clustering is particularly useful when data points may naturally lie on the boundaries between clusters or when you want to account for uncertainty in data classification.

In summary, bulk RNA-seq offers great opportunities and ways to explore and analyze data but unfortunately, this approach hinders the direct examination of individual cells, which are the fundamental units of biology.

Single-cell/nuclei RNA-sequencing

Traditional bulk RNA sequencing measures the average gene expression from a mixture of thousands or millions of cells. However, this approach masks the variability between individual cells within the population (Figure 2). By contrast, single cell RNA-seq (scRNAseq) enables the profiling of gene expression patterns in individual cells, allowing for a more detailed understanding of cellular diversity. By analyzing gene expression at the single-cell level, scRNA-seq provides valuable insights into cell types, cellular heterogeneity, developmental processes, disease mechanisms, and more. It has revolutionized fields like developmental biology, immunology, neurobiology, and cancer research, opening new avenues for understanding complex biological systems [13-15].

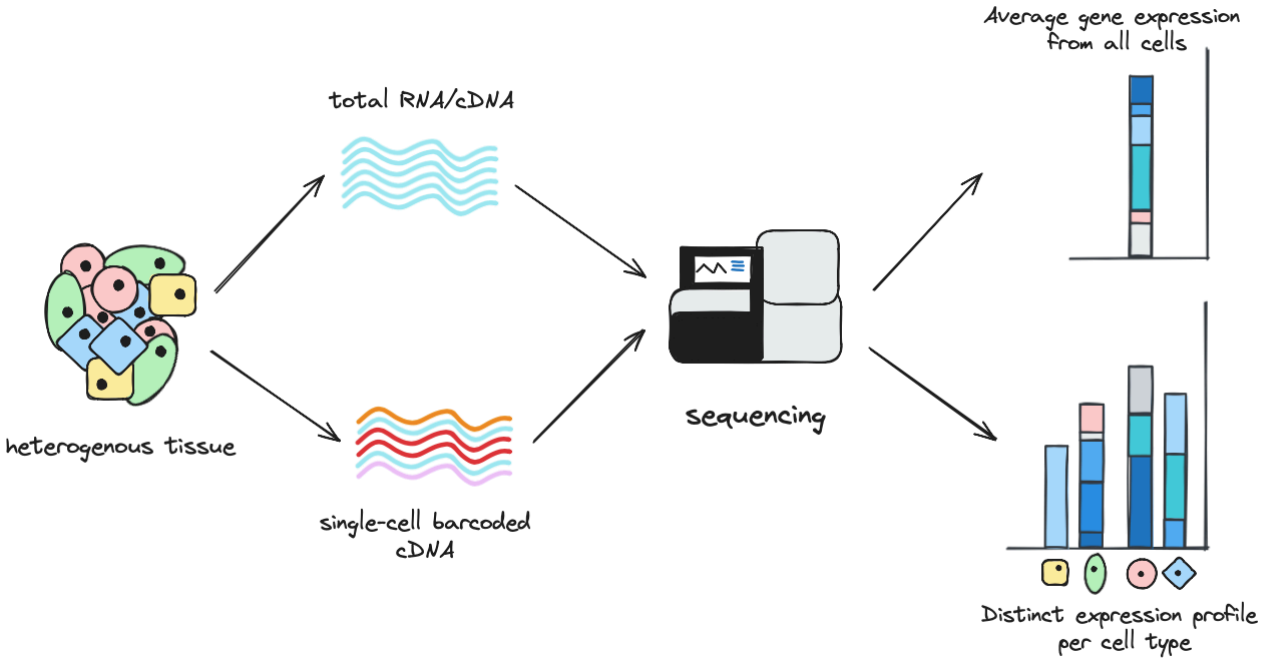


***Figure 2 : bulk RNA-seq vs sc/n-RNA-seq general workflows**. This figure represents the main steps of both pipelines with bulk specificities on top and single cell on bottom. In sample preparation for sequencing, DNA from each cell is barcoded in sc-RNA-seq while DNA is mixed together in bulk. This allows after sequencing to identify populations of cells and obtain a distinct expression profile for each cell type.*

In 2009, the first study utilizing single-cell RNA sequencing was published, paving the way for new investigations [16]. scRNA-seq is nowadays affordable and already frequently used, and projects like Tabula muris [17] and sapiens [18] were a great initiative to explore mouse and human tissues on a single-cell level.

Single-cell RNA sequencing is a cutting-edge technique used in molecular biology and genomics to analyze gene expression at the single-cell level. It provides

researchers with a powerful tool to explore the heterogeneity and complexity of cell populations within tissues or organisms. Single nuclei RNA seq (sn-RNA-seq) is also a powerful technique used to study gene expression, but at the level of individual cell nuclei. While single cell methods require complete dissociation of tissues into individual cells, in single nuclei studies, the nuclei of individual cells are isolated, which can be less damaging for certain tissues [19].

The single cell protocol consists of dissociating the cells from the collected sample and isolating individual cells before sequencing (Figure 3.a). To isolate cells several methods are available :

Fluorescence-Activated Cell Sorting (FACS) [20, 21] is a specialized type of flow cytometry where cells are labeled with fluorescent antibodies or dyes that bind to specific cell surface markers or intracellular components. As cells pass through a laser beam, the fluorescent markers emit light at specific wavelengths. Detectors capture the emitted light, and the data is analyzed to determine various cell characteristics, such as size, granularity, and fluorescence intensity.

Droplet based methods [22,23] involve the encapsulation of cells in microdroplets of a fluid. These droplets act as individual reaction chambers and each of them contains one cell along with reagents for the desired reactions or analyses.

To start a preprocessing pipeline (Figure 3.b), sequenced sc-RNA-seq data must first undergo a quality control check using FastQC[1]. This step includes filtering to remove low-quality reads, reads with adapter contamination, and reads that are too short. Next, the data is preprocessed by assigning reads to individual cells using cell-specific barcodes (demultiplexing) and aligning the reads to a reference genome or transcriptome with aligners such as STAR [4]. Deduplication involves using Unique Molecular Identifiers (UMIs) to identify and eliminate duplicate reads originating from the same mRNA molecule, ensuring more accurate quantification of gene expression. Finally, read counting maps each unique read to its corresponding gene for each cell.
Like preprocessed bulk RNA seq data, preprocessed sc-RNA-seq data also undergo several steps. Beginning with filtering of mt-encoded genes for instance it is important to figure out if a cell is correctly sequenced, and if the cell is not dead (unless it is part of the study). Processing includes next a normalization step of gene expression to account for differences in sequencing depth and capture efficiency between cells. Methods like TPM (Transcripts Per Million), CPM (Counts Per Million), or specialized scRNA-seq normalization methods (e.g., SCTransform provided by Seurat [24,25] or normalize_total from scanpy [26] where data are stored in anndata objects [27]) are used. To help understand data and filter out noise by projecting the data into a lower-dimensional space, one or several reduction of dimensions can be used.

- Principal Component Analysis (PCA) reduces dimensionality by finding the principal components (directions of maximum variance) in the data, thereby emphasizing the most significant variations and reducing noise [28].
- t-SNE (t-Distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) are commonly used to create 2D or 3D plots - even if they do not preserve crucial aspects of the high-dimensional structure [29] - that reveal the relationships between cells, such as clusters of similar cells.

Clustering is a fundamental step that groups cells into clusters based on their gene expression profiles using clustering algorithms like k-means, hierarchical clustering, or graph-based methods (e.g., Louvain or Leiden algorithms [30]).The resulting clusters need then to be annotated based on the expression of known marker genes. Marker genes databases like PanglaoDB [31] provide a precious help identifying cell type and several tools exist to help annotate the clusters like ClustifyR [32].

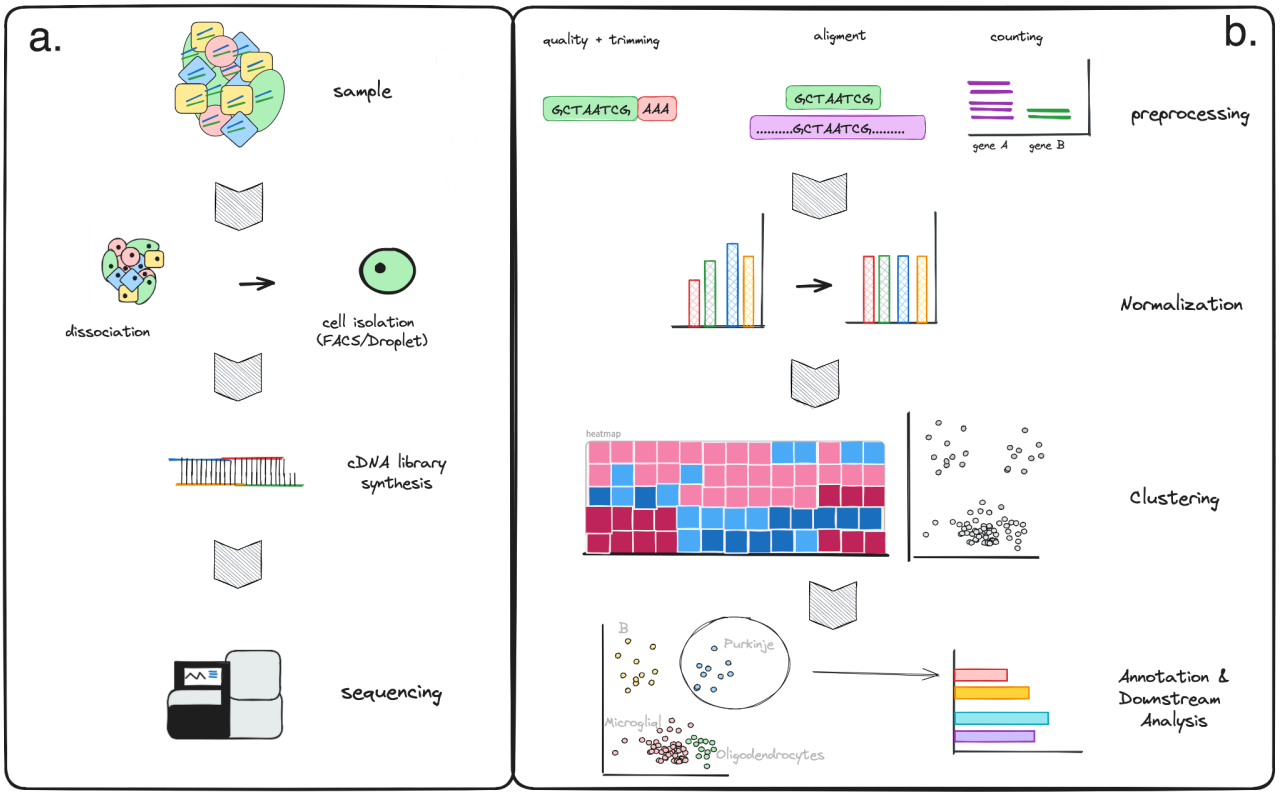


Figure 3 : Single-cell RNA-seq analysis workflow. a. *From data collection to sequencing.* ***b.*** *From preprocessed data, they underdo normalization, followed by the identification of the different cell types as well as their annotation. Finally, some downstream analysis can be further performed like functional enrichment.*

Finally, downstream analysis can be performed like pathway enrichment and functional analysis to understand the biological processes and pathways active in different cell populations. Sub clustering of a specific population is also possible and

could be useful to identify groups of cells having specific roles within one cell population. For instance, Purkinje cells are at the outer layer of the cerebellum but there are 3 lobes in the cerebellum with different functions and it could be possible to identify specificities of each group of cells.

Following clustering and annotation, sc/n-RNA-seq data could then be analyzed in two ways to understand a cell-type specific contribution:
First, a classical analysis of the sc/n RNA-seq data can be performed, for instance by identifying sub-types of cells and looking at gene expression at single-cell level and over cell clusters [33]. Second, cell-type specific data can be extracted, creating pseudo-bulk data for each annotated cell type from sc/nRNA-seq data for subsequent analysis using classical bulk-based methods. This second approach will allow us to compare the different cell types constituting the sequenced tissue.

Mitochondria

Mitochondria are double-membraned, rod-shaped organelles found in the cells of eukaryotic organisms, which include plants, animals, fungi, and protists. They contributed to the origin of eukaryotic cells from evolved free-living prokaryotes in a process called endosymbiosis, a process that also chloroplasts underwent [34,35].
Initially independent prokaryotes, these organelles became intricately integrated with their eukaryotic hosts, eventually evolving into permanent subcellular compartments. They possess their own genetic material, mitochondrial DNA (mtDNA), which is distinct from the nuclear DNA of the host cell. The presence of mtDNA is critical for the synthesis of some mitochondrial proteins and most of the genes encode for key components of aerobic respiration and translation [36-38].

Often referred to as the "powerhouses of the cell," mitochondria play a crucial role in energy production and various other cellular functions. The outer membrane acts as a barrier, while the inner membrane is highly folded into structures called cristae, which significantly increase the surface area available for chemical reactions. The innermost compartment, known as the matrix, is enclosed by the inner membrane and contains mtDNA with only 37 genes, ribosomes, and enzymes [39].

While the exact number of proteins found in mitochondria is currently unknown and may vary slightly between tissues and species [40], proteomic data from various organisms suggests that mitochondria contain approximately 1000 proteins [41]. From these proteins, we can identify a set of corresponding genes, that we will call mito-genes that encompass all genes, both nuclear- and mitochondrial-encoded, whose protein products localize to and carry out their functions within the mitochondria.

Mitochondria maintain cellular health. However, their specific functions can differ between cells based on the varying requirements for energy production, metabolism, and signaling [42-46]. The internal morphology of mitochondria is instrumental in determining the available space for enzymes involved in various functional pathways [47]. The primary function of mitochondria is to generate ATP, which serves as the energy currency of the cell. The entire process, known as oxidative phosphorylation, is highly efficient in extracting energy from nutrients. Mitochondria also play a crucial role in apoptosis, the programmed cell death essential for maintaining tissue homeostasis and eliminating damaged cells.

They release cytochrome c, a protein involved in apoptosis, from the intermembrane space into the cytoplasm, triggering a cascade of events leading to cell death. They are also involved in reactive oxygen species (ROS) defense and regulating intracellular calcium levels as they can sequester and release calcium ions, influencing various cellular processes such as muscle contraction, cell signaling, and metabolism [48] (Figure 4).

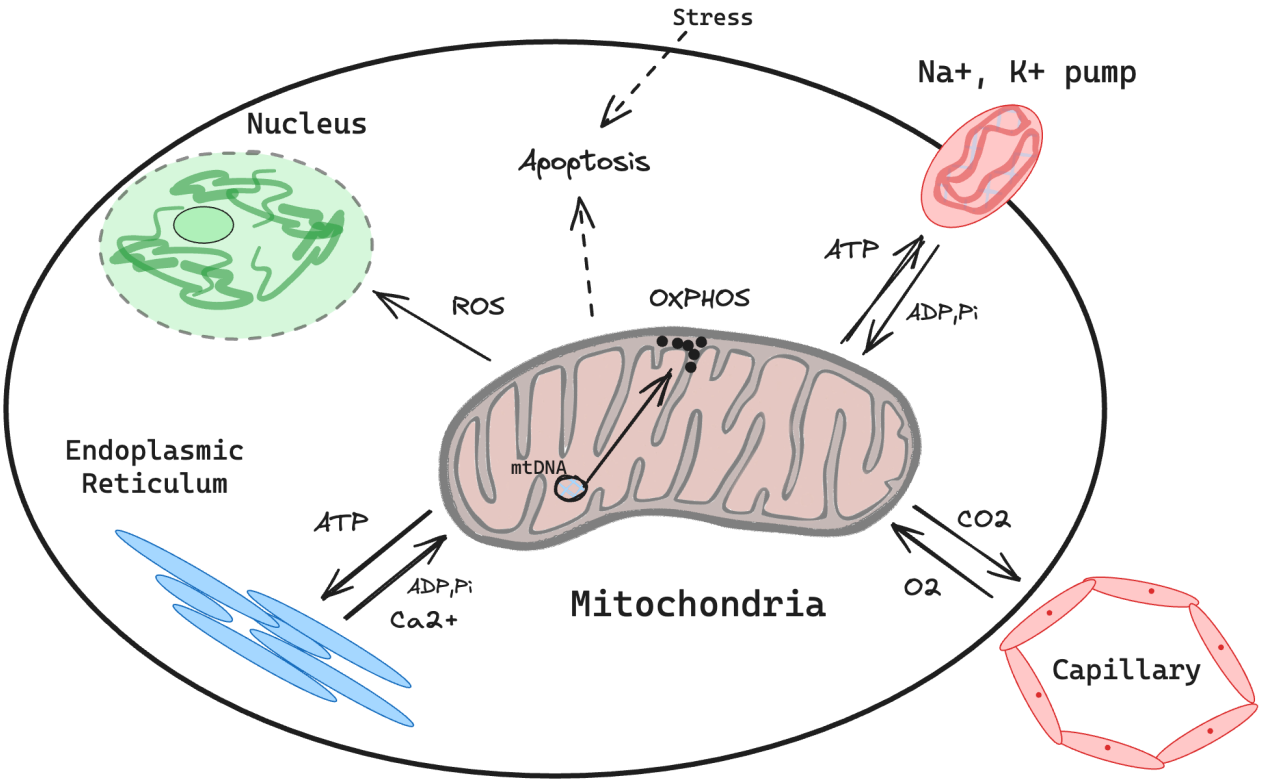


Figure 4 : Mitochondrial functions relying on cellular interactions.
This figure represents some of primary mitochondrial interaction in an eukaryotic cell including the NA+/K+ pumps, maintaining electrochemical gradient, the endoplasmic reticulum as Ca2+ storage site, the intrinsic pathway for apoptosis, the production of ROS and the exchange of oxygen and carbon dioxide with capillaries.

Some examples here in the previous figure are the Na+/K+-ATPase which maintains the electrochemical gradient across the plasma membrane by pumping three sodium ions out of the cell and two potassium ions into the cell, consuming ATP in the process [49]. The endoplasmic reticulum serves as a storage site for intracellular calcium [50]. T The intrinsic apoptotic pathway is one of the main pathways through which programmed cell death can be initiated and is activated by internal cellular stress signals such as DNA damage, oxidative stress, and other forms of cellular injury [51]. Under stress conditions, such as pathogen attack, the production of ROS can increase. ROS act as signaling molecules that activate defense responses, including the expression of defense-related genes and the strengthening of cell walls [52,53]. Capillaries are the smallest blood vessels in the body and are responsible for the exchange of oxygen, nutrients, and waste products between the blood and tissues. Oxygen diffuses from the capillaries into the surrounding tissue cells where it is utilized by mitochondria for oxidative phosphorylation, the process by which ATP is produced [54].

Furthermore, the tricarboxylic acid (TCA) cycle, also known as the citric acid cycle or Krebs cycle, is a series of chemical reactions that occur in the mitochondria of cells. It serves as a central pathway for the oxidation of carbohydrates, fats, and proteins to produce energy in the form of ATP. Through a series of enzymatic reactions, citrate is converted back into oxaloacetate, producing ATP, NADH, and $FADH_2$ along the way [55,56].

In the TCA cycle, various metabolic reactions require NAD+ as a cosubstrate, which is converted to NADH. Additionally, $FADH_2$ is an oxidoreduction cofactor produced during the cycle. NADH and $FADH_2$ both donate electrons to Complex I (I) and Complex II (II) of the electron transport chain (ETC) in the inner mitochondrial membrane [57] (Figure 5). These electrons are subsequently transferred to Complex III (III) via the electron carrier Coenzyme Q (CoQ), and finally to reducing the final electron acceptor Complex IV (IV) with the assistance of cytochrome c. Complexes I and III have been identified as key sites of reactive oxygen species (ROS) production within the ETC [58] and oxygen is used as the final electron acceptor in the ETC. Adequate oxygen supply from capillaries is crucial for efficient ATP production and cellular respiration .

Moreover, coenzyme Q10 (CoQ10) is a vital molecule involved in energy production within mitochondria [59]. It acts as an electron carrier in the electron transport chain, a process crucial for generating ATP. Deficiency in CoQ10 can impair mitochondrial function, leading to decreased energy production and cellular dysfunction.
It is particularly important to speak about iron sulfur (Fe-S) clusters which are essential cofactors for a wide variety of cellular processes, including electron transport in the mitochondrial respiratory chain, as well as DNA replication and repair

[60]. Fe-S clusters are integral components of complexes I, II, and III of the electron transport chain and are found in DNA polymerases and helicases, crucial for DNA synthesis and maintenance.

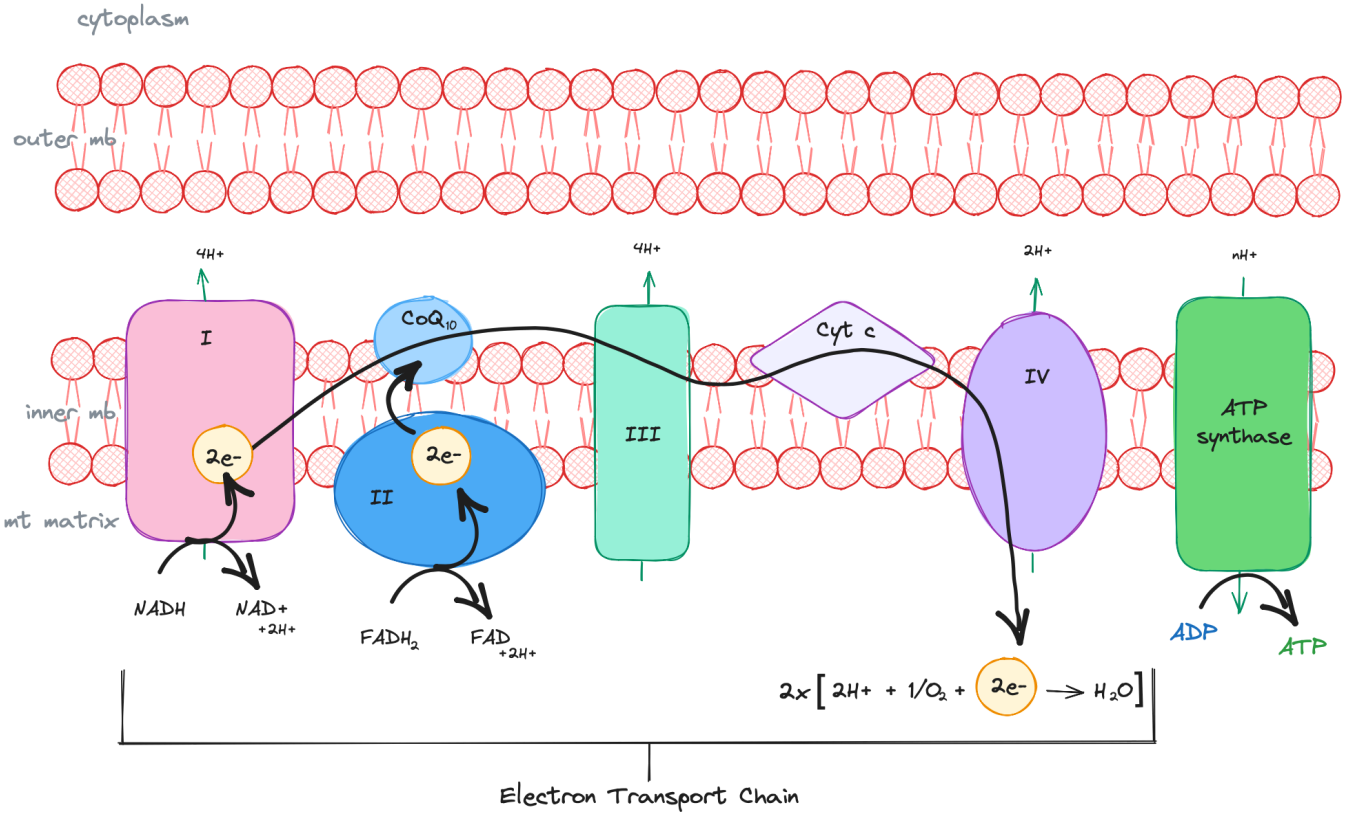


Figure 5 : Electron transport chain. *Transfers electrons from electron donors like NADH and FADH2 to oxygen, the final electron acceptor, through a series of redox reactions. As electrons move through these complexes (I, II, III, and IV), energy is released and used to pump protons from the mitochondrial matrix into the intermembrane space, creating a proton gradient. This gradient drives the synthesis of ATP as protons flow back into the matrix through ATP synthase (Complex V). The process concludes with the reduction of O_2 to form H_2O.*

In summary, mitochondria are multifunctional organelles vital for energy production, apoptosis regulation, calcium homeostasis, and the synthesis of certain cellular components. Their complex structure and functions make them essential contributors to the overall health and functionality of eukaryotic cells. Their dysfunction is increasingly recognized as a key factor in the pathogenesis of various neurodegenerative diseases [61] such as Alzheimer's [62] and Ataxias [63]. With their central role to the cell, they could also be a potential therapeutic target [64].

The cerebellum

The cerebellum is a vital part of the brain located at the back of the skull. It constitutes approximately 10% of the total brain volume but contains more than half of the neurons in the brain. Despite its relatively small size compared to other brain regions, the cerebellum plays a crucial role in motor control, coordination, and certain cognitive functions.

The cerebellum is primarily involved in coordinating voluntary muscle movements and maintaining balance and posture. It is constituted of three lobes. The anterior lobe is involved in fine tuning of body movements and the posterior lobe influences the initiation, planning and coordination of body movements. It also defines the strength, direction and scope of the movement. Finally, the flocculonodular lobe governs eye movements and body equilibrium during stance and gait (Figure 7).

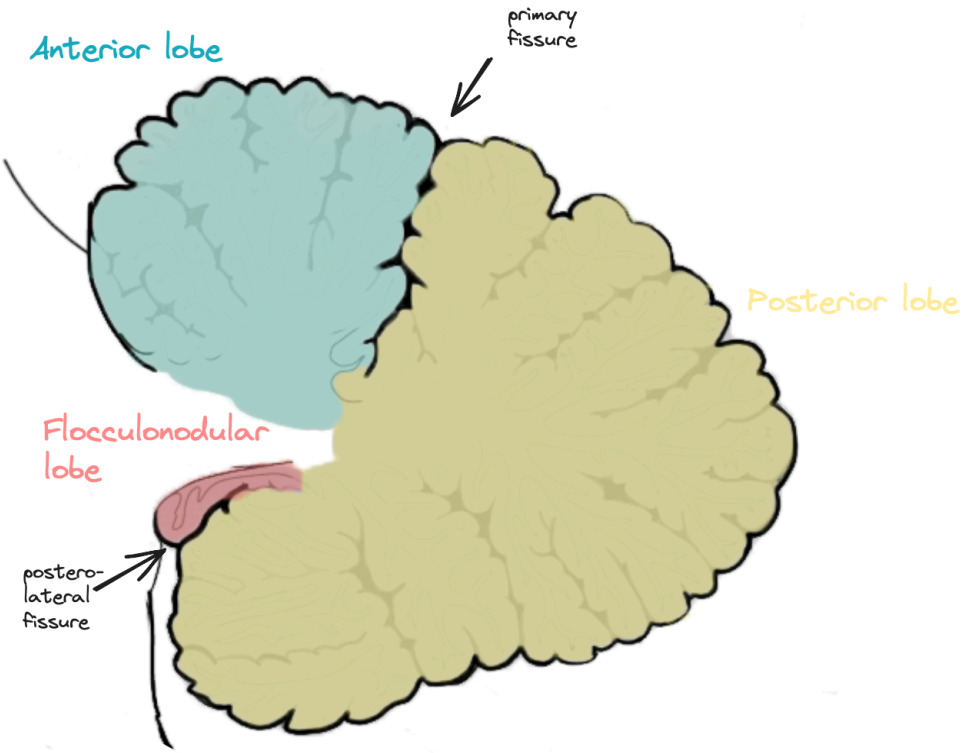


Figure 6 : The 3 lobes of the cerebellum*. Each lobe of the cerebellum is separated by a fissure. The primary fissure cuts the anterior from the posterior lobe, while the posterolateral fissure unveils the flocculonodular lobe from the posterior one.*

The cerebellum communicates with other parts of the brain [65-67], including the cerebral cortex and brainstem, through a network of nerve fibers. Communication

pathways involving the cerebellum are intricate and involve several key structures.
It receives inputs from sensory systems and integrates all information to fine-tune and modulate motor commands from the brain, ensuring smooth and precise movements. It also receives information about the position of different body parts and helps in maintaining balance and posture. It contributes to activities such as walking, running, and other complex motor skills.

It is also essential for motor learning and skill acquisition [68]. It plays a key role in adapting and refining movements based on experience and feedback, allowing individuals to improve their motor skills over time.

The cerebellar cortex has 3 distinct layers, including the granular layer rich in cells, the Purkinje cell monolayer and the molecular layer consisting of interneurons and glia [69] . Together, they contain the seven main types of cells : Purkinje cells, golgi cells, granule cells, Lugaro cells, unipolar brush cells (UBCs), basket cells and stellate cells [70,71] (Figure 8). The molecular layer is characterized by numerous synaptic connections, featuring granule cell axons and Purkinje cell dendrites. This layer also contains stellate and basket cells superficially. Stellate cells typically have short dendrites that connect with a limited number of Purkinje cell dendrites. In contrast, basket cells possess more extensive dendritic networks, allowing them to interact with a greater number of Purkinje cells. Both stellate and basket cells receive stimulating signals from parallel fibers and, in turn, exert an inhibitory effect on Purkinje cells.

The deeper granular layer is densely populated with small granule cells and larger Golgi cells. Granule cells, among the brain's tiniest neurons, typically extend from mossy fibers. Their axons reach into the outer molecular layer, where they split into T-shaped parallel fibers. These fibers then form synapses with the dendrites of Purkinje, basket, and stellate cells.

Finally, the Purkinje cell layer (middle layer) is composed of a single row of large Purkinje cells. These cells have the largest cell bodies in the cerebellum and possess a distinctive appearance. The dendritic branches of Purkinje cells extend into the molecular layer above, while their axons descend through the granular layer beneath, ultimately forming synapses with the deep cerebellar nuclei.

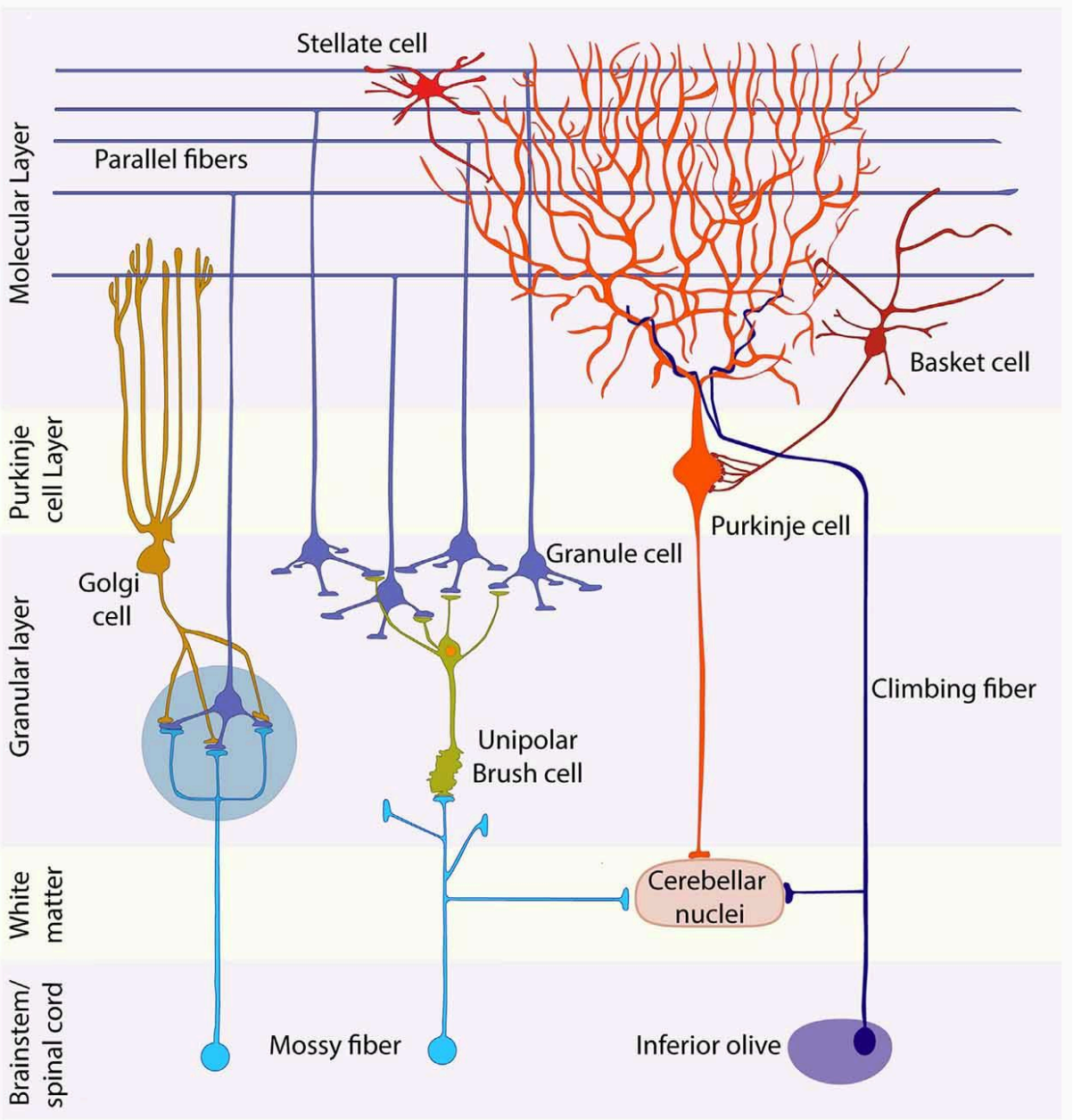


***Figure 7 : The cerebellar architecture**. The cerebellar cortex is organized into three distinct layers vertically followed by the white matter and the brainstem and spinal cord : The outermost molecular layer contains the large dendrites of Purkinje cells, axons of granule cells (parallel fibers), climbing fibers, basket and stellate cells. Below is the Purkinje layer that consists of the cell bodies (somas) of Purkinje neurons. And the innermost granular layer includes granule cells and other interneurons, such as Golgi cells. Figure taken from Conzales et al, 2020 [72].*

Purkinje neurons are a crucial component of the cerebellum [73], they receive inputs from the cerebellar cortex, where various sensory and motor information is

processed. Purkinje neurons also play a significant role in modulating and fine-tuning motor commands but also learning, integrating inputs from the cerebellar cortex and transmitting refined signals to deeper cerebellar nuclei. Most importantly, they are known as the sole output of the cerebellum and this is a reason why they are considered to be the most functionally important neurons of the cerebellum.

The cerebellum is constituted of several primary actors like the red nucleus. The red nucleus is a structure found in the midbrain and is known for its role in coordinating limb movements [74]. It receives inputs from the cerebellum, particularly the deep cerebellar nuclei, through the superior cerebellar peduncle. As the interposed nucleus receives signals from Purkinje neurons [75] and is involved in the regulation of muscle tone and the coordination of limb movements, the inferior olivary nucleus, located in the medulla is a major source of climbing fibers that synapse on Purkinje neurons [76,77]. It forms a loop with the cerebellum, contributing to the adjustment and refinement of motor commands as climbing fibers convey error signals and is important in motor learning and coordination. They work together to integrate sensory information, modulate motor commands, and contribute to the precise control of voluntary movements (Figure 9).

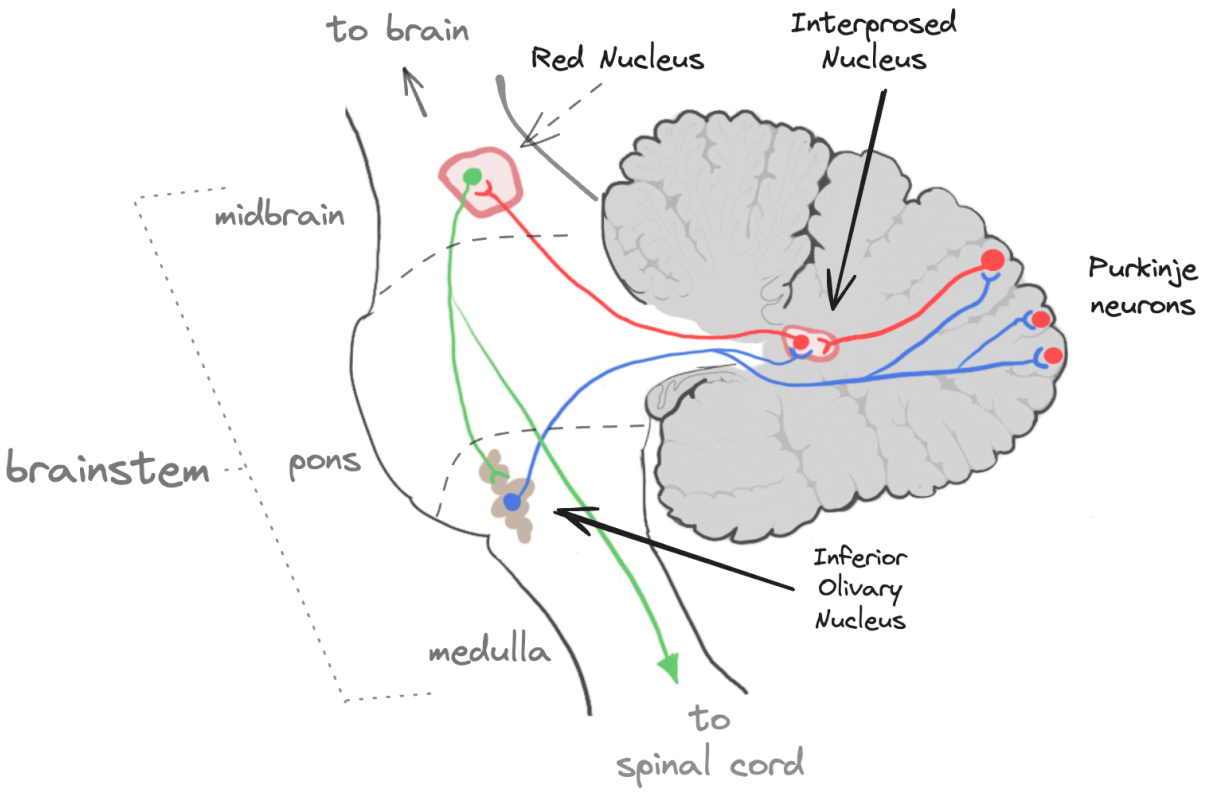


Figure 8 : The cerebellum communication through the brainstem*. This figure depicts the communication pathways of the cerebellum, highlighting its interactions with primary actors and the brainstem.*

Damage or dysfunction of the cerebellum can lead to a range of motor and cognitive impairments, such as ataxia, tremors, and difficulties with balance [78]. Its intricate connections and precise control mechanisms contribute to the overall smooth functioning of the nervous system.

Ataxias and Spinocerebellar Ataxias (SCAs)

Cerebellar ataxia represents a class of neurological disorders characterized by the degeneration, damage, or dysfunction of the cerebellum, the brain region primarily responsible for the coordination of voluntary motor movements [79]. This condition manifests as a disruption in the finely tuned motor control and balance functions mediated by the cerebellum. While ataxias can have diverse causes, they all share a common set of symptoms characterized by uncoordinated movements, instability, and impaired motor skills.

Spinocerebellar ataxias (SCA) constitute a group of cerebellar ataxias with an autosomal dominant inheritance pattern [80], manifesting through the progressive deterioration of coordination in various aspects such as hand movements, mobility, eye function, speech, and balance. With over 40 subgroups falling under the SCA umbrella, the underlying causes can vary, often stemming from degeneration in the cerebellum [81].

Within this diverse spectrum, six subgroups (SCA1, SCA2, SCA3, SCA6, SCA7, and SCA17) share a common mutation, collectively known as polyglutamine spinocerebellar ataxias (polyQ SCA) that can occur at different positions in the genome; as well as other SCA that share other repeat expansions (Figure 6). This distinction arises from a shared genetic mutation involving the expansion of a CAG repeat sequence, which translates into a polyglutamine tract. The resulting polyQ SCA is attributed to the atrophy of Purkinje cells [82]. Expansion of polyQ tracts in a protein typically leads to the misfolding of the protein, and to its aggregation. Next to the loss of function of the polyQ-protein, also the protein aggregates lead to cellular toxicity and neurodegeneration [83].
Notably, the prognosis for individuals with polyQ SCA is influenced by the severity of the disease, with shorter survival rates observed as the condition progresses in its intensity [84].

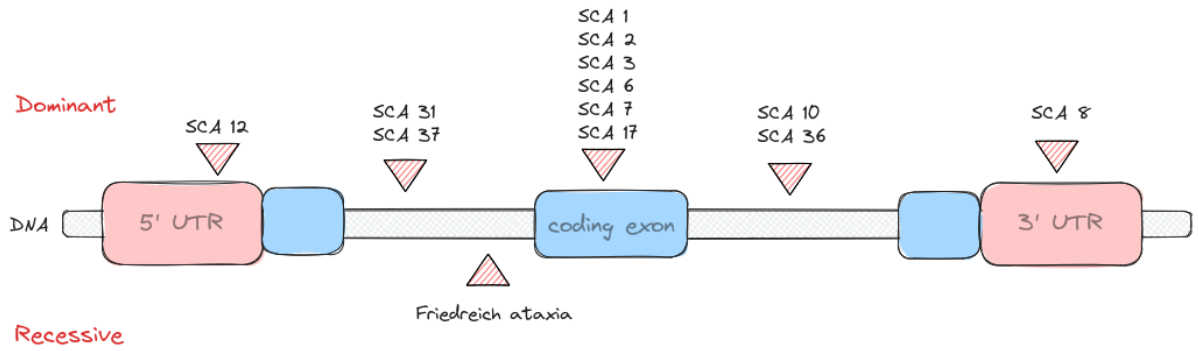


Figure 9 : Repeat expansions causing ataxia. *Friedreich Ataxia is a poly glutamate (polyE) recessive ataxia , about the dominant we can find some in untranslated region like : SCA 12 caused by a CAG repeat in 5' UTR and SCA 8 a CTG repeat in 3' UTR. Some ataxias are caused by intronic repeats like SCA 36, a GGCCTG repeat, SCA 31 a TGGAA repeat insertion, SCA 37 an ATTTC repeat and SCA 10 is an ATTCT repeat. Finally the ataxias that are caused by translated repeats : SCA 1,2,3,6,7,17 have CAG repeats.*

Mutations in genes involved in Fe-S cluster biogenesis and maintenance have been linked to various forms of cerebellar ataxia such as Friedreich's Ataxia (FRDA) [85], a recessive poly glutamate (polyE) form of CA. This type is one of the most frequent forms of ataxias and is caused by mutations in the FXN gene, encoding frataxin, a mitochondrial protein crucial for Fe-S cluster assembly. Frataxin deficiency leads to impaired Fe-S cluster formation, mitochondrial dysfunction, oxidative stress, and eventually, neurodegeneration affecting the cerebellum.
This is also important in other ataxias caused by mutations in genes such as ISCU, NFS1, and IBA57 involved in Fe-S cluster biosynthesis [86]. These mutations can disrupt Fe-S cluster assembly, leading to deficient enzymatic activities in the mitochondria and increased cellular stress.

Impaired Fe-S cluster biogenesis affects the mitochondrial electron transport chain, leading to reduced ATP production and increased reactive oxygen species (ROS), causing cellular damage. The accumulation of free iron due to defective Fe-S cluster assembly catalyzes the formation of ROS, damaging cellular components. Neurons, especially in the cerebellum, are highly susceptible to energy deficits and oxidative stress, resulting in cell death and consequently, in ataxia.

Some studies have reported that CoQ10 deficiency may play a role in certain types of SCAs, including SCA3 [87,88]. Research has shown that CoQ10 supplementation can improve mitochondrial function and reduce oxidative stress, which may potentially benefit individuals with CoQ10 deficiency-related SCAs [89].

Mitochondrial diseases are a group of disorders characterized by dysfunctional mitochondria, leading to impaired energy production and cellular damage. Since mitochondria play a central role in energy metabolism, disorders affecting them often exhibit symptoms similar to those seen in SCAs, including muscle weakness, ataxia, and neurological dysfunction [90-92].

Due to the overlap in symptoms and potential involvement of mitochondrial dysfunction in some forms of SCAs, treatments targeting mitochondrial function are being considered [93-95].

In summary, while CoQ10 deficiency may contribute to certain types of SCAs, treatments targeting mitochondrial function, such as CoQ10 supplementation, are explored due to the shared symptoms and potential involvement of mitochondrial dysfunction in these disorders. However, further research is needed to fully understand the underlying mechanisms and optimize treatment approaches for SCAs.

Methods in web development

From a programming perspective, a web-based tool can be split into two components: the back-end and the front-end (Figure 10).
HTML (Hypertext Markup Language), CSS (Cascading Style Sheets), and JavaScript are the core technologies for building the front end of web applications. Each has a specific role and they work together to create interactive, visually appealing, and well-structured websites. While HTML provides the basic structure of web pages and CSS controls the presentation of HTML elements, JavaScript adds interactivity to web pages. It allows for dynamic content updates, user inputs validation, interactive forms, and other real-time interactions without requiring a page reload. It also allows users to manipulate data and represent the data thanks to graphical libraries like D3 (Data-Driven Documents) [96].

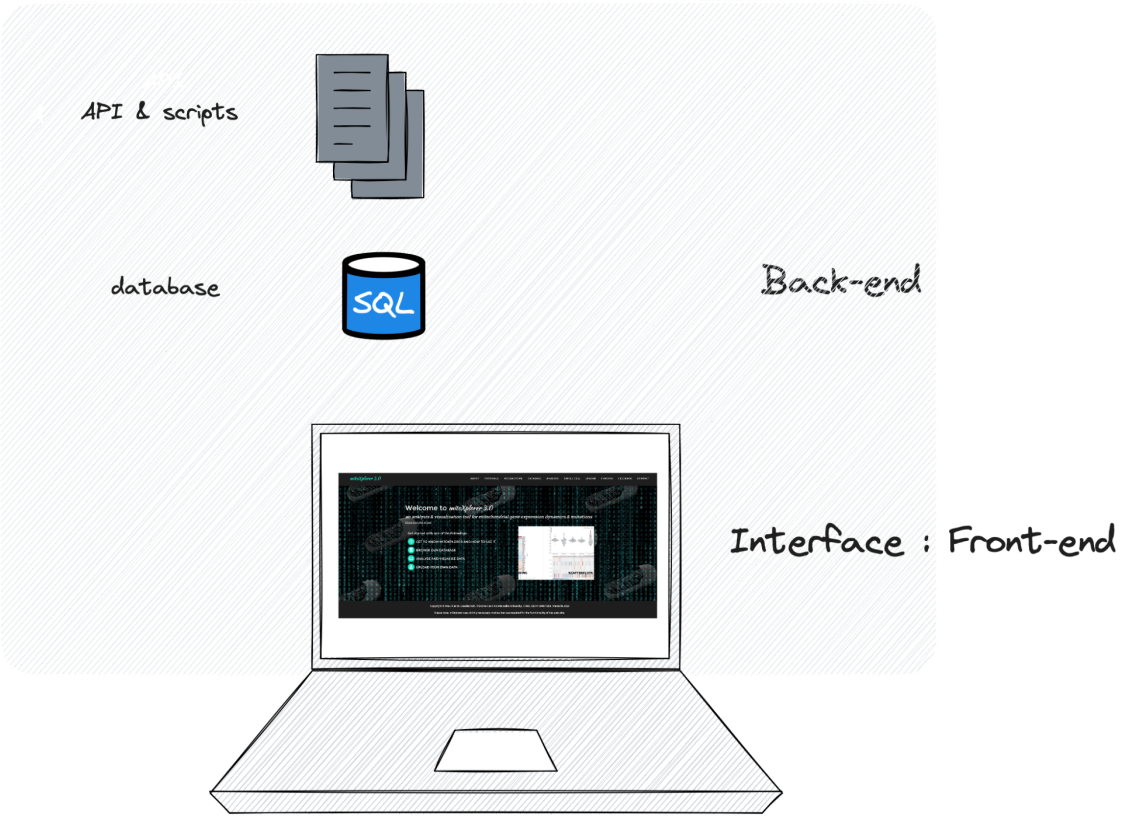


Figure 10 : Web tools structure. *Front-end and back-end communicate. Front end contains the interface, suggests options and displays results while back-end contains databases, application programming interfaces (API) and scripts mandatory to manipulate the data.*

The backend development, or server-side, contains all processes related to calculations and storage. Application programming interfaces (APIs) and related operations allow the website to query the database and process the retrieved data.

Databases can be broadly classified into two types, relational (SQL) and non relational (NoSQL). Structured Query Language (SQL) is used for executing search queries, retrieving data, and manipulating tables, hence these databases are also known as SQL databases. MySQL [97] and SQLite [98] are among the most common relational databases. They are open-source relational database management systems (RDBMS) that are widely used for building web-based applications. It's known for its speed, reliability, and ease of use. For instance, the mitoXplorer platform [99,100] uses a relational database like MySQL due to the small size and tabular organization of its datasets.

Non-relational databases, NoSQL (Not Only SQL) databases, do not use the tabular schema for data storage. Instead, they employ various storage models tailored to the type of data being stored. MongoDB [101] is a popular NoSQL database.

Generally, NoSQL databases are better for handling large datasets, as their structure facilitates faster search queries and allows the database to be distributed across multiple storage units. An important choice about the back-end development is the language. JavaScript is supported by all major web browsers, making it an essential language for web development. Its widespread adoption ensures that web applications built with JavaScript can reach a vast audience without compatibility issues. As discussed earlier, JavaScript enables dynamic content on web pages, allowing developers to create interactive elements such as animations, form validations, and responsive user interfaces that can react to user inputs without requiring page reloads. It also supports asynchronous programming through callbacks, promises, and async/await, which is crucial for performing tasks like API calls and handling real-time updates efficiently. These aspects make JavaScript a powerful and attractive language for developers aiming to build robust, interactive, and scalable web applications thanks to all the available libraries.

PHP is a popular open-source server-side scripting language primarily used for web development. It executes on the server, generating HTML that is then sent to the client's web browser. This allows for dynamic content generation and interaction with databases.

Python 3 was chosen for all calculations related to single cell data. Python is a highly structured programming language, which makes it easy to use and understand. It is a general-purpose programming language and is not limited to data analysis. This means it can be used for a wide range of applications, from web development to automation and scientific computing, providing a unified toolset for different stages of data handling and analysis.

Thesis Aims

The aim of my thesis was to develop 3 user-friendly and interactive tools to analyze bulk- as well as single-cell RNA-sequencing data.

First, due to the lack of easy-to-use and visual tools to analyze bulk RNA-seq data, including such from time-series analysis, I developed the **RNfuzzyApp** tool [102], a R-shiny tool that allows users to analyze bulk RNA-seq data starting from a read count matrix with just a few clicks. RNfuzzyApp offers several methods for data normalization and differential expression analysis; as well as it automates the otherwise manual process of an optimal analysis of time-series data using the Mfuzz algorithm [103]. As it is released as an R-shiny app, it offers an alternative to other, Galaxy [104] -based web-based analysis methods [105] for standard bulk RNA-seq analysis, and represents the first visual interface for fuzzy-clustering based time-series analysis, offering automation of the entire process from cluster number selection to the functional enrichment analysis of results. RNfuzzyApp is freely available at: https://gitlab.com/habermann_lab/rna-seq-analysis-app/-/tree/master

Second, I created the updated version of the visual data mining web platform, **mitoXplorer 3.0** and started the development of **ataxiaXplorer,** which allows to mine single cell data in a context of mitochondria and ataxias, respectively.

These two projects involved a collaboration with the Puccio Lab (INMG Lyon), who are working on spinocerebellar Ataxias (SCAs). Like other neurodegenerative diseases, these are cell-type specific, resulting from the degeneration of the Purkinje cells in the cerebellum. Our aim was to provide a cell type specific solution for data exploration. Based on the previously published mitoXplorer platform [99,100], which allows to analyze bulk RNA-seq data from a mitochondrial perspective, one aim was to develop a sister platform of mitoXplorer, ataxiaXplorer, with a gene-set specific to several types of ataxias. As ataxias are also considered mitochondrial diseases, we decided to first make mitoXplorer single-cell ready, and then to create its sister platform, ataxiaXplorer, with an ataxia-specific gene-set and interactome.

In order to mine scRNA-seq data and understand cell-type specific contributions to disease onset and progression, we first created pseudo-bulk data for each annotated cell type from the scRNA-seq data for subsequent analysis using classical bulk-based methods, including the classical interfaces of mitoXplorer. Secondly, we decided to offer classical scRNA-seq data analysis from a mitochondrial, and then also ataxia-centric point of view, allowing us to identify sub-cell types with specific mitochondrial- or ataxia- features. We opted for including this type of analysis, as cell-type specific mitochondrial variance has been observed previously. Moreover, the cerebellum has 3 lobes and Purkinje cells are found in all three of them. it has been shown previously that these might have different behaviors [106]. Therefore, this

type of analysis offers an interesting new way of better understanding these cell-type specific variants and contributions to disease phenotypes, such as Ataxias. Therefore, we offer new and unique features in mitoXplorer that allow to re-cluster cell types in a mitochondria-centric way, based on the mito-genes stored in the mitoXplorer database. I then developed new interactive interfaces in mitoXplorer 3.0 to analyze each subpopulation with respect to mito-processes, as well as to individual differences in gene expression in each subpopulation for each mito-process.

Data pre-processing is required prior to mitoXplorer 3.0 analysis. For pre-processing of sc/sn-RNA-seq data, I developed a command-line based script, **scXplorer**, that processes and prepares the single cell/nuclei data for MitoXplorer 3.0. scXplorer creates a pseudo-bulk from the original single cell/nuclei data and also creates a gene-set specific sc/snRNA-seq count matrix that only contains mito-genes. As input, the user only needs to provide a pre-analyzed and annotated single-cell sequencing data file in h5ad or rds format.

Finally, together with the team of Helene Puccio, we assembled and annotated the ataxia interactome, based on gene expression data of 4 different types of SCAs, which will serve as the main interactome for the ataxiaXplorer platform.

Results

Chapter 1 : RNfuzzyApp

Bulk RNAseq analysis is very common but can be repetitive as different methods and thresholds can be applied. In order to offer an easy to use analysis tool from count matrix, I developed the RNfuzzyApp as a R-shiny app, so that non-expert users can redo their analyses at will by testing different methods for normalization and differential expression analysis. Next to classical bulk RNA-seq analysis, RNfuzzyApp was also designed to help users to run time-series analysis using the soft-clustering algorithm Mfuzz. This new implementation offers methods to help choose the number of clusters for soft clustering and runs Mfuzz clustering automatically several times to test stability of the clusters, a process that is usually executed manually. Additionally, it offers the possibility to perform enrichment analysis of a chosen cluster. Finally, it allows to search for orthologs and automatically convert gene IDs.
RNfuzzyApp is the first tool that provides a web-based interface for an automated soft-clustering of time-series data using the Mfuzz algorithm. It helps users select cluster numbers, performs iterative Mfuzz clustering steps to test the stability of clusters and automates enrichment analysis of chosen clusters.
In the manuscript, we discuss the algorithm, and show use-cases for classical bulk RNA-seq analysis, as well as time-series analysis with the integrated and automated Mfuzz algorithm.

I was the sole developer, as well as main tester of RNfuzzyApp.

Check for updates

SOFTWARE TOOL ARTICLE

REVISED RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis [version 2; peer review: 3 approved]

Margaux Haering, Bianca H Habermann

Aix-Marseille University, CNRS, IBDM UMR 7288, The Turing Centre for Living systems, Marseille, 13009, France

v2 **First published:** 26 Jul 2021, **10**:654
https://doi.org/10.12688/f1000research.54533.1
Latest published: 12 Nov 2021, **10**:654
https://doi.org/10.12688/f1000research.54533.2

Abstract

RNA sequencing (RNA-seq) is a widely adopted affordable method for large scale gene expression profiling. However, user-friendly and versatile tools for wet-lab biologists to analyse RNA-seq data beyond standard analyses such as differential expression, are rare. Especially, the analysis of time-series data is difficult for wet-lab biologists lacking advanced computational training. Furthermore, most meta-analysis tools are tailored for model organisms and not easily adaptable to other species.

With RNfuzzyApp, we provide a user-friendly, web-based R shiny app for differential expression analysis, as well as time-series analysis of RNA-seq data. RNfuzzyApp offers several methods for normalization and differential expression analysis of RNA-seq data, providing easy-to-use toolboxes, interactive plots and downloadable results. For time-series analysis, RNfuzzyApp presents the first web-based, fully automated pipeline for soft clustering with the Mfuzz R package, including methods to aid in cluster number selection, cluster overlap analysis, Mfuzz loop computations, as well as cluster enrichments. RNfuzzyApp is an intuitive, easy to use and interactive R shiny app for RNA-seq differential expression and time-series analysis, offering a rich selection of interactive plots, providing a quick overview of raw data and generating rapid analysis results. Furthermore, its assignment of orthologs, enrichment analysis, as well as ID conversion functions are accessible to non-model organisms.

Keywords

RNA-seq, data normalization, data visualization, differential expression analysis, time-series analysis, soft clustering, Mfuzz, R shiny

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 12 Nov 2021	✓ view	✓ view	
version 1 26 Jul 2021	? view	? view	✓ view

1. **Oliver Hahn**, Stanford University, Stanford, USA
2. **Rhonda Bacher**, University of Florida, Gainesville, USA
3. **Anita Grigoriadis**, King's College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the RPackage gateway.

Corresponding author: Bianca H Habermann (bianca.habermann@univ-amu.fr)

Author roles: Haering M: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Habermann BH**: Conceptualization, Formal Analysis, Funding Acquisition, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the French National Research Agency with ANR grant ANR-18-CE45-0016-01 MITO-DYNAMICS awarded to BHH.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Haering M and Habermann BH. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Haering M and Habermann BH. **RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis [version 2; peer review: 3 approved]** F1000Research 2021, **10**:654 https://doi.org/10.12688/f1000research.54533.2

First published: 26 Jul 2021, **10**:654 https://doi.org/10.12688/f1000research.54533.1

REVISED **Amendments from Version 1**

We have tried to clarify a few points that seemed unclear in the paper. Furthermore, we have created a detailed user manual, including detailed instructions how to install the software and how to run its different functions.

We now also provide test data files (which are used both, in the paper, as well as the manual) so that researchers can test our software and also verify the correctness of the format of their data.

We have meanwhile also updated RNfuzzyApp, and added new functionalities.

1) It is now possible to compare two datasets, even if more than 2 datasets were initially uploaded. For this, the Filter data function has been introduced.
2) Groups are now assigned automatically. The user has to follow the guideline on how to format the column names (condition1_replicate1, condition1_replicate2, etc.).
3) It is now possible to directly load DEG gene lists into the enrichment function of RNfuzzyApp, making the analysis more streamlined.

Any further responses from the reviewers can be found at the end of the article

Introduction

The development of next generation sequencing (NGS) methods has boosted the rapid generation of large datasets and RNA sequencing (RNA-seq) has become the standard for performing robust transcriptional profiling and thus quantifying gene expression in various contexts. Next to the comparison of two conditions, the generation of time-series RNA-seq data has become amenable and popular, allowing to monitor the gene expression dynamics over a process such as development, ageing or cancerogenesis. While web-based, user-friendly R shiny apps have become available recently for differential expression analysis and data visualization of RNA-seq data,[1–7] the analysis of time-series data within R remains largely command-line based and therefore challenging for bench scientists without programming knowledge.

We here present RNfuzzyApp, a user-friendly, web-based R shiny app with an intuitive user interface for the full workflows of differential expression analysis, as well as time-series analysis of RNA-seq data. RNfuzzyApp provides an interface for easy and fast data normalization and differential analysis using several methods, a variety of interactive plots for a quick overview of data and results, and an easy-to-use interface for the complete pipeline of time-series expression analysis using the fuzzy clustering algorithm Mfuzz.[8] In addition, RNfuzzyApp offers ID conversion, orthology assignment and enrichment analysis using gprofiler2.[9] We show the usability of RNfuzzyApp on two examples: an RNA-seq dataset of the ageing limb muscle of mouse, as well as developmental time-series RNA-seq data of the *Drosophila melanogaster* leg.

Methods

Implementation

RNfuzzyApp was built in R (V.4.0.4) using the Shiny framework. The app currently depends on the following R packages: *shiny, shinydashboard, shinycssloaders, shinythemes, shinyWidgets, shinyBS, rmarkdown, plotly, dplyr, RColorBrewer, utils, tidyr, devtools, cluster, DESeq2*,[10] *edgeR*,[11] *TCC*,[12] including *baySeq*,[13] *heatmaply*,[14] *gprofiler2*,[9] *Mfuzz*,[8] as well as the package *e1071*. As a basic feature and to allow users to upload any identifier for analysis, ID conversion is included, using the *gprofiler2* package.

Operation

RNfuzzyApp can be launched locally from any computer with R (version 4.0.4 or higher) installed and will run in any web-browser. As RNfuzzyApp auto-installs all required R-packages, there exist no additional software requirements. Installation instructions are also available. All interfaces and plots of RNfuzzyApp are highly interactive, allowing users to visualize data in real-time as well as to interact efficiently with the data and plots.

Workflows of RNfuzzyApp

The general workflow of RNfuzzyApp is shown in Figure 1. It can be divided into two independent parts: 1) a complete workflow for differential expression analysis of RNA-seq data (Figure 1a); and 2) a complete workflow for the clustering of RNA-seq using the soft clustering algorithm Mfuzz (Figure 1b).

Differential expression analysis workflow of RNfuzzyApp

The differential expression analysis workflow of RNfuzzyApp can be divided in three main parts: the upload and visualization of the raw data; the normalization of the data and the differential expression analysis; and finally, enrichment

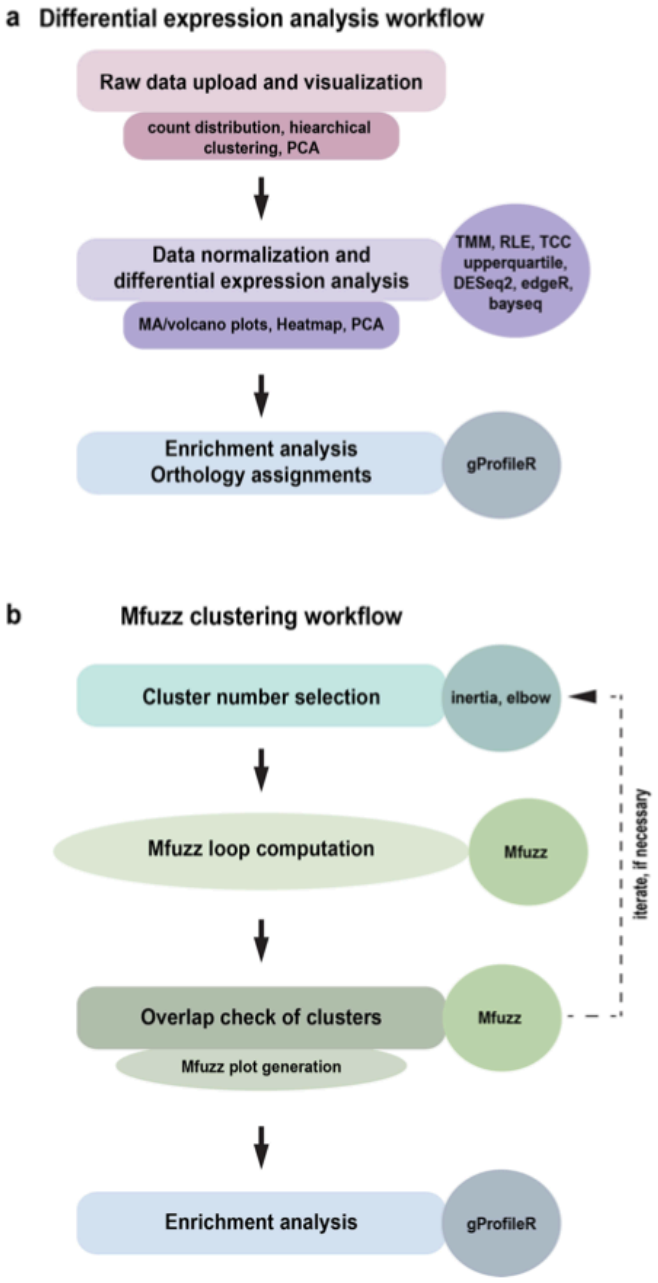


Figure 1. The two RNFuzzyApp analysis pipelines. (a) RNA-seq differential expression analysis workflow with the three main parts: data upload and visualization, data normalization and differential expression analysis, as well as enrichment analysis and the assignment of orthologous genes across species. The types of analyses are shown, as well as the various possible R programs provided for data analysis. (b) Mfuzz workflow for clustering of time-series RNA-seq data. The workflow includes the selection of cluster numbers, checking the overlap of Mfuzz clusters, loop calculations of Mfuzz, Mfuzz plot generation, as well as enrichment analysis of Mfuzz clusters.

analysis of results and orthology assignment (Figure 1a). In each part, several options exist for visualizing the data and thus getting a first-hand impression of the quality of the data, as well as the filters that are applied.

Data upload and visualization of raw gene expression data

Figure 2 shows the RNfuzzyApp start interface, featuring data upload, filtering, as well as raw data visualization possibilities. As a first step, raw read counts need to be uploaded to the app, in the form of a csv count matrix. Data can be filtered for raw read counts (Figure 2c), the resulting summary of the data are interactively updated in the Summary box of

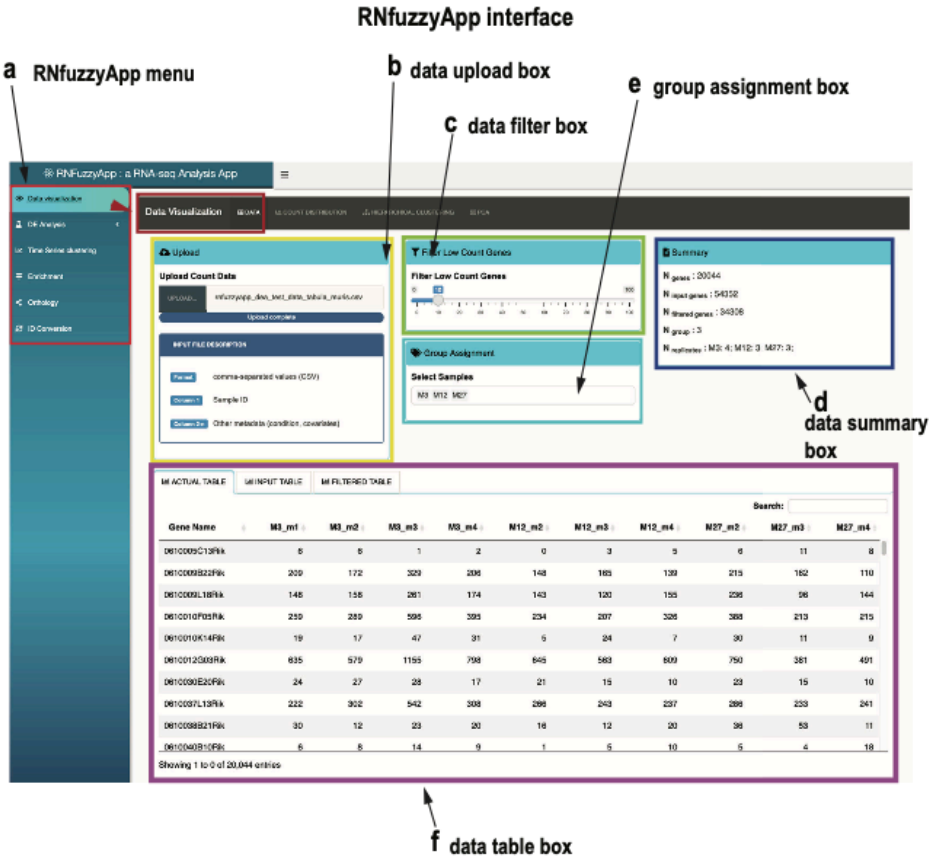


Figure 2. RNfuzzyApp start interface. (a) The RNfuzzyApp menu box is highlighted in red. This box is shown consistently over all interfaces and links to Data visualization, DE analysis, Data normalization and analysis, different visualization possibilities (MA plot, Volcano plot, Heatmap and PCA), Time series clustering (using the pipeline for Mfuzz soft clustering), Enrichment analysis, Orthology assignment, as well as ID-conversion. The main interface shown here belongs to Data visualization (highlighted in dark red). (b) Data upload box, for upload of user-provided data. (c) Data filter box, in which the user can choose to filter out genes with low read counts, (d) in which the identified groups are listed (e.g. wild-type and mutant or different time-points). Group assignment is automatic, so the name format of the samples has to follow a specific pattern. (e) Data summary box. The data shown in this box are updated and in case of filtering and are renewed on the fly. (f) Data table box. Three tables are provided: the actual table, including genes that were not discarded due to filtering; the input table, containing all data uploaded; finally the filtered table, containing genes that were removed due to filtering.

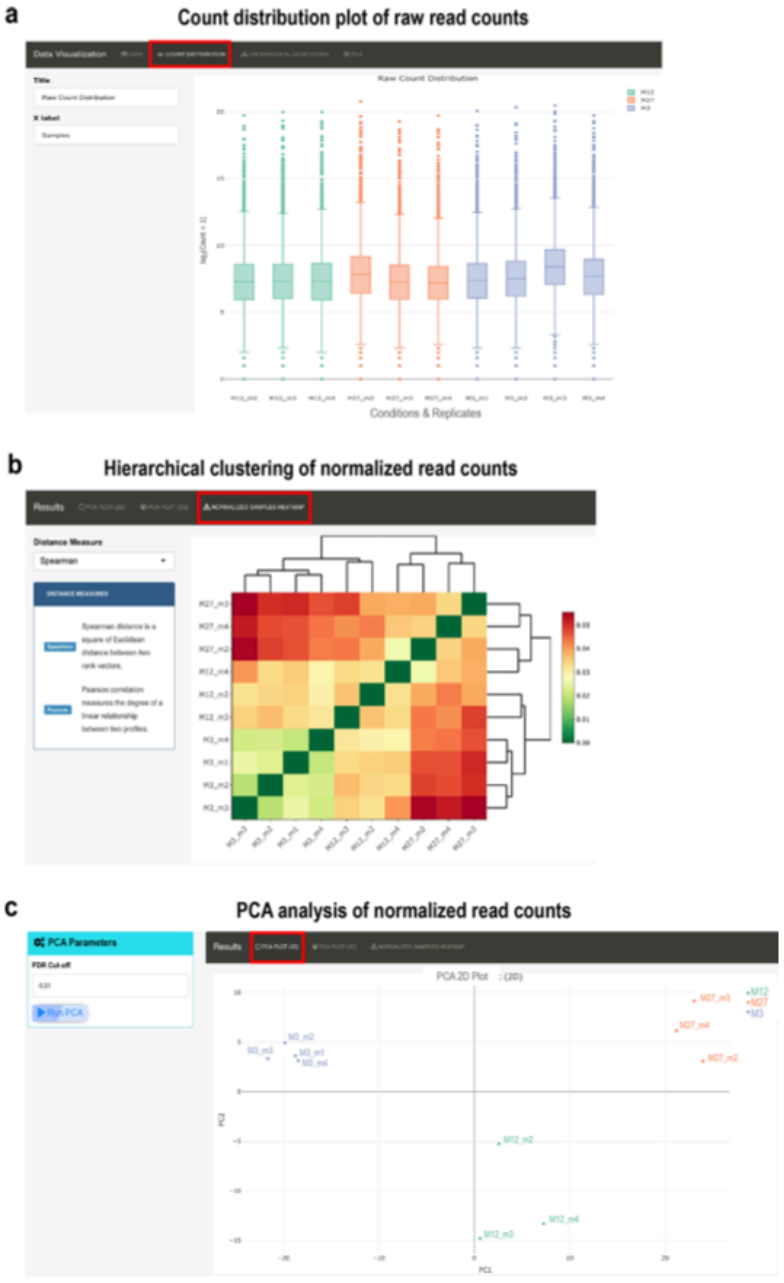


Figure 3. Data visualization plots offered in RNfuzzyApp. We used data from the Tabula muris senis project for demonstration purposes. (a) Count distribution plot of mice from 3 months, 12 months and 27 months. Only replicates from male mice were chosen. Data are grouped by condition. The title of the plot, as well as the X-axis label can be chosen by the user. Raw read counts were chosen for visualization. (b) Hierarchical clustering of normalized read counts. Spearman correlation was used for clustering. (c) PCA plot of normalized read counts. The PCA plot can be visualized in 2D or 3D. DESeq2 was used for normalization of data.

the interface (Figure 2d). Groups can be assigned directly in the interface (Figure 2e). Three tables are available for download: the actual table, containing only the genes that pass the filtering threshold; the original input table; as well as a table containing all genes filtered out due to low read counts (Figure 2f). Raw data can also be visualized (see top of the menu, Figure 2): the count distribution of raw read counts can be visualized (Figure 3a). Moreover, raw read counts can be used for hierarchical clustering, as well as a PCA analysis.

Data normalization and differential expression analysis

RNfuzzyApp offers several packages for data normalization, as well as for differential expression analysis. Normalization can be done using DESeq2; TMM (trimmed mean of M values), RLE (relative log expression) or upperquantile offered by edgeR; finally the TCC package providing TMM or DESeq2 normalization. As for raw read counts, the count distribution, a heatmap for clustering samples (Figure 3b), as well as PCA analysis with a 2D as well as 3D PCA plot (Figure 3c) is available for visualising normalized data. Differential expression analysis can be done using DESeq2, edgeR and bayseq. If more than two conditions are uploaded, normalization and initial differential expression analysis will be done over the entire data set. However, it is often useful to perform pairwise comparison of two conditions or time-points. For pairwise comparisons of two conditions or time-points of larger datasets, a Filter menu is provided. Data resulting from pairwise comparison can be visualized with MA and Volcano plots. All plots are interactive and the user can obtain detailed information about a gene hovering over the dots of the plots. All details on normalization and differential expression analysis can be found in the user manual of RNfuzzyApp.

Clustering of expression data using heatmaply

We wanted to provide a simple way of clustering gene expression data from a limited number of samples, e.g. from a short time-series. To this end, RNfuzzyApp offers a heatmap, generated by *heatmaply*. Figure 4a shows the heatmap of a time-course of 3 time points from the Tabula muris senis project,[15] where the gene expression levels of the replicates per time point are clustered using hierarchical clustering. The user can choose the distance matrix and agglomeration method, as well as the FDR cut-off of genes to include in the clustering. Clustering is done using *hclust* and *cutree* to generate the coloured dendrogram on the heatmap. The genes contained in the different clusters indicated by the colours in the dendrogram are downloadable as a table for further analysis.

Enrichment analysis and orthology assignment.

For enrichment analysis of Gene Ontology (GO-) terms,[16] pathways (Wikipathways,[17] Reactome[18] and KEGG[19]), Human Protein Atlas,[20] CORUM data on protein complexes,[21] and TRANSFAC,[22] the gprofiler2 package is included in RNfuzzyApp. Results are displayed as an image of overall enrichment, as well as a results table. The table together with a bar plot of the enriched term names sorted according to p-value are downloadable by the user (Figure 4b, bar plot of enrichments generated from downloaded table). Gprofiler2 is also used to find orthologs in another species of a user-provided list of genes. To this end, the user simply needs to upload a list of genes, and select the original and the target species.

Complete workflow for fuzzy clustering of time-series data

Fuzzy clustering of time-series expression data is a highly useful technique for analysing temporal data. The Mfuzz package from R was developed for soft clustering of temporal gene expression data.[8] Starting from a count matrix, genes are clustered according to their expression profiles over time. As Mfuzz is a soft clustering algorithm, a gene can in theory be part of more than one cluster. Mfuzz, however, is not straightforward to use for non-experts. First, a number of clusters must be chosen prior to clustering. Second, repeated Mfuzz runs will result in slightly different cluster memberships of genes. A user is therefore well advised to repeat Mfuzz clustering several times to test the robustness of the clustering. The decision, which cluster number is suited for the data then often includes analysis of cluster overlaps, as well as enrichment analyses of clusters and comparative analysis between several chosen numbers of clusters. Several packages exist to help decide on cluster numbers and the entire workflow for a successful Mfuzz clustering can be programmed in R. However, for untrained bench scientists, this is not easily done. We therefore included the complete workflow of Mfuzz soft clustering of time-series expression data in RNfuzzyApp (Figure 1b): first, for choosing the right cluster number, we implemented the *inertia* (using the *hclust* and *dist* packages) and *elbow* (using the *e1071* package) methods. *Inertia* performs hierarchical clustering and plots the dendrogram, indicating the distance steps (height) against the number of clusters. Ideally, a cluster number is chosen when the drop in height gets minimal (Figure 5a). The *elbow* method looks at the total of the within-clusters sums of squares (WCSS) as a function of the number of clusters. The "elbow" shape is formed when WCSS is minimal. These two methods should converge to help choose the right number of clusters. After Mfuzz clustering has been performed, the overlap of clusters can be checked. To do so, the *overlap.plot* function from

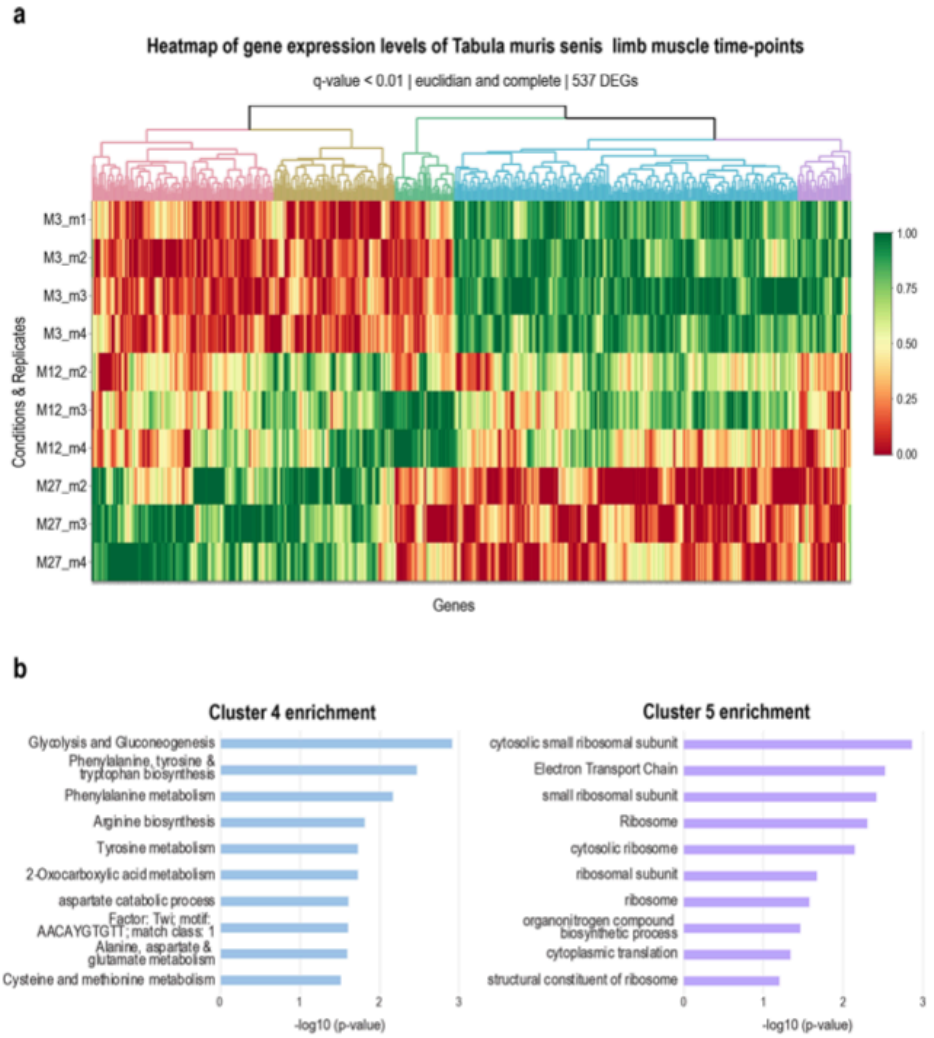


Figure 4. Heatmaply clustering of Tabula muris senis limb muscle data. (a) Heatmap of gene expression levels of the Tabula muris senis limb muscle data. Only significant genes were selected for plotting, with an FDR cutoff of 0.01, resulting in 624 DEGs. These DEGs could be clustered in 5 independent clusters, indicated by different colours in the dendrogram. (b) Enrichment analysis results for cluster 4 and cluster 5 of the limb muscle heat map. In cluster 4, processes related to energy and amino acid metabolism are enriched. Genes belonging to this process have a higher expression level in young versus old mice, suggesting more active metabolism in young muscle cells. In cluster 5, processes related to translation are enriched, whereby associated genes have higher expression levels in young muscle versus old muscle. The minus log10 of the p-value of the enriched process is plotted.

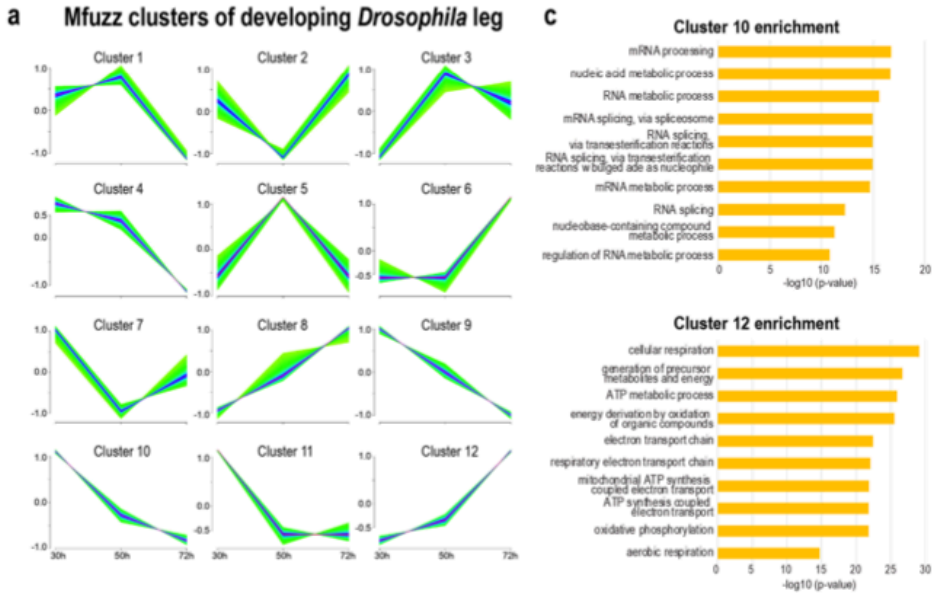
Mfuzz is used and results can be visualized (see Figure 5b). After choosing a suitable number of clusters, Mfuzz is run ten times in a loop to test the robustness of clustering results (Figure 1b). Membership lists of the ten Mfuzz clustering runs can be downloaded and checked for robustness. Plots are generated using the *mfuzz.plot* function and are also downloadable (Figure 6a). Should core clusters be unstable, this entire process can be repeated. Finally, enrichment can be done on Mfuzz cluster gene lists, using gprofiler2 (Figure 6b).



Figure 5. Pre-processing steps required for Mfuzz cluster analysis and cluster number selection. (a) Pre-clustering of data to select the cluster number for Mfuzz time-series clustering. The plot shows the *intertia* drops of the dendrogram. At 12 clusters, the inertia drop was minimal, suggesting that additional clusters would not provide better modelling of the data. (b) Control plot of Mfuzz cluster overlap. A PCA plot is performed with the selected 12 clusters, showing here that no overlap between clusters exists. Data from the developing *Drosophila* leg were chosen for demonstration purposes.

Data preparation

Tabula muris senis[15] limb muscle raw read data (data accessible at NCBI GEO database,[23] accession GSE132040) were taken as is and read into RNfuzzyApp for data processing and differential expression analysis. Raw read data from developing leg muscle[24] (data accessible at NCBI GEO database, accession GSE143430) were first averaged over replicates before reading them into RNfuzzyApp, as Mfuzz does not accept replicates (available as Habermann, Bianca; Haering, Margaux (2021), extended Datatables).



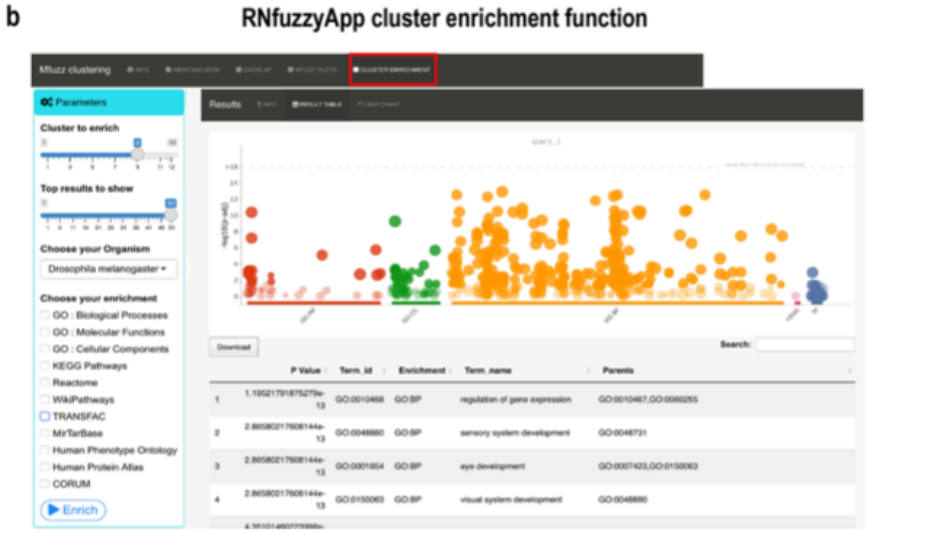


Figure 6. Mfuzz soft clustering analysis. (a) Mfuzz clusters of *Drosophila* leg developmental RNA-seq data. Some similar patterns emerged, with expression profiles peaks early (30 h), mid-phase (50 h) or late (72 h APF (after puparium formation)). (b) Mfuzz clusters have been enriched using gprofiler2. In the parameters box, settings have to be chosen, such as the cluster number submitted for enrichment, the number of results to show, the organism, as well as the databases used for enrichment analysis. The plot on the right-hand side shows the enrichment of cluster 9. (c) Enrichment of the top 10 processes from cluster 10 and cluster 12. In cluster 10, processes related to RNA metabolism, as well as splicing are enriched, associated genes show a decrease in expression over time. In cluster 12, processes related to mitochondrial energy metabolism are enriched, with associated genes showing an increase in expression over time.

Results

RNA-seq analysis of Tabula muris senis bulk RNA-seq data on the ageing limb muscle

We used data from the ageing limb muscle from the Tabula muris senis project (GSE GSE132040[15]). We selected three time-points: 3 months, 12 months and 27 months. We only used samples from male mice. After data upload, we filtered for lowly expressed genes with less than 50 read counts. We then visualized raw read counts of all samples (Figure 3a). After normalization using DESeq2, we compared samples using hierarchical clustering (Figure 3b), which showed that replicates cluster together. We also performed PCA analysis and could confirm that samples from the same time-point cluster together (Figure 3c). We next subjected samples to differential expression analysis using DESeq2, comparing all time-points against each other (see *Extended data*: Tables 1a-c). We found 177 genes differentially regulated between 12 and 3 months, 873 genes differentially regulated between ages 27 and 3 months and 31 genes differentially expressed between ages 12 and 27 months when using an FDR of 0.01 and a log2FC of |0.5|. Enrichment analysis of the lists of differentially expressed genes revealed terms related to translation in young versus adult mice, metabolic and extracellular organisational processes between young and aged mice, as well as between adult and aged mice (*Extended data*: Tables 1d-f).

We next used hierarchical clustering of genes to identify gene groups changing over time. After finding appropriate points to cut the dendrogram from hierarchical clustering using *heatmaply*, we found five different clusters with differing expression levels of genes (Figure 4a): two clusters with high expression levels in aged mice and low in young mice, one with high expression levels in adult mice, and two with high expression levels in young mice and low expression levels in aged mice. We subjected all clusters to enrichment analysis (*Extended data:* Tables 2a-e). For example, cluster 4 and 5, which both show high expression levels in young and low ones in aged mice, had terms related to translation as well as metabolism (energy, amino acids) highly enriched, suggesting more active translation, as well as metabolism in young muscle cells (Figure 4b).

Mfuzz soft clustering of a time-series of RNA-seq data from the developing *Drosophila* leg

We used RNA-seq data from a developmental time course of *Drosophila* leg for soft clustering using the Mfuzz pipeline included in RNfuzzyApp. We used normalized read counts from GEO[25] gene expression dataset GSE143430[24] and uploaded it to RNfuzzyApp. In brief, leg samples had been collected at three stages during pupal development (30, 50 and 72 h APF) and had been subjected to RNA-sequencing. We wanted to analyse the wild-type expression profiles of genes during these three developmental stages and to identify potentially enriched terms and pathways.

We first checked with the *inertia* method the ideal number of clusters (Figure 5a). We chose 12 clusters, as we found no significant change with cluster numbers higher than that. We next tested the overlap of clusters using the *overlap.plot* function of Mfuzz and found good separation of the 12 clusters (Figure 5b). We ran Mfuzz for clustering gene expression profiles and repeated this step 10 times. One of the resulting Mfuzz plots is shown in Figure 6a. We found expression profiles with high expression at 30 h gradually decreasing at 50 h and 70 h (clusters 9 and 10), high expression at 30 h and 50 h, which dereased at 70 h (clusters 1 and 4), expression peaks at 50 h (clusters 3 and 5), low expression levels at 50 h (clusters 2 and 7), low expression levels at 30 h, gradually increasing at 50 h and 70 h (clusters 8 and 12), as well as expression peaks at 70 h (cluster 6). We used all genes of each cluster for enrichment analysis using gprofiler2 within RNfuzzyApp (Figure 6b, *Extended data:* Table 3a-l). We found terms relevant for muscle development enriched in cluster 10 (high expression at 30 h, which gradually decreased at 50 h and 70 h), relating to mRNA metabolic processes and specifically, RNA splicing. RNA splicing has been shown to be essential for muscle cell type specification[26,25] (Figure 6c). Cluster 12, which contained genes with increasing expression levels from 30 h to 70 h was enriched for terms related to mitochondrial function and energy production (Figure 6c). These results are in accordance with earlier observations of increasing electron transport chain components in flight muscle development.[27,26]

Conclusions

We introduced RNfuzzyApp, an intuitive R shiny app for the complete and interactive workflows of RNA-seq, as well as time-course RNA-seq data analysis. RNfuzzyApp includes several algorithms for data normalization and differential expression analysis and offers the possibility for intuitive and interactive data visualization. All data tables and plots are downloadable by the user. While several R shiny apps exist for differential expression analysis, to the best of our knowledge, this is the first web-based, user-friendly R shiny interface for the complete workflow of time-series analysis using the soft clustering app Mfuzz, making RNfuzzyApp the first accessible tool for time-series analysis for wet-lab biologists. We demonstrated the usability of RNfuzzyApp with two examples of RNA-seq data, one from a mouse ageing study of the Tabula muris senis project, and one from the developing leg in *Drosophila melanogaster*.

We chose to offer several packages for normalization, as well as differential expression analysis. This allows the user to exploit several possible combinations of tools for differential expression analysis. Our choice for enrichment analysis in

this version of RNfuzzyApp fell on gprofiler2. There are many software tools available for enrichment analysis. Gprofiler2, however, is available also for non-standard model organisms. Therefore, our app can be used for organisms other than human, mouse, *Drosophila*, *C. elegans* or yeast. Moreover, gprofiler2 allows in addition ID conversion, as well as ortholog assignment and both these functions were made available in RNfuzzyApp. In future releases of RNfuzzyApp, we consider including more enrichment tools, providing a broader spectrum of data to include, such as EnrichR.[28]

To conclude, RNFuzzyApp is an intuitive and easy to use R shiny app that was designed for experimental biologists to enable them to perform RNA-seq and time-series RNA-seq analysis without the need of coding to get a fast overview of their data, results and figures.

Data availability

Underlying data

Gene Expression Omnibus (GEO): Tabula Muris Senis: Bulk sequencing. Accession number GSE132040; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132040.[15]

Gene Expression Omnibus (GEO): Muscle-type specific transcriptomic expression patterns in Drosophila. Accession number GSE143430; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143430.[24]

Extended data

Dryad: Extended data tables to Haering and Habermann, F1000Res, RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. https://doi.org/10.5061/dryad.8pk0p2nnd.

This project contains the following extended data:

- Table 1a: Haering_etal_extendedDatatable_1a_Tabulamurissenis_3vs12m_DEA.txt: results of differential expression analysis (DEA) of Tabula muris senis project (GSE132040), limb muscle, 3 vs 12 months.
- Table 1b: Haering_etal_extendedDatatable_1b_Tabulamurissenis_3vs27m_DEA.txt: results of DEA of Tabula muris senis project (GSE132040), limb muscle, 3 vs 27 months.
- Table 1c: Haering_etal_extendedDatatable_1c_Tabulamurissenis_12vs27m_DEA.txt: results of DEA of Tabula muris senis project (GSE132040), limb muscle, 12 vs 27 months.
- Table 1d: Haering_etal_extendedDatatable_1d_Tabulamurissenis_3vs12m_gpofiler.txt: gprofiler results of Tabula muris senis project (GSE132040), DEA, limb muscle, 3 vs 12 months.
- Table 1e: Haering_etal_extendedDatatable_1e_Tabulamurissenis_3vs27m_gpofiler.txt: gprofiler results of Tabula muris senis project (GSE132040), DEA, limb muscle, 3 vs 12 months.
- Table 1f: Haering_etal_extendedDatatable_1f_Tabulamurissenis_12vs27m_gpofiler.txt: gprofiler results of Tabula muris senis project (GSE132040), DEA, limb muscle, 3 vs 12 months.
- Table 2a: Haering_etal_extendedDatatable_2a_Tabulamurissenis_cluster1_gpofiler.txt: gprofiler results of hierachical clustering of Tabula muris senis project (GSE132040), limb muscle, cluster 1.
- Table 2b: Haering_etal_extendedDatatable_2b_Tabulamurissenis_cluster2_gpofiler.txt: gprofiler results of hierachical clustering of Tabula muris senis project (GSE132040), limb muscle, cluster 2.
- Table 2c: Haering_etal_extendedDatatable_2c_Tabulamurissenis_cluster3_gpofiler.txt: gprofiler results of hierachical clustering of Tabula muris senis project (GSE132040), limb muscle, cluster 3.
- Table 2d: Haering_etal_extendedDatatable_2d_Tabulamurissenis_cluster4_gpofiler.txt: gprofiler results of hierachical clustering of Tabula muris senis project (GSE132040), limb muscle, cluster 4.
- Table 2e: Haering_etal_extendedDatatable_2e_Tabulamurissenis_cluster5_gpofiler.txt: gprofiler results of hierachical clustering of Tabula muris senis project (GSE132040), limb muscle, cluster 5.

- Table 3a: Haering_etal_extendedDatatable_3a_DmLeg_cluster1_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 1.

- Table 3b: Haering_etal_extendedDatatable_3b_DmLeg_cluster2_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 2.

- Table 3c: Haering_etal_extendedDatatable_3c_DmLeg_cluster3_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 3.

- Table 3d: Haering_etal_extendedDatatable_3d_DmLeg_cluster4_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 4.

- Table 3e: Haering_etal_extendedDatatable_3e_DmLeg_cluster5_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 5.

- Table 3f: Haering_etal_extendedDatatable_3f_DmLeg_cluster6_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 6.

- Table 3g: Haering_etal_extendedDatatable_3g_DmLeg_cluster7_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 7.

- Table 3h: Haering_etal_extendedDatatable_3h_DmLeg_cluster8_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 8.

- Table 3i: Haering_etal_extendedDatatable_3i_DmLeg_cluster9_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 9.

- Table 3j: Haering_etal_extendedDatatable_3j_DmLeg_cluster10_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 10.

- Table 3k: Haering_etal_extendedDatatable_3k_DmLeg_cluster11_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 11.

- Table 3l: Haering_etal_extendedDatatable_3l_DmLeg_cluster12_gpofiler.txt: gprofiler resuts of mfuzz clustering of Drosophila leg dataset (GSE143430), cluster 12.

- Table 4: Haering_etal_extendedData_DmdevLeg_GSE143430_mean.txt: mean normalized read counts from GSE143430 to be uplaoded for Mfuzz clustering.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Software availability

Software available from: https://gitlab.com/habermann_lab/rna-seq-analysis-app.

Zenodo: 10.5281/zenodo.5084275 (https://zenodo.org/record/5084275#.YO_e_y0iuik).

Source code available from: https://gitlab.com/habermann_lab/rna-seq-analysis-app.

Archived source code as at time of publication: Zenodo: RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. https://doi.org/10.5281/zenodo.5084275.[29]

License: GNU public license 3.

Author contributions

MA and BHH conceived this project. MA was solely responsible for code implementation, software development and testing. MA and BHH performed data analyses. MA and BHH wrote this manuscript.

Acknowledgements

We want to thank Fanny Chazal, Cedric Maurange for critical input and acting as test user of RNfuzzyApp and Fabio Marchiano for helpful discussions. We thank the CNRS, Aix-Marseille University, as well as the IBDM for supporting this work.

References

1. Su W, Sun J, Shimizu K, *et al.*: **TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data.** *BMC Res Notes.* 2019; **12**: 133–6.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Guo W, *et al.*: **3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists.** *RNA Biol.* 2020: 1–14.
PubMed Abstract | Publisher Full Text

3. Zhang C, *et al.*: **iSeq: Web-Based RNA-seq Data Analysis and Visualization.** *Methods Mol Biol.* 2018; **1754**: 167–81.
PubMed Abstract | Publisher Full Text

4. Gao B, *et al.*: **Quickomics: exploring omics data in an intuitive, interactive and informative manner.** *Bioinformatics.* 2021;
PubMed Abstract | Publisher Full Text

5. Sundararajan Z, *et al.*: **Shiny-Seq: advanced guided transcriptome analysis.** *BMC Res Notes.* 2019; **12**: 432–5.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Gadepalli VS, Ozer HG, Yilmaz AS, *et al.*: **BISR-RNAseq: an efficient and scalable RNAseq analysis workflow with interactive report generation.** *BMC Bioinformatics.* 2019; **20**: 670–7.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Marini F, Linke J, Binder H: **ideal: an R/Bioconductor package for interactive differential expression analysis.** *BMC Bioinformatics.* 2020; **21**: 565–16.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Kumar L, Futschik E, Mfuzz M: **Mfuzz: a software package for soft clustering of microarray data.** *Bioinformation.* 2007; **2**: 5–7.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Kolberg L, Raudvere U, Kuzmin I, *et al.*: **gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.** *F1000Res.* 2020; **9**: 709.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** 2014; **15**: 550–21.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**: 139–40.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Sun J, Nishiyama T, Shimizu K, *et al.*: **TCC: an R package for comparing tag count data with robust normalization strategies.** *BMC Bioinformatics.* 2013; **14**: 219–14.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinformatics.* 2010; **11**: 422–14.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Galili T, O'Callaghan A, Sidi J, *et al.*: **heatmaply: an R package for creating interactive cluster heatmaps for online publishing.** *Bioinformatics.* 2018; **34**: 1600–2.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Schaum N, *et al.*: **Ageing hallmarks exhibit organ-specific temporal signatures.** *Nature.* 2020; **583**: 596–602.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Ashburner M, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** ***Nat.*** *Genet.* 2000; **25**: 25–9.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Martens M, *et al.*: **WikiPathways: connecting communities.** *Nucleic Acids Res.* 2021; **49**: D613–21.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Jassal B, *et al.*: **The reactome pathway knowledgebase.** *Nucleic Acids Res.* 2020; **48**: D498–D503.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Kanehisa M, Furumichi M, Tanabe M, *et al.*: **KEGG: new perspectives on genomes, pathways, diseases and drugs.** *Nucleic Acids Res.* 2017; **45**: D353–61.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Thul PJ, Lindskog C: **The human protein atlas: A spatial map of the human proteome.** *Protein Sci.* 2018; **27**: 233–44.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Giurgiu M, *et al.*: **CORUM: the comprehensive resource of mammalian protein complexes-2019.** *Nucleic Acids Res.* 2019; **47**: D559–63.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Wingender E, Dietze P, Karas H, *et al.*: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res.* 1996; **24**: 238–41.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**: 207–10.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Kao S-Y, Nikonova E, Ravichandran K, *et al.*: **Dissection of Drosophila melanogaster Flight Muscles for Omics Approaches.** *J Vis Exp.* 2019: e60309.
PubMed Abstract | Publisher Full Text

25. Clough E, Barrett T: **The Gene Expression Omnibus Database.** *Methods Mol Biol.* 2016; **1418**: 93–110.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Spletter ML, *et al.*: **The RNA-binding protein Arrest (Bruno) regulates alternative splicing to enable myofibril maturation in Drosophila flight muscle.** *EMBO Rep.* 2015; **16**: 178–91.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Spletter ML, *et al.*: **A transcriptomics resource reveals a transcriptional transition during ordered sarcomere morphogenesis in flight muscle.** *Elife.* 2018; **7**: 1361.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Kuleshov MV, *et al.*: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.** *Nucleic Acids Res.* 2016; **44**: W90–7.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Margaux Hearing M, Habermann BH: **RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis.** *Zenodo.*
PublisherFull Text

Chapter 2 : mitoXplorer 3.0

Mitochondria are important metabolic hubs of almost all eukaryotic cells. Their structure and content will differ depending on their cellular environment and depending on the needs of the host cell. Moreover, many neurodegenerative diseases, such as ataxias, result from perturbation of a specific cell type. It is therefore important to analyze mitochondrial expression dynamics on a cell-type, as well as single-cell basis.
Based on the mitoXplorer 2.0 [100] platform, we developed a novel version of this web-based tool to analyze single-cell or single-nuclei RNA-sequencing data.
We first wanted to be able to create pseudo-bulk data of annotated cell types in sc/sn-RNA-seq data, that allows the user to perform a cell-type specific analysis from a mitochondrial perspective using the classical mitoXplorer 2.0 interfaces. We also wanted to enable a mito-centric analysis of single-cell data and therefore created novel interfaces for the platform to mine mitochondrial expression features on a single-cell level, enabling the user to first identify and then further analyze sub-cell types with different mitochondrial expression profiles. To this end, the sc/snRNA-seq dataset is filtered for mito-genes stored in the mitoXplorer mitochondrial interactome.
For sc/snRNA-seq data pre-processing, I developed a command-line script, scXplorer, that outputs raw pseudo-bulk data based on the annotated cell types, as well as performs classical bulk-based RNA-seq analysis, and formats the output for direct mitoXplorer 3.0 upload. scXplorer also creates the mito-gene based single-cell count matrix that can be uploaded to the web-platform directly.
In mitoXplorer 3.0, several novel interfaces have been developed for single-cell analysis, allowing sub-clustering of a user-chosen cell type, and further analysis of the sub-clusters, including process enrichment, as well as a gene- and process-based analysis, identifying differentially expressed genes between sub-cell types. Data can be visualized with novel, interactive interfaces that allows user-friendly and seamless interaction with the data.
In the manuscript, we demonstrate the usability of the tool using a recently published dataset of Spinocerebellar Ataxia Type 1 (SCA1), which contained data from human patients compared to healthy controls, as well as a time-series study on SCA1 mice from 5 to 30 weeks. When analyzing peudo-bulk data, we identify mito-processes that are perturbed in Purkinje cells in mouse and human. Within Purkinje cells, we identify 3 sub-populations of PCs in mouse wild-type and mutant brains that are distinguished by their mitochondrial expression pattern.

I was the sole developer of all new functionalities of mitoXplorer 3.0, as well as of the scXplorer script for data pre-processing. I prepared, analyzed and interpreted sc/snRNA-seq data and helped write the manuscript.

mitoXplorer 3.0 for exploring mitochondrial dynamics in single-cell RNA-seq data.

Margaux Haering, Bianca H. Habermann

ABSTRACT

Mitochondria are organelles found in most eukaryotic cells, best known for their function in ATP production and their role in cellular metabolism and signaling. It is widely accepted that their structure, composition and function differ across cell types. Moreover, little is known about mitochondrial variability within the same cell type. To truly understand mitochondrial function and dynamics, it is therefore necessary to study individual cell types, as well as variability of cell types on a single-cell level.

Based on our previously developed platform mitoXplorer 2.0, we introduce mitoXplorer 3.0 with new features adapted for analyzing single-cell or single-nuclei sequencing data, focusing only on mitochondria. We provide first a formatting script, scXplorer to generate mitoXplorer 3.0 compatible files for upload. This script creates on the one hand pseudo-bulk transcriptome data of cell types from scRNA-seq data for differential expression analysis and subsequent mitochondria-centric analysis with mitoXplorer 2.0 classical interfaces. Secondly, it creates a single-cell expression matrix only containing mitochondria-associated genes (mito-genes), which can be analyzed for cell-to-cell variability within a cell type, using novel, interactive interfaces created in MitoXplorer 3.0: our new interfaces help identify sub-clusters of cell types based only mito-genes, with the possibility to in-depth analyze mito-process based gene expression variability. We demonstrate the usability and predictive power of mitoXplorer 3.0 using single-cell transcriptome data from a single-cell study of Spinocerebellar Ataxia Type 1, where we perform data mining on the cell type most affected in Cerebellar Ataxias (CAs), the Purkinje neurons. Using pseudo-bulk data, we identify the mito-processes most affected in this cell type in SCA1, which include Transcription, and Fe-S cluster biosynthesis, and we identify 3 sub-cell types of Purkinje cells in samples from human SCA1 patients.

INTRODUCTION

Mitochondria are essential for eukaryotic life and are important for various cellular processes, including bioenergetics, metabolism, apoptosis, and cellular signaling [1]. While found in almost all eukaryotic cells, their structure, as well as their expression profile differs across cell types, mostly reflecting the different requirements of the cells to these organelles [2-6]. Particularly in highly heterogeneous tissues such as the brain, standard bulk-RNA gene expression data can lead to ambiguous results. Moreover, mitochondria are known to exhibit heterogeneity even within seemingly homogeneous cell populations [7]. Some diseases or dysfunctions also target specific cell types [8,9]. Cell-type specificity is particularly important in neurodegenerative disorders like cerebellar ataxias, where cellular heterogeneity can influence disease progression and treatment response, and disease phenotypes are often due to the perturbation or the loss of a single cell type [10].

The use of single-cell (scRNA-seq) or single-nuclei sequencing (snRNA-seq) data has aided in uncovering cellular heterogeneity and functional diversity of individual cells in a cell population, notably by unveiling intra cell line heterogeneity and help characterize specific targets [11-15]. However, for those studies, the entire set of expressed genes was used. Examining the expression dynamics of gene subsets, such as mitochondria-associated genes (mito-genes), at the single-cell level would help uncover variations in mitochondrial activity of individual cells within a population, thus bringing distinct cellular metabolic states to light that might be hidden in bulk analyses [16].

We have previously developed and released mitoXplorer [17] and its first update, mitoXplorer 2.0 [18], a web-based platform to analyze and integrate data in a mitochondria-centric way. At the heart of the mitoXplorer platforms are manually curated and annotated interactomes of 4 different species (human, mouse, fruit fly and budding yeast) which we use for integrating -omics data emanating from bulk RNA-seq or proteomics data. mitoXplorer 2.0 offers 9 interfaces for mining and integrating bulk RNA-seq or proteomics data, allowing to analyze a single dataset, perform comparative analysis, time-series analysis, cross-species analysis, or population-based analyses. Moreover, with the 2nd release, we made it possible to search for transcriptional, or signaling regulators of mito-genes. One limitation of mitoXplorer 2.0 however is the fact that it only accepts bulk -omics data and hence, that it cannot be used for analyzing scRNA-seq or snRNA-seq data. Single-cell based techniques are advancing at an unprecedented speed and will become more and more prominently applied in the future [19-21], opening new perspectives on understanding cellular states. Therefore, mitoXplorer will need to be adapted to account for these new data types.

Here we introduce an update of our mitoXplorer web-tool, mitoXplorer 3.0, which we adapted to work with sc/sn-RNA-seq data to uncover cell-type specific mitochondrial variability, as well as to investigate mitochondrial expression dynamics at a single-cell level. We demonstrate the predictive power of mitoXplorer 3.0 on a time-resolved, high-resolution single-nuclei RNA-sequencing dataset of Spinocerebellar Ataxia 1 (SCA1), where we identify and analyze enriched mito-processes in the disease, such as mito-transcriptionor the Fe-S cluster biosynthesis pathway in Purkinje cells, and where we uncover 3 different mitochondrial sub-populations of Purkinje cells in WT and SCA1 mouse data.

MATERIAL AND METHODS

mitoXplorer 3.0 for analyzing single-cell and single-nuclei transcriptome data

Mitoxplorer 3.0 is a further development of our mitoXplorer 2.0 pipeline allowing mitochondria-centric analysis of single-cell and single-nuclei sequencing data. We had two possible biological scenarios for analysis in mind (Figure 1). First, we wanted to take advantage of cell-type specific information originating from single-cell sequencing data. To this end, we wanted to generate cell-type specific pseudo-bulk data that could be analyzed using standard differential expression analysis pipelines such as DESeq2 [22] or edgeR [23]). And second, to analyze mitochondrial variability within individual cell types and identify mito-gene based cellular sub-types. As single-cell data are typically large R-objects or h5ad files, we developed a pre-processing pipeline, scXplorer, which generates both the pseudo-bulk file of cell types, as well as the single-cell transcriptome data file specific for mito-genes. As input to scXplorer, the user only needs to provide the pre-analyzed single-cell data object, containing the cell annotations and the + UMAP coordinates.

Workflow of mitoXplorer 3.0 for single-cell RNA-seq

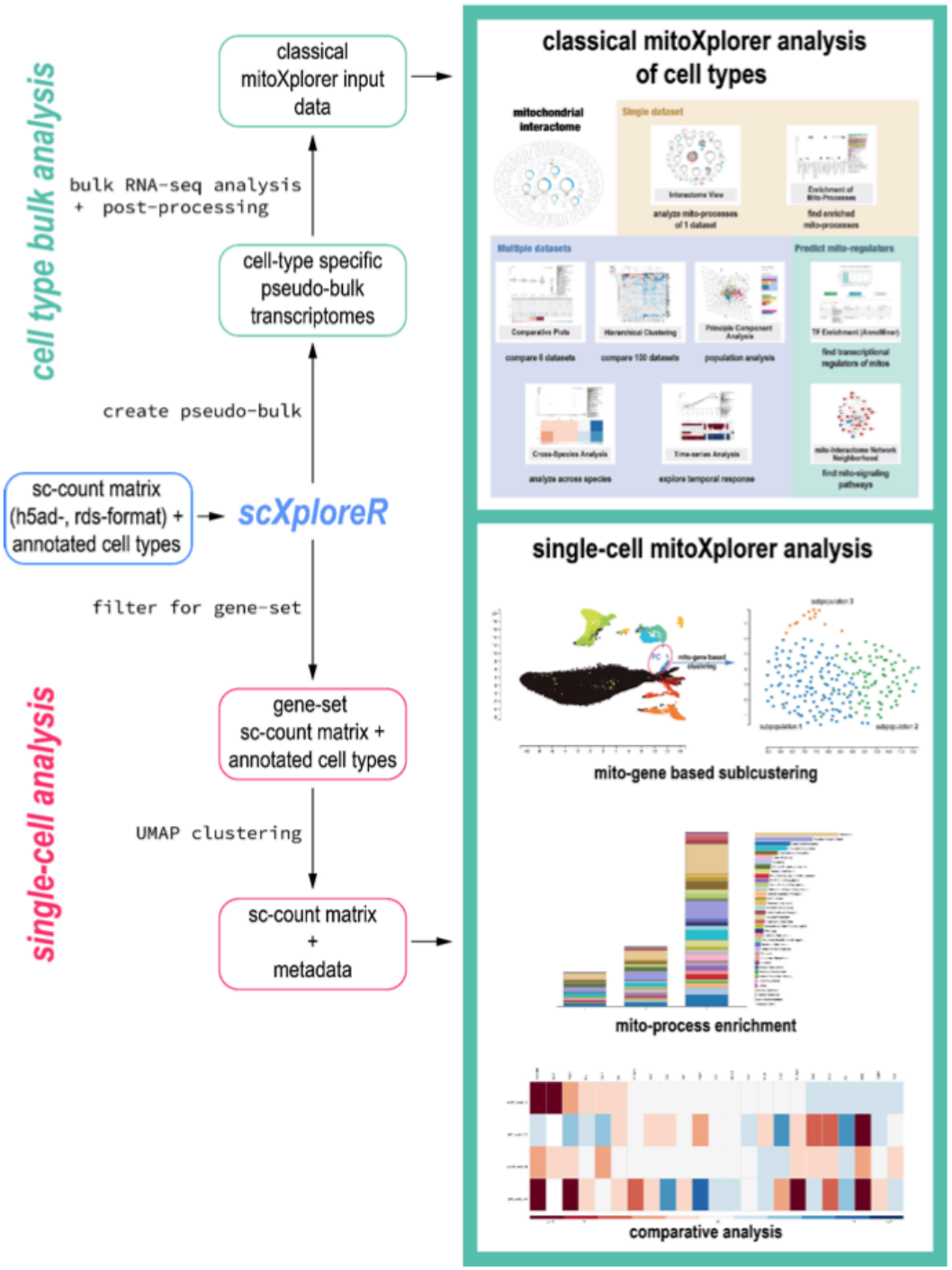


Figure 1 : Workflow of single-cell RNA-seq data pre-processing and analysis with mitoXplorer 3.0. Single-cell RNA-seq data need to be pre-processed prior to mitoXplorer 3.0 analysis. The user needs to provide a scRNA-seq data object with annotated cell types in h5ad or rds format. For pre-processing the data, the scXplorer script was developed, which on the one hand creates pseudo-bulk data for annotated cell types, and performs differential expression analysis using pyDESeq2 [24]; and secondly, mito-genes as stored in the mitoXplorer database are selected from the scRNA-seq data structure for subsequent sub-clustering of cell types. Cell-type subpopulations can be analyzed in the mitoXplorer platform for mito-processes present in the subpopulations, as well as on a mito-process basis for differential gene expression on the level of mito-processes.

The scXplorer pre-processing script for scRNA-seq and snRNA-seq data objects

scXplorer was coded in Python 3.11. It requires an analyzed and annotated single-cell sequencing data object as input, which has already undergone user-specified filtering, cell assignments, as well as generation of UMAP coordinates. We recommend the usage of already existing software packages to analyze scRNA-seq/snRNA-seq data, such as Seurat [25,26] or scanpy [27], (for a typical workflow see tutorial [28]). When using the R-based Seurat package for filtering, cell assignments and UMAP generation, the resulting R-object should be saved as an RDS-file. Alternatively, the h5ad file format from the Anndata python [29] package can be directly used. Both can be fed into the scXplorer script. When the scXplorer pipeline is downloaded for local usage, the mito-gene set for filtering genes for single-cell based mitoXplorer 3.0 analysis is provided as part of the scXplorer package and needs to be localized in the folder of the executable script.

The command line to execute the formatting script is as followed :

```
python3 scXplorer.py -f [h5ad or rds file] -o [organism] -a
[annotation] -n [output name]
```

parameters:

`-f [rds or h5ad file]` : specify the path to input rds or h5ad file(s)

`-o [organism]` : specify the organism of the input file

Choose one of the following 4 organisms : human, mouse, fly and yeast.

`-a [annotation]` : specify the variable that stores the annotation of cells (mandatory for h5ad files).

`-c [control file]`: specify the control (wild-type) inputfile if there are two input files and you want to compare mutant and wild type.

`-g [genotype]`: alternatively, if mutant and wild type are in one file, specify the variable storing the genotype.

`-n [output name]` : specify output file name (optional).

The script generates a folder named "scXplorer", containing the mitoXplorer-ready,

pre-processed and analyzed pseudo-bulk data as well as the pseudo-bulk read counts deduced from the h5ad-formatted data file (in CSV format). In addition, it generates the mito-gene filtered single-cell read count matrix in h5ad format, which can directly be uploaded to mitoXplorer (in zip format).

Pseudo-bulk data pre-processing and analysis: The scXplorer script generates two csv files: The first csv file with the extension "_pseudo_bulk.csv" contains pseudo-bulk data derived from the single-cell sequencing datafile. In this file, each annotated cell type is extracted, summarizing the read counts per gene over all cells of this cell type; at the same time, a pseudo-means is generated, which reflects the mean of the read counts of a gene over all contained cell types. Three pseudo-replicates are generated for each cell type and the pseudo-mean, to allow analysis as bulk RNA-seq data using all standard differential expression analysis packages, such as DESeq2 [22] or edgeR [23]. Pseudo-replicates are calculated by utilizing 33% of cells for each cell type, and the mean value for each gene is computed per replicate. The second csv file with the extension '_bulkXplorer.csv' is the ready-to-use file for upload to mitoXplorer and for mining with the classical mitoXplorer 2.0 interfaces. Depending on the provided input, it will either contain a comparison of each cell type against the mean over all cell types (if only one single-cell RNA-seq dataset is provided), or perform comparative analysis between two submitted RNA-seq datasets (for instance a wild-type and a mutant condition). To generate the mitoXplorer pre-analyzed file, differential expression analysis is performed using the pyDESeq2 package [24], to obtain the log2 fold change (log2FC), and the p-value.

Analysis of mito-gene based subpopulations in mitoXplorer 3.0

All analyses were done using Python v3.11 and scanpy. All visualizations in mitoXplorer 3.0 were developed using the D3.js library [30]. For single-cell analysis, the single-cell count matrix is filtered for mito-genes, which are exclusively used for all downstream analysis steps in the mitoXplorer 3.0 platform.

Clustering: Once the mito-gene filtered single-cell data is uploaded to the webtool, the user first needs to choose a cell type of interest, which will be clustered using the Leiden algorithm to identify subpopulations according to mito-gene expression profiles. The clustering resolution can be chosen by the user (default: 0.8). Clustering can be repeated at any time, choosing for instance another clustering resolution or another cell type. Resulting clusters of subpopulations are displayed in a UMAP for visual inspection by the user.

Differential expression analysis: To identify how much a subpopulation differs from the others, a differential expression analysis is performed using pyDESeq2. One subpopulation is compared to the mean of all other subpopulations excluding itself, calculating the log2FC and the p-value for each gene.

Mito-process based analysis and data visualization: Each subpopulation can be analyzed according to mito-processes contained. Based on a threshold of the p-value as determined by the differential expression analysis that can be set by the user, genes that are below this threshold and are associated with mito-processes are counted for each subpopulation. This gives the user a global overview of the diversity of mito-processes present in each subpopulation.

New Visualisation interfaces for mitoXplorer 3.0 single-cell RNA-seq data analysis

We have developed several novel interactive and user-friendly interfaces to mine single-cell RNA-seq data in mitoXplorer 3.0. To visualize subpopulation gene counts in mito-processes, we offer 3 interfaces: 2 stacked bar charts, and a heatmap colored according to the number of genes in each mito-process for each subpopulation. Hovering over a box in this heatmap will display the number of genes found in the mito-process and subpopulation. Second, we offer to visualize each mito-process with an interactive, and sortable heatmap. Each gene for each subpopulation in the heatmap is colored according to the log2FC. As is standard in all mitoXplorer interfaces, hovering over a gene box will reveal the information of the gene in the info-box at the left panel of the web-tool. The log2FC, the p-value, as well as the raw read counts are displayed in a pop-up box of the chosen cell population.

For human data, we additionally offer to analyze the distribution of subpopulations according to the cell cycle state, using the genes in S and G2M phases from Harmonizome [31] present in the mito-gene set. A prediction of the cell cycle phase for each subpopulation is given by distribution of the sum of scores for each cell cycle phase of the cells of the subpopulation.

MitoXplorer database

Pre-processed pseudo-bulk data were added to the mitoXplorer database and are freely accessible to the users. These currently include data from: the Tabula sapiens [32] and Tabula muris [33] projects, the Fly Cell Atlas [34] and datasets on spinocerebellar ataxia from L.Tejwani et al [35] (GEO accession number GSE246183).

Use case data analysis

We used pre-analyzed and annotated mouse and human single-nuclei RNA-seq data of SCA1 from Tejwani L. et al. [35] (GEO accession number GSE246183). This dataset contained data from 10 human SCA1 patients, and 10 healthy control donors; as well as data from a mouse model of SCA1 knock-in (KI) containing the

mutation *Atxn1*$^{154/+}$, collected at different timepoints : 5, 12,18, 24 and 30 weeks of age. The annotated file in h5ad format was used as input to scXplorer, and pre-processed using 'sub4' as annotation variable. We generated single-cell data (Suppl. File Y), and pseudo-bulk data of SCA1 and WT alone, as well as comparing SCA1 versus WT (Suppl. File Y2). For comparative analysis of pseudo-bulk data we compared human SCA1 versus healthy donor, and equally for mice for each time-point (SCA1 vs WT control). We also performed pseudo-bulk data comparison within each dataset, comparing each cell type against the mean over all cell types.

RESULTS

Use case: cell type- and cellular subpopulation-based mitochondrial variability of the SCA1 and WT data in human and mouse cerebellum.

Neuronal diseases such as ataxias are especially profiting from single-cell or single-nuclei RNA-sequencing data, due to the cellular complexity and heterogeneity of the brain, as well as to often cell-type specific contributions to disease-onset, as is the case for most cerebellar ataxias (CAs). For introducing the novel features and predictive power of mitoXplorer 3.0, we therefore used data from Spinocerebellar Ataxia type 1 (SCA1). SCA1 is a genetic neurological disorder characterized by progressive loss of coordination and balance. It is caused by a polyQ-expansion in the *ATXN1* gene, leading to the production of the defective Ataxin-1 protein. Ataxin-1 is a transcriptional co-repressor and interacts with the fast transcriptional repressor, Capicua. The poly-Q expansion in Ataxin-1 leads to the loss of this interaction and the decrease of Capicua, resulting in the onset of SCA1 [36]. Ataxin-1 is also activating RORa dependent transcription, in conjunction with the transcriptional co-activator Yap/YapDeltaC [37]. Mitochondria play a crucial role in SCA1 and studies suggest that dysfunctional mitochondria may contribute to the degeneration of neurons in SCA1, especially in the Purkinje cells (PCs) [38]: Mitochondrial impairment leads to increased oxidative stress and energy deficits, worsening the neurodegenerative process in SCA1. Understanding these mitochondrial implications could potentially offer new avenues for therapeutic interventions in the future.

Part 1: Cell-type specific pseudo-bulk analysis

Mito-gene expression variability in different cerebellar cell types in the wild-type

We first wanted to investigate the cell-type specific heterogeneity of gene expression in this highly heterogeneous tissue. We therefore chose wild-type mouse data from 24 weeks old mice and used the heatmap function to visualize the differences in gene expression of each cell type against the mean over all cell types. As shown in Figure 2 for the process Fatty acid degradation & beta-oxidation, there was a marked difference in gene expression of the different cell types in the cerebellum. This type of heterogeneity was seen for all mito-processes and thus, it is justified to use snRNA-seq data and perform differential expression analysis on pseudo-bulks of cell types to investigate highly complex and heterogeneous tissues, such as the cerebellum.

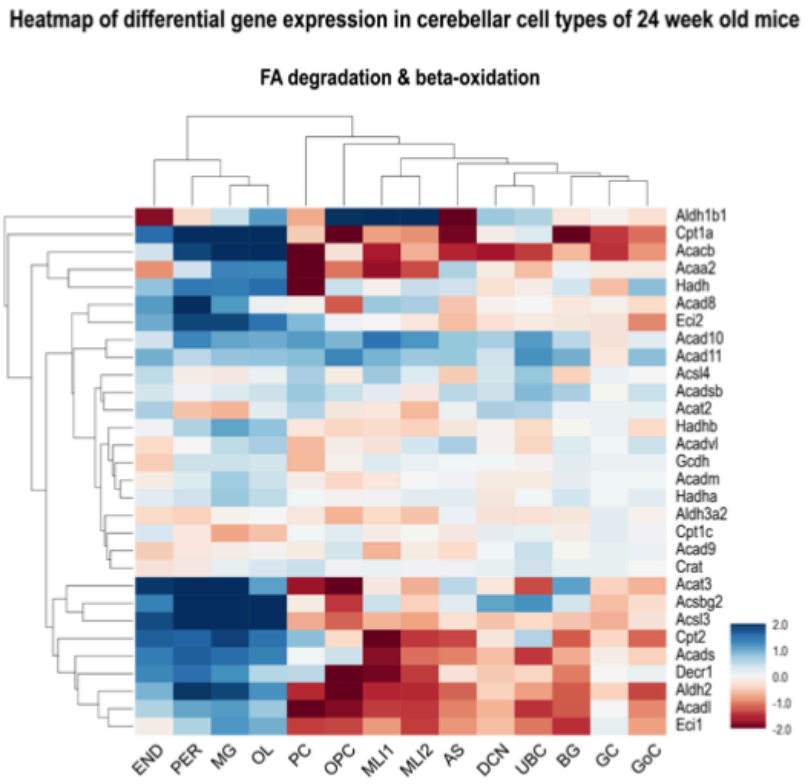


Figure 2: mitoXplorer 3.0 heatmap uncovers gene expression differences in cerebellar cell types. As an example, we use the process Fatty acid degradation and beta-oxidation from 24 week old mice. Pseudo-bulk were generated from single-cell data. Comparison was done of each cell type versus the mean over all cell types.

Nuclear Transcription of mito-genes, mitochondrial dynamics, and Fe-S cluster biosynthesis are enriched mito-processes in late-stage mouse Purkinje cells

We first analyzed mouse SCA1 data and wanted to determine which mito-processes were enriched in late-stage SCA1 mice. We therefore performed mitoXplorer's enrichment analysis on 30 weeks SCA1 Purkinje cells. The most enriched process was, as expected, Transcription (nuclear), followed by TCA cycle, mitochondrial Dynamics, and Fe-S Cluster Biosynthesis (Figure 3a).

We next looked at the differential expression of the mito-genes involved in some of the enriched mito-processes in late-stage mice using comparative plots (Figure 3b,c and Supplementary Figure S1). To look for conservation in human patients, we also analyzed the SCA1 patient data (Figure 3b-d). Several transcription factors (TFs) relevant for mito-genes in the mito-process 'Transcription (nuclear)' were up-regulated in both, SCA1 mice, as well as in human. Amongst those were *Yap1*, *Ppargc1a*, *Tfeb* or *Mitf*. A significantly down-regulated TF was *Tfe3* (Figure 3b). Upregulation of *Yap1* is especially interesting as it has been shown that the *Yap1*

gene product has been implicated in the pathophysiological phenotype of SCA1 as a transcriptional co-activator of RORa and Ataxin-1. It has been shown that mutant *Atxn1* was interfering with the interaction of RORa and YAP/YAPdeltaC, which led to the long-term depletion of RORa. Supplement of YAPdeltaC could restore the complex with RORa and rescue RORa-dependent target gene transcription. This impairment of RORa-Yap interaction in Ataxin-1 mutants was also observed specifically in Purkinje cells [39]. Upregulation of *Yap1* could therefore be a direct result of this broken link.

In the mito-process Fe-S cluster biosynthesis, several genes were down-regulated in mice, as well as in humans. Expression levels of *Iba57* and the *Isca1* homolog AK157302 were not detectable in later time-points, suggesting generally low expression of these genes in mice PCs during these time-points. One gene of the Fe-S cluster biosynthesis pathway that was significantly downregulated already in very early stages of SCA1 in PCs in mouse and in SCA1 patients compared to healthy donors, is *Fdxr*. Fe-S cluster biosynthesis is another mito-process that has been intimately linked with recessive ataxias. Its perturbation is the primary cause of heart failure in Friedreich's Ataxia (FRDA) [40] and also leads to neuronal degradation in FDRA [41,42]. Being central to a wide range of processes including cellular energy production, disruption of Fe-S cluster biosynthesis leads to severe mitochondrial dysfunction. In FRDA, mutation in the Frataxin-gene (*FXN*) leads to dysfunction in Fe-S cluster biosynthesis and a resulting failure in respiratory chain assembly and function [40]. Mutations in ferredoxin reductase (*Fdxr*) have been described to cause neurodegeneration with ataxia-like symptoms [43,44]. Next to this gene, we also found *Bola1* significantly down-regulated at 18 and 24 weeks of age in the SCA1 animals. These are the two time-points which were reported to have the most deregulated genes in the SCA1 animals [35]. Studies in yeast suggested that Bola1 forms a heterodimer with Bola3 and these act as assembly factors for specific iron-sulfur proteins [45].

We further analyzed the mito-process Fe-S cluster biosynthesis in Purkinje cells in the time-resolved mouse SCA1 data. *Fdxr* was down-regulated already early in SCA1 mice, starting from 5 weeks of age. *Bola1* on the other hand, was downregulated at 18 and 24 weeks, when most changes were observed in PCs in SCA1 animals, and upregulated again at 30 weeks (Figure 4a). It is of note that also in post-mortem human SCA1 patients, Bola1 was upregulated (Figure 3c). The differential expression of all Fe-S cluster biosynthesis genes is shown in Figure 4b (temporal plots of selected other mito-genes involved in Fe-S cluster biogenesis are shown in Supplementary Figure S2). These data show that specifically loss of *Fdxr* is an early event in SCA1. Deregulation of Fdxr could be linked to other types of mitochondrial disorders, which could implicate that its loss could potentially contribute to mitochondrial stress and the subsequent degradation of Purkinje cells also in this type of cerebellar ataxia.

Finally, we wanted to know the rate of differential expression of Fe-S cluster biosynthesis genes, and specifically *Fdxr*, in the other cerebellar cell types. To this end, we used the heatmap analysis and compared all cerebellar cell types from comparing SCA1 versus wild-type mice at 30 weeks of age, as well as human SCA1 patients versus healthy donors (Figure 4c,d). *Fdxr* was consistently downregulated in nearly all cerebellar cell types in mice, although strongest downregulation could be observed in PCs. In humans, *FDXR* is more strongly downregulated in Molecular layer interneuron 1, PCs and Bergmann Glia. In other cell types, the de-regulation was weaker, with weak up- or downregulation in the different cell types.

In conclusion, our analyses suggest that next to de-regulated mito-gene transcriptional regulation in the nucleus, deregulation of mito-genes involved in Fe-S cluster biosynthesis are an early event in SCA1 in PCs.

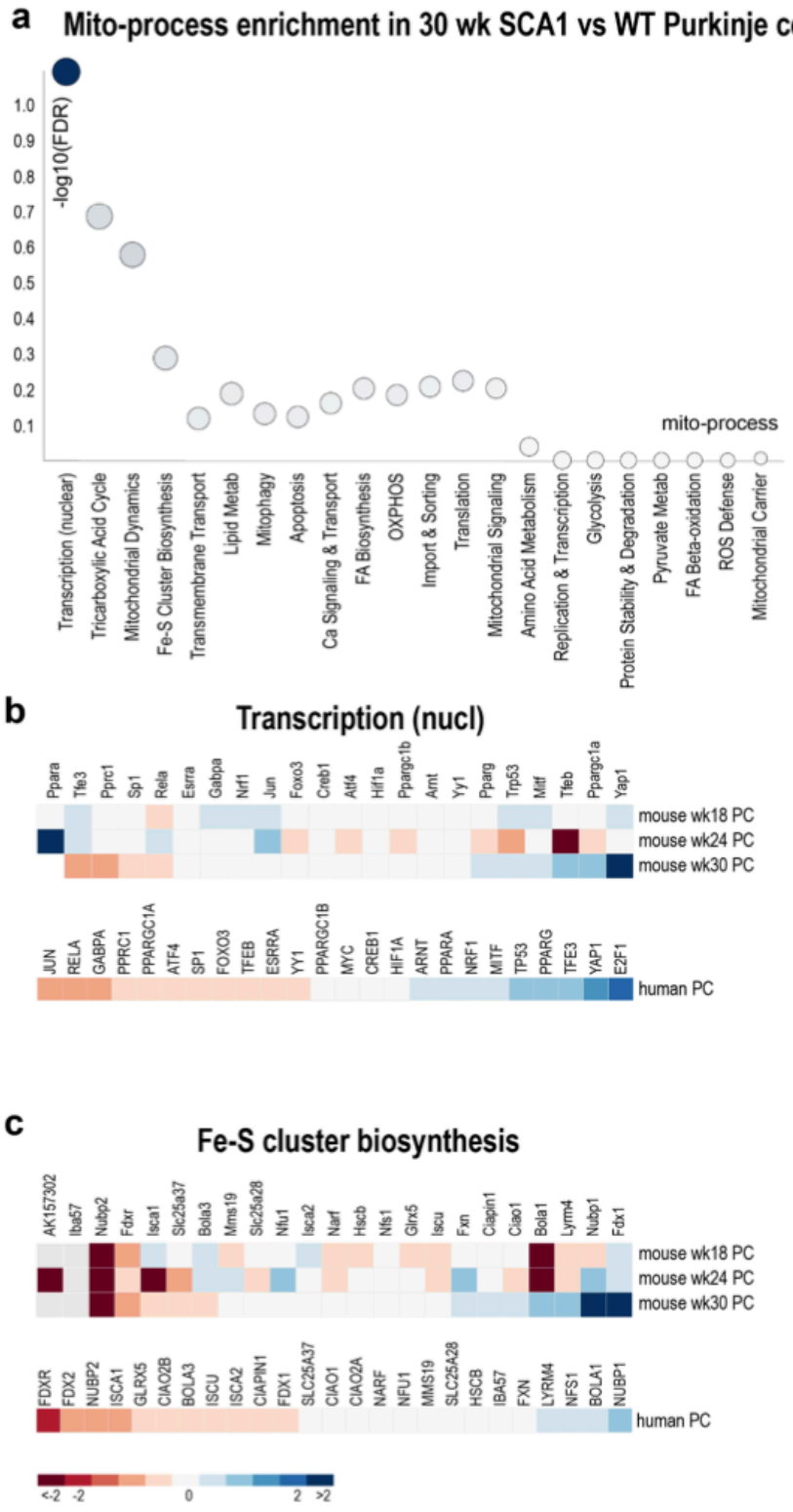

a
Mito-process enrichment in 30 wk SCA1 vs WT Purkinje cells
-log10(FDR)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
mito-process
Transcription (nuclear)
Tricarboxylic Acid Cycle
Mitochondrial Dynamics
Fe-S Cluster Biosynthesis
Transmembrane Transport
Lipid Metab
Mitophagy
Apoptosis
Ca Signaling & Transport
FA Biosynthesis
OXPHOS
Import & Sorting
Translation
Mitochondrial Signaling
Amino Acid Metabolism
Replication & Transcription
Glycolysis
Protein Stability & Degradation
Pyruvate Metab
FA Beta-oxidation
ROS Defense
Mitochondrial Carrier
b
Transcription (nucl)
Ppara
Tfe3
Pprc1
Sp1
Rela
Esrra
Gabpa
Nrf1
Jun
Foxo3
Creb1
Atf4
Hif1a
Ppargc1b
Amt
Yy1
Pparg
Trp53
Mitf
Tfeb
Ppargc1a
Yap1
mouse wk18 PC
mouse wk24 PC
mouse wk30 PC
JUN
RELA
GABPA
PPRC1
PPARGC1A
ATF4
SP1
FOXO3
TFEB
ESRRA
YY1
PPARGC1B
MYC
CREB1
HIF1A
ARNT
PPARA
NRF1
MITF
TP53
PPARG
TFE3
YAP1
E2F1
human PC
c
Fe-S cluster biosynthesis
AK157302
Iba57
Nubp2
Fdxr
Isca1
Slc25a37
Bola3
Mms19
Slc25a28
Nfu1
Isca2
Narf
Hscb
Nfs1
Glrx5
Iscu
Fxn
Ciapin1
Ciao1
Bola1
Lyrm4
Nubp1
Fdx1
mouse wk18 PC
mouse wk24 PC
mouse wk30 PC
FDXR
FDX2
NUBP2
ISCA1
GLRX5
CIAO2B
BOLA3
ISCU
ISCA2
CIAPIN1
FDX1
SLC25A37
CIAO1
CIAO2A
NARF
NFU1
MMS19
SLC25A28
HSCB
IBA57
FXN
LYRM4
NFS1
BOLA1
NUBP1
human PC
<-2
-2
0
2
>2

Figure 3: Mito-process based analysis of SCA1 vs WT Purkinje cells in late stage SCA1. (a) mitoXplorer enrichment analysis of 30 week old SCA1 versus WT Purkinje cells. The most enriched terms are Transcription (nuclear), TCA cycle, mito-Dynamics and Fe-S cluster biosynthesis. (b,c) Heatmap of mouse and human mito-processes Transcription (nucl) (b) and Fe-S cluster biosynthesis (c). Yap1 is strongly induced in PCs in both human and mouse SCA1. Fdxr is reduced in human and mouse SCA1. Down-regulated genes are highlighted in red, up-regulated ones in blue.

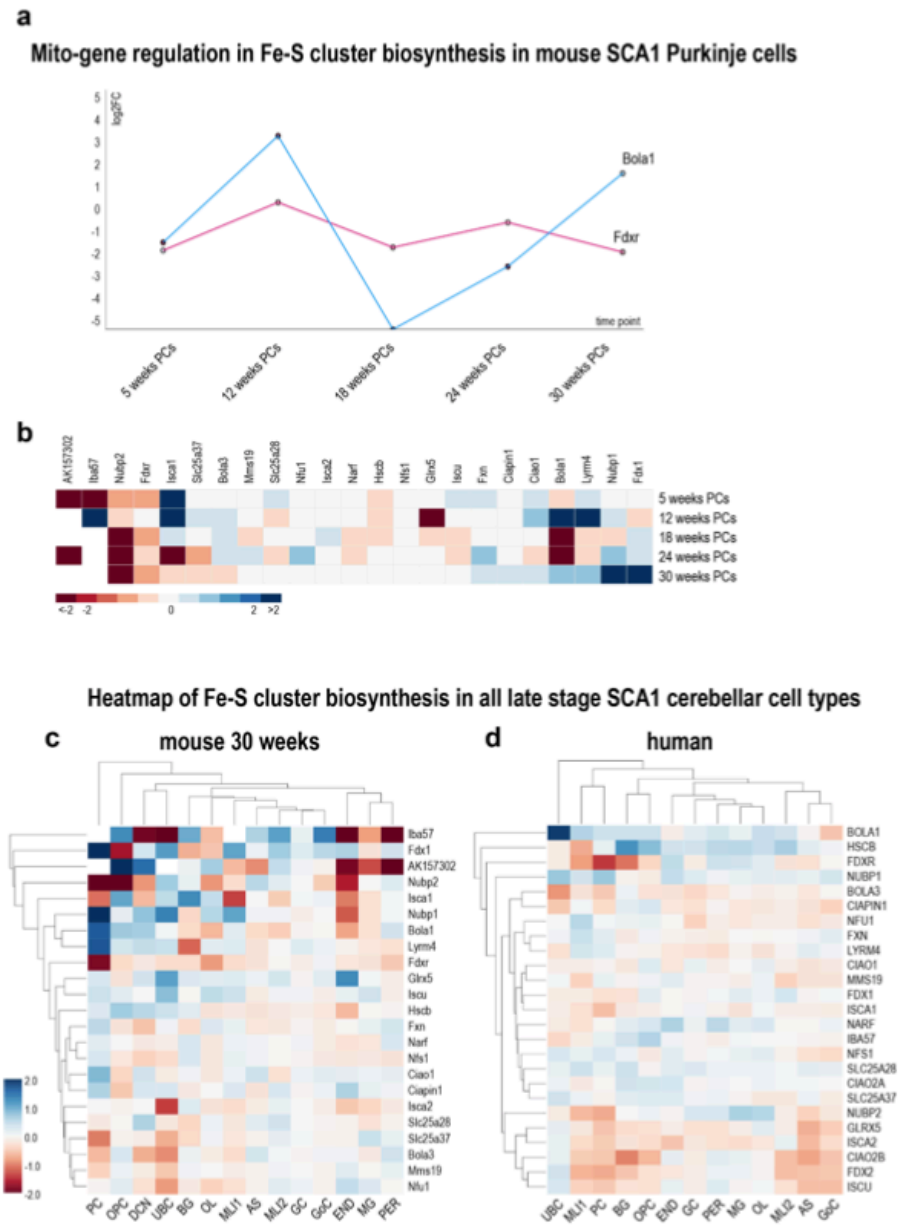


Figure 4: Fe-S cluster biosynthesis is perturbed in SCA1 Purkinje cells. (a) Time Course analysis of Fe-S biosynthesis cluster genes in PCs, comparing SCA1 to WT mice. 2 genes were selected for display. *Fdxr* (red line) is downregulated already at early time-points of SCA1 PCs in mice, while Bola1 (blue line) is downregulated at 18 and 24 weeks in SCA1 PCs. (b) Heatmap time course visualization of mitoXplorer for the mito-process FeS-cluster biosynthesis. All genes belonging to this process are shown. Iba57 and the Isca1 homolog AK157302 have very low expression in PCs. (c,d) Heatmap display of Fe-S cluster biosynthesis genes in all cerebellar cell types for SCA1 versus WT of 30 weeks mice (c) and human SCA1 patients versus healthy donors (d).

Part 2: Single-cell based analysis to uncover cellular subpopulations within cell types

With mitoXplorer 3.0, we also provide the possibility to perform mito-gene centric subclustering and identify subpopulations of cell types based solely on mito-genes. In order to demonstrate the functionality of these novel features, we tried to identify mito-gene based subpopulations of Purkinje cells in cerebellar data.

Analysis of wild-type PC cells of wild-type 30 week mice

We first looked at wild-type mouse cerebellum, and selected PCs with a cluster resolution of 0.6 for mito-gene based sub-clustering. This resulted in 3 subpopulations of PCs (Figure 5a). We next analyzed the presence of mito-processes in the 3 PC populations and found that mito-Dynamics is the process with the highest observed difference between the 3 populations, followed by ROS defense and Ca2+ signaling and transport (Figure 5b). We further looked at Ca2+ signaling and transport, as some mutations in some genes of this process like *Itpr1,* which enables ER to mitochondrial Ca2+ transfer, cause other forms of ataxia, including SCA15 and SCA29 [46-50](Figure 5c). Moreover, *Itpr1* was one of the genes identified in [35] as differentially expressed in PC subpopulations. *Itpr1* was highest in subpopulation 1, while subpopulation 2 showed the highest expression of the mitochondrial calcium uniporters, *Mcu*, as well as *Micu3* and *Micu1* (though the latter two with non-significant p-values).

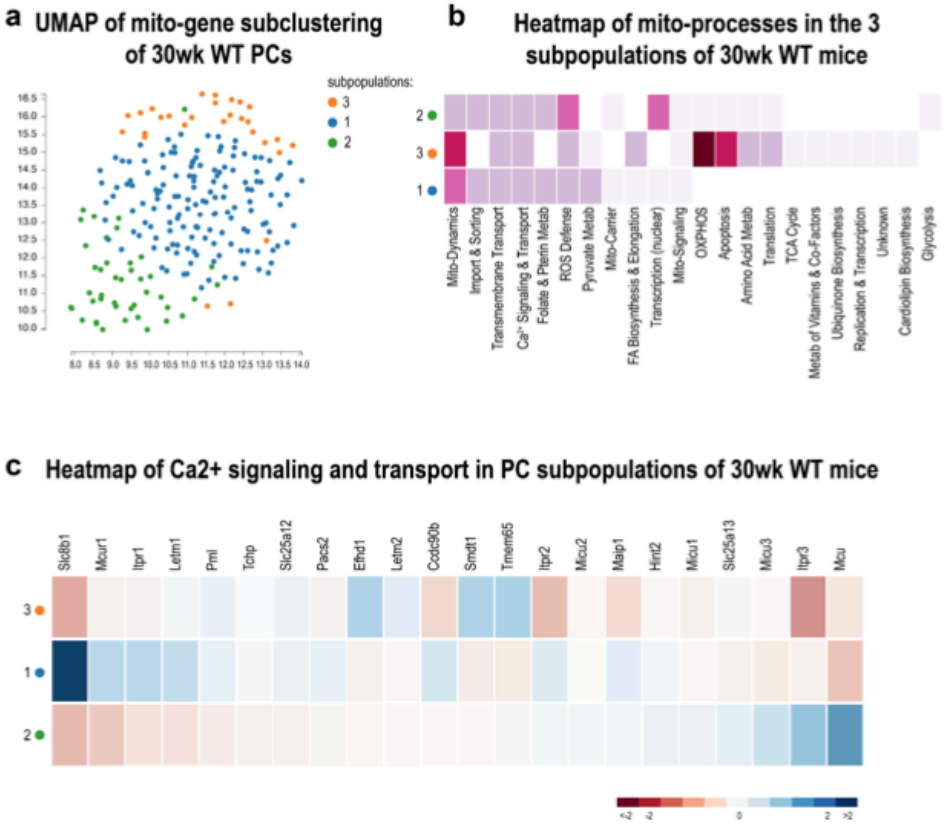


Figure 5: single-cell analysis of PCs of 30 week wild-type mice. (a) UMAP of mito-gene based subclustering of Purkinje cells in wild-type 30 week mice. (b) Heatmap of enriched mito-processes in subpopulation in wild-type PCs of 30 week animals. The darker the color, the more mito-genes of a process are present in a subpopulation. (c) Heatmap of the mito-process Ca^{2+} signaling and transport in PC subpopulations of 30 week wild-type mice. The mitochondrial Ca2+ uniporters *Mcu*, *Micu3* and *Micu1* are highest in subpopulation 2. *Itpr1* is highest in subpopulation 1.

Analysis of PC cells of SCA1 30 week mice

We next analyzed PCs from 30 week SCA1 mice to identify potential PC subpopulations. When using a clustering resolution of 0.6, 3 cell subpopulations were again found (Figure 6a). We then analyzed the mito-processes associated with the subpopulations. Among the most enriched mito-processes were Amino acid metabolism, mito-Dynamics, Ca^{2+} signaling and transport, as well as ROS defense (Figure 6b). We finally analyzed the mito-process Ca^{2+} signaling and transport (c). We also further analyzed Amino acid metabolism (d) in the 3 PC subpopulations in SCA1, as it was the most differential mito-process in the mutant condition. *Itpr1* was again highest in subpopulation 1. *Mcu* was most expressed in cell population 3. In Amino acid metabolism, we see in general a higher expression of genes in

subpopulation 3, with *Gatm*, *Gls*, *Got2* and *Acat1* being significantly differentially expressed in this subpopulation. Subpopulation 3 is also the one with the most enriched mito-processes (Figure 6b).

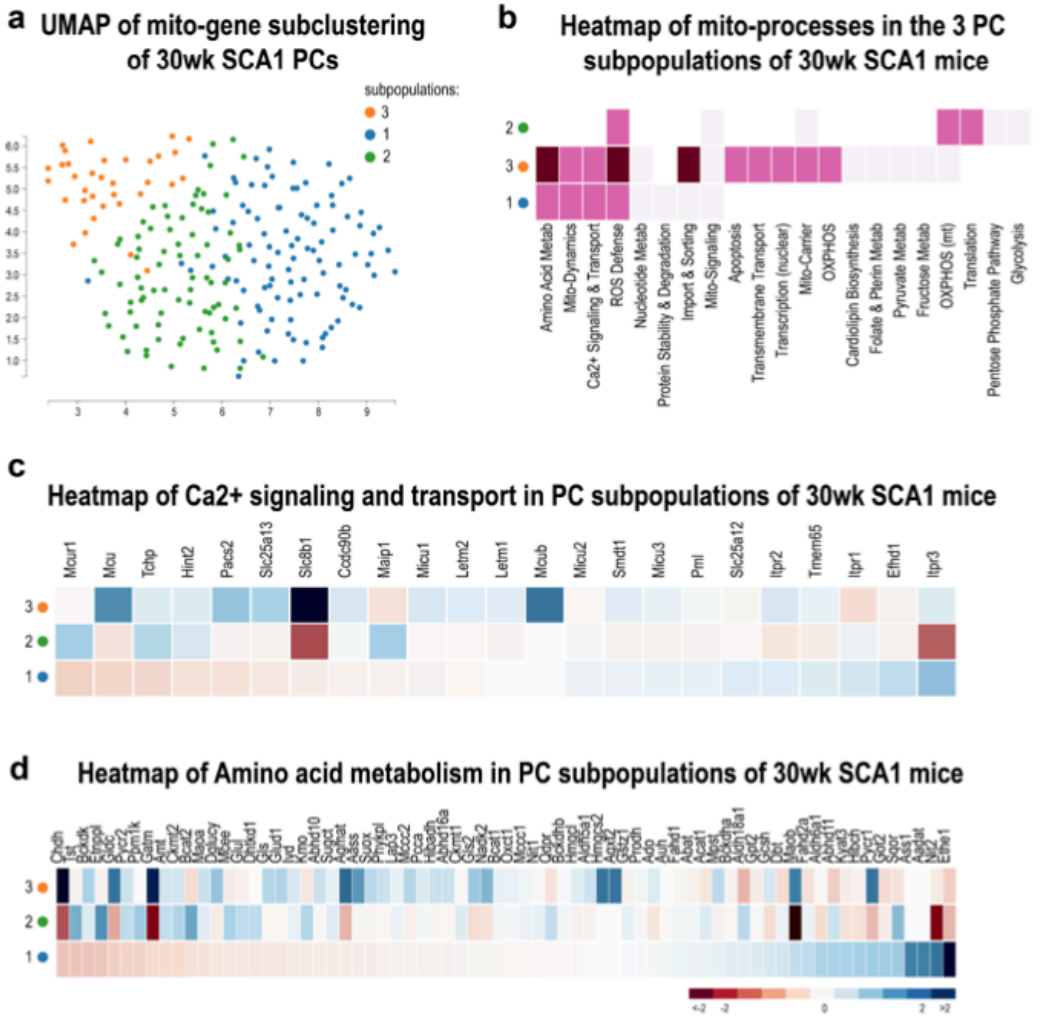


Figure 6: Single-cell analysis of PCs of 30 week SCA1 mice. (a) UMAP of mito-gene based subclustering of Purkinje cells in SCA1 30 weeks old mice. (b) Heatmap of enriched mito-processes in subpopulation in wild-type PCs of 30 week animals. The darker the color, the more mito-genes of a process are present in a subpopulation. (c) Heatmap of the mito-process Ca^{2+} signaling and transport in PC subpopulations of 30 week wild-type mice. The mitochondrial Ca2+ uniporter *Mcu* is higher in subpopulation 3 (orange dots), while *Itpr1* is highest in subpopulation 1 (blue dots). (d) Heatmap of the mito-process Amino acid metabolism of 30 week SCA1 PC subpopulations. Subpopulation 3 (orange dots) has still the highest number of genes more highly expressed than in the other two populations.

DISCUSSION

We here introduce an update of the mitoXplorer 2.0 web-tool that has been adapted to analyze single-cell RNA-seq data from a mitochondria-centric perspective. We have developed and make available a pre-processing script, scXplorer that pre-processes pre-analysed and cell-type annotated scRNA-seq/snRNA-seq data and generates ready-to-be used data for upload and mining with mitoXplorer 3.0.

ScXplorer generates on the one hand cell-type specific pseudo-bulk data from an annotated single-cell or single-nuclei sequencing file. Using scRNA-seq/snRNA-seq data for a cell-type specific pseudo-bulk analysis carries immense power as it can reveal a cell-type expression dynamics [51,52]. This is especially true in highly complex and heterogeneous tissues, such as the brain. However, also other tissues are rarely composed of a single cell type. Focusing on a mito-centric analysis using the mitoXplorer 3.0 web-tool enables users to also mine cell-type specific pseudo-bulk data easily with respect to mitochondrial features. To analyze individual cell types using the mitoXplorer 3.0 platform will therefore help understand better mitochondrial expression dynamics on a cell-type level and will help to identify the cell-type specific differences in heterogeneous tissues.

On the other hand, our script also generates a mito-gene single-cell count matrix which can be used to perform true single-cell based analysis from a purely mitochondrial perspective. This allows to sub-cluster cell types based on mitochondrial features. With novel interactive data mining interfaces, mitoXplorer 3.0 further allows to analyze subpopulations based on mito-processes and identify differentially expressed genes in the different populations.

We have used data from spinocerebellar ataxia type 1 (SCA1) to demonstrate the new features of mitoXplorer 3.0. With the help of pseudo-bulk analysis of Purkinje cells (PCs), comparing SCA1 to wild-type control, we could reveal that next to a strong enrichment in transcription, also Fe-S cluster biogenesis seems to be affected early on in SCA1 PCs. As Ataxin-1 is a transcriptional co-regulator and acts together with transcriptional repressors such as Capicua, as well as activators, such as RORa, an enrichment in the mito-process nuclear transcription is expected and can be seen as a confirmative result. The enrichment of other mito-processes that have been linked to cerebellar ataxias before, such as Fe-S cluster biogenesis, is we believe encouraging. Especially the *Fdxr* gene was downregulated from early to late stages in mouse SCA1 PCs, and was also one of the most downregulated genes in this process in human patients. *Fdxr* is essential for Fe-S cluster biosynthesis, as it enables the transfer of electrons from NAD(P)H to ferredoxin (FDX)[53]. It has been linked with other types of neurological disorders that at least partially have ataxia-like phenotypic manifestations [43,44]. *Bola1* on the other hand, which is

essential for assembly of specific Fe-S cluster proteins, was downregulated at stages with the most differential gene expression observed in the original study. It is tempting to speculate that a potential interference with Fe-S cluster biosynthesis proteins might also contribute to the loss of these cells in SCA1.

We used the same data for mitochondria-centric single-cell data analysis with mitoXplorer 3.0. Our analysis identified 3 PC subtypes in wild-type, as well as SCA1 30 week old mice. When using all genes, the authors of the original study found 5 PC subclusters and could associate a number of genes differentially expressed in these subclusters. One of those was *Itpr1*, responsible together with *Mcu*, for the transfer of Ca^{2+} from the ER to the mitochondria. *Itpr1* was also differentially expressed in our 3 subpopulations of PCs, in both, WT and in SCA1. We further used the new interfaces of mitoXplorer 3.0 to analyze the individual PC subpopulations in the wild-type and the mutant situation. The interpretation of these data is inherently difficult, as we currently lack the knowledge on the mitochondrial expression dynamics in cellular subpopulations. Experimental data coming from e.g. spatial transcriptomics [54] studies on the cellular variance of gene expression are often lacking not only for mito-genes and the cerebellum. However, given predictions coming from snRNA-seq or scRNA-seq could give impulses for novel experimental studies to help verify these results. Moreover, often, few cells and low read counts of subpopulations could skew the results. Purkinje cells, as an example, make up only about 1% of all cerebellar cells [35,55]. Resulting read counts in snRNA-seq experiments are inherently low for such underrepresented cell populations. Using experimental enrichment techniques [56] could help to alleviate this situation. Concomitant with this observation, in generally highly expressed genes such as *Itpr1* were primarily identified as markers for subpopulations [35].

There are some inherent problems when using scRNA-seq/snRNA-seq data that should be further considered. First, these types of data will often result in low read counts for a number of genes per cell, bordering on noise due to sequencing artifacts. As mitoXplorer is focusing on a specific set of genes, this problem is enhanced. We observed this problem not only for single-cell based analysis, but in some cases also for pseudo-bulk analysis. We decided for the moment not to filter based on absolute number of read counts. Such filtering could potentially eliminate entire cell populations due to under-representation. However the read counts, together with the log2FC and the p-value, should be considered by the user when interpreting the results. Second, when performing mito-gene centric sub-clustering single-cell data in mitoXplorer, the user needs to carefully analyze the resulting UMAP and try different clustering resolutions. This is a critical step in further analysis and only when cell populations are well separated, further analyses will provide meaningful information. Finally, the user needs to provide a pre-analysed and annotated scRNA-seq/snRNA-seq data file for subsequent pre-processing and

mitoXplorer 3.0 analysis. This should contain the annotated cell types. This process is highly subjective and the user will influence the assignment of the cell type in these datasets. Some tools exist that can help in cell type assignment (Azimut [57], SciBet [58], ClustifyR [59], SingleR [60]). We recommend performing this process very carefully to avoid later artifacts in analyzing data with tools like mitoXplorer.

REFERENCES

[1] : Friedman, J.R. and Nunnari, J. (2014) Mitochondrial Form and Function. Nature, 505, 335-343. https://doi.org/10.1038/nature12985

[2] : Glancy B, Kim Y, Katti P, Willingham TB. The Functional Impact of Mitochondrial Structure Across Subcellular Scales. Front Physiol. 2020 Nov 11;11:541040. doi: 10.3389/fphys.2020.541040. PMID: 33262702; PMCID: PMC7686514.

[3] : Hansen FM, Kremer LS, Karayel O, Bludau I, Larsson NG, Kühl I, Mann M. Mitochondrial phosphoproteomes are functionally specialized across tissues. Life Sci Alliance. 2023 Nov 20;7(2):e202302147. doi: 10.26508/lsa.202302147. PMID: 37984987; PMCID: PMC10662294.

[4] : Vamsi K. Mootha, Jakob Bunkenborg, Jesper V. Olsen, Majbrit Hjerrild, Jacek R. Wisniewski, Erich Stahl, Marjan S. Bolouri, Heta N. Ray, Smita Sihag, Michael Kamal, Nick Patterson, Eric S. Lander, Matthias Mann,
Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria, Cell, Volume 115, Issue 5, 2003, Pages 629-640,ISSN 0092-8674,https://doi.org/10.1016/S0092-8674(03)00926-7.

[5] : Kuznetsov AV, Hermann M, Saks V, Hengster P, Margreiter R. The cell-type specificity of mitochondrial dynamics. Int J Biochem Cell Biol. 2009 Oct;41(10):1928-39. doi: 10.1016/j.biocel.2009.03.007. Epub 2009 Mar 27. PMID: 19703655.

[6] : Liesa M, Palacín M, Zorzano A. Mitochondrial dynamics in mammalian health and disease. Physiol Rev. 2009 Jul;89(3):799-845. doi: 10.1152/physrev.00030.2008. PMID: 19584314.

[7] : Aryaman J, Johnston IG, Jones NS. Mitochondrial Heterogeneity. Front Genet. 2019 Jan 25;9:718. doi: 10.3389/fgene.2018.00718. PMID: 30740126; PMCID: PMC6355694.

[8] : Ragupathy H, Vukku M, Barodia SK. Cell-Type-Specific Mitochondrial Quality Control in the Brain: A Plausible Mechanism of Neurodegeneration. Int J Mol Sci. 2023 Sep 22;24(19):14421. doi: 10.3390/ijms241914421. PMID: 37833867; PMCID: PMC10572699.

[9] : Kim DK, Mook-Jung I. The role of cell type-specific mitochondrial dysfunction in the pathogenesis of Alzheimer's disease. BMB Rep. 2019 Dec;52(12):679-688. doi: 10.5483/BMBRep.2019.52.12.282. PMID: 31722781; PMCID: PMC6941758.

[10] : Hoxha E, Balbo I, Miniaci MC, Tempia F. Purkinje Cell Signaling Deficits in Animal Models of Ataxia. Front Synaptic Neurosci. 2018 Apr 26;10:6. doi: 10.3389/fnsyn.2018.00006. PMID: 29760657; PMCID: PMC5937225.

[11] : Zhu, Q., Zhao, X., Zhang, Y. *et al.* Single cell multi-omics reveal intra-cell-line heterogeneity across human cancer cell lines. *Nat Commun* 14, 8170 (2023). https://doi.org/10.1038/s41467-023-43991-9

[12] : Zhang, JY., Wang, XM., Xing, X. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol* 21, 1107–1118 (2020). https://doi.org/10.1038/s41590-020-0762-x

[13] : Jiang W, Glaeser JD, Salehi K, Kaneda G, Mathkar P, Wagner A, Ho R, Sheyn D. Single-cell atlas unveils cellular heterogeneity and novel markers in human neonatal and adult intervertebral discs. iScience. 2022 Jun 2;25(7):104504. doi: 10.1016/j.isci.2022.104504. PMID: 35754733; PMCID: PMC9213722.
[14] : Li, C., Wu, H., Guo, L. *et al.* Single-cell transcriptomics reveals cellular heterogeneity and molecular stratification of cervical cancer. *Commun Biol* 5, 1208 (2022). https://doi.org/10.1038/s42003-022-04142-w
[15] : Zheng, M., Hu, Z., Mei, X. *et al.* Single-cell sequencing shows cellular heterogeneity of cutaneous lesions in lupus erythematosus. *Nat Commun* 13, 7489 (2022). https://doi.org/10.1038/s41467-022-35209-1
[16] : Jack R Leary, Yi Xu, Ashley B Morrison, Chong Jin, Emily C Shen, Peyton C Kuhlers, Ye Su, Naim U Rashid, Jen Jen Yeh, Xianlu Laura Peng, Sub-Cluster Identification through Semi-Supervised Optimization of Rare-Cell Silhouettes (SCISSORS) in single-cell RNA-sequencing, *Bioinformatics*, Volume 39, Issue 8, August 2023, btad449, https://doi.org/10.1093/bioinformatics/btad449
[17] : Yim A, Koti P, Bonnard A, Marchiano F, Dürrbaum M, Garcia-Perez C, Villaveces J, Gamal S, Cardone G, Perocchi F, Storchova Z, Habermann BH. mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations. Nucleic Acids Res. 2020 Jan 24;48(2):605-632. doi: 10.1093/nar/gkz1128. PMID: 31799603; PMCID: PMC6954439.
[18] : Fabio Marchiano, Margaux Haering, Bianca Hermine Habermann, The mitoXplorer 2.0 update: integrating and interpreting mitochondrial expression dynamics within a cellular context, *Nucleic Acids Research*, Volume 50, Issue W1, 5 July 2022, Pages W490–W499, https://doi.org/10.1093/nar/gkac306
[19] : Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023 Aug;24(8):494-515. doi: 10.1038/s41576-023-00580-2. Epub 2023 Mar 2. PMID: 36864178; PMCID: PMC9979144.
[20] : Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023 Oct;24(10):695-713. doi: 10.1038/s41580-023-00615-w. Epub 2023 Jun 6. PMID: 37280296; PMCID: PMC10242609.
[21] : Chappell L, Russell AJC, Voet T. Single-Cell (Multi)omics Technologies. Annu Rev Genomics Hum Genet. 2018 Aug 31;19:15-41. doi: 10.1146/annurev-genom-091416-035324. Epub 2018 May 4. PMID: 29727584.
[22] : Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
[23] : Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818.
[24] : Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, Mathieu Andreux, PyDESeq2: a python package for bulk RNA-seq differential expression analysis, *Bioinformatics*, Volume 39, Issue 9, September 2023, btad547, https://doi.org/10.1093/bioinformatics/btad547
[25] : Satija, R., Farrell, J., Gennert, D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502 (2015). https://doi.org/10.1038/nbt.3192
[26] : Hao, Y., Stuart, T., Kowalski, M.H. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* (2023). https://doi.org/10.1038/s41587-023-01767-y
[27] : Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
[28] : Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc. 2021 Jan;16(1):1-9. doi: 10.1038/s41596-020-00409-w. Epub 2020 Dec 7. PMID: 33288955.

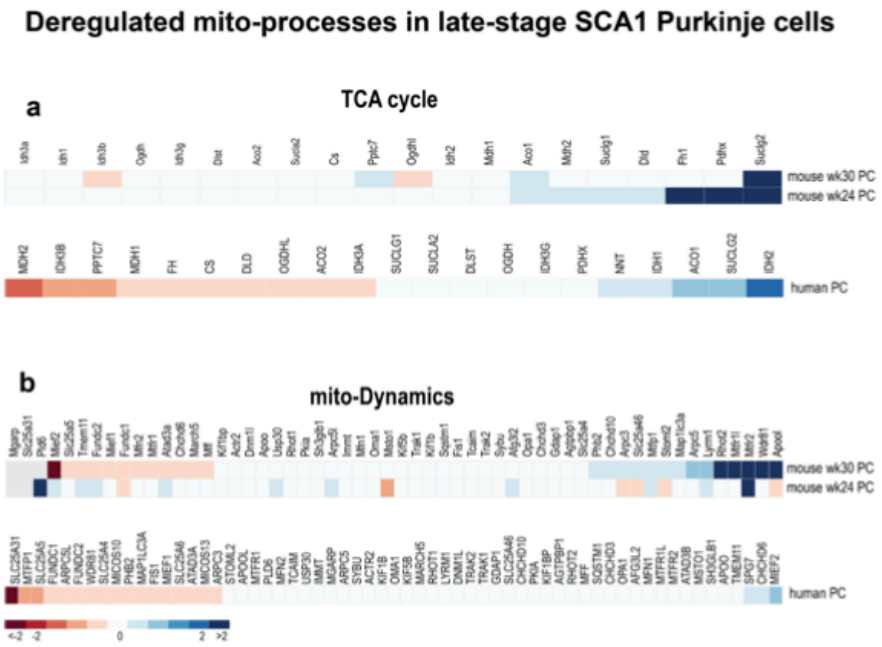
[29] : Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf *bioRxiv* 2021 Dec 19. doi: 10.1101/2021.12.16.473007.
[30] : Bostock M, Ogievetsky V, Heer J. D³: Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011 Dec;17(12):2301-9. doi: 10.1109/TVCG.2011.185. PMID: 22034350.
[31] : Andrew D. Rouillard, Gregory W. Gundersen, Nicolas F. Fernandez, Zichen Wang, Caroline D. Monteiro, Michael G. McDermott, Avi Ma'ayan, The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins, *Database*, Volume 2016, 2016, baw100, https://doi.org/10.1093/database/baw100
[32] : The Tabula Sapiens Consortium*, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.*Science***376**,eabl4896(2022).DOI:10.1126/science.abl4896
[33] : The Tabula Muris Consortium., Overall coordination., Logistical coordination. *et al.* Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* 562, 367–372 (2018). https://doi.org/10.1038/s41586-018-0590-4
[34] : Hongjie Li *et al.*,Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly.*Science***375**,eabk2432(2022).DOI:10.1126/science.abk2432
[35] : Tejwani L, Ravindra NG, Lee C, Cheng Y, Nguyen B, Luttik K, Ni L, Zhang S, Morrison LM, Gionco J, Xiang Y, Yoon J, Ro H, Haidery F, Grijalva RM, Bae E, Kim K, Martuscello RT, Orr HT, Zoghbi HY, McLoughlin HS, Ranum LPW, Shakkottai VG, Faust PL, Wang S, van Dijk D, Lim J. Longitudinal single-cell transcriptional dynamics throughout neurodegeneration in SCA1. Neuron. 2024 Feb 7;112(3):362-383.e15. doi: 10.1016/j.neuron.2023.10.039. Epub 2023 Nov 27. PMID: 38016472; PMCID: PMC10922326.
[36] : Lam YC, Bowman AB, Jafar-Nejad P, Lim J, Richman R, Fryer JD, Hyun ED, Duvick LA, Orr HT, Botas J, Zoghbi HY. ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. Cell. 2006 Dec 29;127(7):1335-47. doi: 10.1016/j.cell.2006.11.038. PMID: 17190598.
[37] : Fujita, K., Mao, Y., Uchida, S. et al. Developmental YAPdeltaC determines adult pathology in a model of spinocerebellar ataxia type 1. Nat Commun 8, 1864 (2017). https://doi.org/10.1038/s41467-017-01790-z
[38] : Stucki DM, Ruegsegger C, Steiner S, Radecke J, Murphy MP, Zuber B, Saxena S. Mitochondrial impairments contribute to Spinocerebellar ataxia type 1 progression and can be ameliorated by the mitochondria-targeted antioxidant MitoQ. Free Radic Biol Med. 2016 Aug;97:427-440. doi: 10.1016/j.freeradbiomed.2016.07.005. Epub 2016 Jul 6. PMID: 27394174.
[39] : Chiabrando D, Bertino F, Tolosano E. Hereditary Ataxia: A Focus on Heme Metabolism and Fe-S Cluster Biogenesis. Int J Mol Sci. 2020 May 26;21(11):3760. doi: 10.3390/ijms21113760. PMID: 32466579; PMCID: PMC7312568.
[40] : Hanson E, Sheldon M, Pacheco B, Alkubeysi M, Raizada V. Heart disease in Friedreich's ataxia. World J Cardiol. 2019 Jan 26;11(1):1-12. doi: 10.4330/wjc.v11.i1.1. PMID: 30705738; PMCID: PMC6354072.
[41] : Gomes CM, Santos R. Neurodegeneration in Friedreich's ataxia: from defective frataxin to oxidative stress. Oxid Med Cell Longev. 2013;2013:487534. doi: 10.1155/2013/487534. Epub 2013 Jul 9. PMID: 23936609; PMCID: PMC3725840.
[42] : Ginestroni A, Diciotti S, Cecchi P, Pesaresi I, Tessa C, Giannelli M, Della Nave R, Salvatore E, Salvi F, Dotti MT, Piacentini S, Soricelli A, Cosottini M, De Stefano N, Mascalchi M. Neurodegeneration in friedreich's ataxia is associated with a mixed activation pattern of the brain. A fMRI study. Hum Brain Mapp. 2012 Aug;33(8):1780-91. doi: 10.1002/hbm.21319. Epub 2011 Jun 14. PMID: 21674694; PMCID: PMC6870007.
[43] : Masnada S, Previtali R, Erba P, Beretta E, Camporesi A, Chiapparini L, Doneda C, Iascone M, Sartorio MUA, Spaccini L, Veggiotti P, Osio M, Tonduti D, Moroni I. FDXR-associated disease: a challenging differential diagnosis with inflammatory peripheral neuropathy. Neurol Sci. 2023

Sep;44(9):3037-3043. doi: 10.1007/s10072-023-06790-0. Epub 2023 Apr 12. PMID: 37046037; PMCID: PMC10096094.
[44]: Slone, J., Peng, Y., Chamberlin, A. *et al.* Biallelic mutations in FDXR cause neurodegeneration associated with inflammation. *J Hum Genet* 63, 1211–1222 (2018). https://doi.org/10.1038/s10038-018-0515-y
[45] : Marta A Uzarska, Veronica Nasta, Benjamin D Weiler, Farah Spantgar, Simone Ciofi-Baffoni, Maria Rosaria Saviello, Leonardo Gonnelli, Ulrich Mühlenhoff, Lucia Banci, Roland Lill (2016) Mitochondrial Bol1 and Bol3 function as assembly factors for specific iron-sulfur proteins eLife 5:e16673;https://doi.org/10.7554/eLife.16673
[46] : Bezprozvanny I, Hayden MR. Deranged neuronal calcium signaling and Huntington disease. Biochem Biophys Res Commun. 2004 Oct 1;322(4):1310-7. doi: 10.1016/j.bbrc.2004.08.035. PMID: 15336977.
[47] : van de Leemput J, Chandran J, Knight MA, Holtzclaw LA, Scholz S, Cookson MR, Houlden H, Gwinn-Hardy K, Fung HC, Lin X, Hernandez D, Simon-Sanchez J, Wood NW, Giunti P, Rafferty I, Hardy J, Storey E, Gardner RJ, Forrest SM, Fisher EM, Russell JT, Cai H, Singleton AB. Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. PLoS Genet. 2007 Jun;3(6):e108. doi: 10.1371/journal.pgen.0030108. Epub 2007 May 16. PMID: 17590087; PMCID: PMC1892049.
[48] : Shimobayashi E, Kapfhammer JP. Calcium Signaling, PKC Gamma, IP3R1 and CAR8 Link Spinocerebellar Ataxias and Purkinje Cell Dendritic Development. Curr Neuropharmacol. 2018 Jan 30;16(2):151-159. doi: 10.2174/1570159X15666170529104000. PMID: 28554312; PMCID: PMC5883377.
[49] : Hisatsune C, Hamada K, Mikoshiba K. Ca2+ signaling and spinocerebellar ataxia. Biochim Biophys Acta Mol Cell Res. 2018 Nov;1865(11 Pt B):1733-1744. doi: 10.1016/j.bbamcr.2018.05.009. Epub 2018 May 16. PMID: 29777722.
[50] : Zambonin JL, Bellomo A, Ben-Pazi H, Everman DB, Frazer LM, Geraghty MT, Harper AD, Jones JR, Kamien B, Kernohan K, Koenig MK, Lines M, Palmer EE, Richardson R, Segel R, Tarnopolsky M, Vanstone JR, Gibbons M, Collins A, Fogel BL; Care4Rare Canada Consortium; Dudding-Byth T, Boycott KM. Spinocerebellar ataxia type 29 due to mutations in ITPR1: a case series and review of this emerging congenital ataxia. Orphanet J Rare Dis. 2017 Jun 28;12(1):121. doi: 10.1186/s13023-017-0672-7. PMID: 28659154; PMCID: PMC5490223.
[51] : Sun, D., Guan, X., Moran, A.E. *et al.* Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nat Biotechnol* 40, 527–538 (2022). https://doi.org/10.1038/s41587-021-01091-3
[52] : Liu, C., Zhang, Y., Gao, X. *et al.* Identification of cell subpopulations associated with disease phenotypes from scRNA-seq data using PACSI. *BMC Biol* 21, 159 (2023). https://doi.org/10.1186/s12915-023-01658-3
[53] : Shi Y, Ghosh M, Kovtunovych G, et al. Both human ferredoxins 1 and 2 and ferredoxin reductase are important for iron–sulfur cluster biogenesis. Biochim Biophys Acta Mol Cell Res. 2012;1823:484–492
[54] : Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78-82. doi: 10.1126/science.aaf2403. PMID: 27365449.
[55] : Borgenheimer E, Hamel K, Sheeler C, Moncada FL, Sbrocco K, Zhang Y, Cvetanovic M. Single nuclei RNA sequencing investigation of the Purkinje cell and glial changes in the cerebellum of transgenic Spinocerebellar ataxia type 1 mice. Front Cell Neurosci. 2022 Nov 15;16:998408. doi: 10.3389/fncel.2022.998408. PMID: 36457352; PMCID: PMC9706545.
[56] : Bartelt LC, Switonski PM, Adamek G, Carvalho J, Duvick LA, Jarrah SI, McLoughlin HS, Scoles DR, Pulst SM, Orr HT, Hull C, Lowe CB, La Spada AR. Purkinje-Enriched snRNA-seq in SCA7 Cerebellum

Reveals Zebrin Identity Loss as a Central Feature of Polyglutamine Ataxias. bioRxiv [Preprint]. 2023 May 11:2023.03.19.533345. doi: 10.1101/2023.03.19.533345. PMID: 37214832; PMCID: PMC10197555.
[57] : Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.
[58] : Li, C., Liu, B., Kang, B. *et al.* SciBet as a portable and fast single cell type identifier. *Nat Commun* **11**, 1818 (2020). https://doi.org/10.1038/s41467-020-15523-2
[59] : Fu R, Gillen AE, Sheridan RM, Tian C, Daya M, Hao Y, Hesselberth JR, Riemondy KA. clustifyr: an R package for automated single-cell RNA sequencing cluster classification. F1000Res. 2020 Apr 1;9:223. doi: 10.12688/f1000research.22969.2. PMID: 32765839; PMCID: PMC7383722.
[60] : Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019). "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." *Nat. Immunol.*, **20**, 163-172. doi:10.1038/s41590-018-0276-y.

Supplementary Information:

Supplementary Figure S1: Heatmap of de-regulated mito-processes in late-stage SCA1 Purkinje cells. (a,b) Heatmap of mouse and human mito-processes TCA cycle (a) and mito-Dynamics (b). Both processes were among the most enriched mito-processes in 30 week PCs of SCA1 mice compared to WT control.



Chapter 3 : ataxiaXplorer

Similar to mitoXplorer, where bulk- and single-cell omics data can be analyzed from a mitochondrial perspective, with ataxiaXplorer, we want to provide a platform that allows to mine these types of data from the point of view of (cerebellar) ataxias. AtaxiaXplorer will be the first sister platform of mitoXplorer 3.0, with a specific ataxia interactome to explore and analyze bulk and sc/n RNA-seq data in an ataxic-centric way.

Introduction

Neurodegenerative disorders are characterized by a progressive loss of specific neurons and inflammation in the central nervous system leading to brain functions deficits that are incurable [107]. Alzheimer's disease and Parkinson's disease as well as ataxias [108-110] are known to be cell type specific and in the case of cerebellar ataxias, mechanisms behind the pathogenesis are not clearly and fully understood [111,112]. The studies of therapeutic targets are going on focusing on Purkinje cells (PC) notably in SpinoCerebellar Ataxia type 1 (SCA1) [113].

Methods

Assembly of the ataxia interactome

Mice public ataxic NGS datasets of SCA1,3,7 and 17 were downloaded from GEO and reanalyzed from the raw sequencing data. We chose similar time points, which are all advanced in the disease progression. These included for :

- SCA1 (Figure 11a):

Atxn1-82Q data from Driessen [114] (GSE122099), Ortiz (Henzler) [115] (GSE114815), Ingram [116] (GSE75778) and Rousseaux [117] (GSE10825), 12 weeks old mice for all and 8-9 weeks old mice from Ortiz [115].

- SCA3 (Figure 11b) :

Atxn1-84Q transgenic mice from Lin [118] (GSE17836) at 18 months time point that we further decided to remove due to lack of common genes, and *Yac*-84Q KI mice from Haas (Hubener)[119] (GSE145613) at 12 months time point.

- SCA7 (Figure 11c) :

Atxn7-92Q mice from Stoyas [120] (GSE139090) at 12 weeks time point and *Atxn7*-140Q/15Q mice from Niewiadowska [121] (GSE138727) at 40 weeks time point .

- SCA17 (Figure 11d) :

Sca17 Knock-In mice from Liu [122] (GSE145067) at 3 months time point.

Data were processed with FastQC [1], mapped using STAR aligner [4] to the mouse reference genome (mm10); reads were counted with featureCounts [5] and differential expression analysis (DEA) were performed using RNfuzzyApp TCC Deseq2 method [102].

After identifying separately the differentially expressed genes (DEGs), genes with a p-value lower or equal 0.01 and a log2fc higher to 0.5 or lower to -0.5 were selected, resulting in 578 DEGs (Figure 11e). As there were few overlapping genes between different types of ataxias, we chose to also include genes specific to only one type of ataxia.

In addition, we added genes that are known to be mutated in hereditary ataxias (provided by our collaboration partners from the Puccio lab). We next annotated the genes with DAVID [123] (Figure 11f) and finally selected 36 terms for manual curation and annotation of the list in ataxia-associated biological processes. Hand-curated the list resulting in 36 ataxia processes (Figure 11g). Orthology was used to find human genes and for the 687 genes in *Mus musculus*, we could annotate 640 orthologs in *Homo sapiens*.

Discussion

We chose only 4 types of SCAs due to the poor amount of transcriptomic data available.. However, new data are emerging and the interactome will evolve in the near future. The choice of these time points in the different types was done looking at common features. We could also integrate the other time points to build a ‘dysregulated genes involved in SCAs’ interactome and would integrate infos about associated form of SCA and time point This could lead to a much richer interactome.

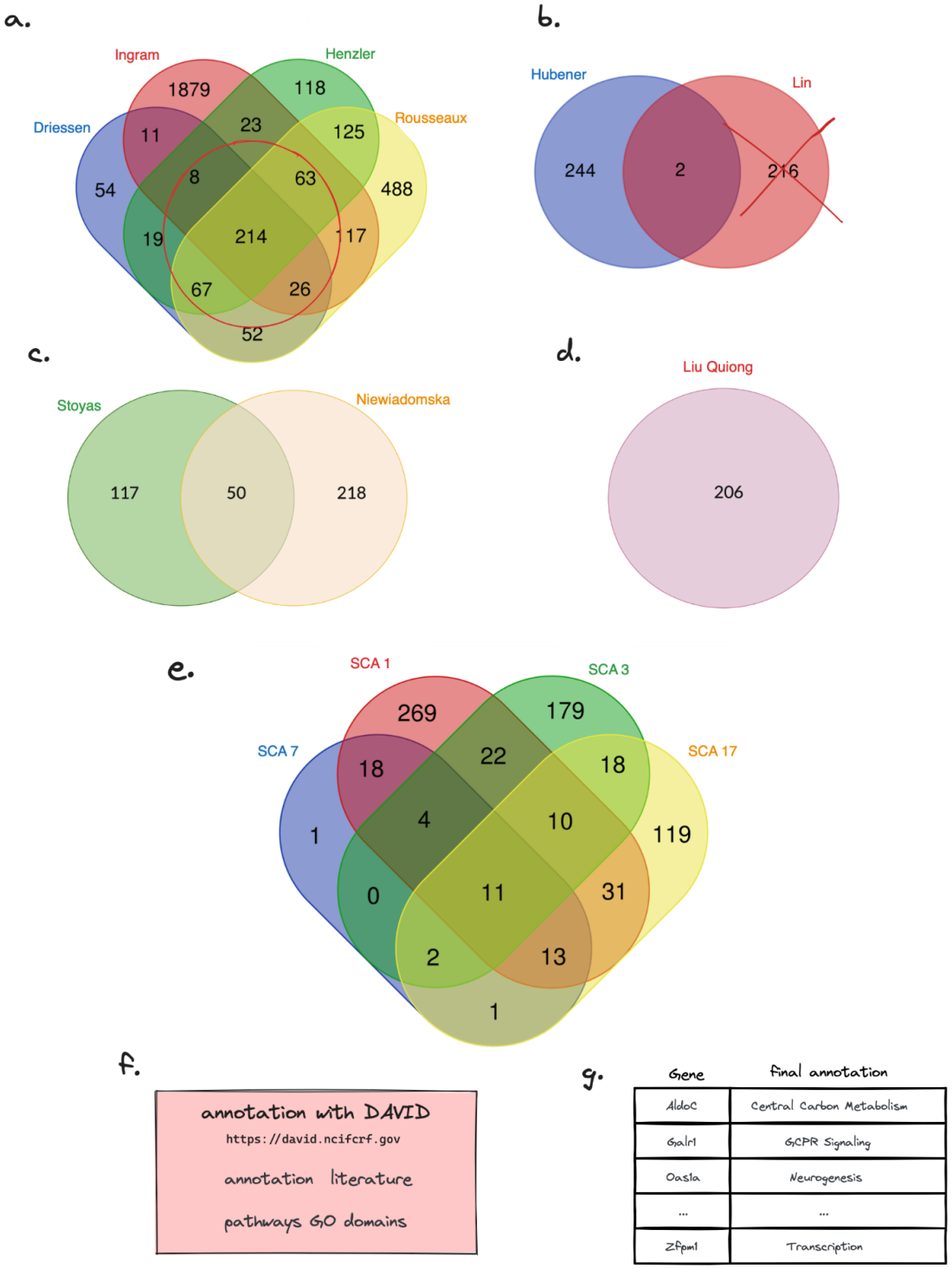


Figure 11 : ataxiaXplorer interactome building. *After DEA, genes have been selected with a p-value < 0.01 and a log2FC< -0.5 and log2FC >0.5.* ***a.****Selection of 378 common genes (present in at least 3 datasets of the 4 selected) in SCA1 condition.* ***b.*** *Selection of 246 SCA3 genes.* ***c.*** *Selection of 50 common genes across SCA7 datasets.* ***d.*** *Selection of 206 dysregulated genes in SCA17.* ***e.*** *Final comparison of 578 selected genes across SCA1,3,7,17 datasets.* ***f.*** *Large annotation with DAVID after adding ataxic responsible genes to the geneset.* ***g.*** *Manual curation of the 687 genes with the 36 chosen processes.*

Manual annotation of a gene list is a difficult task. After deciding on a list of terms we wanted to use, Andrea Del Bondio from the Puccio lab and myself performed manual annotation of the genes with the selected terms independently. Currently, these gene annotations are undergoing control by the group heads, Helene Puccio and Bianca Habermann. Only after this final step of control will we have the final interactome for ataxiaXplorer. Moreover, with novel datasets being published regularly, we need to ensure that this list stays comprehensive and up-to-date. Regular updates of the ataxia interactome are therefore envisioned.

AtaxiaXplorer platform implementation

While I did not arrive at creating the ataxiaXplorer platform, this will be done in the near-future. It requires a duplication of the mitoXplorer instance with replacing the mito-interactome with the ataxia interactome.

Next to developing all new features available in mitoXplorer 3.0, I contributed to this project by performing all the data analysis of the selected bulk RNA-sequencing data, helped select the genes and provided a first list of annotated ataxia genes.

Chapter 4 : MitoXplorer 2.0 and AnnoMiner

MitoXplorer 2.0 [100] is an upgrade of the MitoXplorer platform [99] that analyzes and explores bulk RNA-seq data from a mitochondrial perspective. It is a comprehensive and interactive web-based platform with visualizations that allows to analyze single datasets in the form of enrichment of mitochondrial processes, compare several datasets using comparative plots, heatmaps, time-series analysis or by comparing datasets across species. Two important new features were to analyze mitochondrial gene expression dynamics from a cellular perspective. This includes on the one hand identifying transcriptional regulators, by connecting mitoXplorer to the AnnoMiner [124] tool. AnnoMiner searches epigenetic datasets stored in the ENCODE and modENCODE resources for transcription factor binding events in a list of co-regulated genes. This list of genes can directly be selected in mitoXplorer. Secondly, users are able to search for changes in pathways emanating from mito-genes. This is done by using the viPEr [125] algorithm that allows to identify active subnetworks by searching for deregulated genes in the neighborhood of one gene.
The new functionalities of mitoXplorer 2.0 were demonstrated using published data from Barth syndrome. Moreover, we took data from flight muscle development and identified transcriptional regulators of mito-genes.

AnnoMiner is a specialized data mining tool designed to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators. Moreover, AnnoMiner can be used to annotate peaks from epigenetic data with gene features. It does so in a flexible way and gives the user the ability to choose up- or down-stream association of peaks with gene elements. In the paper, the tool is presented and then tested on a dataset of the developing Drosophila indirect flight muscle. We predicted transcriptional regulators of de-regulated genes during muscle development. Our predictions were experimentally verified by our collaboration partners (the Schnorrer group at the IBDM).

I contributed to mitoXplorer 2.0 and AnnoMiner in analyzing transcriptomics data to test and feed the database. I recently added to mitoXplorer the possibility to download svg figures.

W490–W499 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 7 May 2022*
https://doi.org/10.1093/nar/gkac306

The mitoXplorer 2.0 update: integrating and interpreting mitochondrial expression dynamics within a cellular context

Fabio Marchiano, Margaux Haering and Bianca Hermine Habermann*

Aix-Marseille University, CNRS, IBDM UMR 7288, 13009 Marseille, France

Received January 25, 2022; Revised April 09, 2022; Editorial Decision April 12, 2022; Accepted April 15, 2022

ABSTRACT

Mitochondria are subcellular organelles present in almost all eukaryotic cells, which play a central role in cellular metabolism. Different tissues, health and age conditions are characterized by a difference in mitochondrial structure and composition. The visual data mining platform mitoXplorer 1.0 was developed to explore the expression dynamics of genes associated with mitochondrial functions that could help explain these differences. It, however, lacked functions aimed at integrating mitochondria in the cellular context and thus identifying regulators that help mitochondria adapt to cellular needs. To fill this gap, we upgraded the mitoXplorer platform to version 2.0 (mitoXplorer 2.0). In this upgrade, we implemented two novel integrative functions, network analysis and transcription factor enrichment, to specifically help identify signalling or transcriptional regulators of mitochondrial processes. In addition, we implemented several other novel functions to allow the platform to go beyond simple data visualization, such as an enrichment function for mitochondrial processes, a function to explore time-series data, the possibility to compare datasets across species and an IDconverter to help facilitate data upload. We demonstrate the usefulness of these functions in three specific use cases. mitoXplorer 2.0 is freely available without login at http://mitoxplorer2.ibdm.univ-mrs.fr.

GRAPHICAL ABSTRACT



INTRODUCTION

Mitochondria are essential organelles in most eukaryotic cells and are involved in a multitude of cellular processes, including cellular energy production, metabolism, signalling or apoptosis. The mitochondrial protein content is estimated to be around 1000 proteins, varying slightly between species (1–3). Most of these proteins (mito-proteins) are encoded in the nucleus and imported into mitochondria in a controlled manner. Only few mito-proteins are encoded in the mitochondrial genome and transcribed and translated within mitochondria, encoding proteins of the respiratory chain, ribosomal RNAs required for the mitochondrial ribosome and transfer RNAs. Mitochondria are essential for the functioning of the cell; however, the requirements on mitochondria vary with the cell type. Consequently, mitochondrial structure and protein content differ between cell types (4) and we want to understand how mitochondria adapt to their cellular environment.

mitoXplorer version 1.0 (mitoXplorer 1.0) was developed as a visual data mining platform that allows us to analyse the dynamics of gene expression and mutations of all genes with a mitochondrial function (mito-genes) (1). It contains the most complete and up-to-date mitochondrial interactomes for four different model species (human, mouse, *Drosophila melanogaster* and budding yeast) as well as four different visualization interfaces that enable users to mine expression and mutation data in a comparative manner, as well as from a single dataset. mitoXplorer 1.0 was the first platform available to mine expression dynamics of

*To whom correspondence should be addressed. Tel: +33 649 443263; Email: bianca.habermann@univ-amu.fr

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

mito-genes in different conditions. However, it fails to integrate the mitochondrial interactome into the cellular context and thus to help identify potential signalling and transcriptional regulators that adjust mitochondria and their functions to their cellular environment.

We here introduce mitoXplorer version 2.0 (mitoXplorer 2.0), in which we have added integrative data analysis functions to help address these questions. MitoXplorer 2.0 contains several new functions for data analysis, such as an enrichment function for mito-processes, a function to explore time-series data, the possibility to compare datasets across species and two novel data integration functions that aim at identifying mitochondrial regulators: first, a function to identify potential transcriptional regulators of mito-genes with a similar expression profile (co-regulated mito-genes) by making use of our recently published AnnoMiner web tool (5); and second, a function to identify potential signalling pathways from and to mitochondria, by embedding the mito-interactome in the cellular interactome and subsequently exploring the network neighbourhood of a selected mito-gene based on integrating the network with differential expression data (6). We demonstrate the usefulness of these new functions in three use cases, where we identify potential transcriptional regulators driving the mitochondrial metabolic switch during *Drosophila* flight muscle development (use case 1, Figure 2), identify potential active signalling pathways regulating Ca^{2+} signalling in ataxia (use case 2, Figure 3) and explore conserved mito-gene deregulation in fibroblasts from human and a mouse model of trisomy 21 (use case 3, Supplementary Figure S1).

MATERIALS AND METHODS

Novel functions of mitoXplorer version 2.0, together with those already available in mitoXplorer version 1.0, are shown in Figure 1 as well as Supplementary Figure S2.

Novel functions in mitoXplorer 2.0

Mito-process enrichment analysis. In order to identify enrichment of our 38 manually curated mito-processes in a dataset, we implemented a gene set enrichment analysis (GSEA) function. This function helps guide users to focus on important mito-processes of a dataset for subsequent deeper analysis with other functions available in mitoXplorer 2.0.

Mito-process enrichment analysis was coded using the GSEApy Python library (7,8). When *P*-value and $\log_2$FC are both present, genes are ranked by the combined score as defined by Xiao *et al.* (9):

$$\text{combined score} = \text{abs}(\log_2\text{FC}) \times -\log_{10}(P\text{-value}).$$

If no *P*-value is provided, genes are ranked by the absolute value of $\log_2$FC. The gene set enrichment score is tested against a null distribution of enrichment scores generated from 100 permuted gene sets composed of randomly selected genes from the input dataset. The results of the GSEA analysis are shown in the form of an interactive dot plot, as well as a bar plot (see Figure 2A for an example). In the dot plot, mito-processes are represented by bubbles, whereby the size reflects the normalized enrichment score and the colour represents the combined score. The *Y*-axis shows the significance of the test ($-\log_{10}$FDR), while the *X*-axis shows the different mito-processes. By clicking on one of the mito-process bubbles, all its values associated with the GSEApy analysis are displayed in the information panel. In the bar plot, mito-processes are ranked by combined score, whose value is also given in parentheses next to the mito-process.

Time-course visualization. In mitoXplorer 1.0, only six datasets could be compared to each other with the comparative plots. To overcome this limitation and to effectively visualize an increasing number of available time-course data, we implemented a time-course visualization interface. With this function, we allow users to visualize up to 20 time points of a temporal dataset. The structure of this visualization interface is similar to the comparative plots, with all its interactivity, except that genes can be individually selected from the sortable heatmap for display in the line plot by clicking on the gene name. Time points on the *X*-axis are connected via a line to visualize expression profiles over time. The user can thus visualize the temporal gene expression profile of a single or multiple mito-genes. Colours are assigned automatically to help visualize individual genes (see Figure 2B). This visualization was developed using the D3.js library (10) in order to produce a fully interactive visual interface.

Cross-species analysis. mitoXplorer 2.0 offers the possibility to compare datasets across species. Cross-species analysis is possible with the comparative plots or the heatmap function. This will help unravel similarities and understand differences in mitochondrial dynamics between species.

Orthology information has been included in the backend MySQL database of mitoXplorer by adding orthology ID tables for each model organism. These tables contain dictionaries, having as key the ID of the gene of the model organism of interest and as values the orthologous gene IDs in the other species available in mitoXplorer 2.0. Orthologs between the four model species offered by mitoXplorer were initially collected using the last release of BioMart (11). We manually curated these datasets and only 1:1 relations were kept.

The user has to first choose 'cross-species' from the 'Data Mining' drop-down menu and subsequently select the species to analyse. Only the comparative plot and the heatmap are offered for cross-species analysis. In the resulting plots of both visualization interfaces, the IDs of the first organism selected are used to label genes. However, the user can retrieve the ID of the ortholog of the second organism together with underlying information, by clicking on the respective cell in the heatmap or the bubble in the scatterplot.

Integrative analysis functions. In mitoXplorer 2.0, we also wanted to provide the possibility to identify mitochondrial regulators. More specifically, we wanted users to be able to find potential transcriptional regulators of co-regulated mito-genes, as well as identify potentially active signalling pathways from or to mitochondria. The interactivity of mitoXplorer allows one to directly select the gene(s) for both new integrative analysis functions starting from the com-

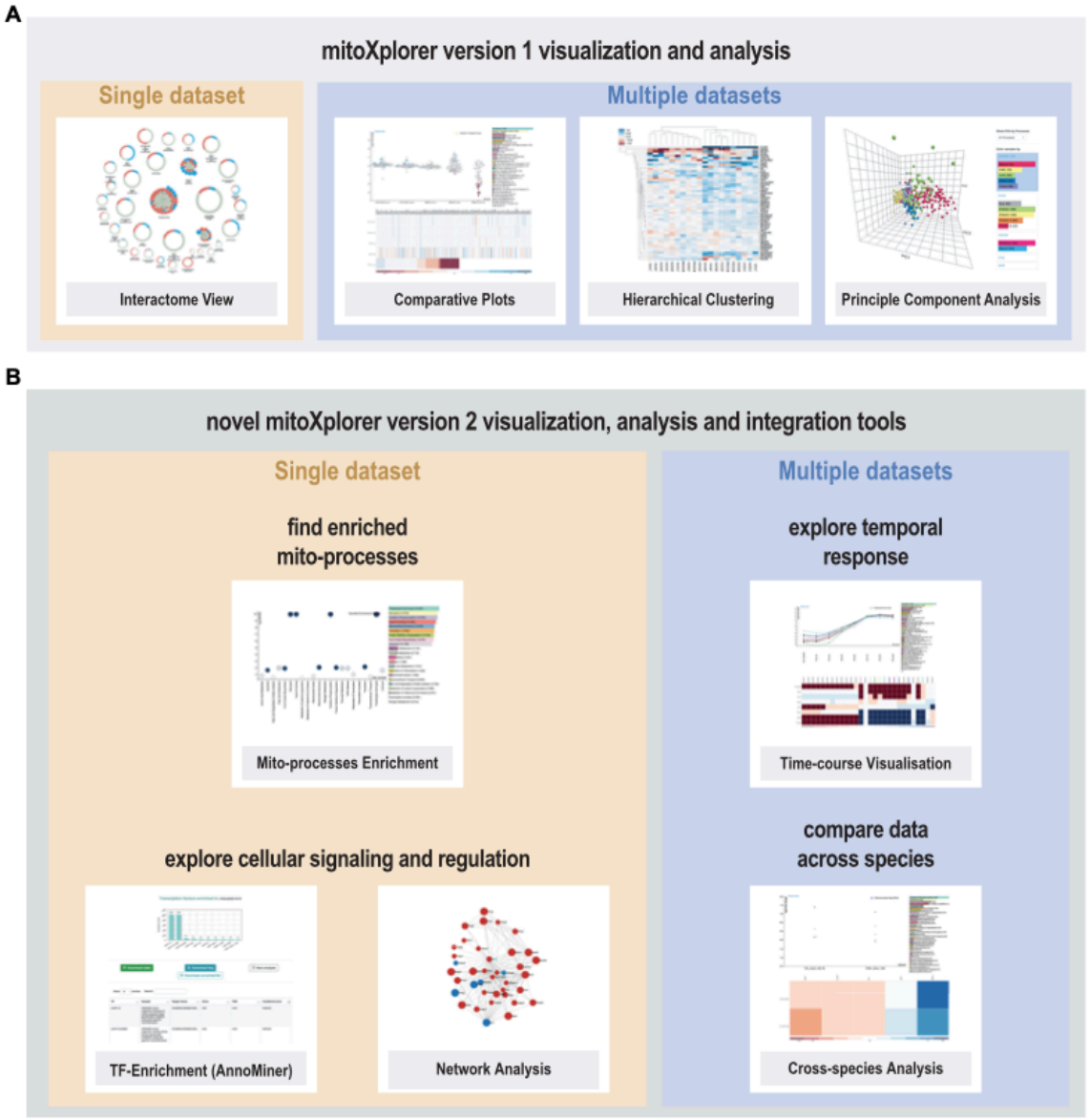


Figure 1. Collection of functions available in mitoXplorer 2.0. **(A)** Visualization and analysis tools available already in mitoXplorer 1.0 include the interactome view, comparative plots, hierarchical clustering and principal component analysis. **(B)** Novel functions available in mitoXplorer 2.0 are enrichment of mito-processes, time-course visualization for up to 20 time points, cross-species analysis, as well as the data integration functions of transcription factor (TF)-enrichment via AnnoMiner and network analysis using the viPEr algorithm.

parative plot, heatmap or time-course visualization functions.

TF-enrichment. Identifying potential transcriptional regulators is done by performing a TF-enrichment analysis via connecting mitoXplorer to the AnnoMiner web server (5) and through direct upload of data from mitoXplorer to AnnoMiner. In brief, AnnoMiner seeks enriched peaks of transcriptional regulators from the ENCODE (12), modENCODE (13) and modERN (14) databases in the promoter regions of a user-provided gene set. This set of genes for an AnnoMiner *TF enrichment analysis* can directly be selected in mitoXplorer 2.0 and uploaded to the AnnoMiner web server via the TF-enrichment function (see Supplementary Figure S2E for the TF-enrichment menu). The selection can be done from three different interfaces: comparative plots, heatmap and time-course visualization. After invoking the 'Integrative Analysis' TF-enrichment, the user needs to select the genes by clicking on the boxes of the genes in the (sortable) heatmap (see Supplementary Figure S3 and our mitoXplorer video tutorials for detailed instructions). The user must then perform TF-enrichment using the AnnoMiner web server (a detailed user tutorial of AnnoMiner can be found at http://chimborazo.ibdm.univ-mrs.fr/AnnoMiner/tutorial.html and Supplementary Figure S3). As a result, the user receives a list of TFs that potentially regulate the selected mito-gene set within the AnnoMiner web server environment.

Network analysis. The integrative network analysis function aims at identifying active subnetworks (representing potential signalling paths) from and to mitochondria, starting from one selected mito-gene. To search for active subnetworks, mitoXplorer 2.0 first embeds the mito-interactome into the entire cellular interactome and then allows to integrate differential expression data with this cellular interactome. The user can then explore the expression dynamics of the network neighbourhood of a single selected mito-gene and, based on differential expression data, identify the active subnetwork surrounding this gene. This function can be invoked from the interfaces: comparative plots, heatmap and time-course visualization. To perform a network analysis, the user must first choose from the 'Integrative Analysis' panel the network analysis function (see Supplementary Figure S2F). Then, they can select a mito-gene to start network analysis by clicking on the box of the gene in the (sortable) heatmap, the maximum number of steps from the starting point (mito-gene), the $\log_2$FC threshold (to define a gene as deregulated or not) and whether unregulated nodes in the results are allowed (see Supplementary Figures S2 and S4 for the network analysis menu; see also our mitoXplorer tutorials for further instructions). Results will be displayed directly in a new window in the mitoXplorer platform.

The network analysis function has been developed using the NetworkX Python library (15). The cellular interactomes were created from the STRING database (v11) (16) using high-confidence physical protein–protein interactions and have been added to the MySQL database. The network neighbourhood of a selected gene is explored by using the 'environment search' algorithm described by the viPEr Cytoscape app (6). Once the analysis has been performed, the active subnetwork is drawn using the D3.js library (10).

IDconverter. In order to facilitate data upload, we added an IDconverter to allow usage of any gene identifier. This function converts input gene IDs (e.g. ENSEMBL, Entrez, RefSeq, Genecode, Flybase) on the fly to gene symbols, without any intervention from the user. The conversion is performed using the latest release of BioMart (11). BioMart data were downloaded and stored in text format. We retrieved them locally for reasons of speed.

Upgrade to Python. In order to enable local deployments and future development of the platform, we upgraded the backend of the web tool from Python 2.7 to Python 3.6.

The new mitoXplorer 2.0 menu is detailed in Supplementary Figure S2A–F; details on how to use new mitoXplorer 2.0 functions are shown in Supplementary Figures S3 and S4 and online at the mitoXplorer 2.0 tutorial pages (http://mitoxplorer2.ibdm.univ-mrs.fr/tutorials).

Data processing for use cases 1 and 2

Flight muscle data (use case 1). Temporal expression data of developing *D. melanogaster* IFM were taken from GSE107247 (17) in the form of raw read counts. In order to obtain differential expression values for the different time points versus the mean over all time points, we created two pseudo-replicates of the mean by first calculating the mean over all time points using one replicate per time point each. Normalization of the raw read counts as well as pairwise differential expression analysis was performed using DESeq2 (18).

Ataxia data (use case 2). Data from the ataxia mouse model ATXN1_82Q_Tg from cerebellum were taken from GSE122099 (19). We downloaded sequencing reads and mapped them to mouse genome version mm10 using STAR (20) with default parameters. FeatureCounts (21) was used to calculate read counts. Normalization and differential expression analysis were done using DESeq2 via RNfuzzyApp (22). We compared all conditions to each other: 5-week wild type (WT) versus ATXN1_82Q_Tg, 12-week WT versus ATXN1_82Q_Tg, 5- versus 12-week WT, and finally 5- versus 12-week ATXN1_82Q_Tg. We uploaded the resulting data to mitoXplorer 2.0. For enrichment analysis of genes found in the active subnetworks extracted by mitoXplorer 2.0, we first downloaded and parsed the respective JSON files to obtain gene lists. These were then submitted to EnrichR (7); for further downstream analysis, we used the KEGG resource (23).

RESULTS

Use case 1: identifying transcriptional regulators of mito-genes during *D. melanogaster* indirect flight muscle development

We analysed temporal mito-gene expression dynamics during *D. melanogaster* IFM development using the new mitoXplorer 2.0 functions. In the developing indirect flight muscle, mitochondria undergo significant structural,

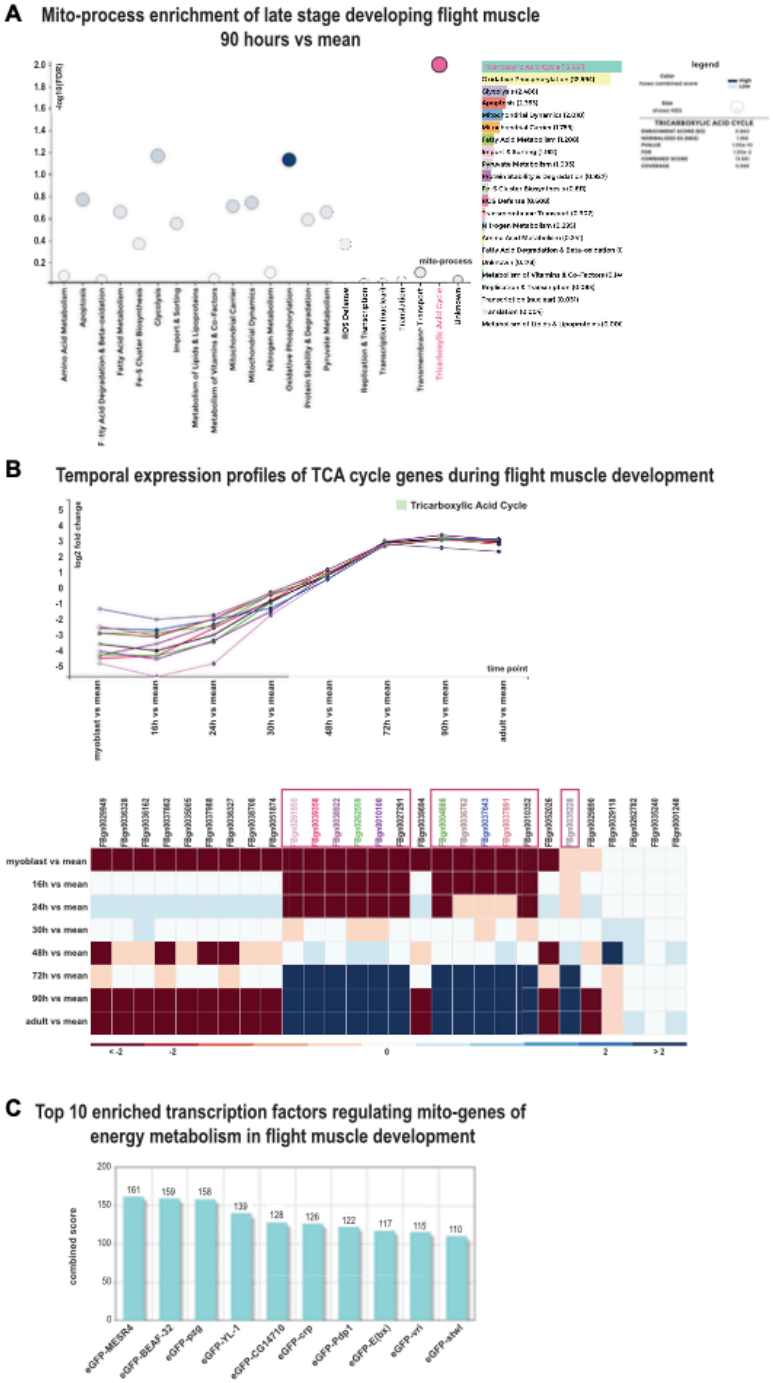


Figure 2. mitoXplorer 2.0 analysis of *D. melanogaster* developing indirect flight muscle (IFM). **(A)** Mito-process enrichment analysis of 90 h IFM after puparium formation (APF) revealed an enrichment of tricarboxylic acid (TCA) cycle, as well as OXPHOS. Mito-processes with a high combined score

but also bioenergetic changes (17,24). We wanted to explore these mitochondrial changes further using mitoXplorer 2.0. We used a transcriptomic resource generated by Spletter *et al.*, consisting of transcriptomic data from *Drosophila* isolated IFM at myoblast stage, 16, 24, 30, 48, 72 and 90 h after puparium formation (APF), and the adult stage (17). To observe temporal changes during IFM development, we compared each time point to the mean over all time points (see the 'Materials and Methods' section) and uploaded the data to mitoXplorer 2.0.

We first investigated which mito-processes are enriched during IFM development and used the new mito-process enrichment function. We used nearly mature IFM at 90 h APF for mito-process enrichment analysis. We found mito-processes related to OXPHOS-dependent energy metabolism enriched at this time point (Figure 2A and Supplementary Figure S3A).

We next investigated the temporal expression profiles of mito-genes from the mito-processes OXPHOS, OXPHOS mt and TCA cycle using the time-course visualization function of mitoXplorer 2.0. As exemplified in Figure 2B, we found several mito-genes within these three mito-processes that showed a specific temporal expression profile, being first downregulated (myoblast stage, 16, 24 and 30 h APF) and subsequently strongly induced from 72 h APF to the adult stage (see also Supplementary Figure S3B).

We were interested in identifying potential TFs responsible for regulating these co-regulated genes. We therefore made use of mitoXplorer 2.0's new integrative analysis function TF-enrichment. We selected all genes from the three mito-processes that showed the above-described temporal expression profile and uploaded them via mitoXplorer's TF-enrichment function to AnnoMiner (see Supplementary Table S1A for the list of genes selected and Supplementary Figure S3C and D). AnnoMiner returned several interesting TFs as enriched in the promoters of the co-regulated mito-genes (Supplementary Table S1B). Among the top 10 hits are *MESR4*, which has been shown to be involved in development and in the cellular response to hypoxia (25); *pzg*, which is known to regulate developmental processes (26,27); *Pdp1*, which is known to regulate muscle genes (28); crp, which controls cell growth and tracheal terminal branching (29); or *vri*, which has been shown to be involved in tracheal development (30,31).

To conclude, this use case demonstrates how the new functions of mitoXplorer 2.0, mito-process enrichment, time-course visualization and the integrative analysis function TF-enrichment could help identify the gene expression dynamics behind the mitochondrial bioenergetic switch in developing *Drosophila* indirect flight muscle and predict potential transcriptional regulators responsible for this switch.

Use case 2: identification of signalling cascades regulating Ca^{2+} signalling in a mouse model of spinocerebellar ataxia type 1

Spinocerebellar ataxias (SCAs) are a group of dominantly inherited neurodegenerative diseases, which are defined by a loss of coordination of body movements and cerebellar degeneration. They can be caused by mutations in nearly 40 genes and are currently untreatable. A disruption of Ca^{2+} signalling in cerebellar cells, and more specifically in the Purkinje neurons, is considered to play a key role in disease onset and progression (32). We wanted to explore the contribution of mitochondria to Ca^{2+} signalling deregulation in SCAs. To this end, we used a mouse model of spinocerebellar ataxia type 1 (SCA1). SCA1 affects the cerebellum as well as the inferior olive and is caused by polyglutamine expansion in the *ATAXIN1* gene, which encodes a transcriptional regulator (33). We used RNA expression data from a mouse model of SCA1 from (19), which looked at cerebellum and inferior olive of 5- and 12-week-old transgenic ATXN1_82Q (ATXN1_82Q_Tg) mice. We focused our analysis on the cerebellum and were mostly interested in the temporal differences of pathways possibly affecting mitochondrial Ca^{2+} signalling and transport.

We uploaded differential expression data of ATXN1_82Q_Tg compared against WT at 5 and 12 weeks, as well as the temporal comparisons of WT (5 and 12 weeks) and ATXN1_82Q_Tg (5 and 12 weeks), to mitoXplorer 2.0. We used comparative plots to visualize differences in gene expression in 'Ca^{2+} signalling and transport' in those four datasets. Among the genes most affected in this mito-process was *Itpr1* (Figure 3A and Supplementary Figure S4A), an intracellular receptor for inositol 1,4,5-trisphosphate that is located at the endoplasmic reticulum (ER). *Itpr1* is one of the main regulators of mitochondrial Ca^{2+} signalling (34) and is itself regulated by mitochondria (35). Furthermore, mutations in *ITPR1* cause SCA15 (36).

We next wanted to know how the network neighbourhood and thus potential signalling pathways change between the two disease stages. We therefore made use of the new integrative network analysis function in mitoXplorer 2.0 (see Supplementary Figure S4B). We extracted active subnetworks surrounding the *Itpr1* gene using differential expression data of 5- and 12-week-old ATXN1 mice, respectively, and then compared them first by visual inspection (Figure 3B and C; no deregulated nodes allowed, $\log_2 FC \geq |1.2|$, step size 3). The extracted active subnetwork surrounding *Itpr1* is substantially larger at 12 weeks with more differentially expressed nodes within a three-step neighbourhood (Figure 3C) compared to the earlier stage (Figure 3B). We then downloaded the two extracted, active subnetworks (Supplementary Figure S4C) for enrichment analysis using EnrichR and KEGG. Many of the en-

←

are highlighted in dark blue; the process TCA cycle has been selected and appears in pink. **(B)** Time-course visualization of developing *Drosophila* IFM, showing TCA cycle as an example. Genes that were downregulated early in development (myoblast, 30 h APF) and induced in late stages (72 h APF, adult) were selected for display in the line plot. Each visualized gene is shown in a different colour. **(C)** Top 10 enriched TFs potentially regulating mito-genes with the described expression profile (down early, up late) of the mito-processes OXPHOS, OXPHOS mt (mitochondrial-encoded genes) and TCA cycle. The plot is directly generated in AnnoMiner upon automatic upload of selected mito-genes within mitoXplorer 2.0's time-course visualization function. See Supplementary Figure S3 as well as mitoXplorer's video tutorials for instructions on how to generate these plots.

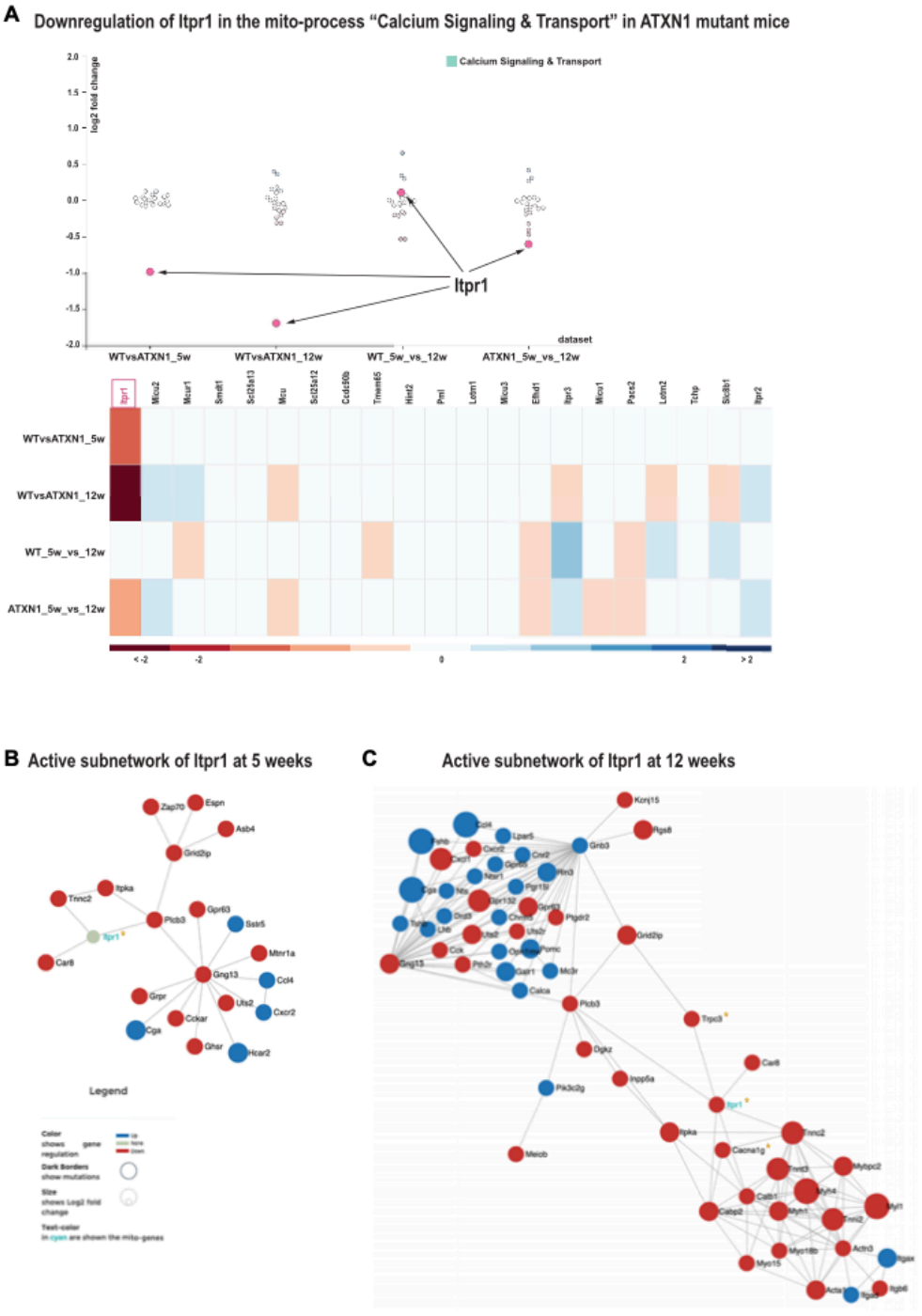


Figure 3. Ca^{2+} signalling and transport is affected in SCA1. **(A)** One of the most affected genes in the ATXN1_Q82 transgene mouse model for SCA1 is *Itpr1*, which localizes to the ER and regulates mitochondrial Ca^{2+} signalling. **(B)** Network neighbourhood of *Itpr1* in 5-week-old ATXN1_Q82_Tg mice.

riched pathways overlap between the two time points, such as calcium signalling, neuroactive ligand–receptor interaction or phosphatidylinositol signalling. At 12 weeks, some more metabolic pathways as well as SCA appear to be enriched (Supplementary Table 2B–D). Moreover, three genes responsible for SCAs are found in the 12-week network, including *Itpr1* itself, together with *Trpc3* and *Cacna1g* (Supplementary Table S2E).

In conclusion, we used the comparative plots and network analysis functions to investigate Ca^{2+} signalling deregulation in an SCA1 mouse model and found several signalling pathways consistently induced from 5 to 12 weeks in the cerebellum. Moreover, two more genes related to SCA were found in the 12-week active subnetwork.

DISCUSSION

Here, we present an important upgrade of the mitoXplorer visual data mining platform. While mitoXplorer version 1.0 (1) allowed to inspect mitochondrial datasets in multiple ways, mitoXplorer 2.0's integrative analysis functions permit the user to go beyond simple data visualization. Our new integrative functions can be used to better understand the interaction between mitochondria and their cellular environment, as we have shown in two of our use cases. In the first use case, we showed how the new functions of mitoXplorer 2.0 (the mito-process enrichment, time-course visualization and TF-enrichment) can be used to identify key transcriptional regulators of important mitochondrial processes. In the second one, we employed network analysis, which, by integrating the mitochondrial interactome into the complete cellular interactome, allows the identification of important non-mitochondrial genes that are connected with and potentially regulate mitochondrial processes. These results demonstrate the power of mitoXplorer 2.0's downstream and integrative analysis functions.

For TF-enrichment, we decided to rely on AnnoMiner (5). By connecting mitoXplorer 2.0 to AnnoMiner, direct data transfer from one to the other allows us to execute AnnoMiner's *TF enrichment analysis*, to identify potential transcriptional regulators of mitochondrial functions. To date, several tools for TF-enrichment are available, but to our knowledge only few of these, namely modEnrichr (37), iCisTarget (38) and AnnoMiner, work on model organisms other than human and mouse and thus can be used for three out of the four model organisms present in mitoXplorer 2.0: human, mouse and *D. melanogaster*. AnnoMiner has also data for *Caenorhabditis elegans*, a model species we plan to offer soon on mitoXplorer 2.0. For network analysis, we implemented a function for active subnetwork extraction starting from a single gene based on the algorithm of the Cytoscape app viPEr (6). The exploration of the network neighbourhood of a gene using viPEr requires only minimal user input: a selected gene; the number of steps from this gene for extracting an active subnetwork; a $\log_2$FC threshold for defining genes as deregulated; and finally, whether to accept connecting genes that are not deregulated. The algorithm thus avoids cumbersome tuning of parameters. The fact that we implemented this algorithm using NetworkX (15) and D3.js (10) permits a fully interactive exploration of the extracted subnetwork, facilitating its interpretation. In the future, we plan to include downstream analysis of the resulting network, such as pathway annotation through enrichment analysis. This should help to better identify potential signalling paths regulating mitochondrial functions. We find these two integrative analysis functions as an important novel upgrade of mitoXplorer, which should help to unravel how mito-gene expression and thus mitochondrial functions adjust to a specific cellular environment.

Additional future developments we are currently working on include mitochondrial metabolic modelling using flux balance analysis as well as logical modelling. Combining mitoXplorer directly with metabolic modelling would allow users to assess the metabolic states of biological samples based on mito-gene expression data directly within the mitoXplorer platform. Starting mitochondrial models are available for human and budding yeast (39–41), which we will adapt to our needs. In addition, we plan to extend mitoXplorer to allow analysis of single-cell RNA-seq data of tissues on cell-type level. All these further implementations are simplified by the modular fashion in which mitoXplorer 2.0 has been developed.

DATA AVAILABILITY

mitoXplorer 2.0 is available as a web server at http://mitoxplorer2.ibdm.univ-mrs.fr. The source code together with installation instructions is available in our GitLab repository (https://gitlab.com/habermann_lab/MitoX2).

Public gene expression data used for this study were obtained from the Gene Expression Omnibus repository at NCBI (42). *Drosophila* developmental IFM time-course data have been downloaded from GSE107247 (17). Mouse ATXN1 raw sequencing data have been downloaded from GSE122099 (19).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alice Carrier, Benoit Giannesini, Frank Schnorrer, Nuno Luis, Jerome Avellaneda, and the members of the IBDM Computational Biology team for helpful discussion on mitoXplorer 2.0. We thank Stephen Chapman for critical reading of the manuscript.

Author contributions: The upgrade of the mitoXplorer web server was developed by F.M. Data analysis was done by

(C) Network neighbourhood of *Itpr1* at 12 weeks. Genes found to cause a form of SCA are marked with an asterisk. Mito-genes are labelled in cyan, while non-mitochondrial genes are labelled in black. The user can interactively change gene positions and labels in order to show different information (gene name or ID, mutation, $\log_2$FC). The visualization is information dense. Node size, colour and border represent magnitude of $\log_2$FC, direction of the deregulation (up or down) and mutation information, when available. See Supplementary Figure S4 as well as mitoXplorer's video tutorials for instructions on how to generate these plots.

B.H.H., F.M. and M.H. The project was conceived by B.H.H. The manuscript was written by F.M. and B.H.H with input from M.H. All authors read and approved the final version of the manuscript.

FUNDING

Agence Nationale de la Recherche [ANR-18-CE45-0016-01 to B.H.H.]; Fondation pour la Recherche Médicale [MND202003011460 to B.H.H.]; Centre National de la Recherche Scientifique; Aix-Marseille University. Funding for open access charge: Agence Nationale de la Recherche [ANR-18-CE45-0016-01]; Fondation pour la Recherche Médicale [MND202003011460].
Conflict of interest statement. None declared.

REFERENCES

1. Yim,A., Koti,P., Bonnard,A., Marchiano,F., Dürrbaum,M., Garcia-Perez,C., Villaveces,J., Gamal,S., Cardone,G., Perocchi,F. *et al.* (2020) mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations. *Nucleic Acids Res.*, **48**, 605–632.
2. Friedman,J.R. and Nunnari,J. (2014) Mitochondrial form and function. *Nature*, **505**, 335–343.
3. Rath,S., Sharma,R., Gupta,R., Ast,T., Chan,C., Durham,T.J., Goodman,R.P., Grabarek,Z., Haas,M.E., Hung,W.H.W. *et al.* (2021) MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.*, **49**, D1541–D1547.
4. Glancy,B. (2020) Visualizing mitochondrial form and function within the cell. *Trends Mol. Med.*, **26**, 58–70.
5. Meiler,A., Marchiano,F., Haering,M., Weitkunat,M., Schnorrer,F. and Habermann,B.H. (2021) AnnoMiner is a new web-tool to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators. *Sci. Rep.*, **11**, 15463.
6. Garmhausen,M., Hofmann,F., Senderov,V., Thomas,M., Kandel,B.A. and Habermann,B.H. (2015) Virtual pathway explorer (viPEr) and pathway enrichment analysis tool (PEANuT): creating and analyzing focus networks to identify cross-talk between molecules and pathways. *BMC Genomics*, **16**, 790.
7. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
8. 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. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 15545–15550.
9. Xiao,Y., Hsiao,T.-H., Suresh,U., Chen,H.-I.H., Wu,X., Wolf,S.E. and Chen,Y. (2014) A novel significance score for gene selection and ranking. *Bioinformatics*, **30**, 801–807.
10. Bostock,M., Ogievetsky,V. and Heer,J. (2011) d^3: data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.
11. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
12. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
13. Muers,M. (2011) Functional genomics: the modENCODE guide to the genome. *Nat. Rev. Genet.*, **12**, 80.
14. Kudron,M.M., Victorsen,A., Gevirtzman,L., Hillier,L.W., Fisher,W.W., Vafeados,D., Kirkey,M., Hammonds,A.S., Gersch,J., Ammouri,H. *et al.* (2018) The modERN resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics*, **208**, 937–949.
15. Hagberg,A., Swart,P. and S Chult,D. (2008) *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab, Los Alamos, NM.
16. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
17. Spletter,M.L., Barz,C., Yeroslaviz,A., Zhang,X., Lemke,S.B., Bonnard,A., Brunner,E., Cardone,G., Basler,K., Habermann,B.H. *et al.* (2018) A transcriptomics resource reveals a transcriptional transition during ordered sarcomere morphogenesis in flight muscle. *eLife*, **7**, e34058.
18. 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.
19. Driessen,T.M., Lee,P.J. and Lim,J. (2018) Molecular pathway analysis towards understanding tissue vulnerability in spinocerebellar ataxia type 1. *eLife*, **7**, e39981.
20. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
21. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
22. Haering,M. and Habermann,B.H. (2021) RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. *F1000Research*, **10**, 654.
23. Kanehisa,M., Sato,Y. and Kawashima,M. (2022) KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.*, **31**, 47–53.
24. Avellaneda,J., Rodier,C., Daian,F., Brouilly,N., Rival,T., Luis,N.M. and Schnorrer,F. (2021) Myofibril and mitochondria morphogenesis are coordinated by a mechanical feedback mechanism in muscle. *Nat. Commun.*, **12**, 2091.
25. Lee,S.-J., Feldman,R. and O'Farrell,P.H. (2008) An RNA interference screen identifies a novel regulator of target of rapamycin that mediates hypoxia suppression of translation in *Drosophila* S2 cells. *Mol. Biol. Cell*, **19**, 4051–4061.
26. Kugler,S.J. and Nagel,A.C. (2007) putzig is required for cell proliferation and regulates notch activity in *Drosophila*. *Mol. Biol. Cell*, **18**, 3733–3740.
27. Zimmermann,M., Kugler,S.J., Schulz,A. and Nagel,A.C. (2015) Loss of putzig activity results in apoptosis during wing imaginal development in *Drosophila*. *PLoS One*, **10**, e0124652.
28. Lin,S.C., Lin,M.H., Horváth,P., Reddy,K.L. and Storti,R.V. (1997) PDP1, a novel *Drosophila* PAR domain bZIP transcription factor expressed in developing mesoderm, endoderm and ectoderm, is a transcriptional regulator of somatic muscle genes. *Development*, **124**, 4685–4696.
29. Wong,M.M.-K., Liu,M.-F. and Chiu,S.K. (2015) Cropped, *Drosophila* transcription factor AP-4, controls tracheal terminal branching and cell growth. *BMC Dev. Biol*, **15**, 20.
30. Gunawardhana,K.L., Rivas,G.B.S., Caster,C. and Hardin,P.E. (2021) Crosstalk between vrille transcripts, proteins, and regulatory elements controlling circadian rhythms and development in *Drosophila*. *iScience*, **24**, 101893.
31. Szuplewski,S., Fraisse-Véron,I., George,H. and Terracol,R. (2010) vrille is required to ensure tracheal integrity in *Drosophila* embryo. *Dev. Growth Differ.*, **52**, 409–418.
32. Paulson,H.L. (2009) The spinocerebellar ataxias. *J. Neuroophthalmol.*, **29**, 227–237.
33. Orr,H.T., Chung,M.Y., Banfi,S., Kwiatkowski,T.J.J., Servadio,A., Beaudet,A.L., McCall,A.E., Duvick,L.A., Ranum,L.P. and Zoghbi,H.Y. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **4**, 221–226.
34. Bezprozvanny,I. and Hayden,M.R. (2004) Deranged neuronal calcium signaling and Huntington disease. *Biochem. Biophys. Res. Commun.*, **322**, 1310–1317.
35. Parekh,A.B. (2003) Mitochondrial regulation of intracellular Ca^{2+} signaling: more than just simple Ca^{2+} buffers. *News Physiol. Sci.*, **18**, 252–256.

36. van de Leemput,J., Chandran,J., Knight,M.A., Holtzclaw,L.A., Scholz,S., Cookson,M.R., Houlden,H., Gwinn-Hardy,K., Fung,H.-C., Lin,X. *et al.* (2007) Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. *PLoS Genet.*, **3**, e108.
37. Kuleshov,M.V., Diaz,J.E.L., Flamholz,Z.N., Keenan,A.B., Lachmann,A., Wojciechowicz,M.L., Cagan,R.L. and Ma'ayan,A. (2019) modEnrichr: a suite of gene set enrichment analysis tools for model organisms. *Nucleic Acids Res.*, **47**, W183–W190.
38. Herrmann,C., Van de Sande,B., Potier,D. and Aerts,S. (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules. *Nucleic Acids Res.*, **40**, e114.
39. Smith,A.C. and Robinson,A.J. (2011) A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle. *BMC Syst. Biol.*, **5**, 102.
40. Smith,A.C., Eyassu,F., Mazat,J.-P. and Robinson,A.J. (2017) MitoCore: a curated constraint-based model for simulating human central metabolism. *BMC Syst. Biol.*, **11**, 114.
41. Malina,C., Di Bartolomeo,F., Kerkhoven,E.J. and Nielsen,J. (2021) Constraint-based modeling of yeast mitochondria reveals the dynamics of protein import and iron–sulfur cluster biogenesis. *iScience*, **24**, 103294.
42. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Methods Mol. Biol.*, **1418**, 93–110.

scientific reports

Check for updates

OPEN

AnnoMiner is a new web-tool to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators

Arno Meiler[1,3], Fabio Marchiano[2,3], Margaux Haering[2], Manuela Weitkunat[1], Frank Schnorrer[1,2] & Bianca H. Habermann[1,2]✉

Gene expression regulation requires precise transcriptional programs, led by transcription factors in combination with epigenetic events. Recent advances in epigenomic and transcriptomic techniques provided insight into different gene regulation mechanisms. However, to date it remains challenging to understand how combinations of transcription factors together with epigenetic events control cell-type specific gene expression. We have developed the AnnoMiner web-server, an innovative and flexible tool to annotate and integrate epigenetic, and transcription factor occupancy data. First, AnnoMiner annotates user-provided peaks with gene features. Second, AnnoMiner can integrate genome binding data from two different transcriptional regulators together with gene features. Third, AnnoMiner offers to explore the transcriptional deregulation of genes nearby, or within a specified genomic region surrounding a user-provided peak. AnnoMiner's fourth function performs transcription factor or histone modification enrichment analysis for user-provided gene lists by utilizing hundreds of public, high-quality datasets from ENCODE for the model organisms human, mouse, *Drosophila* and *C. elegans*. Thus, AnnoMiner can predict transcriptional regulators for a studied process without the strict need for chromatin data from the same process. We compared AnnoMiner to existing tools and experimentally validated several transcriptional regulators predicted by AnnoMiner to indeed contribute to muscle morphogenesis in *Drosophila*. AnnoMiner is freely available at http://chimborazo.ibdm.univ-mrs.fr/AnnoMiner/.

Transcriptional regulation is a highly complex process involving a combination of various molecular players and biochemical mechanisms, such as transcription factors (TFs), histone modifying enzymes, DNA methylases, as well as a structural reorganization of chromatin. Technical advances in analysing the interaction of proteins with DNA (ChIP-seq), to detect open or closed chromatin states (e.g. ATAC-seq, DNase-seq, FAIRE-seq), to detect hypermethylated CpG islands (bisulfite sequencing), or to map higher order chromosomal structural organisation (ChiaPET, Hi-C, 3C-seq) have revolutionized and significantly advanced our understanding of transcriptional regulation during the last decade[1,2]. Among other things, transcriptional enhancers were identified as crucial for regulating spatio-temporal gene expression programs by interacting with target gene promoters, often across large genomic distances (see[3–5] and references therein). Furthermore, the genome sequence in the chromatin is not a simple linear thread but organized in 3D, forming compartments, topologically associated domains (TADs) or chromatin loops that can bring distant elements in proximity, all of which can contribute to transcriptional regulation ([3,6] and references therein). More recent evidence suggests even the presence of dual-action cis-regulatory modules (CRMs) that act as promoters, as well as distal enhancers[7,8]. All these findings suggest that transcriptional regulation is far more complex than initially anticipated and involves a collective effort of specific binding sites in the genome, a complex genome structure and the presence of various transcriptional

[1]Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [2]Aix-Marseille University, CNRS, IBDM UMR 7288, The Turing Centre for Living systems (CENTURI), Aix-Marseille University, Parc Scientifique de Luminy Case 907, 163, Avenue de Luminy, 13009 Marseille, France. [3]These authors contributed equally: Arno Meiler and Fabio Marchiano. ✉email: bianca.habermann@univ-amu.fr

regulators. Hence, we need a tool that can ideally integrate all this information to better understand and predict transcriptional regulation.

Techniques such as ChIP-seq, ATAC-seq, Hi-C seq and others involve an NGS (Next Generation Sequencing) step, resulting typically in paired-end reads of isolated chromosomal fragments. Therefore, the first steps after sequencing consists of read mapping, which is usually done using a software such as BowTie2[9], followed by peak calling. There are a number of tools available for peak calling (reviewed e.g. in[10–13]), which include MACS2[14] for standard peak calling, ChIPdiff[15], EpiCenter[16] or diffReps[17] for differential peak calling. The output of peak callers are the genomic coordinates of epigenetic marks or transcription factor binding sites under study. These coordinates are commonly stored in the Browser Extensible Data (BED) file format, a light-weight, standardized format to share genome coordinates.

The next step to biologically interpret genomic coordinates (also called peaks) is their genomic annotation, a process referred to as peak annotation or gene assignment. A number of peak annotation tools exist. Some of them combine ChIP-seq data analysis (including peak calling) and peak annotation, such as the ChIP-Seq tools and web server ([18], web-based), Sole-Search[19], CIPHER[20], Nebula ([21], web-based), PeakAnalyser[22], BEDTools[23] or HOMER[24]. Some of them are specific to peak annotation and visualization, such as ChIPseeker[25], UROPA[26], annoPeak ([27], web-based), ChIPseek ([28], web-based), PAVIS ([29], web-based), Goldmine[30], GREAT ([31], web-based), or ChIPpeakAnno[32]. While most of the peak annotators assign peaks to the closest TSS (Transcription start site) or genome feature, others consider up- and down-stream gene features, or provide the overlap with gene feature attributes (such as promoter, 5′UTR, 3′UTR, exon, intron) of the nearest gene[18,21,22,25,26,29]. The web-tool GREAT[31] offers peak annotation and gene assignment in larger genomic regions based on Gene Ontology (GO)-term similarity. What appears missing is a web-based, visual tool that helps experimental biologists to explore and integrate peak data with transcriptomics data in a flexible, user-centred and easy way. AnnoMiner is designed to close this gap.

An enormous community effort has been invoked in the past decade to collect, standardize and present annotated genomic data in form of the ENCODE[33–35], modENCODE[36] and modERN[37] resources, with the aim to make sense of the encyclopaedias of genomes. These initiatives have also made it possible to explore available genomic data further and integrate them with each other as well as with user-generated data. These standardized, high-quality data can be used to identify enriched transcription factor binding events in promoters of co-regulated genes, for example from a differential gene expression dataset (e.g. an RNA-seq dataset). This form of data integration predicts possible transcriptional regulators for biological processes under study, and is already widely used in the community. Enrichment analysis is commonly performed by testing for TF overrepresentation in the promoter regions of a user-provided gene list compared to a background list (e.g. considering the entire genome). Most available tools define the promoter region rigidly as a range of upstream and downstream base pairs from the annotated gene transcription start site (TSS) and these parameters are kept fixed for all the TFs tested, without accounting for the differences between individual transcription factors. Following this approach, divers web-based tools have been developed, including VIPER[38], DoRothEA[39], BART[40], oPOSSUM[41], TFEA.ChIP[42], ChEA3[43], EnrichR[44] or i-cisTarget[45]. However, it would be advantageous to have a web-based, flexible and user-friendly solution for TF enrichment, considering promoter boundaries specific to each TF and working for the most widely used species (human, mouse, *Drosophila* and *C. elegans*).

Here we present AnnoMiner, a flexible, web-based, and user-friendly platform for peak annotation and integration, as well as TF and HM (histone modification) enrichment analysis. AnnoMiner allows users to annotate and integrate multiple genomic regions files with gene feature attributes and with transcriptomic data in an interactive and flexible way. AnnoMiner contains three distinct functions with different genomic peak annotation purposes in mind: first, *peak annotation*, which can be used to assign attributes of gene features to peaks, including user-defined upstream and downstream regions, TSS, 5′ and 3′UTRs, as well as the coding region; second, *peak integration* to search for overlapping binding events of up to five transcriptional regulators (e.g. different TMs; different HMs; or combinations of both) with gene feature attributes; these two annotation functions can optionally integrate user provided data, such as results from differential gene expression analysis, allowing to inspect for instance expression data and genomic peak data from multiple transcriptional regulators together. And third, *nearby genes annotation* or *long-range interactions*, which helps to identify long range interaction effects on gene expression of a single genomic region. The function for *nearby genes annotation* and *long-range interactions* requires the upload of data from differential expression analysis. As a fourth function, AnnoMiner performs *TF (Transcription factor) and HM (Histone Modification) enrichment analysis*, using all high-quality filtered TF and HM data from ENCODE, modENCODE and modERN, which are stored in an internal database, and a user defined gene lists as input. AnnoMiner's *TF enrichment function* offers the DynamicRanges option, which uses promoter regions specific to each TF based on pre-calculated binding densities for each individual TF. We tested the predictive power of AnnoMiner's *TF enrichment function* and applied it to flight muscle development in *Drosophila* focusing on the process of myofibril morphogenesis[46]. AnnoMiner predicted two potential transcriptional regulators, Trithorax-like (Trl) and the uncharacterized zinc-finger protein CG14655 to play a role during this process. For both, we provide experimental evidence for an essential function in myofibril morphogenesis in flight muscle. Thus, AnnoMiner correctly predicted a new role for Trl, as well as CG14655 as transcriptional regulators in a specific muscle type in *Drosophila*. Finally, we used AnnoMiner to predict direct transcriptional targets for the likewise enriched TFs Yorkie (Yki) and Scalloped (Sd) required for flight muscle growth[47].

Results

The AnnoMiner web server. The AnnoMiner web server is a new tool for the convenient annotation and integration of epigenetic, and transcription factor occupancy data for the wet-lab researcher. Based on an underlying library of java classes developed for the handling of annotated genomic peak or ranges datasets, the interactive graphical user interface of the web application provides functionalities for the upload of datasets, choice of analysis mode and model organism, visualization and download of analysis results (see Fig. 1 for a schematic of AnnoMiner functions and Supplementary Figure S1 for implementation details).

Genomic peak annotation functions. AnnoMiner's three annotation functions (*peak annotation, peak integration and nearby genes annotation*, Fig. 1) take as input one or more files containing genomic coordinates (in BED format), for example from ChIP-seq, with the aim of finding their associations with annotated gene features. This is done by determining the overlap between the genomic coordinates of a peak and the attributes of annotated gene features. The user can set the following parameters for the search: the minimum required overlap among the gene feature attributes and peak features (in bp or %); whether only the longest isoform of a gene or all its isoforms will be considered; the gene's directionality; the organism and the genome resource. Optionally, a user-provided annotation file can be uploaded, containing for instance differential expression data which will be integrated with the gene lists generated from AnnoMiner annotation. The user-provided dataset is accepted in csv (comma separated values) or tsv (tab separated values) format. It can contain up to 6 columns, without any constraints in content, except the first column has to contain gene IDs to allow integration with AnnoMiner results.

Peak annotation. The *peak annotation* function computes the total coverage of the user-provided genomic regions (representing the peaks) with the attributes of each annotated gene feature in the genome assembly. AnnoMiner considers already annotated attributes of gene features (5′UTR, CDS and 3′UTR) as well as attributes or gene features provided by the user; in particular the promoter region is fully customizable with respect to the upstream and downstream region of the annotated TSS, as is the 5′ flanking region upstream and 3′ flanking region downstream of the gene body (Fig. 2a). The first result shown by AnnoMiner is a coverage plot, visualizing the total base pair coverage of all peaks with the annotated attributes of the gene features (Fig. 3a). The user next chooses a target region (corresponding to a gene feature attribute) by clicking on one of the bars and on the 'Show Genes!' button (Fig. 3b). An interactive, sortable and downloadable table of all genes which overlap with the selected target region with peaks from the BED file is returned to the user (Fig. 3c). If the user provides a gene-based, custom annotation file, for instance containing differential expression data, these data will be integrated and displayed in the resulting table (Fig. 3d). While an annotation file will in most cases contain differential expression data based on RNA-seq analysis, it can contain any numerical or even text data. In summary, *peak annotation* in AnnoMiner annotates genomic regions provided by the user with gene features and their attributes in a flexible and user-centred way.

While we here demonstrate the usability of AnnoMiner's *peak annotation* function with a TF ChIP-seq dataset, which usually comprises narrow peaks in the vicinity or within gene promoters, it can be also used for any type of genomic peak file. In Supplementary Figure S2 (together with Supplementary Table S1), we show peak annotation for activating, as well as repressing histone modifications during early *Drosophila* development.

Peak integration. The *peak integration* function performs the *peak annotation* analysis, but for up to five genomic regions files (representing peaks from independent TFs, TFs together with HMs or independent HMs), allowing the user to integrate peaks from different transcriptional regulators and identify gene features and their attributes that overlap with them (Fig. 2b). The same algorithm as in *peak annotation* is used for the annotation of each individual peak file. With this function, genes co-regulated by TFs can be identified, or TF datasets can be integrated with chromatin structure data defined by histone modifications, or other epigenetic information derived by other experimental techniques. The coverage plots of all chosen genomic region files are returned by AnnoMiner (Fig. 4a), and the user chooses a target region for all and clicks on the 'Show Genes!' button (Fig. 4b). An interactive, sortable and downloadable table is returned with all genes that have a peak of the transcriptional regulators or epigenetic marks in their selected target regions (Fig. 4c). Custom gene annotation, such as differential expression data can again be provided, which is then integrated and displayed in the results table.

To show that AnnoMiner's *peak integration* function can be used to integrate three different datasets in one analysis step, we re-analysed a data series published on STAT3 function in different forms of diffuse large B-cell lymphomas (DLBCL, GEO super-series GSE50724[48]). Two subtypes of DLBCL are known, germinal centre B-cell-like (GCB) and activated B-cell-like (ABC). The ABC type responds only poorly to available therapies and can be often associated with an overexpression of STAT3[48]. The authors had compared STAT3 binding by ChIP-seq analysis between 8 patient-derived cell lines from GCB- and ABC type. They had performed RNA-seq analysis of the same cell lines to retrieve differentially expressed genes between the two subtypes. We made use of these data to identify genes with increased expression levels in the ABC type together with increased STAT3 binding events. We only considered STAT3 peaks that were significantly upregulated (FDR 0.05, fold change 1.25) in ABC-type DLBCL. To demonstrate the added value of Annominer's *peak integration* function, we used H3K4me3 data from ENCODE from one of the ABC cell lines, OCI-Ly3 (accession: ENCFF763KFL), to limit the search to active promoters. Both peak files showed highest coverage with the direct promoter region of associated genes (Fig. 4a,b). We selected the major peaks for further analysis and integrated resulting peaks with differentially expressed genes from the same study. Of the upregulated unique genes, 42 contained an upregulated STAT3 peak in their promoter, as well as a H3K4me3 histone modification (Supplementary Figure S3a, Supplementary Table S2). We submitted the list of 42 genes to the EnrichR web-server[44] and could identify terms strongly related

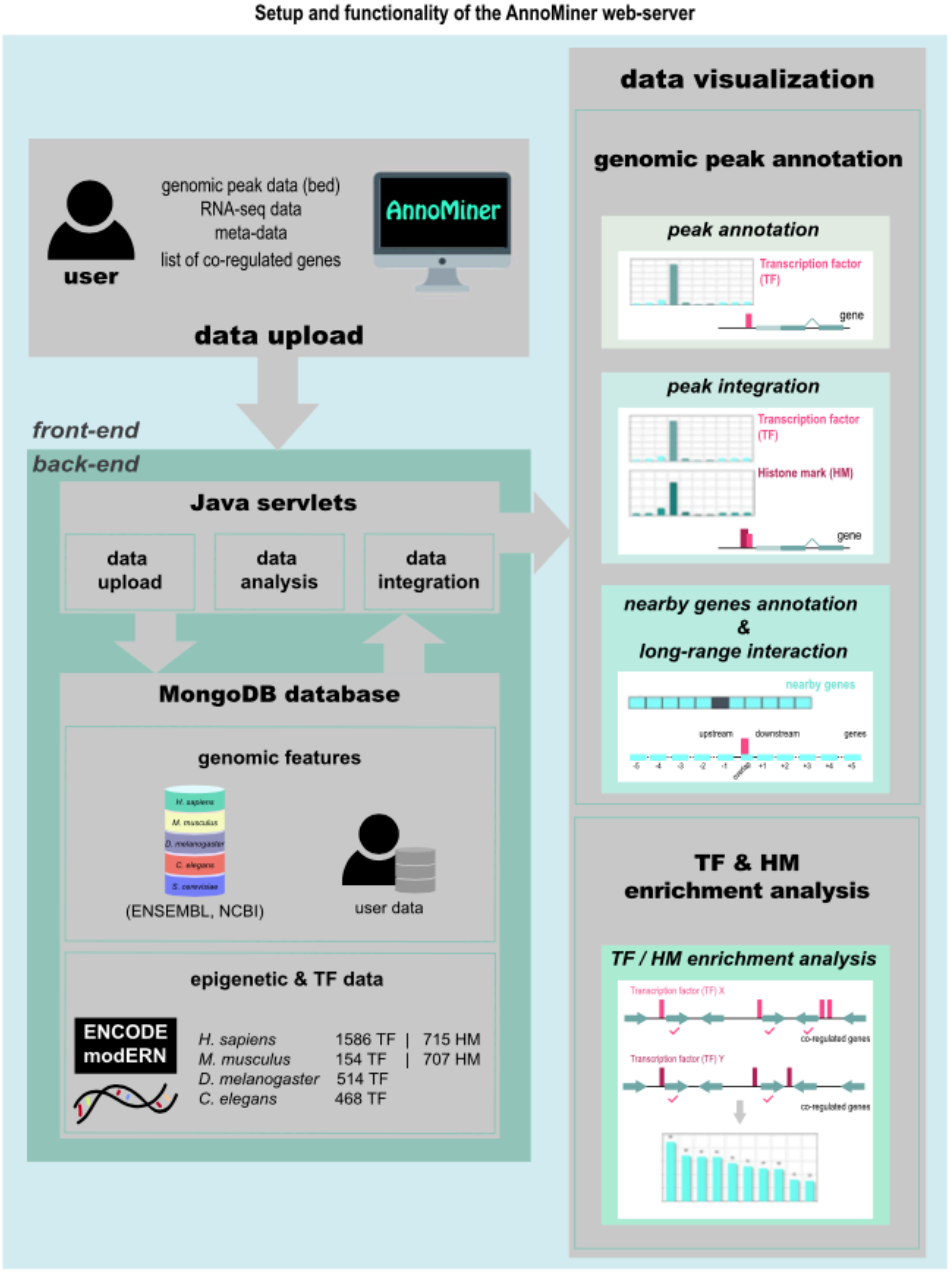


Figure 1. Setup of the AnnoMiner web-server. The user interacts with the web-server via the front-end for data upload and data mining. Genomic peak data (in BED format), annotation data, e.g. from transcriptomic analysis, or a list of co-regulated genes can be uploaded to the server. The user can either annotate peaks (*peak annotation*), integrate peaks (*peak integration*) or annotate and analyse nearby genes (*nearby genes annotation, long-range interactions*) of a peak. AnnoMiner's fourth function allows searching for enriched transcription factor binding or histone modification events in the promoter regions of a list of co-regulated genes (*TF & HM enrichment analysis*). In the back-end, Java servlets are responsible for data upload, analysis, integration and ID conversion. User data, genome feature data from ENSEMBL, UCSC and NCBI, as well as ENCODE, modENCODE and modERN ChIP-seq data are stored in a MongoDB database. Currently, we host genome feature data from *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans* and *S. cerevisiae*. Data from ENCODE, modENCODE and modERN are available in AnnoMiner for human, mouse, *Drosophila* and *C. elegans*.

to cancer, diffuse large B-cell lymphomas of the ABC type, IL10/STAT3 signalling and other relevant terms for the disease under study (Supplementary Figure S3b,c; Supplementary Table S2). Moreover, AnnoMiner identified 28 additional direct targets of STAT3 compared to the ones already described in the original study (Supplementary Table S2). To summarize, AnnoMiner's *peak integration* function was able to identify genes directly associated with ABC-type diffuse large B-cell lymphomas and potential direct targets for the transcription factor STAT3 by using a single user interface and a single AnnoMiner analysis step.

To demonstrate the usability of AnnoMiner's *peak integration* function further, we used four different datasets, as well as different types of genomic data. The results are displayed in Supplemental Material: first, (Supplementary Figure S4, Supplementary Table S1), we integrated H3K4me3 peak data during four stages of *Drosophila* MZT (maternal-to-zygotic transition) and identified its associated genes throughout MZT, which are thus constitutively transcribed from early to late MZT[49]. Second, we integrated ATAC-seq data with ChIP-seq data of an early embryonic TF and genome modifier, GAF/Trl, to identify genes activated by Trl in later stages of MZT[50] (Supplementary Figure S5, Supplementary Table S3).

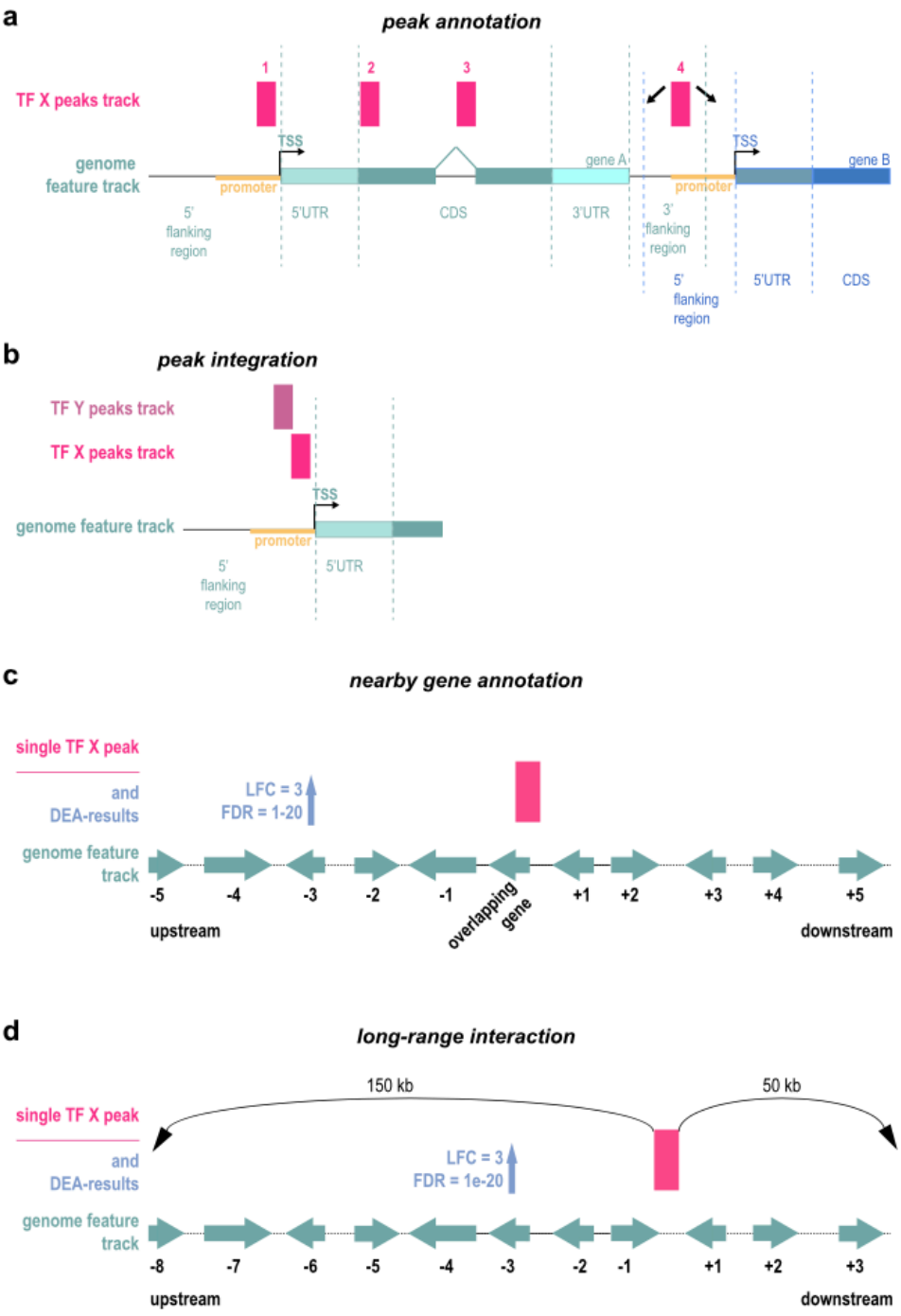
Nearby genes annotation and long-range interactions. The *nearby genes annotation* function allows the user to visualize the differential regulation and retrieve the overlapping, as well as the closest 5 gene features up- and downstream of an individual peak (Fig. 2c). This function is most useful for exploring gene regulation in the vicinity of a genomic mutation or deletion in a non-coding region. Next to a BED file with the region of interest containing the mutation or genomic deletion, the user uploads an annotation file containing significantly differentially expressed genes. The resulting interactive AnnoMiner plot depicts the genomic neighbourhood of the peak, with the differential regulation of the overlapping and the five closest genes up- and downstream (Fig. 5a). The user can choose to visualize only the deregulation, or can discriminate between up- and down-regulation of the genes. In the latter case, the colour of the box reflects the direction of differential expression (either blue for up-, red for downregulated, green if an equal number of genes are up- and downregulated or grey if unchanged, (Fig. 5a). By selecting one or more boxes (Fig. 5b), the selected genes, as well as their log2FC and FDR with the user's annotation file will be returned in a table (Fig. 5c).

Alternative to the *nearby genes annotation*, users can also utilize the *long-range interactions* function of AnnoMiner to explore the neighbourhood of a peak (Fig. 2d). In this case, a range of base-pairs has to be chosen up- and downstream of the peak, which is then decorated with information on differential expression from the user-provided annotation file (Fig. 5d). After selecting the up- or downstream region (Fig. 5e), the user can retrieve the genes within that region of the peak together with their differential expression data (Fig. 5f). The *long-range interactions* function will have its best use whenever Hi-C data are available, as these types of data will give insight about the genomic boundaries and thus help to choose the genomic neighbourhood of a peak.

As a proof of concept for the *nearby genes annotation* function, we used a study that had shown the requirement of long-range enhancers regulating Myc expression for normal facial morphogenesis[51] (GEO dataset GSE52974). In humans, cleft lip or cleft palate (CL/P) is a frequent congenital malformation. This malformation has been associated with risk factors located at a 640 kb noncoding region on chromosome 8. The corresponding region in mouse was studied by Uslu and colleagues and refined to a more specific enhancer region, the medionasal enhancer (MNE[51]). Deletions within the MNE in mouse led to smaller snouts and abnormalities of nasal and frontal bones amongst other defects. Myc was the only gene observed to be differentially expressed in the vicinity of this deletion. We used the CL/P deletion 8–17 (chr15:62668548–63550550) from Uslu et al. to create a single-peak BED file and uploaded it together with the significantly differentially expressed genes from the re-processed RNA-sequencing data from the same strain compared against control (GEO dataset GSE52974) to test the AnnoMiner *nearby genes annotation* function (Supplementary Table S4). Indeed, only a single gene is significantly differentially expressed (pink box, Fig. 5a), which is Myc (Fig. 5c). In principle, this function can also be used to explore the expression dynamics within the gene neighbourhood of multiple peaks (see Supplementary Figure S6). However, large-scale long-range gene regulatory data of this type are very sparse and their interpretation remains too complex to be exhaustively analysed with a tool like AnnoMiner. In summary, AnnoMiner's *nearby genes annotation* or *long-range interaction* functions help in an easy and quick way to identify deregulated genes in the neighbourhood of one genomic position.

Transcription factor & histone modification enrichment analysis. Transcription factor binding sites inferred from experimentally detected TF peaks in the genome can be used to predict TFs, which potentially co-regulate gene-sets. AnnoMiner's *TF & HM enrichment analysis* function identifies enriched peaks in the promoter regions of a user-provided gene list, for instance co-regulated genes from a transcriptomic analysis. Any valid identifier is accepted, as AnnoMiner performs gene ID conversion on-the-fly using BioMart[52]. AnnoMiner considers a gene as a potential target, if its promoter overlaps with a TF peak. The user can either choose the promoter region (up- and downstream number of base-pairs from the TSS) or use the *DynamicRanges* calculated by AnnoMiner, which is based on the distribution of a TF binding event relative to the TSS and therefore specific for each TF (see "Methods", not available for histone modifications). The results of the enrichment analysis are visualized as an interactive bar plot (for the first 10 hits, Fig. 6a), as well as an interactive table. In the table, all available TF ChIP-seq datasets in the AnnoMiner database for the species of interest are ranked according to their *Combined Score*. Along with this value, AnnoMiner also reports information about the experimental condition, cell line or developmental stage, contingency table values, p-value, enrichment score, FDR and the list of potential targets of the TF (Fig. 6b) in the downloadable version of the table.

As a proof of principle for predicting transcriptional regulators we selected a differential expression dataset from *daf-16/FoxO* mutants in *C. elegans*[53]. When uploading the list of all DAF-16A/F targets provided in[53] to AnnoMiner's *TF & HM enrichment analysis* function, daf-16 was the 4th most significantly enriched transcription


a
peak annotation
TF X peaks track
genome feature track
1
2
3
4
TSS
promoter
5' flanking region
5'UTR
CDS
3'UTR
gene A
gene B
3' flanking region
b
peak integration
TF Y peaks track
TF X peaks track
genome feature track
c
nearby gene annotation
single TF X peak
and DEA-results
LFC = 3
FDR = 1-20
genome feature track
-5
-4
-3
-2
-1
overlapping gene
+1
+2
+3
+4
+5
upstream
downstream
d
long-range interaction
150 kb
50 kb
single TF X peak
and DEA-results
LFC = 3
FDR = 1e-20
genome feature track
-8
-7
-6
-5
-4
-3
-2
-1
+1
+2
+3
upstream
downstream

◂**Figure 2.** Different scenarios for the integration of peaks with gene features. (**a**) The peaks from a transcription factor (TF) track can be located in the promoter region or 5′ flanking region (peak 1) of gene A, in the coding region (peaks 2 and 3) or in the 3′ flanking region of gene A (peak 4). Peak 4 is at the same time located in the 5′ flanking region/promoter region of gene B. Its association is therefore not strictly clear, which should be considered during peak annotation. (**b**) The peaks of two transcriptional regulators (two TFs, two Histone Modifications (HMs), or a TF and a HM, etc.) can be integrated based on their overlapping or nearby location in different attributes of a gene feature, as here demonstrated by their common overlap with the promoter region of a gene. (**c**,**d**) Activity of a transcriptional regulator, such as an enhancer, is not necessarily linked to the closest gene, but could affect neighbouring genes up- or downstream of the peak. Data from a differential expression analysis (DEA) of nearby genes can be correlated with a genomic peak harbouring a non-coding mutation or a genomic region harbouring a genome deletion. *Long-range interactions* can also be determined by specific genomic boundaries defining so-called TAD-domains (topologically associating domains).

factor (Fig. 6a). Interestingly, pqm-1, which has been shown to bind to daf-16 response elements[54], was found at 1st and 2nd position by AnnoMiner. Elt-2, a GATA-like transcription factor, which appeared as 3rd most significant hit, has been shown to bind to promoters of some daf-16-regulated genes and to be required for their regulation[55]. To conclude, AnnoMiner's TF & HM enrichment function is a powerful tool for predicting relevant transcription factors co-regulating sets of genes with similar expression patterns.

Performance evaluation of AnnoMiner's *TF & HM enrichment analysis* function. We wanted to compare the performance of AnnoMiner's *TF & HM enrichment analysis* function with other web-tools for TF enrichment analysis. We followed in principle the evaluation protocol proposed by Keenan et al., which used PR-AUCs and ROC-AUCs calculated from the PPROC R-package for estimating performance[43] (for details see also "Methods"). In brief, we took manually curated datasets provided by[43] containing single TF perturbation experiments followed by RNA-seq from Gene Expression Omnibus[56] (GEO). Gene expression data used for benchmarking were restricted to experiments targeting TFs for which AnnoMiner is storing at least one high quality TF ChIP-seq dataset for the human assembly GRCh38, resulting in a total of 75 datasets that we could use for benchmarking. We submitted the list of significantly differentially expressed genes between perturbed TF versus wild-type control (which we hereafter refer to signature gene-sets) from these experiments to perform enrichment analysis using different tools, including AnnoMiner. We used the rank of the perturbed TF in the resulting enrichments of its associated signature gene-set to calculate PR-AUCs and ROC-AUCs. We furthermore calculated the cumulative distribution function for the ranks of each TF across all the experiments it was perturbed in. Only if a TF ranks randomly, the distribution function will be uniform; we performed Anderson–Darling tests to detect deviation from uniformity. We then computed the percentage of perturbed TFs that were correctly ranked within the first percentile to ensure that TFs were ranking high.

We chose the following web-tools for comparison: ChEA3, TFEA.ChIP and EnrichR (Table 1). ChEA3 offers the user 2 different methods to rank predicted TFs (meanRank and topRank) and we evaluated both ranking methods. EnrichR offers different resources for enrichment analysis and we used the resources ARCHS4, ChEA 2016, ENCODE 2015, ENCODE and ChEA Consensus and TRRUST 2019 for enrichment analysis, respectively. The number of datasets included in the benchmarking set did depend on the resource tested (see "Methods"). The Anderson–Darling test returned significant results for all web-tools tested (ADtest in Table 1), except for EnrichR in combination with the ENCODE_and_ChEA_Consensus resource. This highlights the ability of all tools to rank the perturbed TF among the top candidates of the results. AnnoMiner outperformed TFEA.ChIP in all categories including percentage recovered TFs in the 1st percentile (7.0 vs 0.0), ROC AU (0.69 vs 0.63) and PR AUC (0.68 vs 0.60). EnrichR differed in performance depending on the resource used. On average, it outperformed AnnoMiner on the percent recovered TFs (9.5 vs 7.0), while AnnoMiner reached slightly higher values in ROC AUCs (0.69 vs 0.66) and PR AUCs (0.69 vs 0.68). ChEA3 performed similar for both ranking methods used and outperformed AnnoMiner in all categories. To summarize, though it cannot reach the performance of ChEA, AnnoMiner outperforms the other evaluated tools in identifying relevant TFs at a high rank the in gene-sets derived from TF perturbation studies. Other than ChEA, however, AnnoMiner is available for all four major model organisms, including the invertebrates *Drosophila* and *C. elegans*.

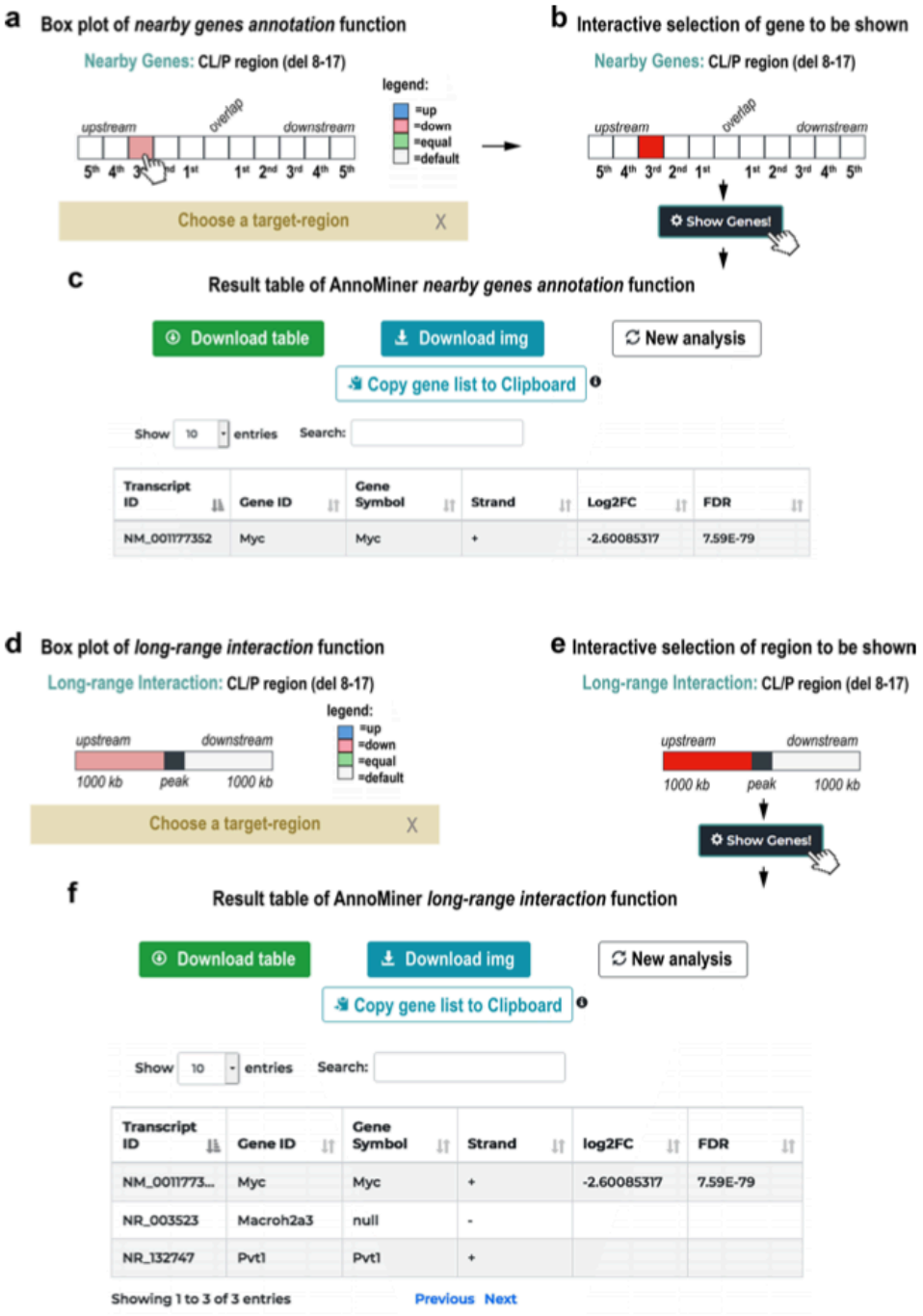
Using AnnoMiner to identify important transcriptional regulators of *Drosophila* flight muscle morphogenesis and growth. The assess the performance of AnnoMiner further, we tested its *TF & HM enrichment analysis* function to predict unknown transcriptional regulators for a list of co-regulated genes. We chose a dataset quantifying gene expression dynamic during development of *Drosophila melanogaster* indirect flight muscles[46], in which mRNA from indirect flight muscles had been isolated at several time-points correlating with key steps during muscle development (Fig. 7a). We focused on the development of the contractile apparatus called myofibrillogenesis and compared gene expression at 30 h after puparium formation (APF), when myofibrils assemble, with 72 h APF, when myofibrils have matured (Fig. 7a). This comparison revealed 2193 differentially expressed genes that were shown in the original study to be strongly enriched for genes relevant for myofibril and mitochondrial development. We submitted this differential gene list to AnnoMiner (Supplementary Table S5) and searched all the 514 modERN and modENCODE TF ChIP-seq datasets from *Drosophila* stored in AnnoMiner for a potential enrichment of peaks. This identified 42 unique TFs significantly enriched with an FDR < 0.05. (Supplementary Table S5). The top 10 enriched datasets included Deaf1, Trl, Hr78 and cwo (Fig. 7b). Hence, these are potential transcriptional regulators of flight muscle development.



Figure 3. *Peak annotation* function of AnnoMiner. In the *peak annotation* function, AnnoMiner allows users to annotate peaks from a BED file. (**a**) After uploading a BED file to the server and choosing the function *peak annotation*, a coverage plot is produced, showing the covered base-pairs/total length relative to attributes extracted from the gene features of an organism's genome. These include the 5′ upstream flanking region, the TSS (− 2000 bp and + 500 bp), the 5′UTR, coding sequence (CDS), 3′UTR, as well as the 3′ flanking region, whereby the boundaries for the TSS and flanking regions can be set by the user. (**b**) The plot is interactive and the user should select the target region (corresponding to the attribute of interest) to be analysed by selecting one or more bars to proceed. (**c**) When clicking on 'Show Genes!', the user retrieves a list of associated genes that fulfils the selected overlap criteria with the peaks. The table can be sorted according to different values, downloaded, browsed online, or the gene list can be copied to the clipboard for further analysis (e.g. enrichment analysis). The user can also select a new target region or start a completely new analysis. Significant differential peaks from STAT3 (GEO dataset GSE50723) were uploaded for producing this Figure. For (**d**), differential expression data were uploaded from the associated GEO dataset GSE50721 from super-series GSE50724.



Figure 4. The *peak integration* function of AnnoMiner. If a user wants to integrate peaks from two transcriptional regulators and annotate them with gene feature attributes, the *peak integration* function must be chosen. (**a**) Two coverage plots of the two selected factors with gene feature attributes are shown to the user, from which a target region for each of them must be chosen (**b**). (**c**) When clicking on 'Show Genes!', the associated genes are displayed as a table. Optionally, the user can upload a custom annotation file that contains for instance differential expression data (log2FC, FDR, etc.). These data are then shown in the resulting table. The table can be sorted according to different values, downloaded, browsed online, and the list of genes can be copied to the clipboard for further analysis. The user can also select a new target region or start a completely new analysis. Significantly upregulated peaks from STAT3 (GEO dataset GSE50723 from super-series GSE50724), as well as H3K4me3 methylation data from a comparable cell line (GEO dataset GSE86718) were uploaded for producing this Figure; differential expression data were taken from the associated GEO dataset GSE50721 from super-series GSE50724.


a
Box plot of nearby genes annotation function
Nearby Genes: CL/P region (del 8-17)
legend:
=up
=down
=equal
=default
upstream
overlap
downstream
5th 4th 3rd 2nd 1st 1st 2nd 3rd 4th 5th
Choose a target-region
X
b
Interactive selection of gene to be shown
Nearby Genes: CL/P region (del 8-17)
upstream
overlap
downstream
5th 4th 3rd 2nd 1st 1st 2nd 3rd 4th 5th
Show Genes!
c
Result table of AnnoMiner nearby genes annotation function
Download table
Download img
New analysis
Copy gene list to Clipboard
Show 10 entries
Search:
Transcript ID
Gene ID
Gene Symbol
Strand
Log2FC
FDR
NM_001177352
Myc
Myc
+
-2.60085317
7.59E-79
d
Box plot of long-range interaction function
Long-range Interaction: CL/P region (del 8-17)
legend:
=up
=down
=equal
=default
upstream
downstream
1000 kb
peak
1000 kb
Choose a target-region
X
e
Interactive selection of region to be shown
Long-range Interaction: CL/P region (del 8-17)
upstream
downstream
1000 kb
peak
1000 kb
Show Genes!
f
Result table of AnnoMiner long-range interaction function
Download table
Download img
New analysis
Copy gene list to Clipboard
Show 10 entries
Search:
Transcript ID
Gene ID
Gene Symbol
Strand
log2FC
FDR
NM_00117773...
Myc
Myc
+
-2.60085317
7.59E-79
NR_003523
Macroh2a3
null
-
NR_132747
Pvt1
Pvt1
+
Showing 1 to 3 of 3 entries
Previous
Next

◂**Figure 5.** The *nearby genes annotation* function of AnnoMiner. (**a**) AnnoMiner's *nearby gene annotation* function shows the differential regulation of 5 up- and down-stream, as well as the overlapping gene of a selected genomic peak. Next to a BED file with a single genomic region (a peak, a SNP or a genomic deletion), the user must upload an annotation file containing information on differential gene expression. Either general differential expression, or up- and down-regulation of neighbouring genes can be displayed. The user can also choose to consider directionality of the gene on the genomic strand. In the chosen example, a single gene is downregulated in the vicinity of the uploaded genomic regions file. (**b**) By selecting one or more boxes and pressing the 'Show Genes!' button, (**c**) the identity, as well as the log2FC and FDR of the selected gene(s) is returned in the table. (**d**) The *long-range interactions* function returns a plot with the up- and downstream regions from a peak, decorated with differential expression information taken from the user-provided annotation file. In this case, 1000 kb up- and downstream was chosen, whereby a down-regulated gene can be found in the upstream region. (**e**) After choosing this region and clicking on the 'Show Genes!' button, (**f**) a table is returned with all genes in the selected region, together with their differential expression data. We chose data from a study that had shown the requirement of long-range enhancers regulating Myc expression for normal facial morphogenesis (GEO dataset GSE52974). Upon deletion of the medionasal enhancer (MNE) in mouse, Myc was the only gene observed to be differentially expressed in the vicinity of this deletion.

To identify which of the 42 transcriptional regulators may have a function during flight muscle development we next integrated data from an RNAi-screen for muscle function[57], which had assayed for viability, flight muscle performance and body locomotion after muscle specific knock-down of individual TFs. From the 42 TFs identified by AnnoMiner, knock-down of two TFs resulted in flightless animals (*CG14655, cwo*), and seven were scored as lethal during development (*Trl, Hr78, lola, Vsx2, Pif1B, salr, Hr51*) (Supplementary Table S5); 21 TFs did not show a phenotype in this assay and the remaining 13 had not been tested (Fig. 7c).

Trithorax-like and the uncharacterized Zinc-finger protein CG14655 are required for flight muscle morphogenesis. For experimental verification we selected two proteins, Trl (Trithorax-like) and an uncharacterized zinc-finger protein called CG14655. We used muscle-specific knock-down to investigate a putative function of both genes in muscle. For *Trl* knock-down we used 4 independent transgenic RNAi lines driven with muscle-specific *Mef2*-GAL4. Two of those resulted in pupal lethality and the other two resulted in viable but flightless flies, demonstrating a function of *Trl* in flight muscle (Supplementary Table S6). For morphological analysis we visualized the myofibrils of flight and leg muscles of mature 90 h APF pupae in wild type and three different *Trl* knock-down lines. We found that knock-down of *Trl* caused disordered and frayed myofibrils in flight muscles, whereas leg muscle myofibrils appeared normal (Fig. 7d, Supplementary Figure S7). This shows that Trl is required for normal myofibril development in flight muscle.

To investigate a role of *CG14655* during muscle development we also applied muscle-specific knock-down with *Mef2*-GAL4 and three different RNAi lines, one of which resulted in viable but flightless animals and two other overlapping hairpins resulted in pupal lethality (Supplementary Table S6). Morphological analysis showed that *CG14655* knock-down flight muscles displayed abnormal actin accumulations between their myofibrils, suggesting a role for *CG14655* in myofibril development of flight muscle. Together, these findings demonstrate the predictive power of AnnoMiner to identify transcriptional regulators by combining chromatin binding and differential expression data.

AnnoMiner helps identify targets co-regulated by Sd and Yki during flight muscle growth in Drosophila. Strikingly, two of the enriched transcriptional regulators in the above comparative flight muscle development dataset were the transcriptional effector of the Hippo pathway in *Drosophila* called Yorkie (Yki) and its essential Tead co-factor Scalloped (Sd)[58] (Supplementary Table S5). Recently, an essential function for the Hippo pathway promoting flight muscle growth by transcriptional up-regulation of mRNAs coding for sarcomeric proteins, which built the myofibrils, was identified[47]. We wanted to know whether we could identify direct transcriptional targets of Yki/Sd during flight muscle growth. To this end, we integrated mRNA BRB-seq data from developing *yki* knock-down flight muscle (*yki-IR*), as well as from flight muscle expressing a constitutive active form of *yki* (*yki-CA)* compared to wild type controls (GEO accession GSE158957) with ChIP-seq data from Yki (modENCODE dataset ENCSR422OTX) and Sd (modERN dataset ENCSR591PRH) obtained in fly embryos using AnnoMiner's *peak integration* function.

Both proteins showed prominent base pair coverage of the TSS regions of their target genes (Fig. 8a). We selected these peaks to retrieve associated genes and then integrated the BRB-seq data for 24 h APF *yki* knock-down (*yki-IR* 24 h), as well as 24 h and 32 h APF constitutively active *yki*, respectively (*yki-CA* 24 h, *yki-CA* 32 h; Supplementary Table S7). Upon knock-down of *yki*, already at 32 h APF, a severe myofibril assembly defect had been observed[47]. Interestingly, AnnoMiner identified Yki and Sd binding sites in the TSS of two genes essential for muscle function and development, which were downregulated in *yki* knock-down muscles at 24 h APF. These genes code for the sarcomeric proteins Tropomyosin 1 (Tm1) and the Nesprin-family protein Muscle-specific protein 300 kDa (Msp300), which is important to link the myofibrils to the nuclei[59] (Fig. 8b). The gain-of-function *yki* phenotype (*yki-CA*) is characterized by premature expression of sarcomeric proteins resulting in muscle fiber hyper-compaction[47]. At 24 h APF *scalloped* (*sd*) itself is the only direct target gene of the Yki/Sd complex which is differentially expressed in *yki-CA* (Fig. 8c). At 32 h AFP, AnnoMiner identified 177 unique genes and in total 541 transcripts as potential direct Yki/Sd targets. Using GO term enrichment analysis by modEnrichR[60] we found many processes and cellular compartments related to muscle development and function among the top enriched terms in these potential direct Yki/Sd targets (Fig. 8d, Supplementary Table S7). To conclude, using AnnoMiner's *peak integration* function, we identified putative direct targets of the Yki/Sd transcriptional complex

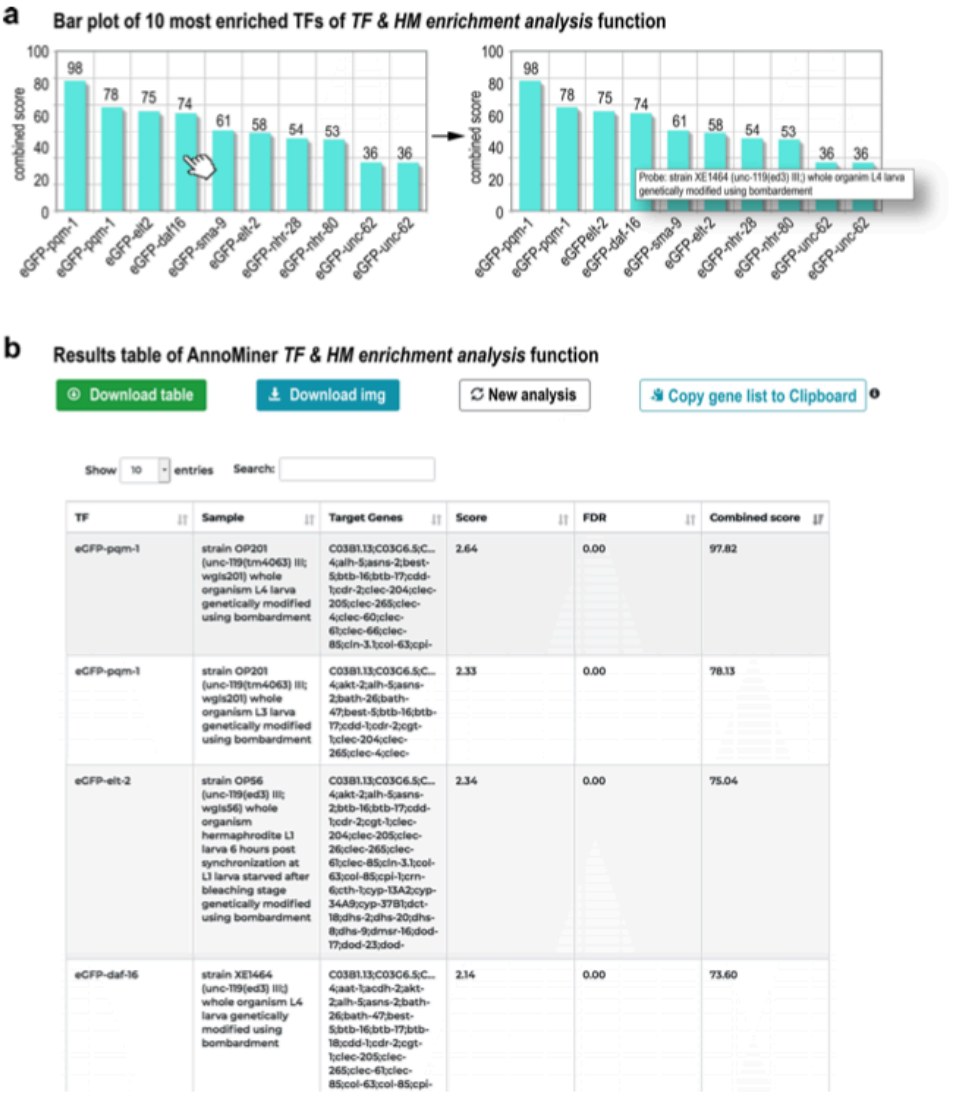


Figure 6. *TF & HM enrichment analysis* function of AnnoMiner. The output of a *TF & HM enrichment analysis* in AnnoMiner is first a plot of the top 10 enriched transcription factors (**a**). The identity of the enriched TF with its combined score is shown. Hovering over the bar of a TF will reveal meta-information associated with the ChIP-seq dataset (**b**). Second and at the same time, a table is returned, which contains the information on all TFs stored in the AnnoMiner database and their enrichments in the submitted gene list. The table is sortable and downloadable. Once downloaded, it includes information on the sample, the target gene for each TF, as well as the score, FDR and combined score. Target gene lists can be copied to the clipboard for further analysis. For demonstrating the *TF enrichment analysis* function of AnnoMiner, we took advantage of a study identifying direct daf-16/FoxO targets from a longevity study ([53] and GEO dataset GSE72426). In this study, the authors looked for isoform-specific targets of daf-16. We combined the list of all DAF-16 isoform targets and submitted it to AnnoMiner for identifying enriched transcription factors. AnnoMiner displays the 10 most enriched TFs, with their combined score indicated in the plot (**a**) as well as the table of all TFs stored for *C. elegans* and including enrichment information for the submitted gene list. Daf-16 is the fourth hit; pqm-1, as well as elt-2, which both precede daf-16 in the hitlist, have been shown to bind to daf-16 response elements.

Tool	Method	ROC AUC	PR AUC	% recovered	AD-test	# of datasets
ChEA3	meanRank	0.83	0.81	6.7	8.00E−06	75
	topRank	0.82	0.81	9.3	8.00E−06	
AnnoMiner		0.63	0.63	4.0	8.10E−05	
TFEA.ChIP		0.63	0.60	0.0	7.90E−04	57
AnnoMiner		0.67	0.67	5.3	1.07E−05	
EnrichR	TRRUST_Transcription_Factors_2019	0.71	0.72	7.8	1.18E−05	51
AnnoMiner		0.63	0.65	5.9	5.57E−04	
EnrichR	ENCODE_TF_ChIP-seq_2015	0.68	0.68	4.6	3.59E−05	44
AnnoMiner		0.72	0.71	6.8	1.36E−05	
EnrichR	ChEA_2016	0.70	0.72	18.2	1.36E−05	44
AnnoMiner		0.67	0.68	4.6	5.12E−05	
EnrichR	ARCHS4_TFs_Coexp	0.65	0.67	6.8	8.11E−06	74
AnnoMiner		0.63	0.64	4.1	6.05E−05	
EnrichR	ENCODE_and_ChEA_Consensus	0.57	0.60	10.3	0.27	29
AnnoMiner		0.76	0.75	6.9	2.07E−05	

Table 1. Performance values of AnnoMiner compared to other TF enrichment web-tools. AnnoMiner's metrics are reported only for RefSeq and the parameters: upstream_tss: DynamicRange, downstream_tss:50 bp and overlap:20 bp.

that showed differential expression upon *yki* knock-down or *yki* constitutive activation. Many of these genes are likely important for flight muscle morphogenesis.

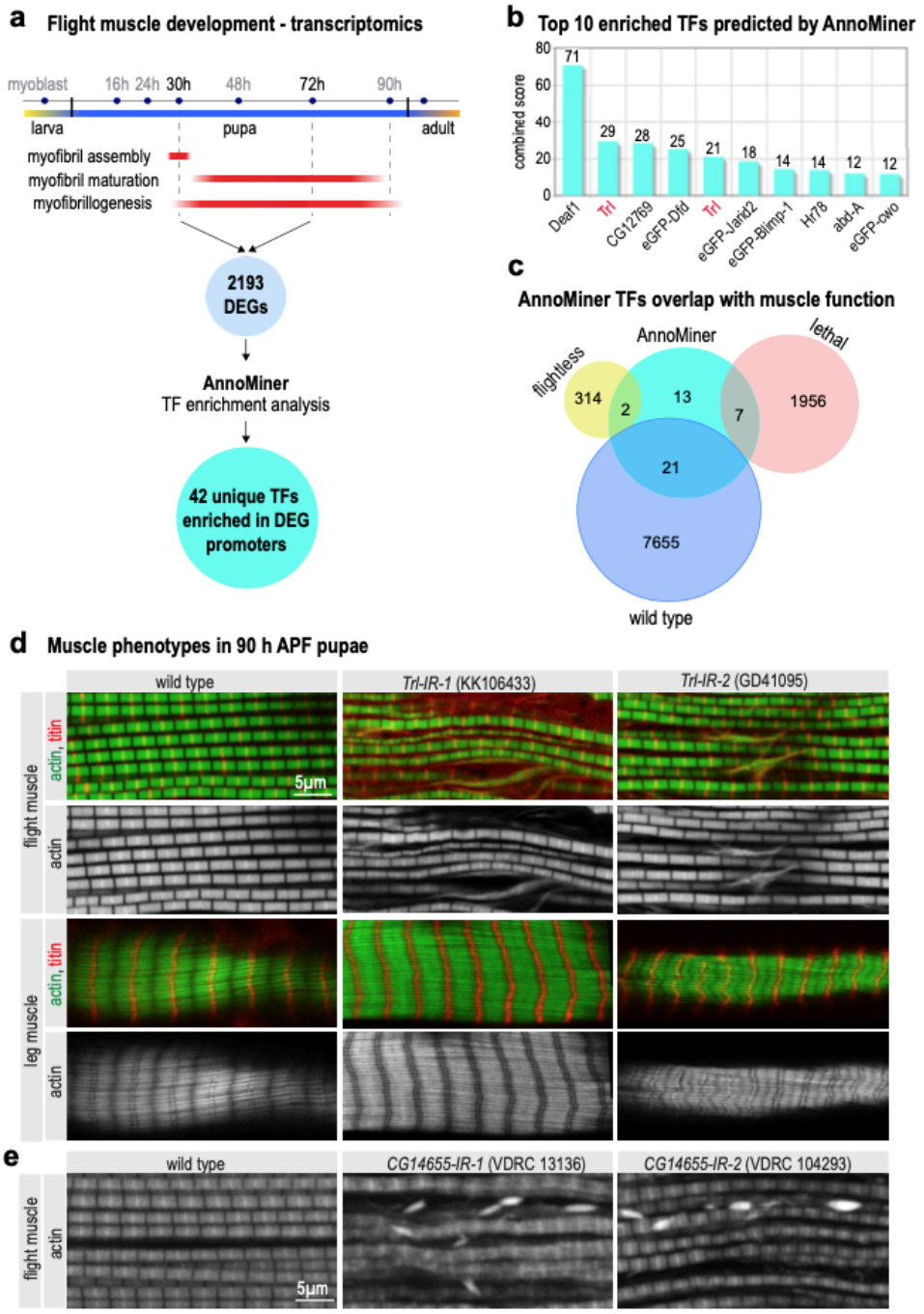
Discussion

Here, we introduced AnnoMiner, a web-based, flexible and user-friendly platform for genomic peak annotation and integration, as well as transcription factor enrichment analysis. We illustrated AnnoMiner's *peak annotation* and *integration*, as well as the *nearby genes annotation* functions with specific examples. We confirmed the predictive power of AnnoMiner's TF enrichment function experimentally by identifying important regulators of *Drosophila* indirect flight muscle development. This was achieved searching for overrepresented TF peaks in promoters of genes differentially regulated during myofibrillogenesis using AnnoMiner.

AnnoMiner distinguishes itself from other peak annotation, as well as TF enrichment tools. AnnoMiner's *peak annotation* and *peak integration* outputs first a bar plot that shows the overlap of a peak with different gene feature attributes, including up- and downstream regions, the TSS, 5′ and 3′ UTRs and the gene body. This has two advantages: first, the user can visualize the distribution of peaks of the uploaded file with respect to all relevant attributes of annotated gene features in the genome. Second, AnnoMiner allows to interactively choose the target region(s) for which the associated genes are returned. While other tools provide statistics on the peak distribution relative to gene feature attributes (e.g.[28]) in the output, to our knowledge, AnnoMiner is the only software that allows to easily retrieve specific gene-sets depending on the distribution of the peak coverage over gene feature attributes. The *peak integration* function offers the same flexibility. Moreover, both functions allow to directly integrate differential gene expression or other numerical data associated to genes with the genomic peak files. The *nearby genes annotation* and *long-range interaction* functions, which integrate expression data with peak data, is novel and for the first time, users can in a web-based manner visualize and retrieve genes that are not the nearest neighbours of a peak. It could be useful to integrate genomic, non-coding variants causative for human genetic diseases with disease-associated gene expression data and explore transcriptional activity within TAD domains. AnnoMiner's *TF enrichment analysis* function offers to treat promoter regions dynamically for each specific TF with its *DynamicRanges* function. Finally, AnnoMiner is independent of the genomic assembly of the source data, as it on-the-fly translates submitted IDs and uses the ID compatible with the database chosen for gene centred peak annotation, as well as TF enrichment.

We compared AnnoMiner's *TF enrichment analysis* function to the best-performing software in the field, which includes ChEA3, TFEA.ChIP and EnrichR. AnnoMiner could not reach the accuracy of ChEA3. One of the reasons could be that the dataset of TF—gene association used by ChEA3 supersedes data we retrieve from ENCODE, modENCODE and modERN, as it includes several additional datasets which are at least partially manually curated or generated. This hypothesis is supported by the fact that EnrichR shows differing performance when using different source data, showing lower performance when using the ENCODE data alone compared to the ones from ChEA3. One possible solution could be to add curated data to AnnoMiner's *TF enrichment analysis* function, for instance from ChIP-Atlas[61] or ReMap[62]. The disadvantage however is the higher cost in curation, as well as the fact that manually curated datasets are typically not available for model organisms such as *Drosophila* or *C. elegans*, but are rather restricted to human or mouse, as is the ChEA3 tool. Yet, users can easily add their own data using the scripts provided to fill the mongoDB present in our gitlab repository, when running AnnoMiner locally (https://gitlab.com/habermann_lab/AnnoMiner/-/tree/master/scripts).

Finally, we predicted and verified potential transcriptional regulators of muscle and myofibril morphogenesis as well as muscle growth using AnnoMiner's *TF & HM enrichment analysis* function. Amongst those is Trl, a


a
Flight muscle development - transcriptomics
myoblast
16h 24h 30h
48h
72h
90h
larva
pupa
adult
myofibril assembly
myofibril maturation
myofibrillogenesis
2193 DEGs
AnnoMiner
TF enrichment analysis
42 unique TFs enriched in DEG promoters
b
Top 10 enriched TFs predicted by AnnoMiner
combined score
Deaf1
Trl
CG12769
eGFP-Dfd
Trl
eGFP-Jarid2
eGFP-Blimp-1
Hr78
abd-A
eGFP-cwo
c
AnnoMiner TFs overlap with muscle function
AnnoMiner
flightless
lethal
wild type
d
Muscle phenotypes in 90 h APF pupae
wild type
Trl-IR-1 (KK106433)
Trl-IR-2 (GD41095)
flight muscle
leg muscle
actin, titin
actin
5µm
e
wild type
CG14655-IR-1 (VDRC 13136)
CG14655-IR-2 (VDRC 104293)

◂**Figure 7.** AnnoMiner identifies transcriptional regulators of flight muscle development. (**a**) Scheme of *Drosophila* indirect flight muscle development with different myofibrillogenesis phases highlighted in red. Comparing 30 h to 72 h APF revealed a list of 2193 differentially expressed genes (taken from GSE107247), which was uploaded to AnnoMiner. The *TF enrichment analysis* function of AnnoMiner identified 42 unique candidate TFs. (**b**) Output of the top-ten enriched TFs from AnnoMiner. Trl was identified from 2 different ChIP-seq datasets and is highlighted in red. (**c**) Phenotypic integration with muscle-specific RNAi knock-down data revealed candidate TFs with a potential function in muscle (flightless or lethal). Data on DEGs, phenotypic data from the RNAi screen, as well as annotated AnnoMiner enrichment results are available in Supplementary Table S5. (**d**) *Trl* has a function during flight muscle myofibrillogenesis. Flight and leg muscles of 90 h APF pupae were fixed and stained for actin (phalloidin in green) and titin homolog Sls (anti-Kettin in red). Note that muscle-specific knock-down of *Trl* using independent hairpins (*Trl-IR-1* and *IR-2*) results in frayed flight muscle myofibrils compared to wild type control but rather normal leg muscle myofibrils. (**e**) *CG14655* has a function in flight muscles. Flight muscles of young adult flies from wild type or *CG14655-IR-1* or of 90 h APF pupae from *CG14655-IR-2* were stained for actin (phalloidin). Note the prominent actin accumulations present in both *CG14655* knock-down flight muscles.

GAGA transcription factor which contains a BTB/POZ domain, as well as a C2H2 zinc-finger that binds to DNA in a sequence-specific manner. Previous studies suggest that Trl is required to keep promoters nucleosome-free, thus allowing Pol-II-access[63]. We showed here that muscle-specific knock-down of *Trl* using four independent hairpins either leads to pupal lethality or flightlessness. Consistently, we find severely perturbed indirect flight muscles upon *Trl* knock-down whereas leg muscles appear largely normal. This indicates a preferential function of Trl in flight muscle, however as two hairpins result in pupal lethality, a role of Trl in other body muscles is also likely.

A second potential direct transcriptional regulator identified by AnnoMiner is the uncharacterized Zinc-finger transcription factor CG14655. Muscle-specific knock-down of *CG14655* either results in pupal lethality or flightlessness and causes abnormal accumulations of actin in flight muscles. This again suggests that CG14655 is important for normal myofibril development in flight muscle.

Lastly, we made use of two other transcriptional regulators, Yorkie and Scalloped, which on DNA act in a complex[58], to identify its direct targets. AnnoMiner identified two direct targets of Yki, which upon loss of *yki* were downregulated. Both code for important muscle structure proteins, constituents of the sarcomere or linking the sarcomere to the nucleus, and hence could contribute to the severe phenotype observed upon *yki* knock-down. Gain-of-function of *yorkie* results in muscle fiber hyper-compaction and premature expression of sarcomeric protein components[47]. Consistently, AnnoMiner identified a number of direct Yki targets with functions related to muscle development and growth. This substantiates a role for Scalloped and its transcriptional co-factor Yorkie during flight muscle growth.

To conclude, the new web-tool AnnoMiner is a user-friendly, intuitive, interactive and highly-flexible platform for genomic *peak annotation* and *peak integration*. It is suitable for identification of *nearby genes* or *long-range interactions* of a genomic peak, as well as to perform *Transcription Factor and Histone Modification enrichment analysis* for a list of genes. This manuscript details all AnnoMiner functions and shows its usefulness for annotating and integrating peaks from two different ChIP-seq experiments together with transcriptomics data. Finally, AnnoMiner helped identify several key regulators of indirect flight muscle development and growth in *Drosophila*, some of which were confirmed experimentally.

Methods

The AnnoMiner software and database. AnnoMiner is a modular software consisting of a library of java classes for retrieval, storage and analysis of annotated genomic peak data together with a web application providing a graphical user interface for a number of predefined analyses.

The AnnoMiner web-sever is a JavaEE web application implemented in the *front-end* as a single page application using Javascript, jQuery and Bootstrap 4. Queries for the provided analyses are processed by individual servlets that send a response back to the front-end in JSON format (Supplementary Figure S1).

The java library used in the *back-end* manages the retrieval of data from remote sources at UCSC and modENCODE[36] or alternatively from files in BED or gff format. Retrieved datasets are stored in a MongoDB database. This document-oriented NoSQL database system has been chosen for its flexible data model and runtime speed benefits benchmarked against an SQL database solution. For optimal performance, a custom database connectivity layer has been developed based on the MongoDB java driver. The overall structure of the application is shown in Fig. 1 and Supplementary Figure S1.

AnnoMiner's database currently holds genomic data from ENCODE[33], modENCODE[36] and modERN[37]. For each model organism, the latest genome assembly is stored. For human, mouse and *Drosophila*, we also provide the second latest release.

Following the Findability, Accessibility, Interoperability and Reusability (FAIR) principle[64], documentation and source code of the tool are available on GitLab: https://gitlab.com/habermann_lab/annominer. Using the java library classes from the code repository developers will be able to define custom analyses on annotated genomic ranges datasets and extend the database towards new data sources. The repository also provides executable java classes and python scripts for local maintenance such as to populate a local database with ENCODE, modENCODE or modERN data or to reproduce our benchmark analysis.

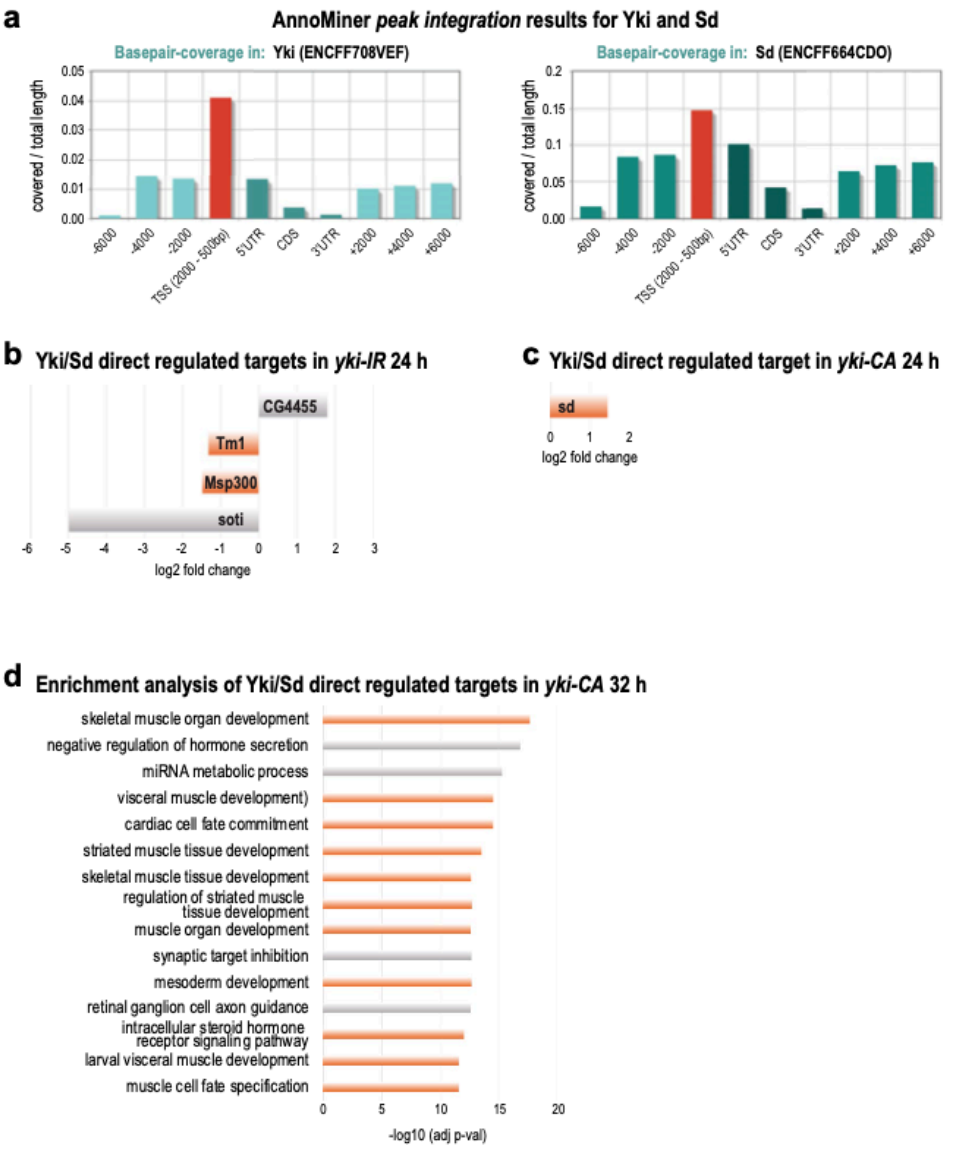


Figure 8. The Yorkie and Scalloped transcriptional complex regulates muscle-specific gene expression during flight muscle growth. (**a**) Using AnnoMiner's *peak integration* function with ChIP-seq data from Yorkie (modENCODE ENCFF708VEF) and Scalloped (modERN ENCFF664CDO) we found high base-pair coverage of both TFs in the TSS (− 2000 to + 500 bp of the Transcription Start Site) of putative target genes. We selected these peaks from both TFs and integrated them with flight muscle specific BRB-seq data ([47] and GEO dataset GSE158957) isolated from a *yorkie* loss-of-function muscles at 24 h APF (**b**) or yorkie gain-of-function muscles at 24 h (**c**) and 32 h APF (**d**). The direct Yki/Sd targets *Tm1* and *Msp300* were downregulated in *yorkie* loss-of-function at 24 h APF. Only *scalloped* was identified as upregulated direct target from 24 h *yorkie* gain of function muscles. At 32 h, 177 genes were identified as potential direct Sd/Yki targets, many of which have muscle-related functions (see also Supplementary Table S7).

Genomic peak annotation functions. *Peak annotation.* For calculating the overlap between peaks and genes, we compare the coordinates of peaks with those of stored gene feature attributes. These include 5′ flanking regions, promoter region, transcription start site (TSS), 5′UTR, coding sequence (CDS, including exons and

introns), 3′UTR, 3′ flanking regions, as well as strandedness of the gene feature. For integrating peaks with gene features, we first calculate the coverage of each peak by testing whether the peak coordinates fall within the coordinates of any attribute of a gene feature. If all isoforms of a gene are chosen, the overlap of the peak with the attributes of each individual isoform will be checked. For calculating the coverage, overlap events are calculated cumulatively and the sums of all overlaps will be divided by the length of the attribute. Depending on which attribute (target region) the user has chosen, only genes whose chosen attribute overlaps (with a minimum % or number of base pairs) with the peak will be displayed in the data table.

Peak integration. The *peak integration* function is based on the same algorithm as the *peak annotation*. However, only genes, which meet the overlap criteria of the user-selected target regions with peaks from all provided BED files will be displayed in the resulting data table.

Nearby genes annotation. For the *nearby genes annotation* function, the 5 neighbouring genes up- and downstream, as well as the overlapping gene of a peak are extracted and displayed to the user. The user can choose whether the directionality of the gene with respect to the peak should be considered. If a gene overlaps with the peak, it will be shown in the *Overlap* box. Data on differential expression, which needs to be provided by the user in form of an annotation file, is displayed on the returned plot. The user can choose whether or not the directionality of the deregulation (up- or downregulated) is shown in the box plot. Depending on the user input, the *nearby genes annotation* function returns neighbouring genes of a single peak, or of many peaks.

Long-range interaction. With the *long-range interactions* function, the user is able to extract and visualize all the genes up- and down-stream the user provided peak (or peaks), which lie within an user-defined genomic window (default 50 kb upstream and 50 kb downstream). The user can choose whether the directionality of the gene with respect to the peak should be considered. Data on differential expression, which needs to be provided by the user in form of an annotation file, is displayed on the returned plot. The user can choose whether or not the directionality of the deregulation (up- or downregulation) is shown in the box plot. In the first case, hovering on the box of interest, the user can extract information regarding the overall number of deregulated genes (coverage). In the second case, the user will extract separate information about the number of up- and down-regulated genes. Depending on the user input, the *long-range interactions* function returns neighbouring genes of a single peak, or of many peaks within the user-defined genomic window.

Transcription factor enrichment analysis. Transcription factor enrichment analysis in AnnoMiner uses publicly available TF ChIP-seq data from the ENCODE, modENCODE and modERN databases. In order to keep only high-quality data, we undertook a strict selection of available TF ChIP-seq datasets. First, we kept only experiments, for which data had already been pre-analysed following the ENCODE processing pipeline. Second, only experiments with replicates and optimal IDR peaks files (narrowPeak) were retained. Third, we manually inspected each narrowPeak file, evaluating the peak distribution around the gene's TSS and discarded the ones with irregular peak distributions in relation to promoter regions of genes, as we considered those as outliers (Supplementary Figure S8c,d; all outliers csv files that were excluded are available in our gitlab repository at https://gitlab.com/habermann_lab/AnnoMiner-paper/-/tree/master/ChIP-seq_experiment_excluded). The number of used TF ChIP-seq datasets for each of the model organisms and assemblies is available in Supplementary Tables S8, S9.

Over-representation analysis. For each TF ChIP-seq dataset stored in the AnnoMiner database, we first compute the contingency table of potential targets in both the uploaded gene list (signature gene-set) and in the background list of all genes of the chosen organism. We then apply a one-tailed hypergeometric test followed by Benjamini–Hochberg correction[65]. Furthermore, we calculate the enrichment score as follows:

$$score = (listHits \div listSize)/(genomeHits \div genomeSize)$$

And a combined score based on the p-value and enrichment score according to[66]:

$$combined\ score = score * -log10(p\text{-}value).$$

Promoter definition. AnnoMiner allows the user to fully customize the promoter region considered for TF enrichment analysis. The user can choose both the upstream and downstream borders to consider without constraints.

DynamicRanges option. The user can also choose dynamic ranges that we pre-calculated for each individual TF based on available ChIP-seq data stored in AnnoMiner. In brief, for each TF ChIP-seq experiment we computed the distances between its peaks and the genes in a range of 20,000 bp upstream of their TSSs. We then binned the closest genes by distances with a resolution of 50 bp. A noise level was computed, representing non-significant bindings, from 5000 to 20,000 bp, as follows:

$$noise = median + 3std$$

Next, we smoothened the binning using the simple moving average approach (SMA) with a sliding window of size = 3. Intersecting the noise threshold with the SMA curve, we obtained the upstream TSS value. This value

is unique for each TF and represents the distance range in which the TF is more often binding than is expected at random (see Supplementary Figure S8e,f).

Choosing the peak-promoter overlap. The user can define the amount of overlap between a TF and a promoter in percentage or in base pairs (bp) to categorize a gene as a potential target.

Histone modification enrichment analysis. The same function that is used for the *Transcription factor enrichment analysis* can be used for enriching histone modifications for a gene list. In this case, a *DynamicRanges*-like option is not meaningful. Thus, the user has to manually define the promoter region (default: 2000 bp upstream and 500 bp downstream of a gene's TSS).

Automatic ID conversion. In order to easily allow comparisons between genome versions (ENSEMBL, RefSeq, UCSC, Genecode, Flybase and Wormbase) and not restrict the user to a specific set of gene identifiers, we developed a geneID converter. This function converts on the fly the user's input gene IDs to match the selected resource, without any intervention from the user. The conversion is performed using the latest release of BioMart[52]. BioMart data were downloaded and are stored locally for reasons of speed and are regularly updated. Users can download the ID conversion table to control for genes that were dropped due to the automatic ID conversion process.

Benchmarking the TF enrichment function of AnnoMiner. *Benchmarking dataset.* Benchmarking data-sets for testing AnnoMiner's enrichment function in the form of TF regulons were obtained from human single TF perturbation experiments (knock-downs, knockouts, over-expression and chemical inhibition) followed by RNA-seq from Gene Expression Omnibus (GEO[56], 160 experiments), manually curated and made available from Keenan et al.[43]. We retained only experiments targeting TFs for which AnnoMiner is storing at least one high quality TF ChIP-seq dataset for the assembly GRCh38 (hg38), leading to a total of 75 RNA-seq experiments left for benchmarking (*benchmarkingSet*).

Benchmarking metrics. Precision-Recall (PR) and Receiver-Operating Characteristic (ROC) areas under the curves (AUC) for the *benchmarkingSet* were computed using the R package PRROC[67], following the benchmarking procedure proposed and applied by[43] and[68]. In brief, for each TF, signature gene-sets were derived based on differentially expressed genes of the perturbed TF versus its control and used for TF enrichment analysis. For each dataset, the rank of the perturbed TF was assigned to the positive (foreground (*fg*)) class, while all other ranks of TFs retrieved for the dataset were assigned to the negative (background (*bg*)) class; using the PPROC package, precision-recall (PR) and Receiver-Operator Curves (ROC) using continuous interpolation were computed and areas under curve (AUCs) for both were calculated (PR-AUC and ROC-AUC). As the negative *bg* class greatly exceeded the positive *fg* class, we down-sampled the negative class to the same size as the positive class. We repeated this procedure 5000 times. We used the *approx* R function to linearly interpolate between all points from the 5000 ROC and the 5000 PR curves and calculated the mean ROC and PR AUCs values of all bootstrap runs. Furthermore, we analysed the cumulative distribution of the rank values for each TF across all experiments it was perturbed in, which is not expected to be uniform, unless the TF would rank randomly in all its associated datasets. We then performed the Anderson–Darling test using the *goftest* R package to test for significant deviations from this distribution, indicating a non-random distribution of the TF. Finally, we computed the percentage of perturbed TFs that were correctly ranked within the first percentile to ensure high ranking of the TF. All values are available in Table 1.

Benchmarking against existing tools. We compared the performance of AnnoMiner's TF enrichment function with the performance of the following web-servers: TFEA.ChIP[42], ChEA3[43] and EnrichR[44]. The same dataset (*benchmarkingSet*) was used for all web-tools tested. Only datasets which could be effectively ranked by the web-tool under study were used. We used 57 signature gene-sets to test TFEA.ChIP and all the 75 signature-gene sets to test ChEA3. For EnrichR we evaluated the performance metrics on 5 different resources: ARCHS4, ChEA 2016, ENCODE 2015, ENCODE and ChEA Consensus and TRRUST 2019 using 74, 44, 44, 29 and 51 signature gene-sets, respectively.

Epigenetic and RNA-seq data used. For analysing STAT3 in acute B-cell lymphomas, we used data from super-series GSE50724[48]. Only peaks which were significantly upregulated in ABC type B-cell lymphoma were selected for AnnoMiner's *peak integration* (FDR cut-off < = 0.05; FC > = 1.25). We used H3K4me3 data from an ABC cell line stored in the ENCODE database (GSE86718_ENCFF763KFL, replicated peaks BED file) for *peak integration* with significantly upregulated STAT3 peaks. Human genome version hg19 was used for analysis. RNA-seq data (GSE50721) from the same super-series was used for integration with STAT3 and H3K4me3 peaks, whereby we used statistical results provided by the authors of the study. For demonstrating the *nearby genes annotation* function, we used data from[51] and GEO dataset GSE52974, describing long-range enhancer elements regulating Myc expression for normal facial morphogenesis, mapped to mouse genome version mm9. For demonstration of the *TF & HM enrichment* function we used RNA-seq data from a longevity study focusing on daf-16/FoxO mutants in *C. elegans* ([53] and GEO dataset GSE72426) and *C. elegans* genome version ce11/WBcel245.

Analysed RNA-seq data from developing indirect flight muscle (IFM) in *Drosophila melanogaster* were taken from GSE107247[46]. We used the DESeq2 comparisons between time-points 72 h and 30 h after puparium

formation (APF) provided by the authors. Genes with log2FC in absolute value greater than 2 and FDR smaller than 0.01 were considered as differentially expressed. For integration with phenotypic data, we made use of data published on muscle morphogenesis and function in *Drosophila*[57] and flight muscle growth in *Drosophila*[47].

For integrating gene expression changes of *yki* knocked-down flight muscles and flight muscles expressing a constitutively active *yki* compared to wild type, with Yki and Sd protein genome occupancy data, we used BRB-seq data from *yki* knock-down and *yki* constitutively active fly strains[47] available from GEO (GSE158957). Optimal IDR threshold narrow peak files for Yorkie (Yki, ENCSR422OTX) and Scalloped (Sd, ENCSR591PRH) ChIP-seq were downloaded from the ENCODE/modERN resource, both for *Drosophila* genome version dm6.

***Drosophila* experimental methods.** *Drosophila* strains were grown under standard conditions at 27 °C to enhance GAL4 activity. All RNAi knock-downs were induced with *Mef2*-GAL4, a GAL4 line specifically expressed during development of all muscle types of the fly[57]. UAS RNAi lines used were from the Vienna[69] and Harvard collections[70] and ordered from VDRC or Bloomington stock centres. See Supplementary Table S6 for all genotypes used. The flight test was done as described in[57].

Indirect flight muscle morphology was analysed in 3 to 5 days adult males or in 90 h APF pupae as previously published[71]. Briefly, 90 h APF pupae or adult flies were fixed with 4% paraformaldehyde in PBT (PBS, 0.5% triton X-100) and cut into half-thoraces using a sharp microtome blade. The half-thoraces were blocked with 3% normal goat serum in PBT for 30 min and F-actin in the flight and leg muscles was visualised with phalloidin (coupled to Alexa488 or rhodamine, Molecular Probes, 1:1000 in PBT for 2 h or overnight) and Sls was stained with anti-Kettin antibody MAC155 (Babraham Institute, 1:100 in PBT overnight). The stained half-thoraces were imbedded in Vectashield (Biozol) and imaged on an LSM780 or LSM880 confocal microscope. Images were processed using Fiji[72].

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Availability of all other data, materials and software: The AnnoMiner web-server is freely available at http://chimborazo.ibdm.univ-mrs.fr/AnnoMiner/. The source code of AnnoMiner, as well as all data necessary to reproduce the study are available at https://gitlab.com/habermann_lab/annominer. Details on software provided in this manuscript: Project name: AnnoMiner, Project home page: http://chimborazo.ibdm.univ-mrs.fr/AnnoMiner/, Archived version: https://gitlab.com/habermann_lab/annominer, Operating system(s): Platform independent, Programming language: JavaScript, Java, PHP, MongoDB, Other requirements: none, License: GNU public license, Any restrictions to use by non-academics: none.

Received: 16 February 2021; Accepted: 14 July 2021
Published online: 29 July 2021

References

1. Zentner, G. E. & Henikoff, S. High-resolution digital profiling of the epigenome. *Nat. Rev. Genet.* **15**, 814–827 (2014).
2. ENCODE Project Consortium *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
3. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
4. Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin Cell Dev Biol* **57**, 57–67 (2016).
5. Krivega, I. & Dean, A. Enhancer and promoter interactions-long distance calls. *Curr. Opin. Genet. Dev.* **22**, 79–85 (2012).
6. Comoglio, F. *et al.* Thrombopoietin signaling to chromatin elicits rapid and pervasive epigenome remodeling within poised chromatin architectures. *Genome Res.* **28**, 295–309 (2018).
7. Mitchelmore, J., Grinberg, N. F., Wallace, C. & Spivakov, M. Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. *Nucleic Acids Res.* **48**, 2866–2879 (2020).
8. Dao, L. T. M. & Spicuglia, S. Transcriptional regulation by promoters with enhancer function. *Transcription* **9**, 307–314 (2018).
9. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
10. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.* **18**, 279–290 (2017).
11. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. *Nat. Rev. Genet.* **13**, 840–852 (2012).
12. Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.* **18**, 441–450 (2017).
13. Wilbanks, E. G. & Facciotti, M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471 (2010).
14. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137–R139 (2008).
15. Xu, H., Wei, C.-L., Lin, F. & Sung, W.-K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* **24**, 2344–2349 (2008).
16. Huang, W. *et al.* Efficiently identifying genome-wide changes with next-generation sequencing data. *Nucleic Acids Res.* **39**, e130 (2011).
17. Shen, L. *et al.* diffReps: Detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE* **8**, e65598 (2013).
18. Ambrosini, G., Dreos, R., Kumar, S. & Bucher, P. The ChIP-Seq tools and web server: A resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics* **17**, 938–1015 (2016).
19. Blahnik, K. R. *et al.* Sole-Search: An integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.* **38**, e13 (2010).
20. Guzman, C. & D'Orso, I. CIPHER: A flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. *BMC Bioinform.* **18**, 363–416 (2017).
21. Boeva, V., Lermine, A., Barette, C., Guillouf, C. & Barillot, E. Nebula—A web-server for advanced ChIP-seq data analysis. *Bioinformatics* **28**, 2517–2519 (2012).
22. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinform.* **11**, 415–512 (2010).

23. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
24. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
25. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
26. Kondili, M. *et al.* UROPA: A tool for Universal RObust Peak Annotation. *Sci. Rep.* **7**, 2593–2612 (2017).
27. Tang, X. *et al.* annoPeak: A web application to annotate and visualize peaks from ChIP-seq/ChIP-exo-seq. *Bioinformatics* **33**, 1570–1571 (2017).
28. Chen, T.-W. *et al.* ChIPseek, a web-based analysis tool for ChIP data. *BMC Genomics* **15**, 539–613 (2014).
29. Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D. & Li, L. PAVIS: A tool for Peak Annotation and Visualization. *Bioinformatics* **29**, 3097–3099 (2013).
30. Bhasin, J. M. & Ting, A. H. Goldmine integrates information placing genomic ranges into meaningful biological contexts. *Nucleic Acids Res.* **44**, 5550–5556 (2016).
31. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
32. Zhu, L. J. *et al.* ChIPpeakAnno: A bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinform.* **11**, 237–310 (2010).
33. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
34. Davis, C. A. *et al.* The encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
35. Jou, J. *et al.* The ENCODE portal as an epigenomics resource. *Curr. Protoc. Bioinform.* **68**, e89 (2019).
36. Gerstein, M. B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
37. Kudron, M. M. *et al.* The ModERN resource: Genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics* **208**, 937–949 (2018).
38. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
39. Garcia-Alonso, L. *et al.* Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* **78**, 769–780 (2018).
40. Wang, Z. *et al.* BART: A transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* **34**, 2867–2869 (2018).
41. Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: Advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3* **2**, 987–1002 (2012).
42. Puente-Santamaria, L., Wasserman, W. W. & Del Peso, L. TFEA.ChIP: A tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets. *Bioinformatics* **35**, 5339–5340 (2019).
43. Keenan, A. B. *et al.* ChEA3: Transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
44. Kuleshov, M. V. *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
45. Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D. & Aerts, S. i-cisTarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* **43**, W57-64 (2015).
46. Spletter, M. L. *et al.* A transcriptomics resource reveals a transcriptional transition during ordered sarcomere morphogenesis in flight muscle. *Elife* **7**, 1361 (2018).
47. Kaya-Çopur, A. *et al.* The Hippo pathway controls myofibril assembly and muscle fiber growth by regulating sarcomeric gene expression. *Elife* **10**, 79 (2021).
48. Hardee, J. *et al.* STAT3 targets suggest mechanisms of aggressive tumorigenesis in diffuse large B-cell lymphoma. *G3* **3**, 2173–2185 (2013).
49. Li, X.-Y., Harrison, M. M., Villalta, J. E., Kaplan, T. & Eisen, M. B. Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. *Elife* **3**, e03737 (2014).
50. Gaskill, M. M., Gibson, T. J., Larson, E. D. & Harrison, M. M. GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo. *Elife* **10**, e66668 (2021).
51. Uslu, V. V. *et al.* Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat. Genet.* **46**, 753–758 (2014).
52. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
53. Chen, A.T.-Y. *et al.* Longevity genes revealed by integrative analysis of isoform-specific daf-16/FoxO mutants of *Caenorhabditis elegans*. *Genetics* **201**, 613–629 (2015).
54. Tepper, R. G. *et al.* PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. *Cell* **154**, 676–690 (2013).
55. Zhang, P., Judy, M., Lee, S.-J. & Kenyon, C. Direct and indirect gene regulation by a life-extending FOXO protein in *C. elegans*: Roles for GATA factors and lipid gene regulators. *Cell Metab.* **17**, 85–100 (2013).
56. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
57. Schnorrer, F. *et al.* Systematic genetic analysis of muscle morphogenesis and function in Drosophila. *Nature* **464**, 287–291 (2010).
58. Wu, S., Liu, Y., Zheng, Y., Dong, J. & Pan, D. The TEAD/TEF family protein Scalloped mediates transcriptional output of the Hippo growth-regulatory pathway. *Dev. Cell* **14**, 388–398 (2008).
59. Wang, S., Reuveny, A. & Volk, T. Nesprin provides elastic properties to muscle nuclei by cooperating with spectraplakin and EB1. *J. Cell Biol.* **209**, 529–538 (2015).
60. Kuleshov, M. V. *et al.* modEnrichr: A suite of gene set enrichment analysis tools for model organisms. *Nucleic Acids Res.* **47**, W183–W190 (2019).
61. Oki, S. *et al.* ChIP-Atlas: A data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255 (2018).
62. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
63. Fuda, N. J. *et al.* GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLoS Genet.* **11**, e1005108 (2015).
64. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018–160019 (2016).
65. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
66. Xiao, Y. *et al.* A novel significance score for gene selection and ranking. *Bioinformatics* **30**, 801–807 (2014).
67. Grau, J., Grosse, I. & Keilwagen, J. PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).

68. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
69. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature* **448**, 151–156 (2007).
70. Ni, J.-Q. *et al.* A genome-scale shRNA resource for transgenic RNAi in Drosophila. *Nat. Methods* **8**, 405–407 (2011).
71. Weitkunat, M. & Schnorrer, F. A guide to study Drosophila muscle biology. *Methods* **68**, 2–14 (2014).
72. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

Acknowledgements
We thank Friedhelm Pfeiffer, Assa Yeroslaviz and all team members of the IBDM Computational Biology team for helpful discussion and critical reading of the manuscript. We thank Aynur Kaya-Çopur for sharing data before publication. Fly stocks obtained from the Bloomington *Drosophila* Stock Center (NIH P40OD018537) and the Vienna *Drosophila* Resource Center (VDRC) were used in this study.

Author contributions
A.M. designed the software and developed the java library and database back-end. A.M. and F.M. were the main developers of the AnnoMiner web-server. F.M. performed all calculations, comparison with other tools and statistical analyses and together with M.H. and B.H.H. performed data analysis. M.W. and F.S. performed all experiments presented in this study. F.M., F.S. and B.H.H. wrote this manuscript, with contributions from A.M. The project was conceived by B.H.H. with contributions from A.M. and F.M. All authors read and approved the final version of the manuscript.

Funding
This work was supported by the French National Research Agency with ANR grant ANR-18-CE45-0016-01 MITO-DYNAMICS awarded to BHH and FS and with ANR-ACHN MUSCLE-FORCES (F.S.); by the excellence initiative Aix-Marseille University A*MIDEX (ANR-11-IDEX-0001-02, F.S.); the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant 310939 (F.S.); the Human Frontiers Science Program (HFSP, RGP0052/2018, F.S.); the France-BioImaging infrastructure supported by the French National Research Agency (ANR–10–INBS-04-01, Investments for the future); by the "Investissements d'Avenir" French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University—A*MIDEX"; and by the Max Planck Society, the Centre National de la Recherche Scientifique (CNRS), Aix-Marseille University and the IBDM UMR 7288.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94805-1.

Correspondence and requests for materials should be addressed to B.H.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Discussion

RNfuzzyApp

RNfuzzyApp is the result of a collaboration with a biological team, where we wanted to enable the experimentalists to perform bulk RNA-sequencing analysis in an easy and user-friendly way, being able to repeat the analysis by choosing different parameters and to generate high-quality figures that can be used for publications.
In addition to the classical bulk RNA-seq analysis pipeline, we chose to include the possibility to analyze temporal data using Mfuzz fuzzy clustering. Fuzzy clustering wasn't widely used because it has several requirements that are not easy to meet by non-computer experts. First, it requires users to select the number of clusters, which is not an easy task. To address this, we integrated fuzzy clustering into our R-shiny App. By incorporating methods like the elbow method, we facilitated the user's choice of the number of clusters. Moreover, to test cluster stability, and as there is a heuristics incorporated in fuzzy clustering, the clustering step needs to be run several times and the resulting clusters have to be compared for stable assignments of genes. For the interpretation, clusters are typically post-processed and analysis steps like enrichment analysis are performed on the list of genes in a cluster. This was also incorporated in RNfuzzyApp. Since no other web-based RNA-seq analysis tools offered fuzzy clustering alongside classical bulk analysis, it was the ideal time to introduce a comprehensive app with a wide array of methods and applications.

A lot of shiny apps are emerging, and offer diverse analysis, RNfuzzyApp remains the only to offer time series data analyses, and one of the few with several DEA methods like TCC-GUI [126], opposite to GENAVi [127] for instance and even so we more recent shiny apps like DEVEA [128] or bulkAnalyseR [129] (Table 1).

Shiny App	Raw data visualizations	several DEA methods	Plots	Enrichment	Time Series
RNfuzzyApp	✔	✔	✔	✔	✔
GENAVi	✔	×	✔	✔	×
TCC-GUI	✔	✔	✔	×	×
DEVEA	×	×	✔	✔	×
bulkAnalyseR	×	×	✔	✔	×

Table 1 : Comparative analysis of different shiny apps in terms of methods offered.
RNfuzzyApp is the only one offering time series data analysis and together with TCC-GUI, one of the few to offer several methods to identify DE genes like DEseq2 [7] or edgeR [130].

RNfuzzyApp could be enhanced by incorporating new DEA methods, clustering techniques, and bulk analysis features. Additionally, the work of Rahma Ait Mahfoud, a Master 2 student in our lab, could further improve the app. She developed a system to automatically annotate all clusters resulting from fuzzy clustering, whereas the current version only allows for one cluster to be annotated at a time.
The biggest limitation of the app is the shiny app that limits the size of a dataset. Too large datasets, more than 100MB can lead to the crash of the App, depending on the memory of the computer used, and the shiny app. A R package without the web-interface could be an alternative to provide users all RNfuzzyApp features but 'on demand', rather than all at once.

Single cell and single nucleus RNA-sequencing

Single-cell based techniques have become very popular in recent years and technological development is advancing at a fast pace also for other single-cell techniques. Despite its potential to reveal cellular heterogeneity, scRNA-seq technology often suffers from noise and sparsity due to technical limitations [131,132].

Not all mRNA molecules present in a cell are captured and sequenced. Dropout events, where transcripts present in the cell are not detected, lead to many zero counts in the data [133]. The efficiency of capturing and converting RNA molecules into cDNA can be low, resulting in only a fraction of the RNA present in the cell being sequenced. These dropout events and inefficiencies contribute to both noise and sparsity.

Variability can also arise from differences in sample processing, library preparation, sequencing runs, and PCR amplification, which can introduce bias as some fragments may be preferentially amplified over others. This can lead to the overrepresentation or underrepresentation of certain transcripts [134]. In many scRNA-seq experiments, the sequencing depth per cell is relatively shallow compared to bulk RNA-seq, meaning fewer reads are available to detect each transcript, increasing the likelihood of missing lowly expressed genes. Variability, PCR amplification bias, and sequencing depth contribute to the overall noise in the data.
Moreover, in tissue dissociation, a crucial step of sample preparation, transcriptional states may be altered due to the stress induced to cells during the process [135-138] and also because of the potential low amount of tissue material [139].

Next to single-cell RNA-seq, single-nuclei RNA sequencing has become very popular. Single nuclei can be isolated from frozen or fixed tissues, which helps preserve RNA integrity better than whole cells, especially from archived samples [140]. This reduces the noise introduced by RNA degradation in single-cell protocols. The

process of dissociating tissues into single cells can induce stress responses and artifacts in gene expression profiles. Isolating nuclei bypasses the need for complete cell dissociation. Moreover, nuclear RNA, particularly pre-mRNA and unspliced transcripts, is more stable and abundant in nuclei compared to cytoplasmic mRNA in whole cells. This can result in better capture efficiency and lower dropout rates, leading to less sparsity in the data. snRNA-seq is particularly useful for fibrous, dense, or highly structured tissues (like brain tissue) where complete dissociation into single cells is difficult. By isolating nuclei, RNA profiles can be obtained from these otherwise challenging samples, enhancing overall data quality and reducing noise associated with incomplete dissociation.

SnRNA-seq addresses many of the noise and biases inherent in single-cell RNA sequencing by improving sample preservation, reducing dropout events, minimizing technical variability, and enhancing the capture of cellular diversity [141]. By exploiting the advantages of nuclei isolation, snRNA-seq provides a valuable tool for studying complex tissues and archived samples, offering a more robust and reliable alternative for transcriptomic analysis.

Once single-cell sequencing data have been processed, the annotation of the cells is a critical step to ensure meaningful follow-up analyses. Cell type annotation is very difficult and hard to automate. Manual annotation is considered the golden standard. It is based on the expression of cell-type specific marker genes. Handmade cell type annotation is done thanks to marker genes databases like PanglaoDB [31] or CellMarker2.0 [142]. Several web tools exist to automate the analysis and annotation of cells like Azimut [143] or SciBet [144] or annotation tools like ClustifyR [32] or SingleR [145]. Automated tools may use different methods like correlation or supervised classification. Ultimately, despite the comprehensive database annotations available, uncertainty can arise from the query side [146]. They cover a wide range of tissues, but their performance declines as the number of features increases in correlation-based methods. Additionally, supervised classification requires a training step, which can be computationally intensive.
There are several inherent problems in cell type annotation: first, marker genes can be present in more than one cell population, making a clear distinction between individual cell types difficult. Second, manual annotation is subjective, as the user can influence the marker genes chosen for annotation of a cell type. Furthermore, manual annotation is inherently slow and needs expert knowledge. Consequently, the knowledge of the user also determines the quality of the annotated cell types.
For post-analysis of single-cell/single-nuclei data, which require annotated cell types, this poses a serious issue. For instance, in mitoXplorer 3.0, the user needs to provide the annotation of the cell types. All subsequent analyses will depend on this annotation, which can lead to misleading and in the worst case, to meaningless results.

Pseudo-bulking of single cell/single-nuclei RNA-seq data

Transforming sc/nRNA-seq data into pseudo-bulk data involves summarizing individual cell measurements to reflect bulk RNA sequencing data. This process is useful for various downstream analyses, such as differential expression, pathway analysis, and comparison with other bulk RNA-seq datasets. Moreover, it provides the potential to investigate cell-type, rather than tissue-specific responses to different conditions, may they be during development, homeostasis, or during the onset or progression of a disease. However as is discussed above, it relies on the cell type annotation of the user. Only a meaningful clustering and annotation of cell types will lead to reasonable results in this type of pseudo-bulking.

The pseudo-bulking of single-cell data in the **scXplorer** script for **mitoXplorer3.0** involves summing gene counts of each cell type. Therefore, accurate annotation is crucial, and users should be mindful of this bias. Additionally, the number of cells per cell type can influence the results. Conversely, the mean expression of all cell types for each gene is intended to reflect bulk RNA-seq counts, but it may vary due to noise.

Finding sub-clusters of cell types

Identifying subpopulations within a cell type can provide insights into the roles of specific genes and cell groups, but it is not sufficient to fully address this question. Classically, all genes are used for sub-clustering and for identifying sub-populations of cell types. We chose to use a specific gene set, the mito-genes (or then the ataxia-genes) to sub-cluster cell types, so the clustering is done based on a limited number of genes with a particular implication in the cell or organism. It is of course interesting to investigate, if clustering of the whole gene site will cluster differently than when using only mito-genes. (Figure 12 a & b). Our data shows that clustering using all genes versus a gene set results sometimes in different cluster numbers, however a deeper analysis will be needed to investigate in how far and in which situations mito-genes are important in contributing to overall sub-clustering of cell types.
In the example of the 24 weeks old mice SCA dataset (Figure 13 a,b) and WT dataset (Figure 13 c,d), we can notice that some highly dysregulated genes are conserved in subpopulations DEA when sub clustering is performed on all genes in the experiment (Figure 13 a,c) compared to sub clustering done on only mitogenes (Figure 13 b,d). However, the expression of AldoC, a protein known to exhibit differences in PCs subpopulations [147,148] is masked when the sub clustering is done on only mitogenes. It would be interesting to dive deeper into the sub clustering of other timepoints, notably those where the number of subpopulations varies and testing other resolutions.
Sub Clustering only mitogenes could lead to discoveries like specific subpopulations

of mitochondria-related genes if matching with a sub clustering of all the genes.

Taking into account both clustering results should be considered when interpreting the results to enhance our knowledge of the system. PCs are still a small number of cells and clustering might not be ideal, even more so in mice. However, methods are emerging to enrich the cells [149] which could be useful to study a small population like PCs in the future.

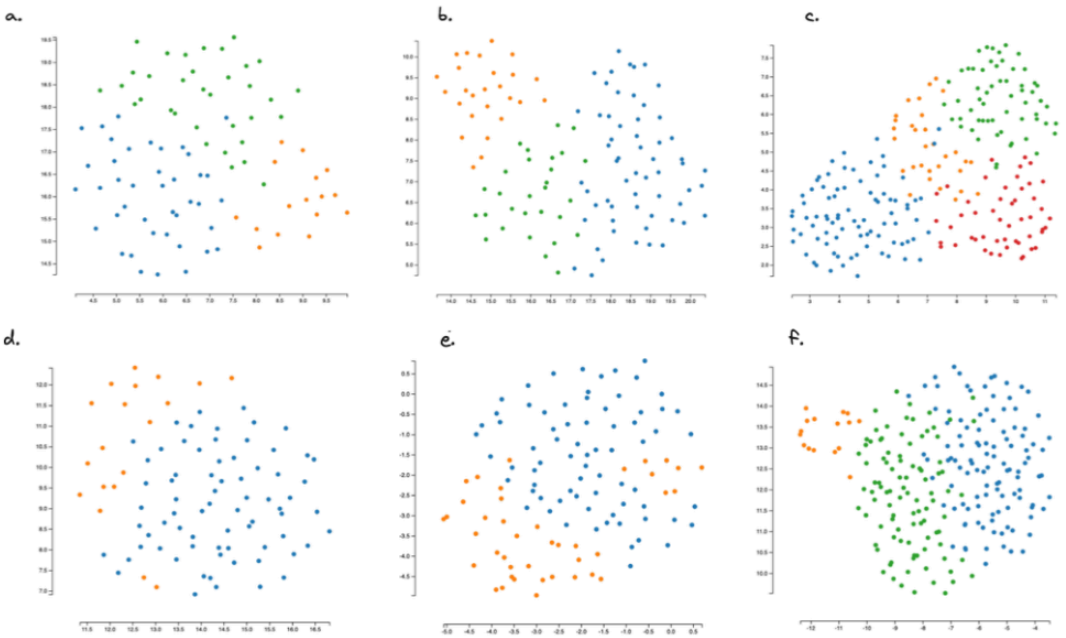


Figure 12a : SCA1 condition : all genes vs mitogenes PC subclustering. *Data of Twejani et al. SCA1 datasets with a resolution of 0.6.* ***a.b.c.*** *Subclustering of PCs using all genes for respectively 12 (a), 24 (b) weeks old mouse and human (c).* ***d.e.f.*** *subclustering of PCs only according to mito-genes in respectively 12 (d), 24 (e) weeks old mouse and human (f).*

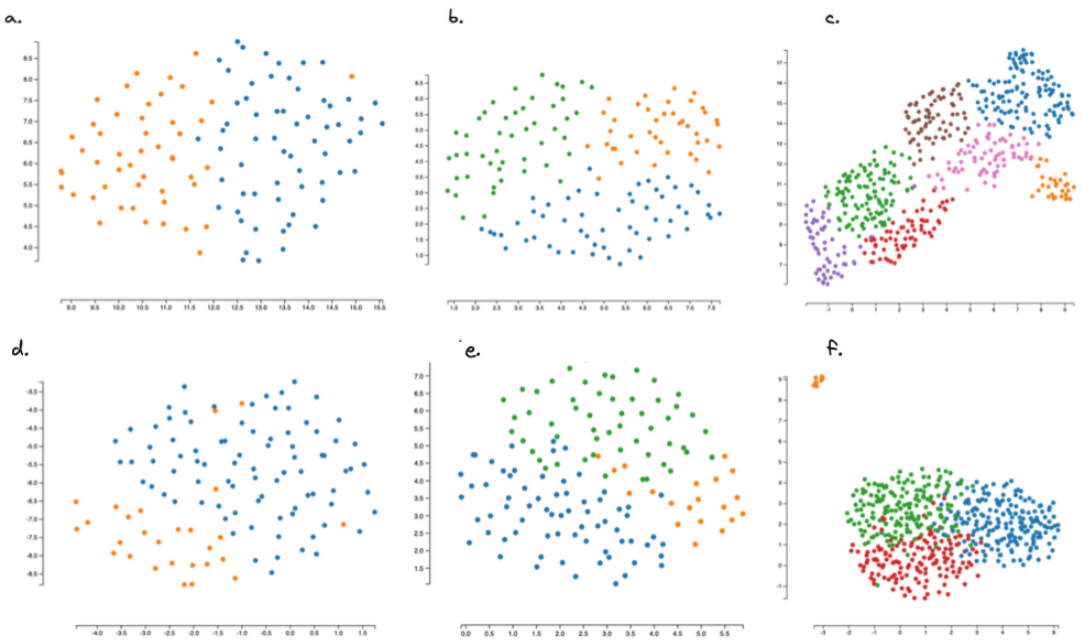


Figure 12b : WT condition : all genes vs mitogenes PC subclustering. *Data of Twejani et al. WT datasets with a resolution of 0.6.* ***a.b.c.*** *Subclustering of PCs using all genes for respectively 12 (a), 24 (b) weeks old mouse and human (c).* ***d.e.f.*** *subclustering of PCs only according to only mitogenes in respectively 12 (d), 24 (e) weeks old mouse and human (f).*

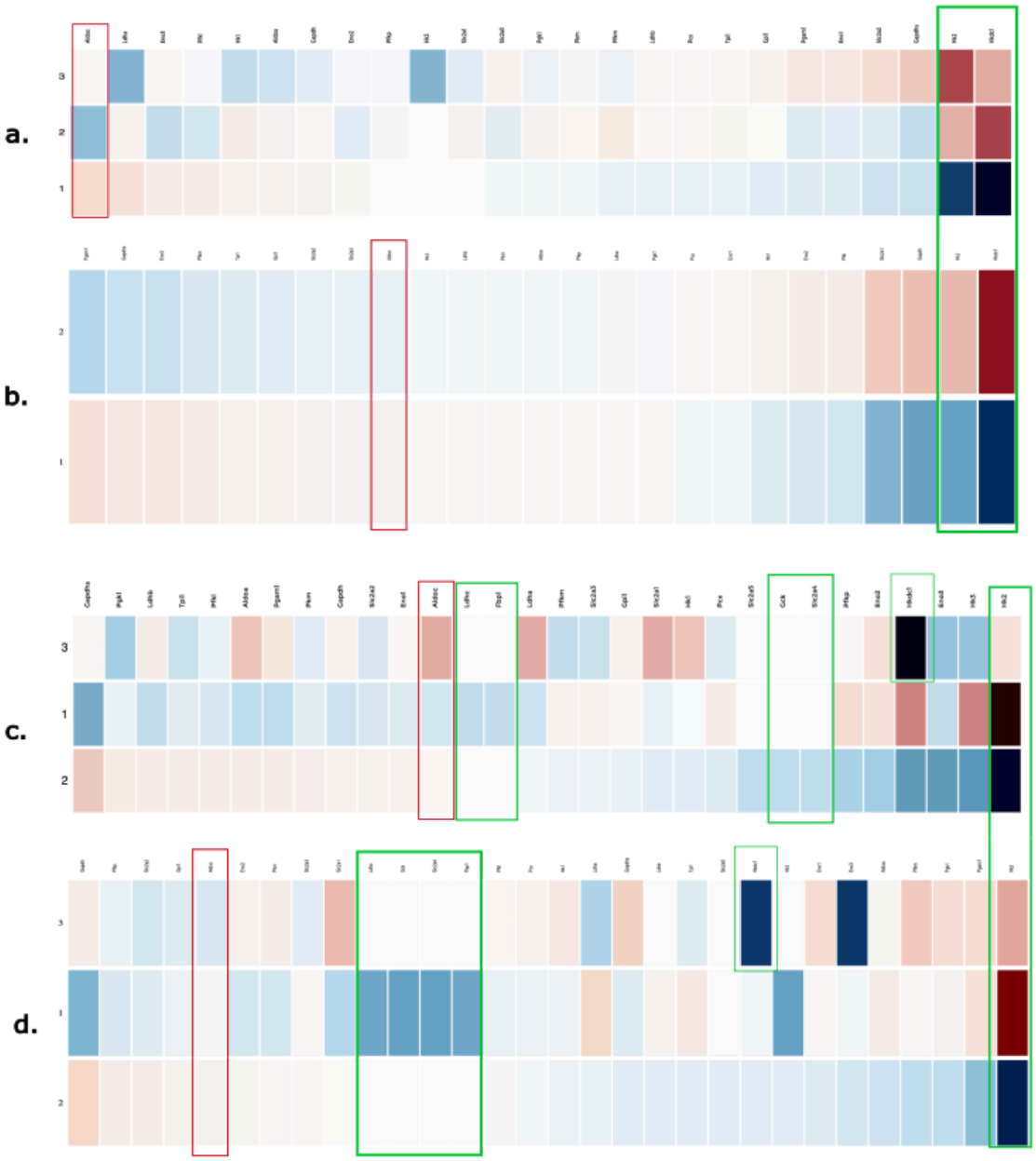


Figure 13 : Subpopulations' gene expression of mitogenes vs all data. *Subclsutering of the PCs according to the expression of all genes in the experiment and showing the difference with cells sub clustered using only mitogenes expressions. Here, genes in the Fe-S cluster biosynthesis process are shown.* ***a.*** *Heatmap of differential analysis between the 3 subpopulations identified (Fig 12a b.) by sub clustering the PCs with all genes from the 24 weeks old mouse SCA1 experiment. Only mitogenes were kept afterwards.* ***b.*** *Heatmap of differential analysis between the 2 subpopulations identified (Fig 12a d.) with only mitogenes from the 24 weeks old mouse SCA1 experiment.*
c. *Heatmap of differential analysis between the 3 subpopulations identified (Fig 12b b.) by sub clustering the PCs with all genes from the 24 weeks old mouse WT experiment.*
Only mitogenes were kept afterwards. ***d.*** *Heatmap of differential analysis between the 3 subpopulations identified (Fig 12b d.) with only mitogenes from the 24 weeks old mouse WT experiment.*

In red, AlodC is highlighted, and in green we can notice similarities between the sub clustering of all genesis vs mitogenes only.

To enhance our understanding, we could combine sub-clustering with spatial transcriptomics [150] to localize and verify these subpopulations. It could be a powerful combination of analysis to identify specific groups of cells with specific functions within a cell population and potentially develop accurate therapeutic targets.

Spatial resolution of gene expression has already been available for a small number of genes with *in situ* hybridization or immunohistochemistry for a long time [151]. However, scRNA-seq or other transcriptomic techniques provide information for all genes in the cell. More high-throughput ready spatial transcriptomic would allow us to better understand the spatial distribution of a large number of genes in different tissue regions and with more resolution [152].

A lot of methods of spatial transcriptomics are emerging as deciphering how cells are arranged - in space within tissues, based solely on the genes expressed by individual cells - is crucial for gaining deeper insights into biological systems. Two approaches of spatial transcriptomics should be especially mentioned : sequencing-based approaches where an addition of spatial barcodes to the RNA fragments is associated. Here, the resolution is not a single cell, but up to 10 cells/spot [153,154]. And imagining-based approaches that allow the detection of individual RNA molecules using fluorescent microscopy, where probes are designed for the targeted genes (<5k) that can then be visualized at the single-cell level [155]. Choosing the right technology and method might be challenging, however guidelines are emerging that could help in making the right choice [156].

The Xplorer platforms and gene-set centric data analysis

The mitoXplorer platform [99] can be used as a blue-print for gene-set centric data analysis and with ataxiaXplorer, we create the first sister platform that harbors a different interactome, the ataxia interactome, for subsequent visual data mining, data analysis and integration from an ataxia-centric point of view. However, this type of analysis can be extended to other gene sets of interest.
Developing an interactome is challenging, gene databases like the Comparative Toxicogenomics Database (CTD) [157] provide a useful starting point with responsible genes. However, it is important to identify dysregulated (hereafter called affected) genes, among the studied disease or phenotype, to be able to explore the data in depth. For ataxiaXplorer, common genes to different forms of the disease were not that frequent. This might be different for other conditions. It is however

crucial to develop a meaningful strategy of how to define a gene set of interest for an Xplorer-like platform.

Annotating such gene sets is the next challenge. For a mitoXplorer-type analysis, that focuses on pathways together with the genes, the number of terms should not be too large. Gene Ontology- based annotation is often not directly useable for such a purpose. To find meaningful terms for ataxiaXplorer, we started a team effort to select terms that are biologically relevant. This is indeed a highly subjective way. However, given user input once the tool is used by the community could ameliorate this list of terms.
The major advantage of using an Xplorer platform would be to use one's own interactome, in a context of choice. An interactome is still bound to evolve with time, like the one of ataxiaXplorer, but also the one of mitoXplorer. The ataxia interactome is currently based on 4 specific types of SCAs and what is generally known as the ataxia gene set (so genes known to cause cerebellar ataxias). There was very little public transcriptomic data on SCAs, explaining the only 4 types of SCAs constituting the current interactome. The ataxia interactome could evolve by integrating dysregulated genes from other types of ataxias, and from other timepoints to build an SCA related dysregulated gene set and offering precision on associated forms of SCAs and timepoints. The interactome could also be reduced, for instance considering only one type of ataxia and thus focusing more specifically at a specific disease phenotype.

Functionality of the platforms could evolve, for instance by implementing other types of data visualization and mining interfaces for single-cell data, as well as integrating other single-cell data types, such as spatial transcriptomics. This would allow to not only identify cellular type subpopulations, but also relate them better to a biological phenotype.

References

[1] : Andrew, S. FastQC: a quality control tool for high throughput sequence data http://www.bioinformatics.babraham.ac.uk/projects/fastqc,2010.

[2] : Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1):10-12, May 2011. DOI:10.14806/ej.17.1.200.

[3] : Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

[4] : Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

[5] : Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, Volume 30, Issue 7, April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656

[6] : Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166-9. doi: 10.1093/bioinformatics/btu638. Epub 2014 Sep 25. PMID: 25260700; PMCID: PMC4287950.

[7] : Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

[8] : Dominguez D, Tsai YH, Gomez N, Jha DK, Davis I, Wang Z. A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. Cell Res. 2016 Aug;26(8):946-62. doi: 10.1038/cr.2016.84. Epub 2016 Jul 1. PMID: 27364684; PMCID: PMC4973334.

[9] : Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol. 2015 Dec 22;16:287. doi: 10.1186/s13059-015-0853-4. PMID: 26694591; PMCID: PMC4699362.

[10] : Oh VS, Li RW. Temporal Dynamic Methods for Bulk RNA-Seq Time Series Data. Genes (Basel). 2021 Feb 27;12(3):352. doi: 10.3390/genes12030352. PMID: 33673721; PMCID: PMC7997275.

[11] : A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. ACM Comput. Surv. 31, 3 (Sept. 1999), 264–323. https://doi.org/10.1145/331499.331504

[12] : Keshava Reddy K, Mrudula K (2017) Hard and fuzzy clustering methods: a comparative study

[13] : Haque, A., Engel, J., Teichmann, S.A. *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 9, 75 (2017). https://doi.org/10.1186/s13073-017-0467-4

[14] : Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med. 2022 Mar;12(3):e694. doi: 10.1002/ctm2.694. PMID: 35352511; PMCID: PMC8964935.

[15] : Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015 May 21;58(4):610-20. doi: 10.1016/j.molcel.2015.04.005. PMID: 26000846.

[16] : Tang, F., Barbacioru, C., Wang, Y. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6, 377–382 (2009).https://doi.org/10.1038/nmeth.1315

[17] : The Tabula Muris Consortium., Overall coordination., Logistical coordination. *et al.* Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* 562, 367–372 (2018). https://doi.org/10.1038/s41586-018-0590-4

[18] : The Tabula Sapiens Consortium*, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.*Science***376**,eabl4896(2022).DOI:10.1126/science.abl4896

[19] : Lake, B.B., Codeluppi, S., Yung, Y.C. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci Rep* 7, 6031 (2017). https://doi.org/10.1038/s41598-017-04426-w

[20] : Bonner WA, Hulett HR, Sweet RG, Herzenberg LA. Fluorescence activated cell sorting. Rev Sci Instrum. 1972 Mar;43(3):404-9. doi: 10.1063/1.1685647. PMID: 5013444.
[21] : Herzenberg LA, Parks D, Sahaf B, Perez O, Roederer M, Herzenberg LA. The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin Chem. 2002 Oct;48(10):1819-27. PMID: 12324512.
[22] : Zilionis, R., Nainys, J., Veres, A. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 12, 44–73 (2017). https://doi.org/10.1038/nprot.2016.154
[23] : Xu, Z., Wang, Y., Sheng, K. *et al.* Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq. *Nat Commun* 14, 5130 (2023). https://doi.org/10.1038/s41467-023-40137-9
[24] : Satija, R., Farrell, J., Gennert, D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502 (2015). https://doi.org/10.1038/nbt.3192
[25] : Hao, Y., Stuart, T., Kowalski, M.H. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* (2023). https://doi.org/10.1038/s41587-023-01767-y
[26] : Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
[27] : Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf *bioRxiv* 2021 Dec 19. doi: 10.1101/2021.12.16.473007.
[28] : Greenacre, M., Groenen, P.J.F., Hastie, T. *et al.* Principal component analysis. *Nat Rev Methods Primers* 2, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w
[29] : Huang, H., Wang, Y., Rudin, C. *et al.* Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun Biol* 5, 719 (2022). https://doi.org/10.1038/s42003-022-03628-x
[30] : Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z
[31] : Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, *PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data*, **Database**, Volume 2019, 2019, baz046, doi:10.1093/database/baz046
[32] : Fu R, Gillen AE, Sheridan RM, Tian C, Daya M, Hao Y, Hesselberth JR, Riemondy KA. clustifyr: an R package for automated single-cell RNA sequencing cluster classification. F1000Res. 2020 Apr 1;9:223. doi: 10.12688/f1000research.22969.2. PMID: 32765839; PMCID: PMC7383722.
[33] : Gao J, Li Z, Lu Q, Zhong J, Pan L, Feng C, Tang S, Wang X, Tao Y, Lin J, Wang Q. Single-cell RNA sequencing reveals cell subpopulations in the tumor microenvironment contributing to hepatocellular carcinoma. Front Cell Dev Biol. 2023 Jun 2;11:1194199. doi: 10.3389/fcell.2023.1194199. PMID: 37333982; PMCID: PMC10272598.
[34] : Margulis L. Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth. Yale University Press; London, UK: 1970.
[35] : Gray MW, Burger G, Lang BF. The origin and early evolution of mitochondria. Genome Biol. 2001;2(6):REVIEWS1018. doi: 10.1186/gb-2001-2-6-reviews1018. Epub 2001 Jun 5. PMID: 11423013; PMCID: PMC138944.
[36] :John P, Whatley FR. Paracoccus denitrificans and the evolutionary origin of the mitochondrion. Nature. 1975 Apr 10;254(5500):495-8. doi: 10.1038/254495a0. PMID: 235742.
[37] : John, P., & Whatley, F. R. (1978). Paracoccus denitrificans Davis (Micrococcus denitrificans Beijerinck) as a mitochondrion. In *Advances in Botanical Research* (Vol. 4, pp. 51-115). Academic Press.
[38] : Andersson SG, Karlberg O, Canbäck B, Kurland CG. On the origin of mitochondria: a genomics perspective. Philos Trans R Soc Lond B Biol Sci. 2003 Jan 29;358(1429):165-77; discussion 177-9. doi: 10.1098/rstb.2002.1193. PMID: 12594925; PMCID: PMC1693097.
[39] : Protasoni M, Zeviani M. Mitochondrial Structure and Bioenergetics in Normal and Disease Conditions. Int J Mol Sci. 2021 Jan 8;22(2):586. doi: 10.3390/ijms22020586. PMID: 33435522; PMCID: PMC7827222.

[40] : Jensen RE, Dunn CD, Youngman MJ, Sesaki H. Mitochondrial building blocks. Trends Cell Biol. 2004 May;14(5):215-8. doi: 10.1016/j.tcb.2004.03.006. PMID: 15130575.
[41] : Friedman, J., Nunnari, J. Mitochondrial form and function. *Nature* 505, 335–343 (2014). https://doi.org/10.1038/nature12985
[42] : Glancy B, Kim Y, Katti P, Willingham TB. The Functional Impact of Mitochondrial Structure Across Subcellular Scales. Front Physiol. 2020 Nov 11;11:541040. doi: 10.3389/fphys.2020.541040. PMID: 33262702; PMCID: PMC7686514.
[43] : Hansen FM, Kremer LS, Karayel O, Bludau I, Larsson NG, Kühl I, Mann M. Mitochondrial phosphoproteomes are functionally specialized across tissues. Life Sci Alliance. 2023 Nov 20;7(2):e202302147. doi: 10.26508/lsa.202302147. PMID: 37984987; PMCID: PMC10662294.
[44] : Vamsi K. Mootha, Jakob Bunkenborg, Jesper V. Olsen, Majbrit Hjerrild, Jacek R. Wisniewski, Erich Stahl, Marjan S. Bolouri, Heta N. Ray, Smita Sihag, Michael Kamal, Nick Patterson, Eric S. Lander, Matthias Mann,
Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria, Cell, Volume 115, Issue 5, 2003, Pages 629-640,ISSN 0092-8674,https://doi.org/10.1016/S0092-8674(03)00926-7.
[45] : Kuznetsov AV, Hermann M, Saks V, Hengster P, Margreiter R. The cell-type specificity of mitochondrial dynamics. Int J Biochem Cell Biol. 2009 Oct;41(10):1928-39. doi: 10.1016/j.biocel.2009.03.007. Epub 2009 Mar 27. PMID: 19703655.
[46] : Liesa M, Palacín M, Zorzano A. Mitochondrial dynamics in mammalian health and disease. Physiol Rev. 2009 Jul;89(3):799-845. doi: 10.1152/physrev.00030.2008. PMID: 19584314.
[47] : Friedman, J.R. and Nunnari, J. (2014) Mitochondrial Form and Function. Nature, 505, 335-343. https://doi.org/10.1038/nature12985
[48] : Glancy B, Kim Y, Katti P, Willingham TB. The Functional Impact of Mitochondrial Structure Across Subcellular Scales. Front Physiol. 2020 Nov 11;11:541040. doi: 10.3389/fphys.2020.541040. PMID: 33262702; PMCID: PMC7686514.
[49] : Nita II, Hershfinkel M, Lewis EC, Sekler I. A crosstalk between Na^+ channels, Na^+/K^+ pump and mitochondrial Na^+ transporters controls glucose-dependent cytosolic and mitochondrial Na^+ signals. Cell Calcium. 2015 Feb;57(2):69-75. doi: 10.1016/j.ceca.2014.12.007. Epub 2014 Dec 18. PMID: 25564413.
[50] : Meldolesi J, Pozzan T. The endoplasmic reticulum Ca2+ store: a view from the lumen. Trends Biochem Sci. 1998 Jan;23(1):10-4. doi: 10.1016/s0968-0004(97)01143-2. PMID: 9478128.
[51] : Elmore S. Apoptosis: a review of programmed cell death. Toxicol Pathol. 2007 Jun;35(4):495-516. doi: 10.1080/01926230701320337. PMID: 17562483; PMCID: PMC2117903.
[52] : Czarna M, Jarmuszkiewicz W. Rola mitochondriów w wytwarzaniu i usuwaniu reaktywnych form tlenu; zwiazek z przesyłaniem sygnałów i programowana smiercia komórki [Role of mitochondria in reactive oxygen species generation and removal; relevance to signaling and programmed cell death]. Postepy Biochem. 2006;52(2):145-56. Polish. PMID: 17078504.
[53] : Napolitano G, Fasciolo G, Venditti P. Mitochondrial Management of Reactive Oxygen Species. Antioxidants (Basel). 2021 Nov 17;10(11):1824. doi: 10.3390/antiox10111824. PMID: 34829696; PMCID: PMC8614740.
[54] : Poole DC, Musch TI, Colburn TD. Oxygen flux from capillary to mitochondria: integration of contemporary discoveries. Eur J Appl Physiol. 2022 Jan;122(1):7-28. doi: 10.1007/s00421-021-04854-7. Epub 2021 Dec 23. PMID: 34940908; PMCID: PMC8890444.
[55] : Martínez-Reyes I, Chandel NS. Mitochondrial TCA cycle metabolites control physiology and disease. Nat Commun. 2020 Jan 3;11(1):102. doi: 10.1038/s41467-019-13668-3. PMID: 31900386; PMCID: PMC6941980.
[56] : Arnold PK, Finley LWS. Regulation and function of the mammalian tricarboxylic acid cycle. J Biol Chem. 2023 Feb;299(2):102838. doi: 10.1016/j.jbc.2022.102838. Epub 2022 Dec 26. PMID: 36581208; PMCID: PMC9871338.
[57] : Hatefi Y. The mitochondrial electron transport and oxidative phosphorylation system. Annu Rev Biochem. 1985;54:1015-69. doi: 10.1146/annurev.bi.54.070185.005055. PMID: 2862839.

[58] : Raimondi G, Mato-Berciano A, Pascual-Sabater S, Rovira-Rigau M, Cuatrecasas M, Fondevila C, Sánchez-Cabús S, Begthel H, Boj SF, Clevers H, Fillat C. Patient-derived pancreatic tumour organoids identify therapeutic responses to oncolytic adenoviruses. EBioMedicine. 2020 Jun;56:102786. doi: 10.1016/j.ebiom.2020.102786. Epub 2020 May 24. PMID: 32460166; PMCID: PMC7251378.
[59] : Hidalgo-Gutiérrez A, González-García P, Díaz-Casado ME, Barriocanal-Casado E, López-Herrador S, Quinzii CM, López LC. Metabolic Targets of Coenzyme Q10 in Mitochondria. Antioxidants (Basel). 2021 Mar 26;10(4):520. doi: 10.3390/antiox10040520. PMID: 33810539; PMCID: PMC8066821.
[60] : Read AD, Bentley RE, Archer SL, Dunham-Snary KJ. Mitochondrial iron-sulfur clusters: Structure, function, and an emerging role in vascular biology. Redox Biol. 2021 Nov;47:102164. doi: 10.1016/j.redox.2021.102164. Epub 2021 Oct 12. PMID: 34656823; PMCID: PMC8577454.
[61] : Lezi E, Swerdlow RH. Mitochondria in neurodegeneration. Adv Exp Med Biol. 2012;942:269-86. doi: 10.1007/978-94-007-2869-1_12. PMID: 22399427; PMCID: PMC3618469.
[62] : Reddy PH. Role of mitochondria in neurodegenerative diseases: mitochondria as a therapeutic target in Alzheimer's disease. CNS Spectr. 2009 Aug;14(8 Suppl 7):8-13; discussion 16-8. doi: 10.1017/s1092852900024901. PMID: 19890241; PMCID: PMC3056539.
[63] : Lopriore P, Ricciarini V, Siciliano G, Mancuso M, Montano V. Mitochondrial Ataxias: Molecular Classification and Clinical Heterogeneity. Neurol Int. 2022 Apr 2;14(2):337-356. doi: 10.3390/neurolint14020028. PMID: 35466209; PMCID: PMC9036286.
[64] : Bustamante-Barrientos, F.A., Luque-Campos, N., Araya, M.J. *et al.* Mitochondrial dysfunction in neurodegenerative disorders: Potential therapeutic application of mitochondrial transfer to central nervous system-residing cells. *J Transl Med* 21, 613 (2023). https://doi.org/10.1186/s12967-023-04493-w
[65] : McAfee SS, Liu Y, Sillitoe RV, Heck DH. Cerebellar Coordination of Neuronal Communication in Cerebral Cortex. Front Syst Neurosci. 2022 Jan 11;15:781527. doi: 10.3389/fnsys.2021.781527. PMID: 35087384; PMCID: PMC8787113.
[66] : Ito M. Control of mental activities by internal models in the cerebellum. Nat Rev Neurosci. 2008 Apr;9(4):304-13. doi: 10.1038/nrn2332. PMID: 18319727.
[67] : Buckner RL. The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. Neuron. 2013 Oct 30;80(3):807-15. doi: 10.1016/j.neuron.2013.10.044. PMID: 24183029.
[68] : De Zeeuw CI, Ten Brinke MM. Motor Learning and the Cerebellum. Cold Spring Harb Perspect Biol. 2015 Sep 1;7(9):a021683. doi: 10.1101/cshperspect.a021683. PMID: 26330521; PMCID: PMC4563713.
[69] : Voogd J, Glickstein M. The anatomy of the cerebellum. Trends Neurosci. 1998 Sep;21(9):370-5. doi: 10.1016/s0166-2236(98)01318-6. PMID: 9735944.
[70] : Eccles JC, Ito M, Szentágothai J. *The Cerebellum as a Neuronal Machine*. Springer-Verlag; Berlin: 1967.
[71] : Hatten ME, Roussel MF. Development and cancer of the cerebellum. *Trends Neurosci.* 2011;34:134–42
[72] : Sekirnjak C, Vissel B, Bollinger J, Faulstich M, du Lac S. Purkinje cell synapses target physiologically unique brainstem neurons. J Neurosci. 2003 Jul 16;23(15):6392-8. doi: 10.1523/JNEUROSCI.23-15-06392.2003. PMID: 12867525; PMCID: PMC6740533.
[73] : Consalez GG, Goldowitz D, Casoni F, Hawkes R. Origins, Development, and Compartmentation of the Granule Cells of the Cerebellum. Front Neural Circuits. 2021 Jan 15;14:611841. doi: 10.3389/fncir.2020.611841. PMID: 33519389; PMCID: PMC7843939.
[74] : Basile, G.A., Quartu, M., Bertino, S. *et al.* Red nucleus structure and function: from anatomy to clinical neurosciences. *Brain Struct Funct* **226**, 69–91 (2021). https://doi.org/10.1007/s00429-020-02171-x
[75] : Judd EN, Lewis SM, Person AL. Diverse inhibitory projections from the cerebellar interposed nucleus. Elife. 2021 Sep 20;10:e66231. doi: 10.7554/eLife.66231. PMID: 34542410; PMCID: PMC8483738.

[76] : Larson B, Miller S, Oscarsson O. Termination and functional organization of the dorsolateral spino-olivocerebellar path. J Physiol. 1969 Aug;203(3):611-40. doi: 10.1113/jphysiol.1969.sp008882. PMID: 5272211; PMCID: PMC1351533.
[77] : Bell CC, Kawasaki T. Relations among climbing fiber responses of nearby Purkinje Cells. J Neurophysiol. 1972 Mar;35(2):155-69. doi: 10.1152/jn.1972.35.2.155. PMID: 4337637.
[78] : Gowen E, Miall RC. The cerebellum and motor dysfunction in neuropsychiatric disorders. Cerebellum. 2007;6(3):268-79. doi: 10.1080/14734220601184821. PMID: 17786823; PMCID: PMC6010149.
[79] : Joyce MR, Nadkarni PA, Kronemer SI, Margron MJ, Slapik MB, Morgan OP, Rosenthal LS, Onyike CU, Marvel CL. Quality of Life Changes Following the Onset of Cerebellar Ataxia: Symptoms and Concerns Self-reported by Ataxia Patients and Informants. Cerebellum. 2022 Aug;21(4):592-605. doi: 10.1007/s12311-022-01393-5. Epub 2022 Mar 25. PMID: 35334077.
[80] :Durr A. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. Lancet Neurol. 2010 Sep;9(9):885-94. doi: 10.1016/S1474-4422(10)70183-6. PMID: 20723845.
[81] : Klockgether T, Mariotti C, Paulson HL. Spinocerebellar ataxia. Nat Rev Dis Primers. 2019 Apr 11;5(1):24. doi: 10.1038/s41572-019-0074-3. PMID: 30975995.
[82] : Koeppen, A.H. The pathogenesis of spinocerebellar ataxia. *Cerebellum* **4**, 62–73 (2005). https://doi.org/10.1080/14734220510007950
[83] : Paulson, H., Shakkottai, V., Clark, H. *et al.* Polyglutamine spinocerebellar ataxias — from genes to potential treatments. *Nat Rev Neurosci* 18, 613–626 (2017). https://doi.org/10.1038/nrn.2017.92
[84] : Monin ML, Tezenas du Montcel S, Marelli C, Cazeneuve C, Charles P, Tallaksen C, Forlani S, Stevanin G, Brice A, Durr A. Survival and severity in dominant cerebellar ataxias. Ann Clin Transl Neurol. 2015 Feb;2(2):202-7. doi: 10.1002/acn3.156. Epub 2015 Jan 7. PMID: 25750924; PMCID: PMC4338960.
[85] : Chiabrando D, Bertino F, Tolosano E. Hereditary Ataxia: A Focus on Heme Metabolism and Fe-S Cluster Biogenesis. Int J Mol Sci. 2020 May 26;21(11):3760. doi: 10.3390/ijms21113760. PMID: 32466579; PMCID: PMC7312568.
[86] : Ajit Bolar N, Vanlander AV, Wilbrecht C, Van der Aa N, Smet J, De Paepe B, Vandeweyer G, Kooy F, Eyskens F, De Latter E, Delanghe G, Govaert P, Leroy JG, Loeys B, Lill R, Van Laer L, Van Coster R. Mutation of the iron-sulfur cluster assembly gene IBA57
causes severe myopathy and encephalopathy. Hum Mol Genet. 2013 Jul 1;22(13):2590-602. doi: 10.1093/hmg/ddt107. Epub 2013 Mar 5. PMID: 23462291.
[87] : Lo RY, Figueroa KP, Pulst SM, Lin CY, Perlman S, Wilmot G, Gomez C, Schmahmann J, Paulson H, Shakkottai VG, Ying S, Zesiewicz T, Bushara K, Geschwind M, Xia G, Subramony SH, Ashizawa T, Kuo SH. Coenzyme Q10 and spinocerebellar ataxias. Mov Disord. 2015 Feb;30(2):214-20. doi: 10.1002/mds.26088. Epub 2014 Dec 1. PMID: 25449974; PMCID: PMC4674072.
[88] : Zhang L, Ashizawa T, Peng D. Primary coenzyme Q10 deficiency due to COQ8A gene mutations. Mol Genet Genomic Med. 2020 Oct;8(10):e1420. doi: 10.1002/mgg3.1420. Epub 2020 Aug 2. PMID: 32743982; PMCID: PMC7549598.
[89] : Wu YL, Chang JC, Sun HL, Cheng WL, Yen YP, Lin YS, Chao YC, Liu KH, Huang CS, Liu KL, Liu CS. Coenzyme Q10 Supplementation Increases Removal of the *ATXN3* Polyglutamine Repeat, Reducing Cerebellar Degeneration and Improving Motor Dysfunction in Murine Spinocerebellar Ataxia Type 3. Nutrients. 2022 Aug 31;14(17):3593. doi: 10.3390/nu14173593. PMID: 36079853; PMCID: PMC9459709.
[90] : Norat, P., Soldozy, S., Sokolowski, J.D. *et al.* Mitochondrial dysfunction in neurological disorders: Exploring mitochondrial transplantation. *npj Regen Med* 5, 22 (2020). https://doi.org/10.1038/s41536-020-00107-x
[91] : McFarland R, Taylor RW, Turnbull DM. A neurological perspective on mitochondrial disease. Lancet Neurol. 2010 Aug;9(8):829-40. doi: 10.1016/S1474-4422(10)70116-2. PMID: 20650404.
[92] : Parikh, S. (2010), The neurologic manifestations of mitochondrial disease. Dev Disabil Res Revs, 16: 120-128. https://doi.org/10.1002/ddrr.110

[93] : Li, L., Conradson, D.M., Bharat, V. *et al.* A mitochondrial membrane-bridging machinery mediates signal transduction of intramitochondrial oxidation. *Nat Metab* 3, 1242–1258 (2021). https://doi.org/10.1038/s42255-021-00443-2
[94] : Stovickova, L., Hansikova, H., Hanzalova, J. *et al.* Exploring mitochondrial biomarkers for Friedreich's ataxia: a multifaceted approach. *J Neurol* **271**, 3439–3454 (2024). https://doi.org/10.1007/s00415-024-12223-5
[95] : Lynch DR, Farmer G. Mitochondrial and metabolic dysfunction in Friedreich ataxia: update on pathophysiological relevance and clinical interventions. Neuronal Signal. 2021 May 17;5(2):NS20200093. doi: 10.1042/NS20200093. PMID: 34046211; PMCID: PMC8132591.
[96] : Bostock M, Ogievetsky V, Heer J. D³: Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011 Dec;17(12):2301-9. doi: 10.1109/TVCG.2011.185. PMID: 22034350.
[97] : Axmark, D., & Widenius, M. (2015). *MySQL 5.7 reference manual.* Redwood Shores, CA: Oracle. Available online at http://dev.mysql.com/doc/refman/5.7/en/index.html
[98] : Hipp, R, et. al. (2015). SQLite (Version 3.8.10.2) [Computer software].
SQLite Development Team. Retrieved June 15, 2015.
Available from <https://www.sqlite.org/download.html>
[99] : Annie Yim, Prasanna Koti, Adrien Bonnard, Fabio Marchiano, Milena Dürrbaum, Cecilia Garcia-Perez, Jose Villaveces, Salma Gamal, Giovanni Cardone, Fabiana Perocchi, Zuzana Storchova, Bianca H Habermann, mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations, *Nucleic Acids Research*, Volume 48, Issue 2, 24 January 2020, Pages 605–632, https://doi.org/10.1093/nar/gkz1128
[100] : Fabio Marchiano, Margaux Haering, Bianca Hermine Habermann, The mitoXplorer 2.0 update: integrating and interpreting mitochondrial expression dynamics within a cellular context, *Nucleic Acids Research*, Volume 50, Issue W1, 5 July 2022, Pages W490–W499, https://doi.org/10.1093/nar/gkac306
[101] : https://www.mongodb.com/docs/
[102] : Haering M, Habermann BH. RNfuzzyApp: an R shiny RNA-seq data analysis app for visualization, differential expression analysis, time-series clustering and enrichment analysis. F1000Res. 2021 Jul 26;10:654. doi: 10.12688/f1000research.54533.2. PMID: 35186266; PMCID: PMC8825645.
[103] : Kumar L, E Futschik M. Mfuzz: a software package for soft clustering of microarray data. Bioinformation. 2007 May 20;2(1):5-7. doi: 10.6026/97320630002005. PMID: 18084642; PMCID: PMC2139991.
[104] : Afgan E, Baker D, Batut B et al (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544
[105] : Batut B, van den Beek M, Doyle MA, Soranzo N. RNA-Seq Data Analysis in Galaxy. Methods Mol Biol. 2021;2284:367-392. doi: 10.1007/978-1-0716-1307-8_20. PMID: 33835453.
[106] : White JJ, Bosman LWJ, Blot FGC, Osório C, Kuppens BW, Krijnen WHJJ, Andriessen C, De Zeeuw CI, Jaarsma D, Schonewille M. Region-specific preservation of Purkinje cell morphology and motor behavior in the ATXN1[82Q] mouse model of spinocerebellar ataxia 1. Brain Pathol. 2021 Sep;31(5):e12946. doi: 10.1111/bpa.12946. Epub 2021 Mar 16. PMID: 33724582; PMCID: PMC8412070.
[107] : Gao HM, Hong JS. Why neurodegenerative diseases are progressive: uncontrolled inflammation drives disease progression. Trends Immunol. 2008 Aug;29(8):357-65. doi: 10.1016/j.it.2008.05.002. Epub 2008 Jul 1. PMID: 18599350; PMCID: PMC4794280.
[108] : Yang, HS., Teng, L., Kang, D. *et al.* Cell-type-specific Alzheimer's disease polygenic risk scores are associated with distinct disease processes in Alzheimer's disease. *Nat Commun* **14**, 7659 (2023). https://doi.org/10.1038/s41467-023-43132-2
[109] : Reynolds, R.H., Botía, J., Nalls, M.A. *et al.* Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. *npj Parkinsons Dis.* **5**, 6 (2019). https://doi.org/10.1038/s41531-019-0076-6
[110] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937225/

[111] : Zoghbi H., Orr H. Pathogenic Mechanisms of a Polyglutamine-mediated Neurodegenerative Disease, Spinocerebellar Ataxia Type 1. *J. Biol. Chem.* 2009;284:7425–7429. doi: 10.1074/jbc.R800041200.
[112] : Tejwani L., Lim J. Pathogenic mechanisms underlying spinocerebellar ataxia type 1. *Cell. Mol. Life Sci.* 2020;77:4015–4029. doi: 10.1007/s00018-020-03520-z.
[113] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10216181/#B24-biomolecules-13-00788
[114] : Driessen TM, Lee PJ, Lim J. Molecular pathway analysis towards understanding tissue vulnerability in spinocerebellar ataxia type 1. Elife. 2018 Dec 3;7:e39981. doi: 10.7554/eLife.39981. PMID: 30507379; PMCID: PMC6292693.
[115] : Pérez Ortiz JM, Mollema N, Toker N, Adamski CJ, O'Callaghan B, Duvick L, Friedrich J, Walters MA, Strasser J, Hawkinson JE, Zoghbi HY, Henzler C, Orr HT, Lagalwar S. Reduction of protein kinase A-mediated phosphorylation of ATXN1-S776 in Purkinje cells delays onset of Ataxia in a SCA1 mouse model. Neurobiol Dis. 2018 Aug;116:93-105. doi: 10.1016/j.nbd.2018.05.002. Epub 2018 May 11. PMID: 29758256; PMCID: PMC6028938.
[116] : Ingram M, Wozniak EAL, Duvick L, Yang R, Bergmann P, Carson R, O'Callaghan B, Zoghbi HY, Henzler C, Orr HT. Cerebellar Transcriptome Profiles of ATXN1 Transgenic Mice Reveal SCA1 Disease Progression and Protection Pathways. Neuron. 2016 Mar 16;89(6):1194-1207. doi: 10.1016/j.neuron.2016.02.011. Epub 2016 Mar 3. PMID: 26948890; PMCID: PMC4795980.
[117] : Rousseaux MWC, Tschumperlin T, Lu HC, Lackey EP, Bondar VV, Wan YW, Tan Q, Adamski CJ, Friedrich J, Twaroski K, Chen W, Tolar J, Henzler C, Sharma A, Bajić A, Lin T, Duvick L, Liu Z, Sillitoe RV, Zoghbi HY, Orr HT. ATXN1-CIC Complex Is the Primary Driver of Cerebellar Pathology in Spinocerebellar Ataxia Type 1 through a Gain-of-Function Mechanism. Neuron. 2018 Mar 21;97(6):1235-1243.e5. doi: 10.1016/j.neuron.2018.02.013. Epub 2018 Mar 8. PMID: 29526553; PMCID: PMC6422678.
[118] : Lin YT, Lin YS, Cheng WL, Chang JC, Chao YC, Liu CS, Wei AC. Transcriptomic and Metabolic Network Analysis of Metabolic Reprogramming and IGF-1 Modulation in SCA3 Transgenic Mice. Int J Mol Sci. 2021 Jul 26;22(15):7974. doi: 10.3390/ijms22157974. PMID: 34360740; PMCID: PMC8348158.
[119] : Haas E, Incebacak RD, Hentrich T, Huridou C, Schmidt T, Casadei N, Maringer Y, Bahl C, Zimmermann F, Mills JD, Aronica E, Riess O, Schulze-Hentrich JM, Hübener-Schmid J. A Novel SCA3 Knock-in Mouse Model Mimics the Human SCA3 Disease Phenotype Including Neuropathological, Behavioral, and Transcriptional Abnormalities Especially in Oligodendrocytes. Mol Neurobiol. 2022 Jan;59(1):495-522. doi: 10.1007/s12035-021-02610-8. Epub 2021 Oct 30. PMID: 34716557; PMCID: PMC8786755.
[120] : Stoyas CA, Bushart DD, Switonski PM, Ward JM, Alaghatta A, Tang MB, Niu C, Wadhwa M, Huang H, Savchenko A, Gariani K, Xie F, Delaney JR, Gaasterland T, Auwerx J, Shakkottai VG, La Spada AR. Nicotinamide Pathway-Dependent Sirt1 Activation Restores Calcium Homeostasis to Achieve Neuroprotection in Spinocerebellar Ataxia Type 7. Neuron. 2020 Feb 19;105(4):630-644.e9. doi: 10.1016/j.neuron.2019.11.019. Epub 2019 Dec 16. PMID: 31859031; PMCID: PMC7147995.
[121] : Niewiadomska-Cimicka A, Doussau F, Perot JB, Roux MJ, Keime C, Hache A, Piguet F, Novati A, Weber C, Yalcin B, Meziane H, Champy MF, Grandgirard E, Karam A, Messaddeq N, Eisenmann A, Brouillet E, Nguyen HHP, Flament J, Isope P, Trottier Y. SCA7 Mouse Cerebellar Pathology Reveals Preferential Downregulation of Key Purkinje Cell-Identity Genes and Shared Disease Signature with SCA1 and SCA2. J Neurosci. 2021 Jun 2;41(22):4910-4936. doi: 10.1523/JNEUROSCI.1882-20.2021. Epub 2021 Apr 22. PMID: 33888607; PMCID: PMC8260160.
[122] : Liu Q, Huang S, Yin P, Yang S, Zhang J, Jing L, Cheng S, Tang B, Li XJ, Pan Y, Li S. Cerebellum-enriched protein INPP5A contributes to selective neuropathology in mouse model of spinocerebellar ataxias type 17. Nat Commun. 2020 Feb 27;11(1):1101. doi: 10.1038/s41467-020-14931-8. PMID: 32107387; PMCID: PMC7046734.
[123] : Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update).

Nucleic Acids Res. 2022 Jul 5;50(W1):W216-W221. doi: 10.1093/nar/gkac194. PMID: 35325185; PMCID: PMC9252805.
[124] : Meiler A, Marchiano F, Haering M, Weitkunat M, Schnorrer F, Habermann BH. AnnoMiner is a new web-tool to integrate epigenetics, transcription factor occupancy and transcriptomics data to predict transcriptional regulators. Sci Rep. 2021 Jul 29;11(1):15463. doi: 10.1038/s41598-021-94805-1. PMID: 34326396; PMCID: PMC8322331.
[125] : Garmhausen, M., Hofmann, F., Senderov, V. *et al.* Virtual pathway explorer (viPEr) and pathway enrichment analysis tool (PEANuT): creating and analyzing focus networks to identify cross-talk between molecules and pathways. *BMC Genomics* 16, 790 (2015). https://doi.org/10.1186/s12864-015-2017-z
[126] : Su, W., Sun, J., Shimizu, K. *et al.* TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Res Notes* 12, 133 (2019). https://doi.org/10.1186/s13104-019-4179-2
[127] : Reyes, A.L.P., Silva, T.C., Coetzee, S.G. *et al.* GENAVi: a shiny web application for gene expression normalization, analysis and visualization. *BMC Genomics* 20, 745 (2019). https://doi.org/10.1186/s12864-019-6073-7
[128] : Riquelme-Perez M, Perez-Sanz F, Deleuze JF, Escartin C, Bonnet E, Brohard S. DEVEA: an interactive shiny application for Differential Expression analysis, data Visualization and Enrichment Analysis of transcriptomics data. F1000Res. 2022 Jun 28;11:711. doi: 10.12688/f1000research.122949.2. PMID: 36999088; PMCID: PMC10043628.
[129] : Ilias Moutsopoulos, Eleanor C Williams, Irina I Mohorianu, bulkAnalyseR: an accessible, interactive pipeline for analysing and sharing bulk multi-modal sequencing data, *Briefings in Bioinformatics*, Volume 24, Issue 1, January 2023, bbac591, https://doi.org/10.1093/bib/bbac591
[130] : Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818.
[131] : Lähnemann, D., Köster, J., Szczurek, E. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020). https://doi.org/10.1186/s13059-020-1926-6
[132] : Adil A, Kumar V, Jan AT, Asger M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. Front Neurosci. 2021 Apr 22;15:591122. doi: 10.3389/fnins.2021.591122. PMID: 33967674; PMCID: PMC8100238.
[133] : Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun* **11**, 1169 (2020). https://doi.org/10.1038/s41467-020-14976-9
[134] : Phipson B, Zappia L, Oshlack A. Gene length and detection bias in single cell RNA sequencing protocols. F1000Res. 2017 Apr 28;6:595. doi: 10.12688/f1000research.11290.1. PMID: 28529717; PMCID: PMC5428526.
[135] : Ascensión AM, Araúzo-Bravo MJ, Izeta A. The need to reassess single-cell RNA sequencing datasets: the importance of biological sample processing. F1000Res. 2022 Mar 8;10:767. doi: 10.12688/f1000research.54864.2. PMCID: PMC8984215.
[136] : Gisele M. Hodges, D. C. Livingston, L. M. Franks; The localization of trypsin in cultured mammalian cells. *J Cell Sci* 1 May 1973; 12 (3): 887–902. doi: https://doi.org/10.1242/jcs.12.3.887
[137] : Khan, S. J., Abidi, S. N. F., Tian, Y., Skinner, A., & Smith-Bolton, R. K. (2016). A rapid, gentle and scalable method for dissociation and fluorescent sorting of imaginal disc cells for mRNA sequencing. *Fly*, *10*(2), 73–80. https://doi.org/10.1080/19336934.2016.1173296
[138] : Neuschulz A, Bakina O, Badillo-Lisakowski V, Olivares-Chauvet P, Conrad T, Gotthardt M, Kettenmann H, Junker JP. A single-cell RNA labeling strategy for measuring stress response upon tissue dissociation. Mol Syst Biol. 2023 Feb 10;19(2):e11147. doi: 10.15252/msb.202211147. Epub 2022 Dec 27. PMID: 36573354; PMCID: PMC9912023.
[139] : Wiegleb G, Reinhardt S, Dahl A, Posnien N. Tissue dissociation for single-cell and single-nuclei RNA sequencing for low amounts of input material. Front Zool. 2022 Nov 12;19(1):27. doi: 10.1186/s12983-022-00472-x. PMID: 36371195; PMCID: PMC9652833.

[140] : Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, Kwon JYH, Barak B, Ge W, Kedaigle AJ, Carroll S, Li S, Hacohen N, Rozenblatt-Rosen O, Shalek AK, Villani AC, Regev A, Levin JZ. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020 Jun;38(6):737-746. doi: 10.1038/s41587-020-0465-8. Epub 2020 Apr 6. Erratum in: Nat Biotechnol. 2020 Apr 27;: PMID: 32341560; PMCID: PMC7289686.
[141] : Kim N, Kang H, Jo A, Yoo SA, Lee HO. Perspectives on single-nucleus RNA sequencing in different cell types and tissues. J Pathol Transl Med. 2023 Jan;57(1):52-59. doi: 10.4132/jptm.2022.12.19. Epub 2023 Jan 10. PMID: 36623812; PMCID: PMC9846005.
[142] : Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, Xia Li, Peng Wang, Yunpeng Zhang, CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data, *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D870–D876, https://doi.org/10.1093/nar/gkac947
[143] : Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.
[144] : Li, C., Liu, B., Kang, B. *et al*. SciBet as a portable and fast single cell type identifier. *Nat Commun* **11**, 1818 (2020). https://doi.org/10.1038/s41467-020-15523-2
[145] : Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019). "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." *Nat. Immunol.*, **20**, 163-172. doi:10.1038/s41590-018-0276-y.
[146] :Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J. 2021 Jan 19;19:961-969. doi: 10.1016/j.csbj.2021.01.015. PMID: 33613863; PMCID: PMC7873570.
[147] : Apps, R., Hawkes, R. Cerebellar cortical organization: a one-map hypothesis. *Nat Rev Neurosci* 10, 670–681 (2009). https://doi.org/10.1038/nrn2698
[148] : Blot, F.G.C., White, J.J., van Hattem, A. *et al*. Purkinje cell microzones mediate distinct kinematics of a single movement. *Nat Commun* 14, 4358 (2023).
[149] : Bartelt LC, Switonski PM, Adamek G, Carvalho J, Duvick LA, Jarrah SI, McLoughlin HS, Scoles DR, Pulst SM, Orr HT, Hull C, Lowe CB, La Spada AR. Purkinje-Enriched snRNA-seq in SCA7 Cerebellum Reveals Zebrin Identity Loss as a Central Feature of Polyglutamine Ataxias. bioRxiv [Preprint]. 2023 May 11:2023.03.19.533345. doi: 10.1101/2023.03.19.533345. PMID: 37214832; PMCID: PMC10197555.
[150] : Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78-82. doi: 10.1126/science.aaf2403. PMID: 27365449.
[151] : Armit C, Richardson L, Venkataraman S, Graham L, Burton N, Hill B, Yang Y, Baldock RA. eMouseAtlas: An atlas-based resource for understanding mammalian embryogenesis. Dev Biol. 2017 Mar 1;423(1):1-11. doi: 10.1016/j.ydbio.2017.01.023. Epub 2017 Feb 2. PMID: 28161522; PMCID: PMC5442644.
[152] : Salmén F, Ståhl PL, Mollbrink A, Navarro JF, Vickovic S, Frisén J, Lundeberg J. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat Protoc. 2018 Nov;13(11):2501-2534. doi: 10.1038/s41596-018-0045-2. PMID: 30353172.
[153] : *JACS Au* 2024, 4, 5, 1723−1743
Publication Date:April 15, 2024https://doi.org/10.1021/jacsau.4c00118
[154] : Systematic comparison of sequencing-based spatial transcriptomic methods

Yue You, Yuting Fu, Lanxiang Li, Zhongming Zhang, Shikai Jia, Shihong Lu, Wenle Ren, Yifang Liu, Yang Xu, Xiaojing Liu, Fuqing Jiang, Guangdun Peng, Abhishek SampathKumar, Matthew E. Ritchie, Xiaodong Liu, Luyi Tian

bioRxiv 2023.12.03.569744; doi: https://doi.org/10.1101/2023.12.03.569744

[155] : Strack, R. Imaging-based spatial transcriptomics goes electric. *Nat Methods* 19, 1338 (2022). https://doi.org/10.1038/s41592-022-01678-x

[156] : Zhou Y, Jia E, Pan M, Zhao X, Ge Q. Encoding Method of Single-cell Spatial Transcriptomics Sequencing. Int J Biol Sci. 2020 Jul 30;16(14):2663-2674. doi: 10.7150/ijbs.43887. PMID: 32792863; PMCID: PMC7415427.

[157] : Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res. 2022 Sep 28.