Year : 2023

A systems genetics approach for sleep regulation

Gobet Nastassia

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Centre Intégratif de Génomique (CIG)

A systems genetics approach for sleep regulation

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Nastassia GOBET

Master de l'Université de Lausanne

Jury

Prof. Esat Mahmut Ozsahin, Président
Prof. Ioannis Xenarios, Directeur de thèse
Prof. Paul Franken, Co-directeur de thèse
Dr. Karsten Borgwardt, Expert
Prof. Olivier Delaneau, Expert

Lausanne 2023

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Esat Mahmut	**Ozsahin**
Directeur·trice de thèse	Monsieur	Prof.	Ioannis	**Xenarios**
Co-directeur·trice	Monsieur	Prof.	Paul	**Franken**
Expert·e·s	Monsieur	Dr	Karsten	**Borgwardt**
	Monsieur	Prof.	Olivier	**Delaneau**

le Conseil de Faculté autorise l'impression de la thèse de

Nastassia Gobet

Maîtrise universitaire ès Sciences en sciences moléculaires du vivant, Université de Lausanne

intitulée

A systems genetics approach for sleep regulation

Lausanne, le 26 avril 2023

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Esat Mahmut Ozsahin

University of Lausanne (UNIL)
Faculty of Biology and Medicine (FBM)
Center for integrative genomics (CIG)

A systems genetics approach for sleep regulation

Nastassia Gobet

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Biology of the University of Lausanne, April 2023

Table of contents

Acknowledgements

I would like to express my thanks to:

My supervisor Ioannis Xenarios, for his big ideas, his optimism, and his support.

My co-supervisor Paul Franken, for his attention to details, his questions, and his support.

My experts Karsten Borgwardt and Olivier Delaneau, for helpful suggestions and discussions.

My colleagues, for the scientific and non-scientific interactions: guidance, discussions about incomprehensible papers, hugs, advices, lunchs, chats, aperos, swimming in the lake, ice skating, and dreaming of a more ideal world. Special thanks to Maxime Jan for his calm, Nina Đukanović for her kindness and her cakes, Mathilde Goullieux and Alessandro Cuozzo for reaching out. The colleagues from the Franken lab: Carlos Neves, Jeff Hubbard, Shanaz Diessler, Marieke Hoekstra, Andrey Lazopulo, Kostas Kompotis, Sonia Jimenez, Yann Emmenegger and Charlotte Hor. The colleagues at vital-it for taking me under their wing although they just lost theirs. Particularly the Analysts: Florence Mehl, Flavia Marzetta, Josefina Lascano Maillard, Leonore Wigger, Thuong Van Du Tran, Marco Pagni, Frédéric Burdet. And also Olivier Martin, Sébastien Moretti, Robin Engler, Dmitry Kuznetsov, Orlin Topalov, Anne Niknejad, Laith Abu-Nawwas, Diana Marek, Vassilios Ioannidis, Etienne de Rham, Brian Stevenson, and Mark Ibberson. My mentor, Liliane Michalik, for her precious advice, and the nice moments shared together.

The secretaries for guiding me through the administrative mysteries: Iris Marouani, Corinne Dentan, Nathalie Clerc, Julie Papet, and Muriel Metrailler Lenoir.

Last but not least, I warmly thank my family and friends, for the games, the meals, the discussions, the poetry, and the essential in life.

Dedication

I dedicate this thesis to my father, who showed me that courage is not to not have fears and who knew the words to speak to my heart.

Abstract

Sleep is a daily behavior important for health. Many people studied sleep with more or less sophisticated technologies over time, and yet it has not revealed all its mysteries. To help uncover the molecular consequences of sleep deprivation, the Franken group have assembled a systems genetics resource interrogating the BXD mouse panel. The genotypes and sleep-wake phenome were characterized, along with intermediate phenotypes: the transcriptome in brain and in liver, and the targeted metabolome in the blood plasma. I have used this rich multi-omics BXD dataset for computational investigation and development of analytical methods for data and knowledge integration to expand the current understanding of sleep regulation. First, in collaboration with Maxime Jan we used this real-world example of data and bioinformatic analysis management to highlight multi-omics challenges and solutions used to help internal or external reusability. This includes more details on the quality check and validations of the methods, the use of Rmarkdown reports for more higher levels parts of the analyses, a metadata workflow document illustrating and referencing the different code and data files, and a web site for exploration of the results. The robustness of the results was also assessed through the change to the newest version of the mouse genome reference assembly used. Then, the classical pipeline to analyse RNA-sequencing reads uses one mouse reference for all samples, irrespective of the strain of the samples, which is potentially creates a reference bias. Therefore, to improve the genetic-specificity of the read mapping, I customized the standard assembly based on one parental strain with variants from the BXD population. An important step was adding a tailored imputation of the population genetic variants using haplotypes blocks/regions to achieve a sufficient resolution for each line-specific reference. This strategy alleviated the reference bias and allowed to detect proportionally more eQTLs with the custom BXD-specific references than with the standard reference. Lastly, I assembled a multi-layer prior knowledge network and integrated the gene expression sleep-specific on it. This integration of data-driven and knowledge driven approach sets the basis for a way to generate hypotheses based on multiple genes to explain the genetic and environmental interactions culminating in the different sleep phenotypes.

Résumé

Le sommeil est un comportement quotidien important pour la santé. De nombreuses personnes ont étudié le sommeil avec des technologies plus ou moins sophistiquées au fil du temps, et il n'a cependant pas encore révélé tous ses mystères. Pour aider à découvrir les conséquences moléculaires de la privation de sommeil, le groupe Franken a assemblé une ressource de génétique des systèmes relative aux lignées de souris BXD. Les génotypes et le phénome de sommeil-éveil ont été charactérisés, ainsi que des phénotypes intermédiaires : d'une part le transcriptome dans le cerveau et le foie, d'autre part le métabolome ciblé dans le plasma sanguin. J'ai utilisé ce riche jeu de données multi-omics sur les BXD pour le développement de méthodes analytiques pour l'intégration de données et de connaissances afin d'étendre la compréhension actuelle de la régulation du sommeil. D'abord, en collaboration avec Maxime Jan, nous avons utilisé cet exemple réel de la gestion des données et de l'analyse bioinformatique pour mettre en évidence les défis multi-omics et les solutions utilisées pour que le travail puisse être réutilisé à l'interne ou à l'externe. Cela inclut plus de détails sur le contrôle de qualité et les validations des méthodes, l'utilisation de rapports Rmarkdown pour les parties de plus haut niveau d'abstraction des analyses, un document concernant les méta-données du flux de travail pour illustrer et référencer les différents scripts et fichiers de données et un site web pour l'exploration des résultats. La stabilité des résultats a également été évaluée au travers du changement de version de l'assemblée de réference utilisée. Puis, la pipeline traditionnelle pour analyser des reads de séquençage d'ARN utilise une référence murine pour tous les échantillons, quelle que soit leur souche. Afin d'améliorer la spécificité génétique du mapping des reads, j'ai utilisé et personnalisé l'assemblée standard basée sur une souche parentale avec les variants de la population BXD. L'imputation des variants génétiques en utilisant les blocs/régions haplotypes était importante pour obtenir une résolution suffisante pour chacune des lignées. Cette stratégie a diminué le biais de référence et a permis de détecter proportionnellement plus d'eQTLs avec les références spécifiques aux BXD qu'avec la référence traditionnelle. Finalement, j'ai assemblé un réseau à plusieurs couches de connaissances préalables et y ait intégré l'expression des gènes contenant la composante spécique au sommeil. L'intégration des approches basées sur les données et les connaissances préalables met en place la base pour un moyen de générer des hypothèses basées sur plusieurs gènes pour expliquer les interactions génétiques et environmentales provoquant les différents phénotypes du sommeil.

List of Figures

List of Tables

Chapter 1

Introduction

Sleep is an essential behavior, and disturbed sleep or sleep loss has many short and long term consequences on concentration, mood, and health [Tariq et al., 2020]. Inadequate sleep seem to increase risks to develop different diseases such as type 2 diabetes [Spiegel et al., 2009] and some types of cancer [Spiegel et al., 2005, Lu et al., 2017, Chen et al., 2018, Szkiela et al., 2020, Manouchehri et al., 2021, Szkiela et al., 2021]. Additionally, theses conditions also seem responsible for sleep disturbances [Atef et al., 2022]. It remains unclear exactly what mechanisms could link sleep to these diseases [Everson et al., 2014]. It appears that sleep is an important and vast subject with many unknowns.

In my thesis, by using the systems genetics resource interrogating the BXD mouse panel assembled by the Franken lab [Diessler et al., 2018], I first assessed the reusability of the analyses that were already performed in a formal way. This also allowed to identify directions for further explorations. One is addressing the gap of standard RNA-sequencing analyses that use one unique reference for all genetically diverse samples. The other is to combine the dependencies between biological items and their variation in different genetic backgrounds to reveal multi-variate subnetworks of sleep.

1.1 Background

1.1.1 Sleep, molecules, and mouse

To clearly distinguish sleep from other states (for examples: coma, hibernation, or hypnosis) five criteria can be used for mammals. 1) It has to be a period of reduced activity, 2) the responsiveness to stimuli from environment is reduced, 3) it is quickly reversible, 4) it is homeostatically regulated, including a need to compensate after being deprived of it, 5) it is often linked to circadian (daily) rhythms.

Outside mammals, other species are consider to sleep like flies (drosophila), or to have a sleep-like state such as worms (nematodes) or jellyfishs [Nath et al., 2017]. The reticence to call it sleep may somehow depends on the fact they do not have a brain and monitoring brain activity is often considered to be the gold standard for measuring sleep. However, having a brain is not in the criteria. It was observed that the sleep activity is not homogeneous in the brain and dolphins are a famous example of how brain asymmetry of sleep [Schulz, 2022].

The two process model is a framework where the propensity to sleep comes from 2 processes [Borbély, 1982, Borbély et al., 2016]. There is an homeostatic process (S) which increases when awake and decreases while asleep and there is a circadian process (C) which determine the time of the day more suitable to sleep, based on internal oscillator in the suprachiasmatic nucleus (SCN). Process S can be measured with specific frequency of activity during the sleep and wake states. Process C can be measured through core body temperature and melatonin levels.

The circadian rhythms are maintained in the cells by transcriptional activation-repression loops of negative feedback [Franken and Dijk, 2009]. For example, a protein heterodimer of Circadian Locomotor Output Cycles Kaput (CLOCK) and Brain and Muscle Arnt-Like 1 (BMAL1) activates the transcription of Period Circadian Regulator 2 (Per2). The PER2 protein then repress the transcription of Clock and Bmal1 gene, which reduces the amount of CLOCK and BMAL1 proteins available and stops the transcription of Period Circadian Regulator 2 (Per2). Multiple nested loops are forming this system and the genes are called clock genes. The SCN is responsible to coordinate the oscillation in the organs, but not to drive them.

It is difficult to separate circadian and sleep homeostasis at the molecular level, as mutations in some clock genes (Clock, Bmal1, Npas2, Per1, Per2, Cry1, Cry2) have an impact on the homeostatic response [Franken, 2013]. Homer1a gene expression correlates with process S in mouse [Schulz, 2022], a point mutation in the mouse gene Cacna1a reduces wakefulness by about one hour per day [Jan et al., 2020], and mutations in a few genes were found to cause sleep disorders such as hypocretin (orexin) cause narcolepsy in dogs [Schulz, 2022] but these remain anecdotal to explain much of the heritability in the sleep characteristics.

The mouse is an experimental model of choice for the study of sleep. The sleep characteristics in different strains have already been described extensively [Franken et al., 1998, Franken et al., 1999]. As for humans, the gold standard to measure sleep is electroencephalography (EEG) coupled with electromyography (EMG) which provide a high temporal resolution [Schulz, 2022] (Figure 1.1A,B). There are 2 sleep stages for mouse: the rapid-eye-movement sleep (REM) sleep and the non-REM (NREM) sleep. The REM sleep is also called paradoxical sleep because of the brain activity similar to the awake state while the body is not active, except for characteristic eye movements. Brain activity during NREM sleep is characterized by slower amplitude waves: delta (1-4.25 Hz), which is thought to be indicative of sleep pressure. Recent work has shown however that this binning groups heterogeneous waves that could more accurately be divided into slow delta or $\delta1$ (1-2.25 Hz) and fast delta or $\delta2$ (2.25-4 Hz) [Hubbard et al., 2020]. From the brain and muscle activity is derived hundred of quantitative sleep phenotypes describing the sleep-wake duration, distribution, and architecture (Figure 1.1C). To challenge the sleep homeostat, the perturbation is often a sleep deprivation (SD), done by gentle handling to avoid unnecessary increase of stress. It has to be noted though that the deprivation of sleep is in itself a cause of stress. The effect of sleep deprivation is however not only caused by stress [Mongrain et al., 2010]. Further precision on the recording of sleep in mouse can be found in the protocol [Mang and Franken, 2012].

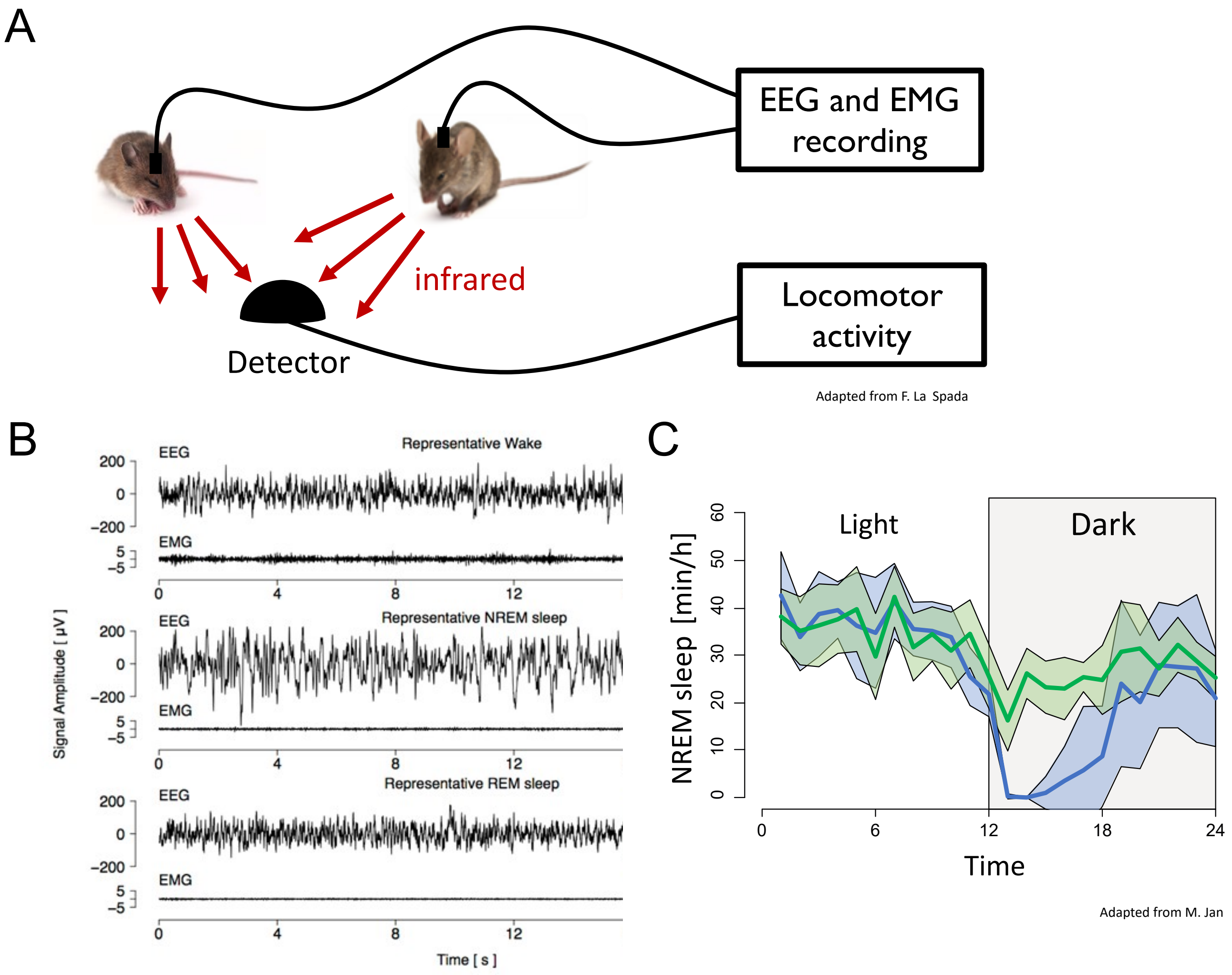


Figure 1.1: Recording sleep in mouse

A) Recording devices B) Typical EEG/EMG signal in the mouse C) Example of quantitative sleep phenotypes for two different BXD lines.

Sleep combines neuronal and molecular interactions and is a complex trait involving environmental and genetic factors [O'Callaghan et al., 2019]. The sleep phenotypes recorded in a strictly controlled environment show a variability greatly heritable and the parental lines do not necessarily have more extreme phenotypes than their crosses (Figure 1.2). Ranking phenotypes values by strain/line is done elsewhere and identification of extreme value help to figure out one BXD actually had a mutation that makes it now a substrain [Cook et al., 2006].

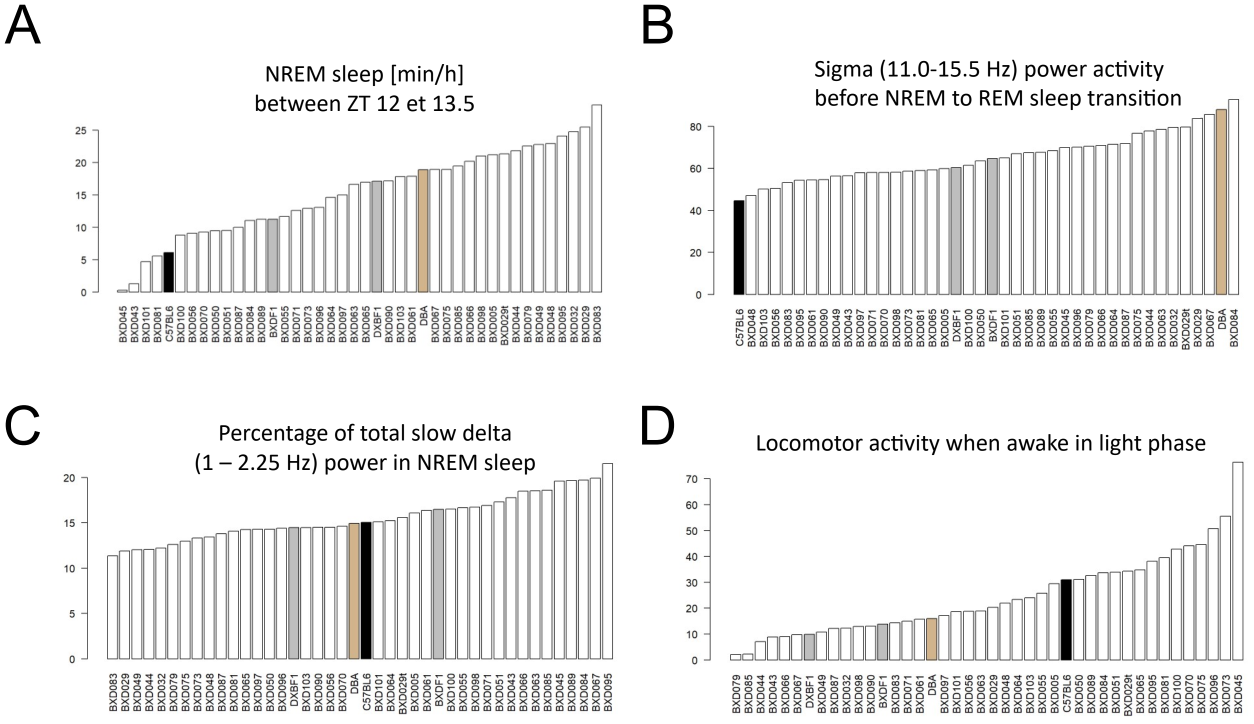


Figure 1.2: Sleep phenotypes ranked by mouse lines
Four examples of sleep phenotypes ranked by strain/BXD line with different patterns of whether the parental strains have more extreme phenotypes than the crosses. ZT: Zeitgeber time (here ZT0 is when the light is switched on, ZT12 is when the light is switched off, on a 12 hours light-12 hours dark pattern), BXD: recombinant inbred derived from a cross between B6 and D2 strains, REM: rapid-eye-movement sleep, NREM: non-REM.

1.1.2 Reproducibility, methodology, and RNA mapping

Scientists produce always more experiments and publications over the years and recently more and more are coming to the realization that many of these are not actually ideal scientific productions because of their lack of reproducibility, which questions their validity and usefulness. We can make the distinction between two types of reproducibility: the first is that repeating the same experiment using the same methods would obtain the same results, the second that the code and data are available to redo the same analysis [noa, 2016]. [Freedman et al., 2015] argues that the first definition is valid for confirmatory analyses whereas exploratory analyses would only require the second one. The FAIR principles is an initiative to guide researchers to sharing their data [Wilkinson et al., 2016]. It can be used as a theoretical checklist to help to consider the different aspects. The main idea are that the data and meta-data should be: i) Findable: which means they are indexed, can be searched for with keywords and have a universal identifier for example a Digital Object Identifier (DOI) to refer to. ii) Accessible: The data can be retrieved either freely, either with an authorisation procedure if necessary because the data are sensitive or under a particular regulation. And that meta-data is available even if the data is not. iii) Interoperable: they are in a format and language broadly used and they follow standard vocabularies when existing. iv) Reusable: which has to do with legal license, to clarify who can use the data, whether or not it can be modified or used for commercial applications. This sets ideal guidelines for more transparency and sharing of the (intermediate) data and analyses, not just the final results. However, what limits the reusability of previous research is not always the access to the data or analyses but sometimes simply the access to the knowledge to full understand them, which argues in favor of more documentation and transmission during the entire process. It is important to avoid isolation since already having another person's look can prevent conscious or unconscious bias [noa, 2016] but also as it allows to not reinvent the wheel while still exploring further the world.

1.1.3 Multi-omics, Databases, and Network

With so many datasets available, the trend is to somehow merge the information for different biological layers or different methods with the idea that it will allow to strengthen the biological signal of interest (present across multiple layers and methods) while reducing the experimental noise (different between omics). This data-driven approach assumes the data tables can be summarized into a lower number of variables, as the Principal Component Analysis (PCA) does with one data table. There is a wide range of algorithms to perform multi-omics dimensionality reduction [Dubin et al., 2016, Cantini et al., 2021], which arise from different fields having developed, sometimes in parallel, their methods. As a result, different vocabularies are found for the same concepts (Table 1.1). Notably, multi-omics factor analysis (MOFA) [Argelaguet et al., 2018, Argelaguet et al., 2020] is modelling in a probabilistic Bayesian framework and allows to retrieve non-linear patterns, which most other methods would miss.

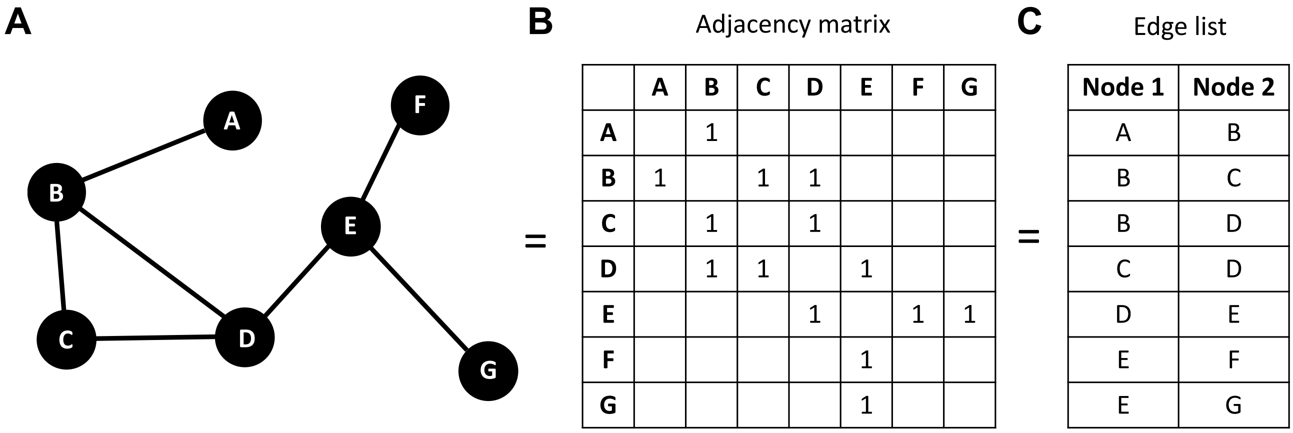
The data-driven approach needs eventually to compare the particular information extracted with the scientific literature and knowledge. A part of this knowledge is stored in formalized

Term in MOFA	Term in CCSWA	Meaning
View (samples as columns, features as rows)	Block (samples as rows, features as columns)	Data matrix or table with features values (genes, metabolites, phenotypes, genotypes, . . .) for the samples
Latent Factor (LF)	Common dimension (dim)	Multi-block equivalent of principal component in a PCA
Loadings/weights	Saliences	Contribution of each feature to a LF or dim (no equivalent in PCA)

Table 1.1: Comparison of multi-omics dimensionality reduction terminology. The nomenclature differs between different methods to reduce the dimensionality of multi-blocks data. MOFA: multi-omics factor analysis, CCSWA: Common Components and Specific Weights Analysis

form of databases. PubTator stores publication-based text occurrences of different biological elements [Wei et al., 2019], Rhea stores reactions [Bansal et al., 2022], STRING stores protein-protein known and predicted interactions [von Mering et al., 2005, Szklarczyk et al., 2019]. Different databases were built for different purposes, and because of that they focus on different types of interactions: atomic, molecular, complex, cell [Xenarios and Eisenberg, 2001]. Some are specific to a species [Kim et al., 2016], tissue, or condition [Sügis et al., 2019] whereas some aim to be generalist like Gene Ontology (GO) [Ashburner et al., 2000, Blake, 2013]. Even when focus on the same biological objects, databases tend to differ in how they will identify each object, and some databases are agglomerations of multiple other (primary) databases which have been processed to form coherent ensembles [Hermjakob et al., 2004, Orchard et al., 2013, Türei et al., 2021]. Mapping identifiers to make entries somehow comparable is an unavoidable but rarely perfect process where information is loss or incorrectly transmitted, whether done manually (curated), automatically, or a mix of both [Krassowski et al., 2020].

If interactions are often stored as list of entries for further analysis, for visualisation they may be under another form (Figure 1.3). Graphs and networks are synonyms (except when graph is used to describe a visual representation) where the nodes (or vertices) are the objects and the edges are the interactions between the objects. The ball-and-stick (Figure 1.3A) type of representation is the most commonly used for visualization with many variations to show different aspects of the state of the interactors (node properties) and the characteristics of the interactions (edges properties) [Shannon et al., 2003, Marai et al., 2019]. There are many different possibilities to spatially place the nodes in the 2D (or 3D) space and different layouts are algorithms that will prioritize different aspect in the choice of where the nodes are displayed.



	A	B	C	D	E	F	G
A		1					
B	1		1	1			
C		1		1			
D		1	1		1		
E				1		1	1
F					1		
G					1		

Node 1	Node 2
A	B
B	C
B	D
C	D
D	E
E	F
E	G

Figure 1.3: Different network representations
A) Ball-and-stick model B) Adjacency matrix C) Edge list.

A few basic metrics can help characterise the network and its components [Koutrouli et al., 2020]. The degree of a node is the number of its connections. The notion of shortest path between two nodes (if a path exists) is also important, and allows for example to define the diameter of the network: generally the network is considered to be the longest of all shortest paths. The density is the ratio between the existing edges and the edges that could exist considering the number of nodes, it can be calculated for the complete network or for a subpart (subgraph) of it. From these few measures, networks across different fields have been put in a few categories. The random networks have a small diameter but are poorly clustered, the regular lattices have a large diameter and are highly clustered, and the small-world networks are intermediate with a small diameter but highly clustered [Watts and Strogatz, 1998]. Scale-free networks are often encountered for biological processes, which is thought to be because that structure is more adaptable to environment and robust to change [Jeong et al., 2000]. For example, a mutation in a protein may suppress its interaction with another protein. A computational way to address robustness is by removing or adding edges from prior knowledge [Badia-i Mompel et al., 2022].

Graph Neural Networks (GNN) are neural networks that use graphs as input for their model

[Sanchez-Lengeling et al., 2021]. The goal is to predict information, which can work well in cases with very large number of input data and can perform well even with relatively simple architecture [Xu* et al., 2018]. However the disadvantage is that the predictions are seen as coming from a black box, whereas often the biological questions require to interpret and explain what the model is. Some advances are made to make neuronal networks more interpretable but progress is still needed to overcome this limitation [Ying et al., 2019].

1.2 Objectives

Sleep deprivation causes short-term discomfort and is associated with many long-term health problems. Sleep regulation has important genetic and environmental factors, but many aspects remain poorly understood. To uncover the molecular pathways underlying sleep regulation, The Franken group has assembled a systems genetics resource interrogating the BXD mouse panel [Diessler et al., 2018]. The genotypes and sleep-wake phenome were characterized, along with intermediate phenotypes: the transcriptome in brain and in liver, and the targeted metabolome in the blood plasma. My role in this project is to inherit this rich multi-omics BXD dataset to expand the current understanding of sleep regulation by computationally investigating and developing analytical methods for data and knowledge integration. My objectives for my thesis are:

- The assessment of the reproducibility and robustness of previous computational analyses on this dataset. (Chapter 2)
- The contribution to continuous documentation of the project through organisation of the data and metadata, presentations at lab meetings, redaction of scientific literature. (Chapter 2)
- The assessment of parental reference bias in RNA-seq. (Chapter 3)
- The implementation of solutions for better references for the RNA-seq of BXD lines. (Chapter 3)

- The building of a knowledge graph through the mining and the assimilation of publicly available databases, and the mapping of various identifiers. (Chapter 4)
- The integration of the data and knowledge parts for the identification of multi-gene regulation subnetworks of the sleep phenotypes. (Chapter 4)

1.3 Side projects

During my time as a PhD candidate I also participate in the following side projects.

- Ongoing project on the molecular time-dependent effects of sleep deprivation on transcription factor BMAL1 (ChIP seq).

  This project aims to see the impact of sleep deprivation on BMAL1 [Mongrain et al., 2011] on the entire genome and at different time points [Rey et al., 2011]. I contributed to data analysis of preliminary sequencing tests to optimize the protocol and experimental design.
- Published review on structural variant calling [Mahmoud et al., 2019]. Included in this thesis as Annex 1. I contributed to articles mining, table/figure preparation, and manuscript drafting and revisions. I also wrote a short blog post for outreach to a larger public.

Chapter 2

Reproducibility

Many research groups are working around the world and generating an always increasing amount of data and analyses. Is this work resusable? The goal of this chapter is to assess the reproducibility and robustness of previous computational analyses on the BXD dataset produced in the Franken lab and participate to the continuous documentation of the project through organisation of the data and metadata.

2.1 Results summary

In this paper, we are presenting a real-world concrete example of the data and bioinformatic analysis management. We give insights about multi-omics challenges and solutions we used to help the work to be reused internally or externally. This includes more details on the quality check and validations of the methods, the use of Rmarkdown reports for more higher levels parts of the analyses, a metadata workflow document illustrating and referencing the different code and data files, and a web site for exploration of the results. We demonstrated assessment of the robustness of the results through the change of version of mouse genome reference assembly used (from mm9 to mm10).

2.2 My contribution

In this research article, I have reproduced previous analyses and updated the mouse genome reference from mm9 to mm10 in collaboration with Maxime Jan. I participated in data and metadata organization. I produced the figures 2 and 5 and helped with manuscript writing.

2.3 Publication

This article was published in a peer-reviewed journal [Jan et al., 2019]. For outreach, I did a short SIB *in silico* talk available as a video on youtube. The publication is included below.

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

A multi-omics digital research object for the genetics of sleep regulation

Received: 5 March 2019
Accepted: 26 July 2019
Published online: 31 October 2019

Maxime Jan[1], Nastassia Gobet[1,2], Shanaz Diessler[1], Paul Franken[1] & Ioannis Xenarios[3,4]

With the aim to uncover the molecular pathways underlying the regulation of sleep, we recently assembled an extensive and comprehensive systems genetics dataset interrogating a genetic reference population of mice at the levels of the genome, the brain and liver transcriptomes, the plasma metabolome, and the sleep-wake phenome. To facilitate a meaningful and efficient re-use of this public resource by others we designed, describe in detail, and made available a Digital Research Object (DRO), embedding data, documentation, and analytics. We present and discuss both the advantages and limitations of our multi-modal resource and analytic pipeline. The reproducibility of the results was tested by a bioinformatician not implicated in the original project and the robustness of results was assessed by re-annotating genetic and transcriptome data from the mm9 to the mm10 mouse genome assembly.

Background & Summary

A good night's sleep is essential for optimal performance, wellbeing and health. Chronically disturbed or curtailed sleep can have long-lasting adverse effects on health with associated increased risk for obesity and type-2 diabetes[1].

To gain insight into the molecular signaling pathways regulating undisturbed sleep and the response to sleep restriction in the mouse, we performed a population-based multi-level screening known as *systems genetics*[2]. This approach allows to chart the molecular pathways connecting genetic variants to complex traits through the integration of multiple *omics datasets such as transcriptomics, proteomics, metabolomics or microbiomes[3].

We built a systems genetics resource based on the BXD panel, a population of recombinant inbred lines of mice[4], that has been used for a number of complex traits and *omics screening such as brain slow-waves during NREM sleep[5], glucose regulation[6], cognitive aging[7] and mitochondria proteomics[8].

We phenotyped 34 BXD/RwwJ inbred lines, 4 BXD/TyJ, 2 parental strains C57BL6/J and DBA/2 J and their reciprocal F1 offspring. Mice of these 42 lines were challenged with 6 h of sleep deprivation (SD) to evaluate the effects of insufficient sleep on sleep-wake behavior and brain activity (electroencephalogram or EEG; Fig. 1, Experiment 1) and, on gene expression and metabolites (Fig. 1, Experiment 2). For Experiment 1 we recorded the EEG together with muscle tone (electromyogram or EMG) and locomotor activity (LMA) continuously for 4 days. Based on the EEG/EMG signals we determined sleep-wake state [wakefulness, rapid-eye movement (REM) sleep, and non-REM (NREM) sleep] as well as the spectral composition of the EEG signal as end phenotypes. For Experiment 2 we quantified mRNA levels in cerebral cortex and liver using illumina RNA-sequencing and performed a targeted metabolomics screen on blood using Biocrates p180 liquid chromatography (LC-) and Flow injection analysis (FIA-) coupled with mass spectrometry (MS). These transcriptome and metabolome data are regarded as intermediate phenotypes linking genome information to the sleep-wake related end phenotypes.

The keystone of systems genetics is data integration. Accordingly, the scientific community can benefit from data sharing strategies that facilitate the integration of datasets among research groups. However, reliable methods for data integration are needed and require a broad range of expertise such as in mathematical and statistical models[9], computational methods[10], visualization strategies[11], and deep understanding of complex phenotypes. Therefore, data sharing should not be limited to the dataset *per se* but also to analytics in the form

[1]Centre for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [2]Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Ludwig Cancer Research/CHUV-UNIL, Lausanne, Switzerland. [4]Health 2030 Genome Center, Geneva, Switzerland. These authors contributed equally: Maxime Jan and Nastassia Gobet. Correspondence and requests for materials should be addressed to I.X. (email: ioannis.xenarios@unil.ch)

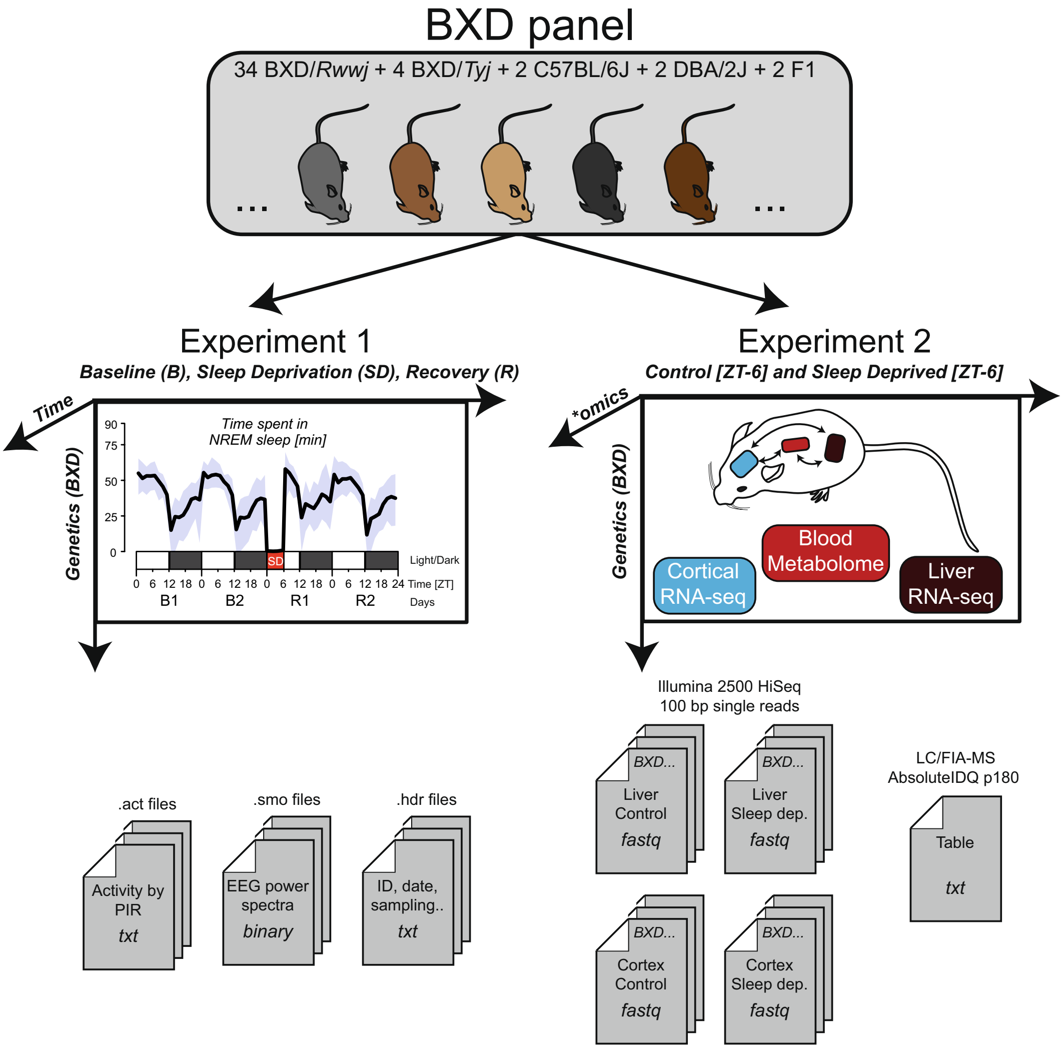


Fig. 1 Data generation. The behavioral/EEG end-phenotypes of the BXD mouse panel were quantified in Experiment 1. Mice were recorded for 4 days: 2 days of baseline (B1 & B2), followed by 6 h of sleep deprivation (SD) and 2 days of recovery (R1 & R2). EEG spectral composition was written in *.smo* files, activity in *.act* files and meta-data in *.hdr* files. Blood metabolomics, liver transcriptomics and cortical transcriptomics were quantified in Experiment 2. 'Control' and 'Sleep deprived' batches were sampled at a single time point: ZT6 (i.e. directly after sleep deprivation for the 'sleep deprived' batch). Transcriptomics was performed on pooled sampled per BXD strains. For blood metabolomics, metabolite quantification was performed for each BXD replicates. Adapted from[2].

of analysis workflows, code, interpretation of results, and meta-data[12]. The concept of a Digital Research Object (DRO) was proposed to group dataset and analytics into one united package[13]. Various guidelines have been suggested to address the challenges of sharing such DRO with the goal to improve and promote the human and computer knowledge sharing, like the FAIR (Findable, Accessible, Interoperable, Reusable) principles proposed by FORCE 11[14] or by the DB2K (Big Data to Knowledge) framework. These guidelines concern biomedical workflow, meta-data structures and computer infrastructures facilitating the reusability and interoperability of digital resources[15]. Although such guidelines are often described and applied in the context of single data-type assays, they can be challenging to achieve for trans-disciplinary research projects such as systems genetics, in which multiple data types, computer programs, references and novel methodologies need to be combined[16]. Moreover, applying these principles can also be discouraging because of the time required for new working routines to become fully reproducible[17] and because only few biomedical journals have standardized and explicit data-sharing[18] or reproducibility[19] policies. Nonetheless, DROs are essential for scientific reliability[20], and can save time if a dataset or methods specific to a study need to be reused or improved by different users such as colleagues at other institutes, new comers to the lab, or at long-term yourself.

We here complement our previous publication[2] by improving the raw and processed data availability. We describe in more details the different bioinformatics steps that were applied to analyze this resource and improve the analytical pipeline reproducibility by generating *R* reports and provide code. Finally, we assess the reproducibility of our bioinformatic pipeline from the perspective of a new student in bioinformatics that recently joined the group, and the robustness of the results by changing both the mouse reference genome and the RNA-seq reads alignment to new standards.

Methods

The methods detailed below are an expanded version of the methods described in our related paper[2]. Appreciable portions are reproduced verbatim to deliver a complete description of the data and analytics with the aim to enhance reproducibility.

Experiment 1 and Experiment 2 (Fig. 1) were approved by the veterinary authorities of the state of Vaud, Switzerland (SCAV authorization #2534).

Animals, breeding, and housing conditions. 34 BXD lines originating from the University of Tennessee Health Science Center (Memphis, TN, United States of America) were selected for Experiment 1 and Experiment 2. These lines were randomly chosen from the newly generated advanced recombinant inbred line (ARIL) RwwJ panel[4], although lines with documented poor breeding performance were not considered. 4 additional BXD RI strains were chosen from the older *TyJ* panel for reproducibility purposes and were obtained directly from the Jackson Laboratory (JAX, Bar Harbor, Maine). The names used for some of the BXD lines have been modified over time to reflect genetic proximity. Online-only Table 1 lists the BXD line names we used in our files alongside the corresponding current JAX names and IDs. In our analyses, we discarded the BXD63/RwwJ line for quality reasons (see Technical Validation) as well as the 4 older BXD strains that were derived from a different DBA/2 sub-strain, i.e. DBA/2Rj instead of DBA/2 J for RwwJ lines[21]. The methods below describe the remaining 33 BXD lines, F1 and parental strains.

Two breeding trios per BXD strain were purchased from a local facility (EPFL-SV, Lausanne, Switzerland) and bred in-house until sufficient offspring was obtained. The parental strains DBA/2 J (D2), C57BL6/J (B6) and their reciprocal F1 offspring (B6D2F1 [BD-F1] and D2B6F1 [DB-F1]) were bred and phenotyped alongside. Suitable (age and sex) offspring was transferred to our sleep-recording facility, where they were singly housed, with food and water available *ad libitum*, at a constant temperature of 25 °C and under a 12 h light/12 h dark cycle (LD12:12, fluorescent lights, intensity 6.6 cds/m^2, with Zeitgeber time 0 (ZT0) and ZT12 designating light and dark onset, respectively). Male mice aged 11–14 week at the time of experiment were used for phenotyping, with a mean of 12 animals per BXD line among all experiments. Note that 3 BXD lines had a lower replicate number (n), with respectively BXD79 ($n = 6$), BXD85 ($n = 5$), and BXD101 ($n = 4$) because of poor breeding success. For the remaining 30 BXD lines, replicates were distributed as follows: for EEG/behavioral phenotyping (Experiment 1 in Fig. 1; mean = 6.2/line; $5 \leq n \leq 7$) and for molecular phenotyping (Experiment 2 in Fig. 1; mean = 6.8/line; $6 \leq n \leq 9$). Additionally, to control for the reproducibility of the outcome variables over the course of the experiment, parental lines were phenotyped twice—i.e., at the start (labeled in files as B61 and DB1) and end (labeled B62 and DB2) of the breeding and data-collecting phase, which spanned 2 years (March 2012–December 2013). To summarize, distributed over 32 experimental cohorts, 227 individual mice were used for behavioral/EEG phenotyping (Experiment 1) and 263 mice for tissue collection for transcriptome and metabolome analyses (Experiment 2), the latter being divided into sleep deprived (SD) and controls ("Ctr"; see Study design section below). We put in an effort to distribute the lines across the experimental cohorts so that biological replicates of 1 line were collected/recorded on more than 1 occasion while also ensuring that an even number of mice per line was included for tissue collection so as to pair SD and "Ctr" individuals within each cohort (for behavioral/EEG phenotyping, each mouse serves as its own control).

Study design and sleep deprivation. The study consisted of 2 experiments, i.e., Experiments 1 and 2 (Fig. 1). Animals of both experiments were maintained under the same housing conditions. Animals in Experiment 1 underwent surgery and, after a > 10 days recovery period, electroencephalography (EEG), electromyography (EMG) and locomotor activity (LMA) were recorded continuously for a 4-day period starting at ZT0. The first 2 days were considered Baseline (B1 and B2). The first 6 hours of Day 3 (ZT0–6), animals were sleep deprived (SD) in their home cage by "gentle handling" referring to preventing sleep by changing litter, introducing paper tissue, presenting a pipet near the animal, or gently tapping the cage. Experimenters performing the SD rotated every 1 or 2 hours (for more information, see[22]). The remaining 18 h of Day 3 and the entire Day 4 were considered Recovery (R1 and R2).

Half of the animals included in Experiment 2 underwent SD alongside the animals of Experiment 1. The other half was left undisturbed in another room (i.e., control or Ctr, also referred as Non Sleep Deprived or NSD). Both SD and "Ctr" mice of Experiment 2 were sacrificed at ZT6 (i.e., immediately after the end of the SD) for sampling of liver and cerebral cortex tissue as well as trunk blood. All mice were left undisturbed for at least 2 days prior to SD.

Experiment 1: EEG/EMG and LMA recording and signal pre-processing. EEG/EMG surgery was performed under deep anesthesia. IP injection of Xylazine/Ketamine mixture (91/14.5 mg/kg, respectively) ensures a deep plane of anesthesia for the duration of the surgery (i.e., around 30 min). Analgesia was provided the evening prior and the 3 days after surgery with Dafalgan in the drinking water (200–300 mg/kg). Six holes were drilled into the cranium, 4 for screws to fix the connector with Adhesive Resin Cement, 2 for EEG electrodes. The caudal electrode was placed over the hippocampal structure and the rostral electrode was placed over the frontal cerebral cortex. Two gold-wire electrodes were inserted into the neck muscle for EMG recording (for details, see[22]). Mice were allowed to recover for at least 10 days prior to baseline recordings. EEG and EMG signals were amplified, filtered, digitized, and stored using EMBLA (Medcare Flaga, Thornton, CO, USA) hardware (A10 recorder) and software (Somnologica). Digitalization of the signal was performed as followed: the analog-to-digital conversion of the signal was performed at a rate of 2000 Hz, the signal was down sampled at 200 Hz, high-pass filter at 0.0625 Hz was applied to reject DC offset of the signal and a 50-Hz notch filter applied to reduce line artefacts. Signals were transformed by Discrete Fourier Transform (DFT) to yield power spectra between 0 and 100 Hz with a 0.25 frequency resolution using a 4-seconds time resolution (referred to as a 4 s "epoch"). EEG frequency bins with artefacts of known (line artefacts between 45–55 Hz) and unknown (75–77 Hz) source were removed from the average EEG spectra of all mice. Other specific 0.25 Hz bins containing artefacts (notably the 8.0, 16.0 and 32.0 Hz bins) of unknown source, were removed from individual mice based on the visual inspection of individual EEG spectra in each of the three sleep-wake states (i.e. wakefulness, REM sleep and NREM sleep). Power density in frequency bins deemed artefacted were estimated by linear interpolation. For details, see Pascal scripts in https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation.

LMA was recorded by passive infrared (PIR) sensors (Visonic, Tel Aviv, Israel) at 1-min resolution for the duration of the 4-day experiment, using ClockLab (ActiMetrics, IL, USA). Activity data were made available as *.act* files at Figshare[23].

Offline, the sleep-wake states wakefulness, REM sleep, and NREM sleep were annotated on consecutive 4-second epochs, based on the EEG and EMG pattern (see Sleep-wake state annotation section). EEG/EMG power spectra and sleep-wake state annotation were made available as binary (*.smo*) files at Figshare[23].

Experiment 2: Tissue collection and preparation. Mice were sacrificed by decapitation after being anesthetized with isoflurane, and blood, cerebral cortex, and liver were collected immediately. The whole procedure took no more than 5 min per mouse. Blood was collected at the decapitation site into tubes containing 10 ml heparin (2 U/μl) and centrifuged at 4000 rpm during 5 min at 4 °C. Plasma was collected by pipetting, flash-frozen in liquid nitrogen, and stored at −80 °C until further use. Cortex and liver were flash-frozen in liquid nitrogen immediately after dissection and were stored at −140 °C until further use.

RNA extraction and pooling. For RNA extraction, frozen samples were homogenized for 45 seconds in 1 ml of QIAzol Lysis Reagent (Qiagen; Hilden, Germany) in a gentleMACS M tube using the gentleMACS Dissociator (Miltenyi Biotec; Bergisch Gladbach, Germany). Homogenates were stored at −80 °C until RNA extraction. Total RNA was isolated and purified from cortex using the automated nucleic acid extraction system QIAcube (Qiagen; Hilden, Germany) with the RNeasy Plus Universal Tissue mini kit (Qiagen; Hilden, Germany) and were treated with DNAse. Total RNA from liver was isolated and purified manually using the Qiagen RNeasy Plus mini kit (Qiagen; Hilden, Germany), which includes a step for effective elimination of genomic DNA. RNA quantity, quality, and integrity were assessed utilizing the NanoDrop ND-1000 spectrophotometer (Thermo scientific; Waltham, Massachusetts, USA) and the Fragment Analyzer (Advanced Analytical). The 263 mice initially sacrificed for tissue collection yielded 222 cortex and 222 liver samples of good quality.

Equal amounts of RNA from biological replicates (3 samples per strain, tissue, and experimental condition, except for BXD79, BXD85, and BXD101; see above under Animals, breeding, and housing conditions) were pooled, yielding 156 samples for library preparation. RNA-seq libraries were prepared from 500 ng of pooled RNA using the Illumina TruSeq Stranded mRNA reagents (Illumina; San Diego, California, USA) on a Caliper Sciclone liquid handling robot (PerkinElmer; Waltham, Massachusetts, USA).

RNA sequencing. Libraries were sequenced on the Illumina HiSeq. 2500 using HiSeq SBS Kit v3 reagents, with cluster generation using the Illumina HiSeq PE Cluster Kit v3 reagents. A mean of 41 M 100 bp single-end reads were obtained (29 M $\leq$ n $\leq$ 63 M). Quality of sequences were evaluated using FastQC software (version 0.10.1) and reports made available here https://bxd.vital-it.ch/#/dataset/1. Figure 2 (a, b, c and d) shows the median Phred quality score per base among all samples reads for 'Cortex Control', 'Cortex SD', 'Liver Control' and 'Liver SD' respectively. Fastq files were made available at NCBI Gene Expression Omnibus[24].

Targeted LC-MS metabolomics. Targeted metabolomics analysis was performed using flow injection analysis (FIA) and liquid chromatography/mass spectrometry (LC/MS) as described in[25,26]. To identify metabolites and measure their concentrations, plasma samples were analyzed using the AbsoluteIDQ p180 targeted metabolomics kit (Biocrates Life Sciences AG, Innsbruck, Austria) and a Waters Xevo TQ-S mass spectrometer coupled to an Acquity UPLC liquid chromatography system (Waters Corporation, Milford, MA, USA). The kit provided absolute concentrations for 188 endogenous compounds from 6 different classes, namely acyl carnitines, amino acids, biogenic amines, hexoses, glycerophospholipids, and sphingolipids. Plasma samples were prepared according to the manufacturer's instructions. Sample order was randomized, and 3 levels of quality controls (QCs) were run on each 96-well plate. Data were normalized between batches, using the results of quality control level 2 (QC2) repeats across the plate (n = 4) and between plates (n = 4) using Biocrates METIDQ software (QC2 correction). Metabolites below the lower limit of quantification or the limit of detection, as well as above the upper limit of quantification, or with standards out of limits, were discarded from the analysis[26]. Out of the 188 metabolites assayed, 124 passed these criteria across samples and were used in subsequent analyses. No hexoses were present among the 124 metabolites. Out of the 256 mice sacrificed for tissue collection, 249 plasma samples were used for this analysis. An average of 3.5 animals (3 $\leq$ n $\leq$ 6) per line and experimental condition were used (except for BXD79, BXD85, and BXD101 with respectively 2, 1, and 1 animal/condition used; see above under Animals, breeding, and housing conditions). Note that in contrast to the RNA-seq experiment, samples were not pooled but analyzed individually. Mean metabolite levels per BXD line were made available at https://bxd.vital-it.ch/#/dataset/1 for details see intermediate files[27].

Corticosterone quantification. In the same plasma samples, we determined corticosterone levels using an enzyme immunoassay (corticosterone EIA kit; Enzo Life Sciences, Lausanne, Switzerland) according to the manufacturer's instructions. All samples were diluted 40 times in the provided buffer, kept on ice during the

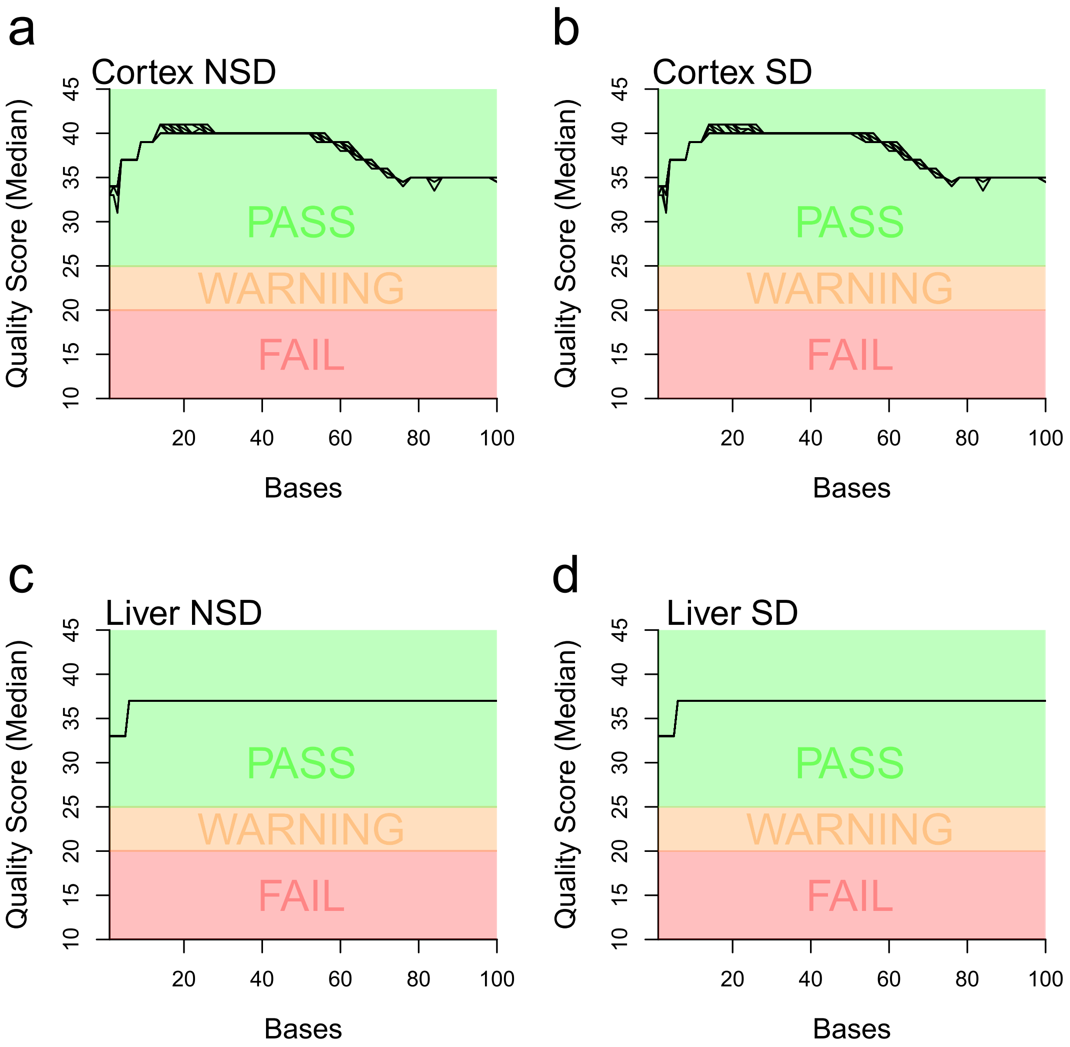


Fig. 2 Median PHRED read quality per base for BXD RNA-sequencing. PHRED quality score based on illumina 1.9. (**a**) Samples from Cortex during control (NSD). (**b**) Samples from Cortex after sleep deprivation (SD). (**c**) Samples from Liver during control (NSD). (**d**) Samples from Liver after sleep deprivation (SD). Median score was computed using MultiQC[69].

manipulation, and tested in duplicate. BXD lines were spread over multiple 96-well plates in an attempt to control for possible batch effects. In addition, a "control" sample was prepared by pooling plasma from 5 C57BL6/J mice. Aliquots of this control were measured along with each plate to assess plate-to-plate variability. The concentration was calculated in pg/ml based on the average net optical density (at $\lambda = 405$ nm) for each standard and sample.

Corticosterone level were made available on Figshare[27].

Bioinformatics pipeline. To facilitate the interpretation of the complete bioinformatic workflow that was performed on this dataset, we here describe first our general strategy to construct an analytics pipeline with which we hope to improve reproducibility (Fig. 3). This strategy has some similarities with the recently published tool Qresp[28] that facilitates the visualization of paper workflow. We then describe the specific methods used to analyze this dataset.

The analytics and input datasets were separated into 3 layers according to an increasing level of data abstraction (Fig. 3). This hierarchical structure of the workflow was particularly useful to identify steps downstream novel versions of a script or data (e.g. Figure 3, red) and simplify workflow description. The first *low-level* layer contains the procedures needed to reduce and transform the raw-data (i.e. RNA-seq reads, EEG/EMG signals) into an exploitable signal such as sleep phenotypes, gene expression, or mice genotypes by further analytical steps. This layer is characterized by long and computationally intensive procedures which required the expertise of different persons, each with their own working environment and preferred informatics language.

The *intermediate-level* layer contains some established analyses that could be performed on the data such as gene expression normalization followed by differential expression or Quantitative Trait Locus (QTL) mapping. With the scripts of this layer we explored the effects of sleep deprivation, genetic variations, as well as their interaction on EEG/behavioral phenotypes and intermediate phenotypes.

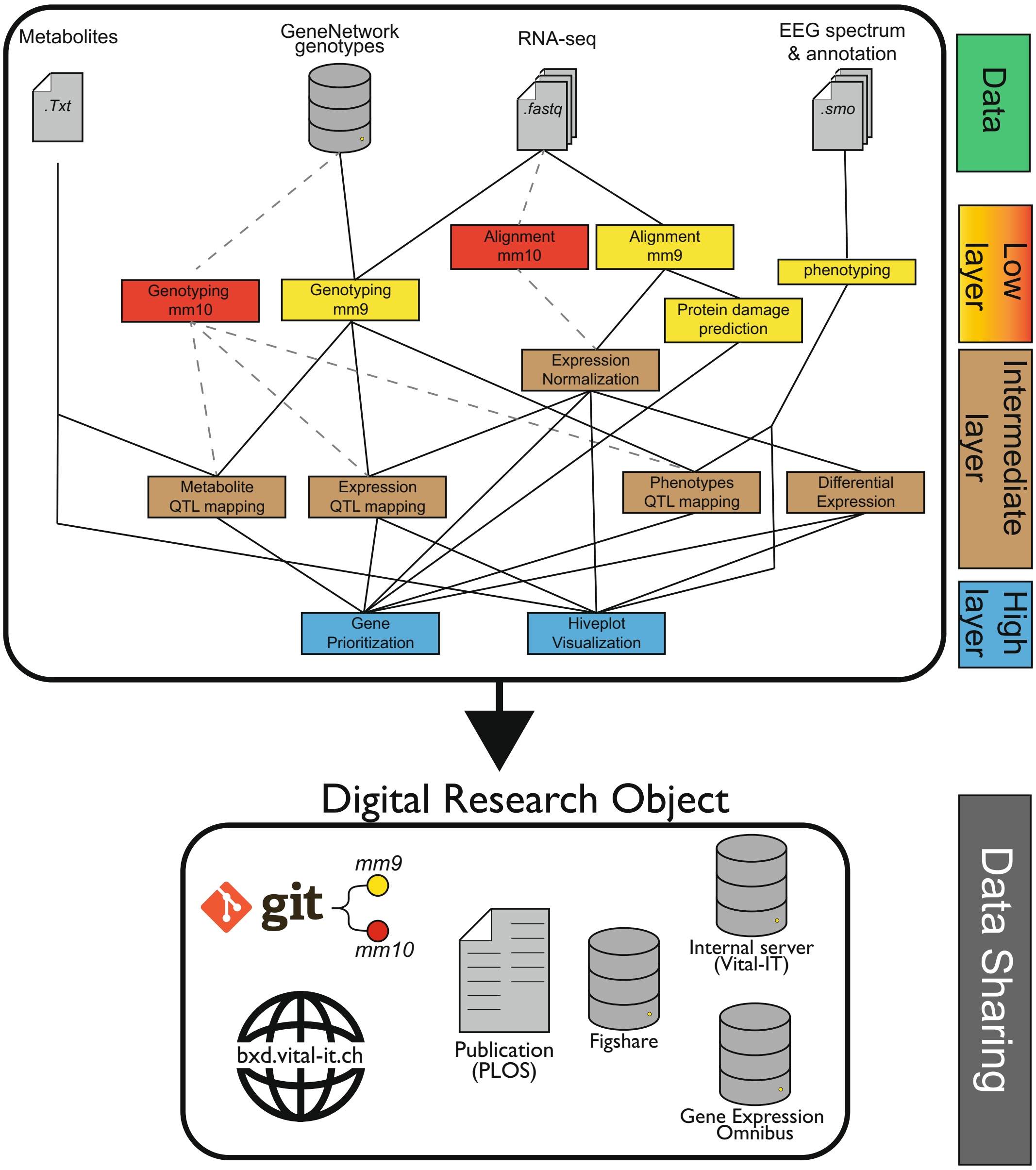


Fig. 3 Summary of the bioinformatic analytical pipeline. Representation of the main bioinformatics methods used. Original analyses were performed using the mm9 mouse assembly (yellow). Results were also reproduced using the mm10 mouse assembly (red) and all downstream analyses. Layers represent the scripts organization on gitlab and available intermediate files.

The *high-level* layer contains the novel integrative methods that we developed to prioritize genes driving sleep regulation and to visually represent the meta-dimensional multi-omics networks underlying sleep phenotypes.

Standard and non-standard semantics. To improve the reproducibility and reusability of our workflow, we tried to prioritize standard semantics and established pipelines when applicable, such as the RNA-seq processing by STAR and htseq-count[29]. The use of curated symbols for genes nomenclature by RefSeq allowed a better semantic interoperability with other resources such as Uniprot protein ID using solutions like biomaRt[30]. We provided some of the references files used in these scripts, like the RefSeq.*gtf* reference file (see the Exome/RefSeq_20140129.gtf file in the DataSystemsGeneticsOfSleep_mm9.tar.gz file[27], this file comes from UCSC table browser and was generated using RefSeq Reflat database on the 2014/01/29).

These annotations can be updated and possibly change the gene quantification with updated version or different genome reference.

However, the EEG/behavioral phenotyping procedure could not be performed by any standard computational workflow or common semantics as none exist. The nomenclature that was chosen in this case to generate unique phenotypic ID was a combination of the phenotype observed (e.g. EEG power during NREM sleep) and the features observed in this phenotype (e.g. delta band 1–4 Hz). These phenotypes were also present as file name and column name in our dataset[27]. Nevertheless, we mapped our phenotypes to the Human Phenotype Ontology (HPO)[31] to help non-specialists to explore these traits and facilitate human-mouse data integration. These associations are not exact matches as most of the terms available in the HPO are disease oriented while our phenotypes should be considered as normal traits for inbred lines. The mapping can be found in the *General_Information.xlsx* file (https://bxd.vital-it.ch/#/dataset/1).

Favor *R* and *Rmarkdown* reports for reproducible results. After data processing within the *low-level* layer, the effect of sleep deprivation, genotype, and their interaction were measured using various statistical models and computational methods. We chose to prioritize the programming language *R* as it was the best suited tool for the statistical analyses and for the generation of figures. Beside the advantages of a license-free and portable language, *R* was already recommended as main tool for systems genetics analysis[32]. Many available packages were particularly adapted for the systems genetics design, involving phenotype-genotype association (*r/qtl*), network analysis (*WGCNA*, *SANTA*, *igraph*), differential expression (*EdgeR*, *DESeq*, *limma*), bayesian network learning (*bnlearn*), visualization (*ggplot2*, *grid*), enrichment (*topGO*, *topAnat*) and parallel computing (*parallel*). Only a few analyses were performed using other softwares, principally for efficiency reasons in cis-/trans-eQTL analysis where the number of models to test was quite large[33,34]. R is one of the flagships of open science and reproducibility[35] with a reviewable source code and the possibility of generating reports known as '*Rmarkdown*' with 2 packages: *knitr*[36] and *rmarkdown*[37]. This report format contains combination of code, figures, and comments within a single *markdown* document that can be easily converted into *pdf* or *html* format. Rmarkdown scripts were made available (https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation) and the reports in the form of.html document were made available together with the data[27]. To avoid the need to copy/paste some functions shared between *Rmarkdowns* but still display them in our reports, we used the *readLines()* function within Rmarkdown chunks. Finally, the use of the *sessionInfo()* function at the end of the document allowed to keep track of the packages versions and the environment variable used. Some of these Rmarkdown reports were generated on a remote cluster instead of the more traditional Rstudio environment, for more information on how to generate these Rmarkdown, see the Usage Notes.

Workflow documentation. This systems genetics approach was an integrative project that implicated multiple collaborators, that each contributed to the final results, with their own working habit related to their area of expertise. For better reproducibility of the generated files, a critical goal was to keep track of the different files created, associated documents or analytical steps that were produced. For example, EEG/behavioral phenotypes could be found within many files and reports, from *low-level* to *high-level* layers, but their nomenclatures were still hard to interpret as mentioned above, for those not directly related to this project. A newcomer in this project should be able to easily recover the metadata document containing all the physiological phenotypes information (i.e. understand that a metadata document was created and where to find it or who to ask for it) and understand which scripts were used to produce these phenotypes. To establish what was exactly performed, we generated a documentation file containing the essential information and relationships between all the files, scripts, Rmarkdown, small workflow or database used in this project. This document describes the inputs/outputs needed and where to locate the information distributed among different persons or different directories on a digital infrastructure as presented in Fig. 3 but with more details to improve the reproducibility of the DRO[38].

The markdown format was kept as it was easy to write/read by a human or to generate via a python script. This file was formatted into a simplified RDF-like triples structure, were each file-object (subject) was linked to information (object) by a property. This format allowed to use the following properties to describe each file-objects we had: The file-object name or identification, a brief description (i.e. about the software used or the data content), the file-object version, the input(s)/output(s), the associated documents, hyperlink(s) to remote database or citation, the location of the file-object on the project directory or archiving system, and the author(s) to contact for questions. These associations could be viewed as a graph to display the important files and pipelines used. This document was useful to understand how exactly the different files were generated, and to recover the scripts and input/output used, even after prolonged periods and to use them again, which permits for example, to reproduce data with novel or updated annotation files. Furthermore, if an error was detected within a script, the results and figures downstream that needed to be recomputed could be easily found. This documentation file (Documentation.html) was made available on gitlab (https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation).

Data mining website. The DRO built for this systems genetics resource is constituted of the following collection: raw-data, processed data, Rmarkdown reports, results & interpretation, workflow, scripts, and metadata. To improve the reproducibility of our integrative visualization method (see HivePlots below), we provided some data-mining tools, a server to store some intermediate results, and a web application[39,40]. The home page of the web application displays the information for the NREM sleep gain during the 24 hours (in four 6-hour intervals) after sleep deprivation. Three data-mining tutorials were described on the website the web interface to: (i) mine a single phenotype, (ii) search for a gene, and (iii) compare hiveplots. Currently, no centralized repository exists containing all types of phenotypic data that were extracted within this project. This web-interface can, however be viewed as a hub for this DRO that became findable and accessible with a web-browser. With this web resource, we provided an advanced interactive interface for EEG/behavioral end-phenotypes and their associated intermediate phenotypes (variants, metabolites, gene expression). Compared to other web-resources for systems genetics like GeneNetwork where the principal focus is QTL mining, this interface provides an integrative view of this one dataset, with also data files and link to code to reproduce some of our analyses in the form of Rmarkdown, like the prioritization strategy.

Low-level layer analyses. *Sleep-wake state annotation.* To assist the annotation of this extensive dataset (around 20 million 4 s epochs), we developed a semiautomated scoring system. The 4-day recordings of 43 mice (19% of all recordings), representing animals from 12 strains, were fully annotated visually by an expert according to established criteria[22]. Due to large between-line variability in EEG signals, even after normalization, a partial overlap of the different sleep-wake states remained, as evidenced by the absolute position of the center of each state cluster, which differed even among individuals of the same line (precluding the use of 1 "reference" mouse), even per line, to reliably annotate sleep-wake states for the others. To overcome this problem, 1 day out of 4 (i.e., Day 3 or R1, which includes the SD) was visually annotated for each mouse. These 4 seconds sleep-wake scores were used to train the semiautomatic scoring algorithm, which took as input 82 numerical variables derived from the analyses of EEG and EMG signals using frequency- (discrete Fourier transform [DFT]) and time-domain analyses performed at 1 second resolution. We then used these data to train a series of support vector machines (SVMs)[41] specifically tailored for each mouse, using combinations of the 5 or 6 most informative variables out of the 82 input variables. The best-performing SVMs for a given mouse were then selected based on the upper-quartile performance for global classification accuracy and sensitivity for REM sleep (the sleep-wake state with the lowest prevalence) and used to predict sleep-wake states in the remaining 3 days of the recording. The predictions for 4 consecutive 1-s epochs were converted into 1 four-second epoch. Next, the results of the distinct SVMs were collapsed into a consensus prediction, using a majority vote. In case of ties, epochs were annotated according to the consensus prediction of their neighboring epochs. To prevent overfitting and assess the expected performance of the predictor, only 50% of the R1 manually annotated data from each mouse were used for training (randomly selected). The classification performance was assessed by comparing the automatic and visual scoring of the fully manually annotated 4 d recordings of 43 mice. The global accuracy was computed using a confusion matrix[42] of the completely predicted days (B1, B2, and R2). For all subsequent analyses, the visually annotated Day 3 (R1) recording and the algorithmically annotated days (B1, B2, and R2) were used for all mice, including those for which these days were visually annotated. The resulting sleep-wake state annotation together with EEG power spectra and EMG levels were saved as binary files (.smo) with their corresponding metadata files (.hdr) and deposited at Figshare[23]. For more information on *.smo* and *.hdr* files, see Usage Notes.

EEG/Behavioral Phenotyping. We quantified 341 phenotypes based on the sleep-wake states, LMA, and the spectral composition of the EEG, constituting 3 broad phenotypic categories. For the first phenotypic category ("State"), the 96 hours sleep-wake sequence of each animal was used to directly assess traits in 3 "state"-related phenotypic subcategories: (i) duration (e.g., time spent in wakefulness, NREM sleep, and REM sleep, both absolute and relative to each other, such as the ratio of time spent in REM versus NREM sleep); (ii) aspects of their distribution over the 24 h cycle (e.g., time course of hourly values, midpoint of the 12 h interval with highest time spent awake, and differences between the light and dark periods); and (iii) sleep-wake architecture (e.g., number and duration of sleep-wake bouts, sleep fragmentation, and sleep-wake state transition probabilities). Similarly, for the second phenotypic category ("LMA") overall activity counts per day, as well as per unit of time spent awake, and the distribution of activity over the 24 h cycle was extracted from the LMA data. As final phenotypic category ("EEG"), EEG signals of the 4 different sleep-wake states (wakefulness, NREM sleep, REM sleep, and theta-dominated waking [TDW], see below) were quantified within the 4-s epochs matching the sleep-wake states using DFT (0.25 Hz resolution, range 0.75–90 Hz, window function Hamming). Signal power was calculated in discrete EEG frequency bands—i.e., delta (1.0–4.25 Hz, δ), slow delta (1.0–2.25 Hz; $\delta1$), fast delta (2.5–4.25; $\delta2$), theta (5.0–9.0 Hz during sleep and 6.0–10.0 Hz during TDW); θ), sigma (11–16 Hz; σ), beta (18–30 Hz; β), slow gamma (32–55 Hz; $\gamma1$), and fast gamma (55–80 Hz; $\gamma2$). Power in each frequency band was referenced to total EEG power over all frequencies (0.75–90 Hz) and all sleep-wake states in days B1 and B2 to account for interindividual variability in absolute power. The contribution of each sleep-wake state to this reference was weighted such that, e.g., animals spending more time in NREM sleep (during which total EEG power is higher) do not have a higher reference as a result[43]. Moreover, the frequency of dominant EEG rhythms was extracted as phenotypes, specifically that of the theta rhythm characteristic of REM sleep and TDW. The latter state, a substate of wakefulness, defined by the prevalence of theta activity in the EEG during waking[44,45], was quantified according to the algorithm described in[46]. We assessed the time spent in this state, the fraction of total wakefulness it represents, and its distribution over 24 h. Finally, discrete, paroxysmal events were counted, such as sporadic spontaneous seizures and neocortical spindling, which are known features of D2 mice[47], which we also found in some BXD lines.

All phenotypes were quantified in baseline and recovery separately, and the effect of SD on all variables was computed as recovery versus baseline differences or ratios. Pascal source code used for EEG/behavioral phenotyping was made available on gitlab (https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation). Processed phenotypes and descriptions were made available at https://bxd.vital-it.ch/#/dataset/1 and were submitted the Mouse Phenome Database[48].

Read alignment. For gene expression quantification, we used a standard pipeline that was already applied in a previous study[6]. Bad quality reads tagged by Casava 1.82 were filtered from fastq files and reads were mapped to MGSCv37/mm9 using the STAR splice aligner (v 2.4.0 g) with the *2pass* pipeline[49].

Genotyping. The RNA-seq dataset was also used to complement the publicly available GeneNetwork genetic map (www.genenetwork.org), thus increasing its resolution. RNA-seq variant calling was performed using the Genome Analysis ToolKit (GATK) from the Broad Institute, using the recommended workflow for RNA-seq data[50]. To improve coverage depth, 2 additional RNA-seq datasets from other projects using the same BXD lines were added[6]. In total, 6 BXD datasets from 4 different tissues (cortex, hypothalamus, brainstem, and liver) were used. A hard filtering procedure was applied as suggested by the GATK pipeline[50–52]. Furthermore, genotypes with more than 10% missing information, low quality (<5000), and redundant information were removed. GeneNetwork genotypes, which were discrepant with our RNA-seq experiment, were tagged as "unknown" (mean of 1% of the GeneNetwork genotypes/strain [$0.05\% \leq n \leq 8\%$]). Finally, GeneNetwork and our RNA-seq genotypes were merged into a unique set of around 11000 genotypes, which was used for all subsequent analyses. This set of genotypes was already used successfully in a previous study of BXD lines[6] and is available through our "Swiss-BXD" web interface (https://bxd.vital-it.ch/#/dataset/1).

Protein damage prediction. Variants detected by our RNA-seq variant calling were annotated using Annovar[53] with the RefSeq annotation dataset. Nonsynonymous variations were further investigated for protein disruption using Polyphen-2 version 2.2.2[54], which was adapted for use in the mouse according to recommended configuration. Variant annotation file and polyphen2 scores were made available here[27].

Gene expression quantification. Count data was generated using htseq-count from the HTseq (v0.5.4p3) package using parameters "stranded = reverse" and "mode = union"[55]. Gene boundaries were extracted from the mm9/refseq/reflat dataset of the UCSC table browser (extracted the 29th Jan. 2014). Raw counts were made available[27].

Intermediate-level Layer Analyses. *Gene expression normalization.* EdgeR (v3.22) was then used to normalize read counts by library size. Genes with with low expression value were excluded from the analysis, and the raw read counts were normalized using the TMM normalization[56] and converted to log counts per million (CPM). Although for both tissues, the RNA-seq samples passed all quality thresholds, and among-strain variability was small, more reads were mapped in cortex than in liver, and we observed a somewhat higher coefficient of variation in the raw gene read count in liver than in cortex. Genes expression as CPM or log2 CPM were made available[27].

Differential expression. To assess the gene differential expression between the sleep-deprived and control conditions, we used the R package limma[57] (v3.36) with the voom weighting function followed by the limma empirical Bayes method[58]. Differential expression tables were made available[27].

QTL mapping. The R package qtl/r[33] (version 1.41) was used for interval mapping of behavioral/EEG phenotypes (phQTLs) and metabolites (mQTLs). Pseudomarkers were imputed every cM, and genome-wide associations were calculated using the Expected-Maximization (EM) algorithm. p-values were corrected for FDR using permutation tests with 1000 random shuffles. The significance threshold was set to 0.05 FDR, a suggestive threshold to 0.63 FDR, and a highly suggestive threshold to 0.10 FDR according to[59,60]. QTL boundaries were determined using a 1.5 LOD support interval. To preserve sensitivity in QTL detection, we did not apply further p-value correction for the many phenotypes tested. Effect size of single QTLs was estimated using 2 methods. Method 1 does not consider eventual other QTLs present and computes effect size according to $1 - 10\text{\^{}}(-(2/n)*\text{LOD})$. Method 2 does consider multi-QTL effects and computes effect size by each contributing QTL by calculating first the full, additive model for all QTLs identified and, subsequently, estimating the effects of each contributing QTL by computing the variance lost when removing that QTL from the full model ("drop-one-term" analysis). For Method 2, the additive effect of multiple suggestive, highly suggestive, and significant QTLs was calculated using the fitqtl function of the qtl/r package[61]. With this method, the sum of single QTL effect estimation can be lower than the full model because of association between genotypes. In the Results section, Method 1 was used to estimate effect size, unless specified otherwise. It is important to note that the effect size estimated for a QTL represents the variance explained of the genetic portion of the variance (between-strain variability) quantified as heritability and not of the total variance observed for a given phenotype (i.e., within- plus between-strain variability).

For detection of eQTLs, cis-eQTLs were mapped using FastQTL[33] within a 2 Mb window for which adjusted p-values were computed with 1000 permutations and beta distribution fitting. The R package qvalue[62] (version 2.12) was then used for multiple-testing correction as proposed by[33]. Only the q-values are reported for each cis-eQTL in the text. Trans-eQTL detection was performed using a modified version of FastEpistasis[34], on several million associations (approximately 15000 genes × 11000 markers), applying a global, hard p-value threshold of 1E−4.

List of ph-QTLs, cis-eQTL, trans-eQTL and m-QTLs were made available[27].

High-level layer Analyses. *Hiveplot visualization.* Hiveplots were constructed with the R package HiveR[63] for each phenotype. Gene expression and metabolite levels represented in the hiveplots come from either the "Ctr" (control) or SD molecular datasets according to the phenotype represented in the hiveplot; i.e., the "Ctr" dataset is represented for phenotypes related to the baseline ("bsl") condition, while the SD dataset is shown for phenotypes related to recovery ("rec" and "rec/bsl"). For a given hiveplot, only those genes and metabolites were included (depicted as nodes on the axes) for which the Pearson correlation coefficient between the phenotype concerned and the molecule passed a data-driven threshold set to the top 0.5% of all absolute correlations between all phenotypes on the one hand and all molecular (gene expression and metabolites) on the other. This threshold was calculated separately for "Bsl" phenotypes and for "Rec" and "Rec/Bsl" phenotypes and amounted to absolute correlation thresholds of 0.510 and 0.485, respectively. The latter was used for the recovery phenotypes. Associations between gene expressions and metabolites represented by the edges in the hiveplot were filtered using quantile thresholds (top 0.05% gene–gene associations, top 0.5% gene–metabolite associations). We corrected for cis-eQTL confounding effects by computing partial correlations between all possible pairs of genes. Hiveplots figures and Rmarkdowns reports were made available[27].

Candidate-gene prioritization strategy. In order to prioritize genes in identified QTL regions, we chose to combine the results of the following analyses: (i) QTL mapping (phQTL or mQTL), (ii) correlation analysis, (iii) expression QTL (eQTL), (iv) protein damaging–variation prediction, and (v) DE. Each result was transformed into an "analysis score" using a min/max normalization, in which the contribution of extreme values was reduced by a winsorization of the results. These analysis scores were first associated with each gene (see below) and then integrated into a single "integrated score" computed separately for each tissue, yielding 1 integrated score in cortex and 1 in liver. The correlation analysis score, eQTL score, DE score, and protein damaging–variation score are already associated to genes, and these values were therefore attributed to the corresponding gene. To associate a gene with the ph-/m-QTL analysis score (which is associated to markers), we used the central position of the gene to infer the associated ph-/m-QTL analysis score at that position. In case of a cis-eQTL linked to a gene or a damaging variation within the gene, we used the position of the associated marker instead. To emphasize diversity and reduce analysis score information redundancy, we weighted each analysis score using the Henikoff algorithm. The individual scores were discretized before using the Henikoff algorithm, which was applied on all the genes within the ph-/mQTL region associated with each phenotype. The integrated score was calculated separately for cortex and liver. We performed a 10000-permutation procedure to compute an FDR for the integrated scores. For each permutation procedure, all 5 analysis scores were permutated, and a novel integrated score was computed again. The maximal integrated score for each permutation procedure was kept, and a significance threshold was set at quantile 95. Applying the Henikoff weighting improved the sensitivity of the gene prioritization. E.g., among the 91 behavioral/EEG phenotypes associated with 1 or more suggestive/significant QTLs after SD, 40 had at least 1 gene significantly prioritized with Henikoff weighting, against 32 without. Gene prioritization figures and Rmarkdown reports were made available[27].

Reproducibility of the pipeline. *Technical reproducibility of the pipeline.* To assess the reproducibility of our analytical pipeline, we asked a bioinformatician that was not involved in the data collection and analysis to reanalyze some of the results. A relatively short computational time as well as importance in the published results were taken as selection criteria of analyses to be replicated. The TMM normalisation of RNA-seq counts, differential gene expression, cis-eQTL detection, and the ph-/m-QTL mapping for 4 sleep phenotypes (slow delta power gain after SD, fast delta power after SD, theta peak frequency shift after SD and NREM sleep gain in the dark after SD) and 2 metabolites (Phosphatidylcholine ae C38:2 and alpha amino-adipic acid) used as main examples in our previous publication were all re-analyzed. Finally, gene prioritization and hiveplot visualization of these 4 examples were replicated. Originally, ties in the nodes ranking function on the hiveplots axis was solved using the "random" method, but this function was modified in the hiveplot code and set as "first" to remain deterministic (see Technical Validation for results).

Reanalysis with mm10. To quantify the effect of new standards and robustness of our end-results and interpretation we changed some analyses within our low-level layer. The mm10 genome assembly was set as our new reference and the gene expression was reanalysed from the raw fastq files with the BioJupies reproducible pipeline[64,65] that use kallisto pseudo-alignement[66]. The gene positions were retrieved from the headers of the ENSEMBL fasta file used by BioJupies (Mus_musculus.GRCm38.cdna.all.fa.gz). Genotypes were downloaded from GeneNetwork database and our annovar/polyphen2 variations positions based on mm9 were adapted to mm10 using CrossMap version 0.2.4[67]. The analyses performed to assess the technical reproducibility of our pipeline (see above) were finally replicated using these new files. (see Technical Validation for results).

Data Records

EEG/EMG power spectra and locomotor activity files were submitted to Figshare[23]. Raw data of RNA-sequencing were submitted to Gene Expression Omnibus[24]. Processed phenotypes files as gene expression, metabolites level and mean EEG/behavioral phenotypes per lines, as well as phenotypes descriptions, were submitted to our data-mining web-site (https://bxd.vital-it.ch/#/dataset/1) on the 'Downloads' panel. Scripts and code were submitted to gitlab (https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation). Intermediate files required to run these scripts were submitted to *Figshare*[27]. The data hosted on our server and the data we used from external repositories like *GeneNetwork* original genotypes[40] and RefSeq transcripts[68] were also copied on Figshare[27] for reproducibility purpose. Please cite R.Williams or the NCBI if you use these two files.

Technical Validation

Compare genotype RNA-seq vs GeneNetwork. To verify the genetic background of each mice we phenotyped, we analyzed the correspondence between GeneNetwork genotypes and RNA-seq variants detected by GATK. Of the 3811 GeneNetwork (2005) genotypes, 1289 could be recalled in our RNA-seq variant calling pipeline. Figure 4 shows the similarity proportion between RNA-seq variants and GeneNetwork genotypes, for each pair of BXD lines. Our BXD63 was more similar with the GeneNetwork BXD67 than with the BXD63, probably due to mislabeling. We therefore chose to exclude this line. The matrix also shows the genetic similarity between BXD73 and BXD103 (now renamed as BXD73b), between BXD48 and BXD96 (now BXD48a) and between BXD65 and BXD97 (now BXD65a), which confirmed the renaming of these BXD lines on GeneNetwork.

Reproducibility of the pipeline. *Technical reproducibility of the pipeline.* To assess the technical reproducibility of the pipeline, a bioinformatics student (NG) new to the project, reproduced selected steps of the bioinformatic pipeline. The results (Fig. 5, upper part) were consistent with previous analyses (PLOS Biology publication figures: 2c, 4c left, 7d, and 7c bottom). The robustness of the pipeline was verified because the same conclusions could be drawn. For examples, the same 3 genes showed the largest differential expression after SD in the cortex (*Arc*, *Plin4*, and *Egr2* in Fig. 5b). Moreover, the *Acot11* gene was prioritized by gene prioritization (Fig. 5

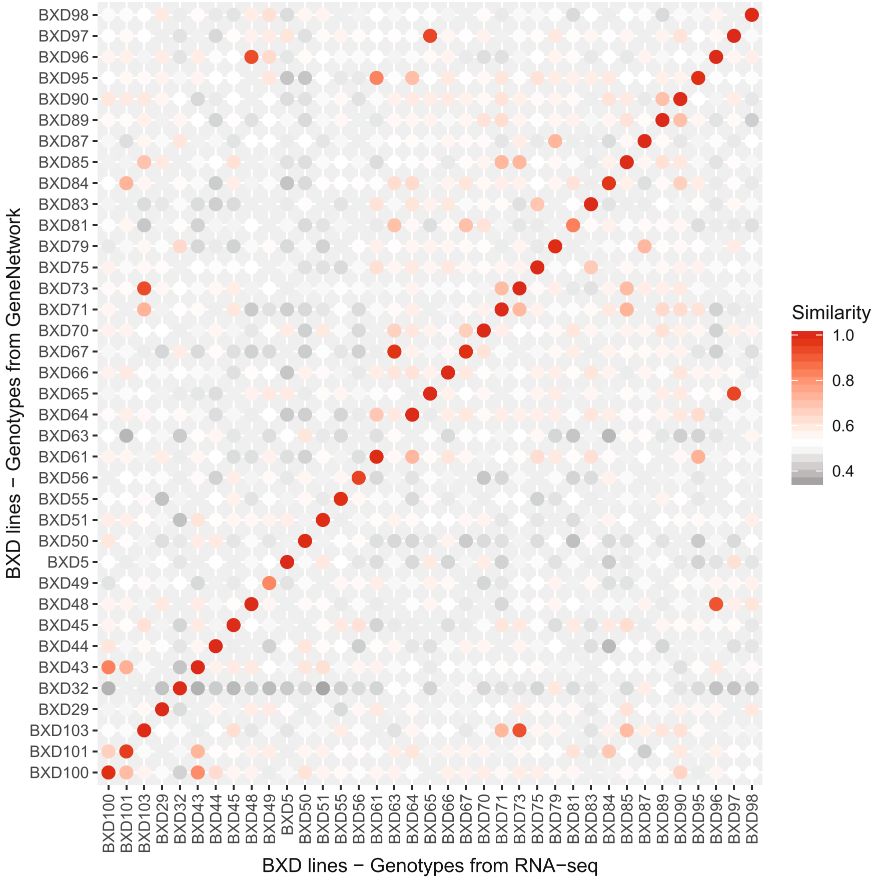


Fig. 4 Similarity matrix [in %] between RNA-seq variant calling and GeneNetwork genotypes. A similarity of 1 indicates that all common genotypes are similar. We here compare only genotypes that were labeled as 'B' or 'D' and excluded unknown 'U' or heterozygous 'H' genotypes. BXD63 genotypic similarity in our dataset was low and could indicate mislabeling.

d and e). Nevertheless, the numbers of significant genes of cis-eQTL showed variations compared to previous analysis[2] due to use of a hard significance threshold for visualization. For example, the number of genes with significant QTL unique to Cortex SD changed from 870 (PLOS Biology publication Fig. 2c) to 872 (Fig. 5a). Genes were considered as significant if their FDR-adjusted p-value was below or equal to 0.05, which was obtained by estimating the β-distribution fitting of random permutations tests. Changing the fastqtl version (version 1.165 to version 2.184) seems to change the pseudo-random number generation, even when using the concept of fixed seed. Consequently, the number of genes considered as significant varies because their FDR-adjusted p-value passed just above or below the threshold (FDR in the range of 0.04864 to 0.05054). This confirms that looking at the order of magnitude is important, though the use of significance threshold is convenient.

Moreover, the reanalysis process helped to improve the code documentation by explicitly writing project-related knowledge, such as common abbreviations. Having another perspective on the code also allowed to improve its structure. Indeed, a retrospective overview helped improve the organisation of files, which was more difficult to do within the implementation phase of the project because the code was incrementally created and adapted. The process allowed to catch and correct minor mistakes or make improvements to readability and consistency. For example, it was highlighted that the ranking function used in hiveplot to order nodes in the axes was using the "random" argument for differentiating ties. As a key concept of the hiveplots was to be fully reproducible in the sense of "perpetual uniformity"[63], we changed the ties.method parameter to "first" so that the same input always gives the same result, without having to fix a seed for the pseudo-random generation. Another example was the ranking of the x-axis in the gene DE volcano plot and the colouring that were based on log-odds values (B statistic according to in limma R package) instead of FDR-adjusted p-values. However, this reproducibility 'experiment' was performed internal to the group, which facilitated communication such as which steps to focus on and whether to run them locally or on a high-performance computing (HPC) structure. An assessment of the computational requirements for each step, such as computing time, memory, software, and libraries used may be interesting to provide to facilitate external reproducibility.

Reanalysis with mm10. To assess the influence of the reference genome used in the analyses, we reproduced selected parts of bioinformatic pipeline using the updated mm10 version (instead of mm9). The results (Fig. 5, lower part, Tables 1 and 2) were consistent with previous analyses but presented also some substantial variations. The cis-eQTL detection revealed differences in the number of significant associations found, as showed in

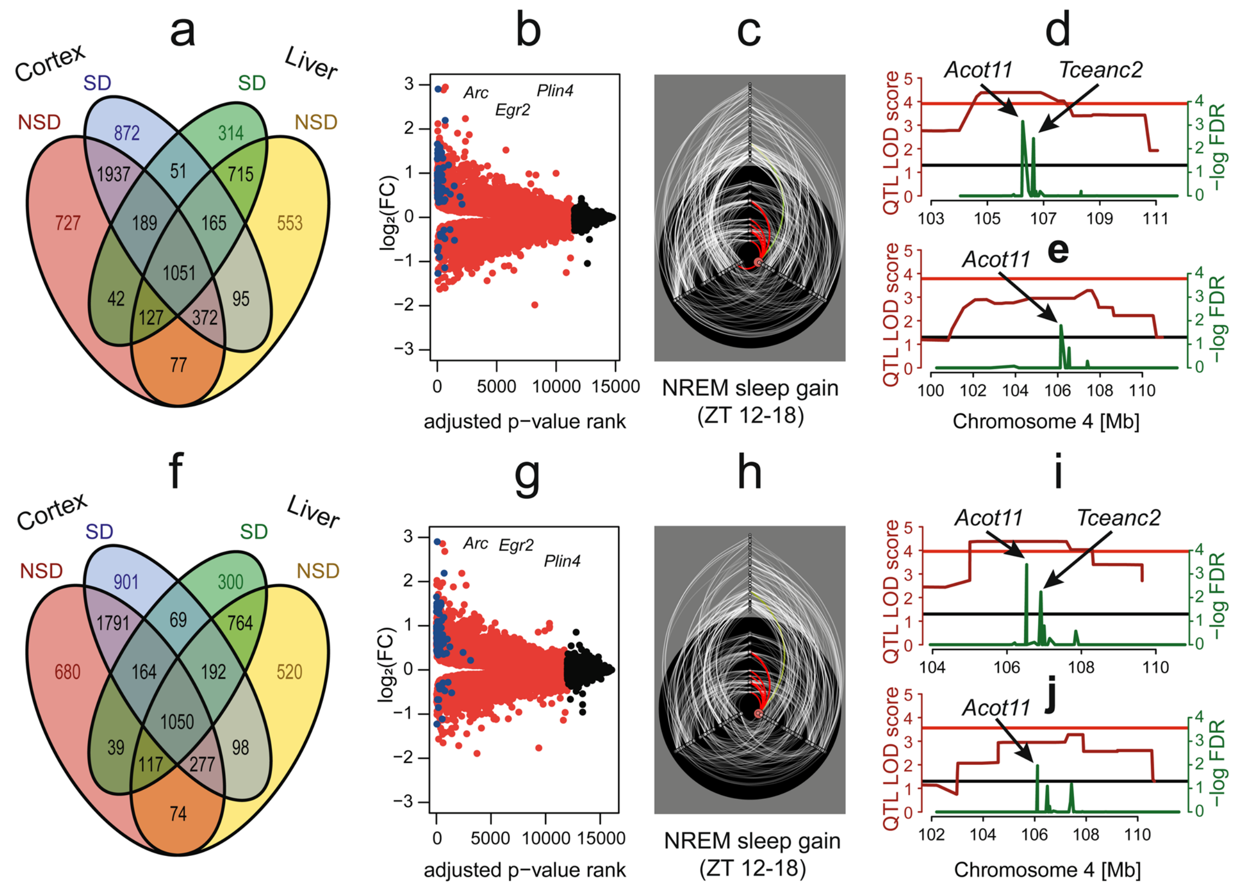


Fig. 5 Robustness of the analysis pipeline. (**a to e**) Technical reanalysis with mm9 reference genome. (**f to j**) Reanalysis with mm10 reference genome. (**a and f**) Venn diagram of significant cis-eQTL. (**b and g**) Volcano plot of differential gene expression in cortex. (**c and h**): Hiveplot for NREM sleep gain during recovery with highlight on *Acot11*. (**d,e,i,j**) Gene prioritization for NREM sleep gain during recovery (**d and i**) or phosphatidylcholine acyl-alkyl C38:2 levels (**e and j**). recovery = first 6 hours of dark period after sleep deprivation (ZT 12–18), SD = sleep deprivation, NSD = not sleep deprivation (control), FC = fold-change, NREM = non-rapid eye movement, LOD = logarithm of odds, FDR = false discovery rate.

	Liver NSD		Liver SD		Cortex NSD		Cortex SD	
Assembly	**mm9**	**mm10**	**mm9**	**mm10**	**mm9**	**mm10**	**mm9**	**mm10**
Total genes	14103	12647	14103	12647	14889	15734	14889	15734
Unique genes	2405	949	2405	949	1043	1888	1043	1888
Genes with significant cis-eQTL	3155	3092	2654	2695	4522	4192	4732	4542
Proportion of genes with significant cis-eQTL	0.22	0.24	0.19	0.21	0.30	0.27	0.32	0.29
Genes with significant cis-eQTL overlapping	2255		1911		3204		3483	
Genes with not significant cis-eQTL overlapping	8375		8857		9062		8801	
Genes with significant cis-eQTL not overlapping	900	837	743	784	1318	988	1249	1059
Genes with significant cis-eQTL almost overlapping	2995	2898	2535	2505	4201	4019	4441	4350

Table 1. Comparison of cis-eQTL summary statistics using mm9 vs mm10. 'Unique' is defined as specific to an assembly (mm9 or mm10). Significance is defined as a q-value below or equal to 0.05. 'Overlapping' is defined as common between mm9 and mm10 reanalyses. 'Almost overlapping' is defined as common between mm9 and mm10 at a threshold of 0.1 but not as the 0.05 threshold.

Table 1. These differences could be mainly explained by small q-value variation around the significant threshold. Nevertheless, around 5% of cis-eQTLs did not reproduce even at a more permissive significant threshold (0.1 FDR), which affected some of our end results. For example, *Wrn* was no longer prioritized for the gain of slow EEG delta power (δ1) after SD compared to previous results on mm9. Although the cis-eQTL for *Wrn* was present in both assemblies for the 'Cortex Control' samples, it disappeared for 'Cortex SD' samples using mm10. A number of factors could have contributed to this discrepancy among which i) the variations between mm9 and mm10 could change the mappability of some transcripts, although this did not seem to be the case for the *Wrn*

	Liver		Cortex	
	mm9	mm10	mm9	mm10
Total genes	12539	13264	14754	16057
DE genes	7573	8754	11534	11980
Proportion of DE genes	0.6040	0.6600	0.7818	0.7461
Suggestive DE genes	8253	9392	12069	12580
Proportion of suggestive DE genes	0.6582	0.7081	0.8180	0.7835

Table 2. Comparison of gene DE in mm9 and mm10 reanalyses. Suggestive is defined as a q-value below or equal to 0.1.

sequence, ii) pseudo-alignment (Kallisto) was used instead of alignment (STAR), which may have influenced the quantification, iii) bad quality reads were filtered with our STAR pipeline according to Casava 1.82 but not with Kallisto, and iv) variant calling on RNA-seq data to add markers was not done for mm10, so only markers from GeneNetwork (2017) were used. Specifically to the latter factor, the marker closest to the *Wrn* gene in mm9 merged (GeneNetwork 2005 + RNA-seq) genotypes (rs51740715) is not present in mm10. The change in the number of genetic markers could have therefore influenced the cis-eQTL detection, which is an important factor in the gene prioritization that resulted in the identification of *Wrn* as candidate underlying the EEG delta power (δ1) trait under mm9.

Usage Notes

SMO files. Binary *.smo* files were structured as follows: Each file contains a 4-day recording or precisely 86400 consecutive 4 s epochs. Each 4 s epoch contains the following information: one byte character and 404 single precision floating-points, which represent, respectively: sleep-wake state of the 4 s epoch as a character (wake = 'w', NREM sleep = 'n', REM sleep = 'r', wake w/ EEG artifact = '1', NREM sleep w/ EEG artifact = '2', REM sleep w/ EEG artifact = '3', wake w/ spindle-like EEG activity = '4', NREM sleep w/ spindle-like EEG activity = '5', REM sleep w/ spindle-like EEG activity = '6', Paroxysmal EEG activity in wake = '7', Paroxysmal EEG activity in NREM sleep = '8', Paroxysmal EEG activity in REM sleep = '9'), EEG power density from the full DFT spectrum of the 4 s epoch from 0.00 Hz to 100.00 Hz (401 values at 0.25-Hz resolution), the EEG variance, the EMG variance, and temperature. Temperature was not measured and was set to 0.0.

HDR files. Text *.hdr* files are generated alongside their corresponding *.smo* file and contain among other information, the mouse ID (*Patient*) and recording date.

Rmarkdown scripts. Some of the Rmarkdown scripts were created for a remote cluster environment on a CentOS distribution which required the use of a second script that generated the document with the *rmarkdown::render()* function and pass the expected function arguments. Therefore some functions that use the parallel package in R are only executable on a linux environment (i.e. mclapply()). These functions can be modified with the *doSNOW* R library to be applicable on a windows environment. The author can set many option in the YAML (Yet Another Markup Language) header to: create dynamic and readable table that contains multiple rows, hide/show source code or integrated CSS style and table of contents. The reports can be visualized using any web-browser.

Code Availability

The scripts used for analytics were made available on gitlab (https://gitlab.unil.ch/mjan/Systems_Genetics_of_Sleep_Regulation). The master branch contains the scripts used for our publication and mm9 analysis. A second branch was created for analysis performed on a mm10 mouse references (see Technical Validation). The intermediate files required to run these scripts were made available at Figshare[27]. Finally, a documentation file was generated documenting the hierarchical relationship between the scripts and datasets in a form of a dynamic html document (see Workflow documentation).

References

1. Schmid, S. M., Hallschmid, M. & Schultes, B. The metabolic burden of sleep loss. *The Lancet Diabetes & Endocrinology* **3**, 52–62 (2015).
2. Diessler, S. *et al.* A systems genetics resource and analysis of sleep regulation in the mouse. *PLOS Biology* **16**, e2005750 (2018).
3. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* **15**, 34–48 (2014).
4. Peirce, J. L., Lu, L., Gu, J., Silver, L. M. & Williams, R. W. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genetics* **5**, 7 (2004).
5. Franken, P., D. Chollet and M. Tafti. The homeostatic regulation of sleep need is under genetic control. *The Journal of neuroscience: the official journal of the Society for Neuroscience* **21**, 2610–2621 (2001).
6. Picard, A. *et al.* A Genetic Screen Identifies Hypothalamic Fgf15 as a Regulator of Glucagon Secretion. *Cell Reports* **17**, 1795–1806 (2016).
7. Neuner, S. M. *et al.* Systems genetics identifies Hp1bp3 as a novel modulator of cognitive aging. *Neurobiology of Aging* **46**, 58–67 (2016).
8. Andreux, P. A. A. *et al.* Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. *Cell* **150**, 1287–1299 (2012).
9. Baliga, N. S. *et al.* The State of Systems Genetics in 2017. *Cell Systems* **4**, 7–15 (2017).
10. Gligorijević, V. & Pržulj, N. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface* **12**, 20150571 (2015).

11. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* **13**, 667–672 (2012).
12. Figueiredo, A. S. Data Sharing: Convert Challenges into Opportunities. *Front Public Health* **5**, 327 (2017).
13. Bechhofer, S. *et al*. Why linked data is not enough for scientists. *Future Generation Computer Systems* **29**, 599–611 (2013).
14. Wilkinson, M. D. *et al*. Interoperability and FAIRness through a novel combination of Web technologies. *Peerj Computer Science* **3**, e110 (2017).
15. Jagodnik, K. M. *et al*. Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. *Journal of Biomedical Informatics* **71**, 49–57 (2017).
16. Sansone, S. A. *et al*. Toward interoperable bioscience data. *Nature Genetics* **44**, 121–126 (2012).
17. Lowndes, J. S. S. *et al*. Our path to better science in less time using open data science tools. *Nature Ecology &. Evolution* **1**, 160 (2017).
18. Vasilevsky, N. A., Minnier, J., Haendel, M. A. & Champieux, R. E. Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ* **5**, e3208 (2017).
19. Wallach, J. D., Boyack, K. W. & Ioannidis, J. P. A. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biology* **16**, e2006930 (2018).
20. Munafo, M. R. *et al*. A manifesto for reproducible science. *Nature Human Behaviour* **1**, 21–21 (2017).
21. Shin, D.-L. L. *et al*. Segregation of a spontaneous Klrd1 (CD94) mutation in DBA/2 mouse substrains. *G3: Genes*, *Genomes*. *Genetics* **5**, 235–239 (2014).
22. Mang, G. M. M. & Franken, P. Sleep and EEG Phenotyping in Mice. *Current Protocols in Mouse Biology* **2**, 55–74 (2012).
23. Jan, M. *et al*. A multi-omics digital research object for the genetics of sleep regulation. *figshare*. https://doi.org/10.6084/m9.figshare.c.4421327 (2019).
24. Diessler, S. *et al*. Systems genetics of sleep regulation. *Gene Expression Omnibus*, http://identifiers.org/geo:GSE114845 (2018).
25. Davies, S. K. *et al*. Effect of sleep deprivation on the human metabolome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 10761–10766 (2014).
26. Isherwood, C. M., Van der Veen, D. R., Johnston, J. D. & Skene, D. J. Twenty-four-hour rhythmicity of circulating metabolites: effect of body mass and type 2 diabetes. *The FASEB Journal* **31**, 5557–5567 (2017).
27. Jan, M., Gobet, N., Diessler, S., Franken, P. & Xenarios, I. A multi-omics digital research object for the genetics of sleep regulation: Input-data and code. *figshare*. https://doi.org/10.6084/m9.figshare.7797434 (2019).
28. Govoni, M. *et al*. Qresp, a tool for curating, discovering and exploring reproducible scientific papers. *Scientific Data* **6**, 190002 (2019).
29. Conesa, A. *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).
30. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 1184–1191 (2009).
31. Kohler, S. *et al*. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* **47**, D1018–D1027 (2019).
32. Durrant, C. *et al*. Bioinformatics tools and database resources for systems genetics analysis in mice–a short review and an evaluation of future needs. *Briefings in Bioinformatics* **13**, 135–142 (2012).
33. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **15**, 1479–1485 (2015).
34. Schupbach, T., Xenarios, I., Bergmann, S. & Kapur, K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* **26**, 1468–1469 (2010).
35. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLOS Computational Biology* **9**, e1003285 (2013).
36. Xie, Y. *Dynamic Documents with {R} and knitr*. **2nd edition** (Chapman and Hall/CRC, 2015).
37. Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L. & Horton, N. J. R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics. *Technology Innovations in Statistics Education* **8**, (2014).
38. Cohen-Boulakia, S. *et al*. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems* **75**, 284–298 (2017).
39. Neff, E. P. A mouse sleep database for systems genetics. *Lab Animal* **47**, 272 (2018).
40. Williams, R. W., Ingels, J., Lu, L., Arends, D. & Broman, K. W. *BXD Genotype Database*, http://genenetwork.org/webqtl/main.py?FormID=sharinginfo&GN_AccessionId=600 (2018).
41. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. *e1071: Misc Functions of the Department of Statistics (E1071)*, *TU Wien*., https://CRAN.R-project.org/package=e1071 (2014).
42. Kuhn., M. *et al. caret: Classification and Regression Training*, http://CRAN.R-project.org/package=caret (2014).
43. Franken, P., Malafosse, A. & Tafti, M. Genetic variation in EEG activity during sleep in inbred mice. *The American Journal of Physiology* **275**, 37 (1998).
44. Buzsáki, G. Theta oscillations in the hippocampus. *Neuron* **33**, 325–340 (2002).
45. Welsh, D. K., Richardson, G. S. & Dement, W. C. A circadian rhythm of hippocampal theta activity in the mouse. *Physiology & Behavior* **35**, 533–538 (1985).
46. Vassalli, A. & Franken, P. Hypocretin (orexin) is critical in sustaining theta/gamma-rich waking behaviors that drive sleep need. *Proceedings of the National Academy of Sciences* **114**, E5464–E5473 (2017).
47. Ryan, L. J. Characterization of cortical spindles in DBA/2 and C57BL/6 inbred mice. *Brain Research Bulletin* **13**, 549–558 (1984).
48. Bogue, M. A. *et al*. Mouse Phenome Database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice. *Nucleic Acids Research* **46**, D843–D850 (2018).
49. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, (15–21 (2013).
50. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).
51. DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).
52. Van der Auwera, G. A. *et al*. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.11–33 (2013).
53. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164 (2010).
54. Adzhubei, I. A. *et al*. A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
55. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
56. Robinson, M. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25 (2010).
57. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).
58. Law, C. W., Chen, J. C., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, R29 (2014).

59. Burgess-Herbert, S. L., Cox, A., Tsaih, S.-W. W. & Paigen, B. Practical applications of the bioinformatics toolbox for narrowing quantitative trait loci. *Genetics* **180**, 2227–2235 (2008).
60. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241–247 (1995).
61. Broman, K. W. & Sen, S. *A Guide to QTL Mapping with R/qtl*. Vol. 46 (Springer, 2009).
62. Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. *qvalue: Q-value estimation for false discovery rate control*, http://github.com/jdstorey/qvalue (2019).
63. Krzywinski, M., Birol, I., Jones, S. J. & Marra, M. A. Hive plots–rational approach to visualizing networks. *Briefings in Bioinformatics* **13**, 627–644 (2012).
64. Lachmann, A. *et al*. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications* **9**, 1366 (2018).
65. Torre, D., Lachmann, A. & Ma'ayan, A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Systems* **7**, 556–561 e553 (2018).
66. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016).
67. Zhao, H. *et al*. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
68. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–745 (2016).
69. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

Acknowledgements

We thank Yann Emmenegger, Mathieu Piguet and Josselin Soyer for the help organizing and handling the BXD mice; Benita Middleton and Debra J. Skene at the university of Surrey for metabolomics quantification; the experts at the Lausanne Genomics Technologies Facility (GTF) for their help with RNA-sequencing; Nicolas Guex, Mark Ibberson, Frederic Burdet, Alan Bridge, Lou Götz, Marco Pagni, Martial Sankar and Robin Liechti for their bioinformatics expertise and help. Computations and analyses were performed at Vital-IT (http://www.vital-it.ch). P.F., S.D. and M.J. were funded by the University of Lausanne (Etat de Vaud). M.J. and N.G. were funded by the Swiss National Science Foundation grant to PF (31003A_173182). M.J. and S.D. were funded by the Swiss National Science Foundation grant to PF and IX (CRSII3_13620).

Author Contributions

M.J. Data analysis, data organization, manuscript writing. N.G. Pipeline reproduction, data organization, manuscript writing. S.D. Data generation and supervision for experiment 1 and 2. P.F. Data analysis, design experiments 1 and 2, project supervision, manuscript writing. I.X. Conceptualization, methodology, project supervision, manuscript writing.

Additional Information

Competing Interests: The authors declare no competing interests.

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

Chapter 3

Towards mouse genetic-specific RNA-sequencing read mapping

The traditional pipeline for RNA-sequencing uses the GRC reference which is predominantly based on one mouse strain: C57BL/6J (B6) which is one of the parental strain that the BXD recombinant inbred panel were derived from. Therefore, the DBA/2J (D2) regions risk to be less well represented. How is this impacting the results and can we find a better solution to this short-comming? The goal of this chapter is to assess the parental reference bias in RNA-seq and to implement a solution to improve references for BXD lines.

3.1 Results summary

To tackle the genetic-specificity of the reads, the first strategy was to use two different genome assemblies to map the reads, one per parental strain to balance. However the difference in quality between the assemblies was prohibitive to used them as co-contributors. The second strategy was to use the standard assembly based on B6 and customize it with BXD variants. Imputing the D2 genetic variants in D2 blocks/regions was important achieve enough resolution in the BXD-specific references. The impact can be seen at different levels: the mapping, the gene expression and the eQTLs. The reference bias was alleviated and we detected proportionally more eQTLs with the custom BXD-specific references than with the standard reference.

3.2 My contribution

With the help of co-authors, I planned and performed all the analyses. I wrote the draft, produced all the figures and tables, and edited the manuscript with input from co-authors.

3.3 Publication

This article was published in a peer-reviewed journal [Gobet et al., 2022]. The publication is included below.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Towards mouse genetic-specific RNA-sequencing read mapping

Nastassia Gobet[1,2], **Maxime Jan**[1,3], **Paul Franken**[1], **Ioannis Xenarios**[4,5]*

[1] Centre for Integrative Genomics, University of Lausanne, Lausanne, Switzerland, [2] Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3] Bioinformatics Competence Center, University of Lausanne, Lausanne, Switzerland, [4] Ludwig Cancer Research/CHUV-UNIL, Lausanne, Switzerland, [5] Health 2030 Genome Center, Geneva, Switzerland

* ioannis.xenarios@unil.ch

OPEN ACCESS

Citation: Gobet N, Jan M, Franken P, Xenarios I (2022) Towards mouse genetic-specific RNA-sequencing read mapping. PLoS Comput Biol 18(9): e1010552. https://doi.org/10.1371/journal.pcbi.1010552

Editor: Sonika Tyagi, Monash University, AUSTRALIA

Received: January 19, 2022

Accepted: September 7, 2022

Published: September 26, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010552

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. Raw reads files are accessible on NCBI Gene Expression Omnibus under accession id GSE114845. Given that data sources are diverse

Abstract

Genetic variations affect behavior and cause disease but understanding how these variants drive complex traits is still an open question. A common approach is to link the genetic variants to intermediate molecular phenotypes such as the transcriptome using RNA-sequencing (RNA-seq). Paradoxically, these variants between the samples are usually ignored at the beginning of RNA-seq analyses of many model organisms. This can skew the transcriptome estimates that are used later for downstream analyses, such as expression quantitative trait locus (eQTL) detection. Here, we assessed the impact of reference-based analysis on the transcriptome and eQTLs in a widely-used mouse genetic population: the BXD panel of recombinant inbred lines. We highlight existing reference bias in the transcriptome data analysis and propose practical solutions which combine available genetic variants, genotypes, and genome reference sequence. The use of custom BXD line references improved downstream analysis compared to classical genome reference. These insights would likely benefit genetic studies with a transcriptomic component and demonstrate that genome references need to be reassessed and improved.

Author summary

To understand how genetic variations affect behavior and cause disease it is common to quantify expression of transcripts by sequencing. Transcripts are extracted, fragmented, and the sequence of the fragments read. An important step for their quantification is to virtually assign the different fragments to the transcript they originate from using a reference genome. Reference genomes are costly to build, so usually only one high-quality reference per animal model species is available. When comparing genetically different individuals, using a single reference may introduce a bias because it might be more similar to some individuals than to others. Paradoxically, the variations at the core of genetic studies are thus ignored at the start of the analysis. We built customized references with known genetic variants for each of the mouse lines we had and quantified the impact of the reference at different levels of the bioinformatic analysis. We found that using customized references reduced the bias compared to using a single reference. Our study uses

and some do not have a version or identifier, we grouped data so that it is easier to find (10.5281/zenodo.5513980). This repository contains input data: D2-specific variants from dbSNP (vcf), genotypes from GeneNetwork (tab); intermediate and produced data: D2 blocks (bed), imputed variants (vcf), BXD-specific genome sequences (fasta), BXD-specific transcriptome annotation (gtf), gene counts, normalized gene expression, local eQTLs. The code used for the analyses is on https://github.com/nagobet/BXDmapping.

Funding: P.F. and M.J. were funded by the University of Lausanne (Etat de Vaud). N.G. was funded by the Swiss National Science Foundation grant to P.F. (31003A_173182 and 310030B_192805, https://www.snf.ch/en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

publicly available data and tools, so others can easily implement this improvement in their analyses.

Introduction

To decipher how genome leads to phenome, measuring gene expression by RNA-sequencing (RNA-seq) is widely used. Fragments of RNA are read and then virtually mapped back onto a reference genome to determine the transcriptomic location of origin. Read mapping is often regarded as trivial but relies on many choices. Indeed, the user decides for example which reference to use and how exact the alignments are required to be. Most of the time, little information is published on how these choices are made. The mapping needs to account for amplification and sequencing errors, and for repeated sequences within the genome. It is important that the reference precisely represents the samples to guide the mapping. However, generating a reference assembly is complex and expensive and it is common practice to map all samples of a model organism to a single assembly provided by the genome reference consortium (GRC) [1,2]. The expression of non-reference alleles may be altered compared to that of reference alleles. This reference bias on the transcriptome can then spread to downstream analyses such as expression quantitative trait loci (eQTL) detection, where gene expression is associated to genomic variants. The genomic variations between individuals at the core of genetic studies are thus paradoxically often ignored at the start of the analysis and may alter interpretations and conclusions.

The genetic characteristics of humans have been widely studied and reference bias is known to alter DNA-seq, RNA-seq [3], and chromatin immunoprecipitation (ChIP)-seq analyses [4,5]. The ideal solution would be to use a sample-specific genome assembly. Since this is currently too costly, many methods to reduce reference bias were proposed [3,6,7]. One strategy notably used by the Genotype-Tissue Expression (GTEx) consortium is to tailor the analysis to each individual as initially implemented in the *WASP* suite of tools [8]. The WASP-correction proposes to map reads to the GRC assembly and identify mapped reads that overlap SNVs, then re-map these reads after replacing the reference alleles by variant alleles in the assembly and discard reads that change mapping loci. Although this strategy removes reference bias it also discards reads that are potentially informative of a genetic effect. Nevertheless, the idea of modifying the GRC reference assembly with variants specific to the individual or sample is used by many tools, with the difference that all the reads are mapped to the customized references only. For example, the AlleleSeq pipeline was developed for human trios where the variants for the two parents are known [9]. One of its tools, *vcf2diploid*, constructs two haplotype-specific references from one reference assembly and a list of genomic variants which can include single nucleotides variants (SNVs), indels, and structural variants (SVs). The authors proposed to map the offspring sample separately onto its two parental references, and to retain for each read the alignment with the highest alignment score. In case of equality the alignment is randomly taken from that of either parent to avoid systematic bias. *RefEditor* offers a similar approach, but adds a genotype imputation option [10]. Many tools aim at making the best use of large-scale variants and genotypes databases by genotype imputation to have for each individual a more complete set of alleles. However, these genotype imputation strategies cannot be applied to mouse or other animal models because of a lack of genetic characterization at the individual level.

Mouse genetic research mostly uses inbred lines, in which individuals are presumed isogenic. Therefore, it seems logical to aim for reference customization for mouse strains rather than for individuals. The GRC mouse assembly is mainly based on the inbred strain C57BL/6J (B6) [11] and short genomic variants for many other inbred strains are available at dbSNP [12]. To compare retinal transcriptomes in two inbred strains (DBA/2J (D2) and B6), Wang et al. modified the GRCm38 reference genome with D2-specific variants from dbSNP to map the D2 samples [13]. This improved slightly the mappability by reducing the fraction of unmapped reads. *Seqnature* software aims at producing individualized diploid references for RNA-seq analysis and was used on simulated and real world data of Diversity Outbred (DO) mice, in which each mouse is a unique combination of 8 founder strains [14]. It shows improvement of the number of reads mapped, of the accuracy of transcript expression estimates, and of the number of eQTLs detected. Unfortunately, this type of study is very rare and the R package (DOQTL) used for the QTL analyses is specific to this mouse population, which renders comparisons with studies on different populations challenging. The Mouse Genome Project tries a more global approach to characterize the genetic variation among mouse strains. Many genetic variants were discovered and strain-specific genome assemblies for sixteen mouse strains were released [15]. However, it remains unclear how to use these resources for mice that are intercrossed.

The BXD panel of recombinant inbred lines is a well-studied and genetically simple population derived from the B6 and D2 strains [16]. Each BXD line has genetic markers (genotypes) available on the GeneNetwork website (http://genenetwork.org). Although this panel has been used in hundreds of studies, nobody to our knowledge has performed neither BXD-specific read mapping, nor BXD genome assembly. Here, we explored different strategies using publicly available resources to accurately represent the genetic diversity of the samples. We assessed the influence of the reference used for read mapping in this panel and how it impacts read mappability, gene expression, and eQTLs. We also measured how various parameter settings would influence the number of eQTLs found. We evaluated the use of the two parental genome assemblies and found this strategy not adequate. We implemented an alternative strategy which enhanced the GRC assembly with known variants. This improved the quality of BXD transcriptomics analyses. Our approach reduces reference bias in the BXD transcriptomics, and raises awareness about pitfalls of RNA-seq analyses.

Methods

Ethics statement

The authorization was given by the veterinary authorities of the state of Vaud, Switzerland (SCAV authorization #2534) as previously described in [17]. No sequencing data were collected specifically for this study.

Samples and RNA-sequencing

We used RNA-seq samples obtained from the liver and the cerebral cortex of male mice from 33 BXD/RwwJ lines, the two parental strains C57BL/6J (B6), DBA/2J (D2), and their reciprocal F1 offspring (Fig 1A). The two tissues were collected at zeitgeber time (ZT) 6 (i.e., 6h after light onset) in mice that were either sleep deprived (SD) for the preceding 6 hours (ZT0-6) or mice that could sleep ad libitum (i.e. non-sleep deprived or NSD) [17,18]. Prior to sequencing, the RNA was pooled by mouse line and experimental condition (NSD or SD), such that material for a maximum of 3 mice contributed to each sample. Single-end reads of 100 bp were obtained using Illumina HiSeq 2500 system. A list of samples, including which BXD lines were used, is available (S1 Table). All samples passed quality control [17]. The eventuality of a mix-up of samples between strains, was tested previously by comparing the similarity between RNA-seq variant calling and GeneNetwork genotypes (see Fig 4 in [18]).

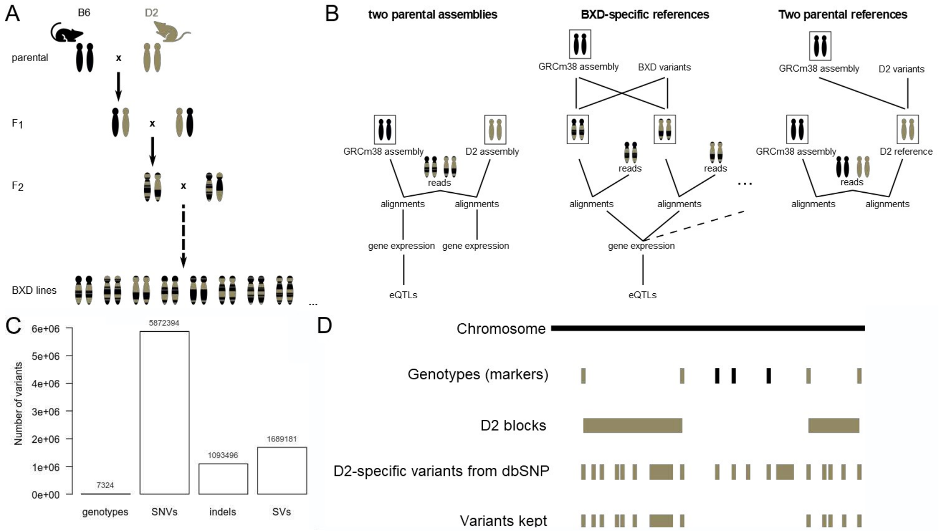


Fig 1. Overview of strategies to utilize genomic variants in transcriptome read mapping in inbred mouse lines. A. BXD mouse recombinant inbred panel. Samples came from mice that are: BXD advanced recombinant inbred lines, their parental inbred strains; i.e., C57BL/6J (B6) and DBA/2J (D2), and first generation cross between the parental strains (F1). B. The 3 RNA-seq read mapping strategies used in this study. In the 'two parental assemblies' strategy (left), the reads of all samples were mapped to the classical mouse genome assembly (GRCm38 or mm10) and to the D2 assembly. The 'BXD-specific references' (middle) were made from GRCm38 and BXD-specific variants. There is one reference for each BXD line, and the reads of each sample were mapped to the corresponding reference. The 'two parental references' (right panel) is an intermediate strategy in which the D2-specific reference was built from GRCm38 assembly and D2-specific variants. C. BXD genotypes available from GeneNetwork (genotypes) and D2-specific genomic variants (SNVs, indels, SVs) available from dbSNP. D. Genotypes imputation workflow. D2 haplotype blocks were delineated based on available genotypes in the BXD lines. D2-specific variants within these D2 blocks were included in the BXD-specific references. B6 regions or alleles are in black, D2 regions or alleles are in brown.

https://doi.org/10.1371/journal.pcbi.1010552.g001

Genome assemblies and transcriptome annotation download

Two genome assemblies and two transcriptome annotations were downloaded from Ensembl release 94 (ftp://ftp.ensembl.org/pub/release-94). The classical genome sequence GRCm38, also referred to as mm10, is based on the B6 strain (Mus_musculus.GRCm38.dna_sm.primary_assembly.fa). As D2 assembly the DBA/2J v1 genome sequence was used (Mus_musculus_dba2j.DBA_2J_v1.dna_sm.toplevel.fa) [15]. Both genome sequences do not contain alternative haplotypes, and repeats or low complexity regions are soft-masked (sm), which means they are represented as lowercase letters. Summary statistics of the assemblies were calculated in GAAS toolkit (https://github.com/NBISweden/GAAS, S2 Table). The transcriptome annotations correspond to these two assemblies (Mus_musculus.GRCm38.94.gtf and Mus_musculus_dba2j.DBA_2J_v1.94.gtf).

Variants download and genotype imputation

The BXD genotypes were downloaded from GeneNetwork. These are the alleles for each BXD line for 7324 genetic markers, which are variants selected to be indicative of recombination events between the parental genomes (BXD_Geno-19Jan2017_forGN.xlsx, Fig 1C). The D2-specific variants, which are 5'872'394 SNVs (DBA_2J.mgp.v5.snps.dbSNP142.vcf.gz) and 1'093'496 indels (DBA_2J.mgp.v5.indels.dbSNP142.normed.vcf.gz) from dbSNP (version 142, variants version 5), were downloaded (Fig 1C). To have a more complete set of genetic variants specific to each of the BXD lines, we performed genotype imputation as follows (Figs 1D and S3):

1. With the GeneNetwork BXD genotypes we defined for each BXD line the D2 haplotype blocks, as sets of at least 2 consecutive genotypes with D2 alleles without B6 or heterozygous alleles in between.
2. We checked which dbSNP D2-specific variants overlapped with the D2 blocks using bedtools (version v2.28.0).
3. We imputed the D2-specific variants overlapping with D2 blocks of a BXD line to be D2 alleles for this specific BXD line.

During this study, we noticed that genotypes from GeneNetwork for BXD100 (based on GRCm38 genome assembly also called mm10) had multiple chromosomes without any D2 alleles, which was unexpected considering this was not the case in the previous version of genotypes (based on MGSCv37 genome assembly also called mm9). GeneNetwork has been informed and the error was thought to have occurred during lift-over of the genotypes. We did not try to correct this mistake and kept the erroneous BXD100 genotypes in the current analysis. The effect, if any, on the eQTL analysis should be small, as this concerned only one BXD line out of 33, and not all chromosomes were affected.

Customization of references

We built a customized reference genome for each BXD based on the GRCm38 assembly and BXD-specific genotypes (from GeneNetwork and imputed). For this, the reference genome sequence GRCm38 was customized for each BXD line with BXD-specific genotypes (from GeneNetwork and imputed) using vcf2diploid software (version 0.2.6) with a slight modification to change the software's behaviour with unphased heterozygous variants. Prior to compiling the software according to the installation instructions, we removed the function that randomizes unphased heterozygous variants (to determine whether there are included in the paternal or maternal genome) and the call to this function (see Table 1). It is, however, entirely possible to use the software without these modifications.

In the modified software, all unphased heterozygous variants were included in the genome sequence called "maternal" but ignored in the one called "paternal". We used the paternal sequence so that heterozygous genotypes were ignored. Note that all D2-specific variants from dbSNP and the BXD markers from GeneNetwork are unphased and heterozygous labels may be indicative of low or uncertain quality. On GeneNetwork (January 2017), genotypes are defined as "H (heterozygous) if the genotype was uncertain". In the 33 BXD lines we used, 1449 loci had H alleles out of the 7320 genotypes (20%) and on average 60 H alleles out of 7320 loci or a total 1967 H alleles out of 241560 alleles (0.82%). For D2-specific variants from dbSNP (version 142), Het means "Genotype call is heterozygous (low quality)". Of the

Table 1. Modifications to vcf2diploid software.

File name	Code removed

Variant.java	public void randomizeHaplotype()
	{
	   if (_rand.nextDouble() > 0.5) return;
	   int tmp = _paternal;
	   _paternal = _maternal;
	   _maternal = tmp;
	   return;
	}

VCF2diploid.java	if (\!var.isPhased()) var.randomizeHaplotype();

https://doi.org/10.1371/journal.pcbi.1010552.t001

5’872’394 SNPs 481’158 SNPs were “Het” (8%) and of the 1’093’496 indels 80’075 were Het (7%).

We also built a D2-specific reference genome based on the GRCm38 assembly and D2-specific SNPs and indels from dbSNP (Fig 1B “2 parental references”). We refer to this modified version as D2 reference, which differs from the D2 assembly in that the D2 assembly was assembled from DNA reads obtained in the D2 strain, whereas the D2 reference was a modified version of the assembly based on the B6 strain. We adapted the coordinates of the transcriptome annotation to the new coordinates for BXD and D2 references using the chain files generated by vcf2diploid and the liftOver tool (version 8.28) from UCSC (http://genome.ucsc.edu).

Read mapping and setting of mapping parameters

We performed read mapping with STAR (version 2.7.0e) [19] with different values for the parameters. The default values were used for permissive alignment, whereas more restrictive alignments were obtained by varying the settings of the parameters: “*--scoreDelOpen -40”* to prevent deletions, “*--scoreInsOpen -40”* to prevent insertions, “*--alignIntronMax 1”* to prevent introns (splicing), “*--alignEndsType EndToEnd”* to prevent partial alignment of the read, and “*--outFilterMismatchNmax 0”* to prevent mismatches (the value is the maximal number of mismatches allowed). We also varied the inclusion (with annotation) or exclusion (without annotation) of the transcriptome annotation in the genome index.

To count the uniquely mapped reads per gene after STAR, HTseq (version 0.6.1p1) was used with samtools (version 1.9) to convert alignments from bam to sam format. Only the alignments with a quality score of 10 or above were kept (default). The command used was:

```
samtools view -h alignment.bam | htseq-count -s reverse -t exon -m union-reference.gtf
```

The “-s reverse” parameter was used for the stranded library which is specific to the library preparation and sequencing protocol. Alternatively, for mapping using transcriptome annotation, the HTseq counting implemented in STAR was used with--quantMode GeneCounts.

Filtering and normalization of gene counts

Lowly expressed genes were filtered by tissue keeping only genes with counts per million (cpm) above 0.5 (min_cpm) for at least 20 samples (on a total of 66 BXD samples/tissue). Counts were normalized with the edgeR package (version 3.24.3) which uses the weighted trimmed mean of M-values (TMM) method to take into account the variation in library size and in RNA population [20] and log transformed (log2).

Differential mapping analysis of genes

To compare the impact of the reference on the gene expression, duplicated gene names were removed, and only gene names common to both GRCm38 and D2 or GRCm38 and BXD references transcriptome annotation were kept. A differential expression analysis was performed on the BXD samples using the voom function from R package limma (version 3.38.3). Notice that in each case, the two groups compared had exactly the same samples, so only the reference used for read mapping differed.

Local eQTL detection and comparison

QTL detection is sometimes referred to as QTL mapping, but we will avoid this terminology to avoid confusion with read mapping. Local eQTLs (often referred to as cis eQTLs) were detected using FastQTL (version 2.184) using 2 Mbps above and below transcription start site

(TSS). 1000 permutations were used to adjust p-values for multiple markers tested and seed 1 was chosen to help reproducibility. Correction for multiple gene testing was performed with R (version 3.4.2) package qvalue (version 2.10.0). The percentage of expressed genes that have a significant local eQTL serves to measures the improvement between the different values used for the mapping parameters tested in the evaluation. The slope (allelic mean difference, representing the direction and strength of allele-specific gene expression) of the linear regression and qvalue of eQTLs from different references were considered similar (unaffected) if they are within less than 5%:

$$\left|\frac{(X_{BXD} - X_{GRCm38})}{average(X_{BXD}, X_{GRCm38})}\right| < 0.05,$$

where X refer to slope or qvalue.

Reference bias

Assuming that D2 alleles are on average as much expressed as B6 alleles, significant eQTLs are as likely to have one or the other allele more expressed. The percentage of skewness of local eQTLs is thus calculated as follows to reflect reference bias (0% indicates no reference bias):

$$\frac{significant\ eQTLs\ with\ negative\ slope - significant\ eQTLs\ with\ positive\ slope}{expressed\ genes} \cdot 100,$$

where significant is defined as FDR < 5%.

Computational requirements

Some computations were performed on the Wally cluster of the University of Lausanne with the Vital-IT software stack (https://www.vital-it.ch) of the Swiss Institute of Bioinformatics for speed (parallelizing multiple mapping runs) and convenience (having a functional installation of FastQTL software). However, none of the steps require unreasonable memory or computational power, and all softwares used in this study are freely available for reproducibility purpose.

Results

To improve genetic coherency of RNA-seq read mapping, we explored two alternative strategies to exploit available data in the BXD panel. The first strategy used the two parental strains (B6 & D2) assemblies (Fig 1B "2 parental assemblies"). The second strategy used BXD-specific references obtained from the GRC assembly modified with BXD known and imputed variants (Fig 1B "BXD-specific references"). An intermediate between these two strategies was used for comparison: a D2 reference built from the GRC assembly modified with known D2 variants (Fig 1B "2 parental references"). For each strategy, we evaluated the impact on various downstream steps of the analytical pipeline by quantifying how the strategies affected mappability of the RNA reads (Figs 2A–2C and 3A and 3C), gene expression estimates (Figs 2D and 3B), and eQTLs (Fig 4). In addition, we have evaluated how key mapping parameters influence these results (Figs 2B and 2C and 5).

Mapping strategy with two parental strains assemblies

To explore the impact of using one reference for all samples despite their genetic differences, we mapped all samples on the classical GRCm38 (B6) genome assembly and on the more recent D2 assembly. We expected that more reads from D2 samples would be uniquely

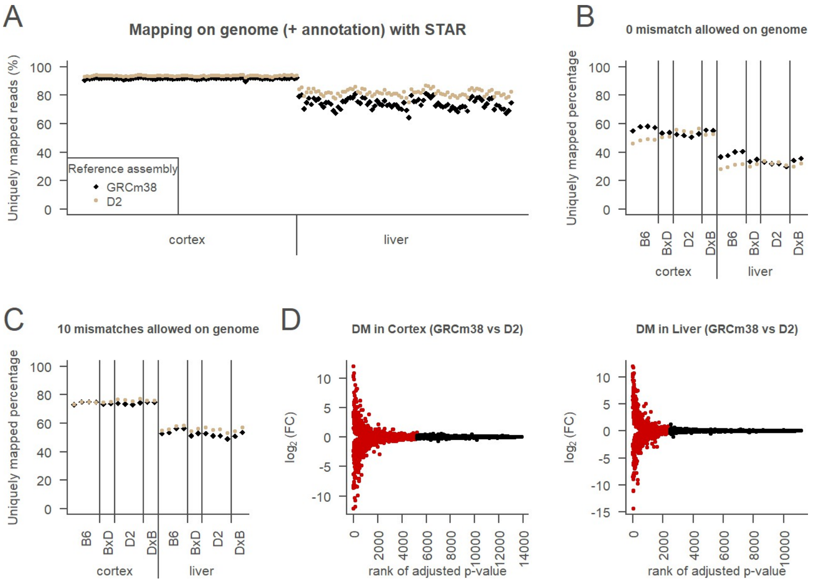


Fig 2. Two parental assemblies strategy. A. Mappability of all samples on 2 parental assemblies (samples are mapped on GRCm38: black symbols and on D2 assembly: brown symbols) using permissive mapping setting (STAR default) in cortex (left) and liver (right). Mappability was estimated as the number of uniquely mapped reads expressed as the % of all reads. B. Mappability in samples from the parental strains and their reciprocal F1 offspring (BxD and DxB) on the 2 parental assemblies using restrictive mapping setting allowing 0 mismatches. Same legend than in A. C. Mappability of parental and F1 samples on 2 parental assemblies using restrictive mapping setting but allowing up to 10 mismatches. Same legend than in A. D. Differential mapping (DM) analysis of D2 assembly compared to GRCm38 in the cortex (left) or in the liver (right). Genes are classified as DM genes if FDR adjusted p-value < 0.05 (red) or non DM genes otherwise (black).

https://doi.org/10.1371/journal.pcbi.1010552.g002

mapped on D2 assembly than on the GRCm38 assembly, and that reads from BXD samples would map approximately equally on both parental assemblies. Surprisingly, we observed that the percentage of uniquely mapped reads, used to estimate mappability, was higher for all samples when mapped to the D2 assembly compared to the GRCm38 assembly (Fig 2A), even for B6 samples. We also noticed that mappability differed between the liver and the cortex both in amplitude and in variance. This difference might relate to differences in the preparation of the two tissues for reasons inherent to the tissues, but all the samples passed the quality tests [17]. The liver samples had on average more raw reads than the cortex samples, but the sequencing depth did not seem to explain differences in mappability. It may be that there were more PCR artefacts in the liver reads, so they were either unmapped or multi-mapped which means there were less uniquely mapped reads than in the cortex. Another possibility is that the liver expresses more genes that have regions that are not unique, so more reads are multi-mapped. To further explore the bias for the D2 assembly, we allowed only exact matches. We now observed the expected strain-specificity as B6 samples mapped higher on B6 assembly and D2 samples mapped higher on D2 assembly (Fig 2B). Using up to 10 mismatches (the STAR default for 100 bp reads) but no insertions, deletions and trimming, we lost strain-specificity (Fig 2C). However, the more restrictive mapping setting also importantly reduced the number of uniquely mapped reads (Fig 2A–2C). This raised the question what choice of parameter settings ensures both high read yield and strain specificity (see part "Mapping parameters evaluation" below).

To determine the impact of mapping reference on gene expression, we performed a differential mapping (DM) analysis. The principle is the same as differential expression analysis, but the mapping references are compared instead of different groups or perturbations. Note that the reads and the values used for the mapping parameters are identical for both references, so differences observed are caused strictly by the reference. More than one third (38%) of genes were affected by the mapping reference (GRCm38 vs. D2) in the cortex and about a quarter (22%) in the liver (Fig 2D). Alignments of the top 4 highly affected genes were visually inspected and revealed variation in the quality of the assemblies and their transcriptome annotation at these precise places. Thus what appeared as differences in gene expression between the two assemblies could in some cases be artefacts and not consequences of genetic variants (examples of artefacts in S2 Fig and S3 Table).

Customizing reference for D2 and BXD lines

To avoid differences of quality and completeness between B6 and D2 assemblies (S2 Table), we modified the B6 reference assembly using SNPs and indels specific to the D2 strain from dbSNP. We mapped parental and F1 samples with exact matches on both the mm10 assembly and the mm10 assembly modified for D2. The percentage of uniquely mapped reads was increased when the samples were mapped to their corresponding strain reference, compared to the other parental strain reference (Fig 3A). Indeed, D2 samples gained between 4.6 and 5.7% when mapped to the customized D2 reference, whereas B6 samples lost between 3.4 and 5.8%. In contrast, when mapped on the D2 assembly (Fig 2B) D2 samples gained between 0.03% and 3.4% whereas B6 samples lost between 8.5% and 9.8%. The D2 customized reference thus appears more balanced as the gain for D2 samples is closer to the loss for B6 samples. In both cases, the difference between the two parental references was smaller for F1 samples, which is expected for a mix between the two parental strains. To apply the same strategy to

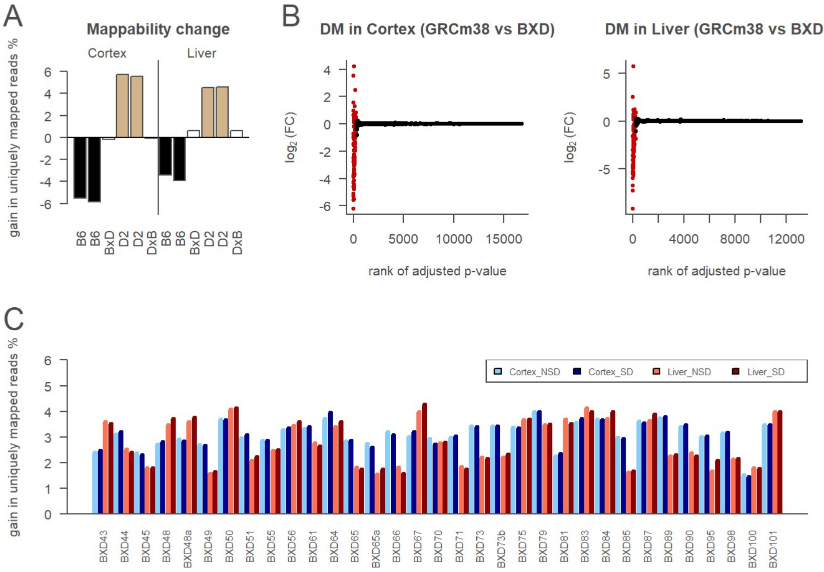


Fig 3. Line-specific references strategy. A. Relative mappability of customized D2-specific reference (GRCm38 modified with D2-specific indels and SNVs from dbSNP) compared to GRCm38 on parental and F1 samples with exact matches. Samples are all NSD. Colors indicate genetic of the samples: B6 (black), D2 (light brown), and F1 (white) between B6 and D2 strains. The F1 samples are BxD if the mother is B6 and the father is D2 (as for the BXD lines), or the reverse for DxB. B. Differential mapping (DM) analysis of BXD-specific references compared to GRCm38, in the cortex (left) or in the liver (right). Genes are classified as DM genes if the FDR adjusted p-value < 0.05 (red) or non DM genes otherwise (black). C. Relative mappability of BXD-specific references (GRCm38 modified for each BXD line with GeneNetwork genotypes and imputed variants) compared to GRCm38 on BXD samples with exact matches.

https://doi.org/10.1371/journal.pcbi.1010552.g003

BXD samples, genotypes were imputed using the large amount of D2-specific variants from dbSNP and the BXD specificity of genotypes from GeneNetwork. All BXD samples gained between 1.4 and 4.3% in mappability from having a customized reference by BXD line (Fig 3C). The amplitude of the gain varied among BXD lines and between tissues with cortex samples having globally higher values than liver samples. Note that it was expected the gain to be lower for BXD samples than for D2 samples, since in BXD lines approximately half of the alleles are D2.

To determine the impact of reference customization on gene expression in the BXD lines, we performed a differential mapping analysis. Around 2% of genes are affected by the mapping reference in the cortex and in the liver (Fig 3B).

Consequences of customization on local eQTL detection

To evaluate the effect of reference customization on estimated biological phenotypes by downstream analysis, we detected local eQTLs using gene expression estimated with the B6 reference or the BXD-specific references. The eQTLs are particularly likely to be influenced since they link gene expression to genetic variants. Significant eQTLs can be seen as a signal-to-noise measure of genetically structured gene expression. The percentage of significant eQTLs was slightly higher (0.1% difference) when using BXD-specific references than when using B6 assembly (Fig 4A). However, this does not necessarily mean that the same eQTLs were detected when using the two different references. The results can differ by the genetic marker associated with the gene expression, the direction of gene expression (whether the gene expression is higher with B6 or D2 allele), or change in the q-value (Figs 4C and S4). When

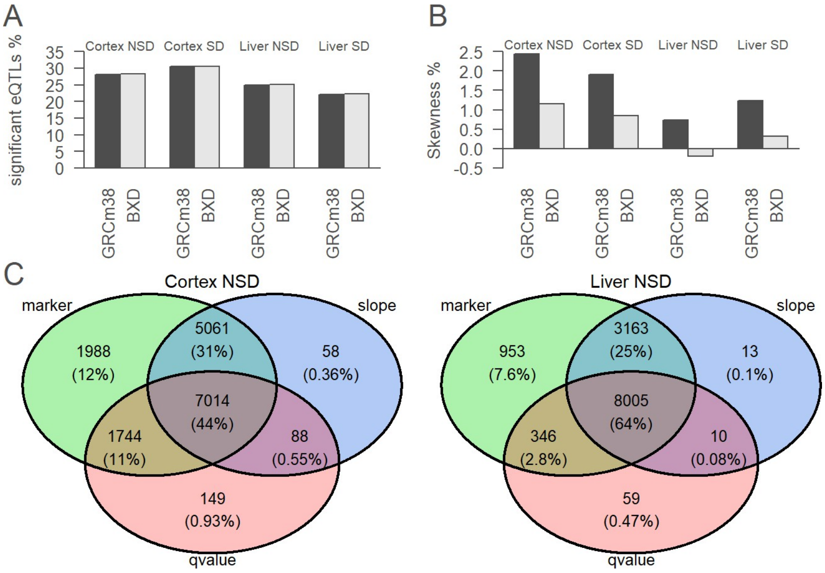


Fig 4. Consequences of mapping reference at local eQTL level. A. Percentage of significant (FDR 5%) local eQTLs over all expressed genes with GRCm38 or BXD-specific references. B. Percentage of skewness of significant (FDR 5%) local eQTLs slope over all expressed genes with GRCm38 or BXD-specific references. C. For all expressed genes, the best local genetic marker to explain gene expression was selected. The Venn diagrams represent the overlap of this analysis between GRCm38 and BXD-specific references for the three criteria in cortex NSD (left) or in the liver NSD (right). The marker (in green) indicates changing the reference result in the same genetic marker associated with gene expression. The slope (in blue) is the direction and strength of allele-specific gene expression, it is considered to be overlapping between the references if it varies less than 5%. The qvalue (in pink) is the statistical significance of the marker to gene expression association, it is considered to be overlapping between the references if it varies less than 5%.

https://doi.org/10.1371/journal.pcbi.1010552.g004 considering these 3 variables, mapping reference did not affect 44% of local genetic marker to gene expression association in the cortex and 64% in the liver.

Reference bias

Next, we wanted to detect a potential reference bias, where reads containing B6 alleles get more easily mapped than those containing D2 alleles or the contrary. In DNA-seq studies, this can be achieved by checking the symmetry of the distribution of allelic ratios at heterozygous loci. In RNA-seq, allele-specific expression can also modify allelic ratio. However, we assumed that globally genes with B6 or D2 alleles are equally expressed. Moreover, since our samples are inbred lines, heterozygous sites are rarer than in other populations, so we compared homozygous alleles of genetically different samples, rather than heterozygous alleles from one sample. The percentage of skewness represents how many local eQTLs deviate from a situation without reference bias (skewness 0%). A positive percentage indicates a B6 bias: more eQTLs with the B6 allele increasing gene expression, whereas a negative percentage indicates a D2 bias: more eQTLs with the D2 allele increasing gene expression. Using the B6 reference shows a reference bias for B6 alleles in all tested tissues and conditions while using BXD-specific references decreased bias (Fig 4B).

Mapping parameters evaluation

The reference used is not the only thing influencing read mapping. We used the BXD-specific references, and to test which values to use for the more critical mapping parameters of STAR we varied: i) the number of mismatches allowed, ii) the possibility to trim end of reads, iii) the possibility to splice reads, and iv) the use of known transcriptome annotation. The ratio of significant eQTLs to expressed genes (Fig 5) was used for performance optimization as done previously [14,21]. A higher ratio indicates a higher proportion of genetically structured gene expression versus random variations in gene expression. The best settings in both tissues are to use trimming, splicing, and transcriptome annotation and to allow mismatches. The exact maximal number of mismatches differed: 10 mismatches in the cortex (Fig 5A), but only 2 in the liver (Fig 5B). All top settings use existing transcriptome annotation and thus appears to be the more important parameter.

Discussion

Genomic variations among individuals are the core of genetic studies. Yet it is common practice in the field to use one assembly as reference for all genetically different samples. Here, we improved genetic specificity of read mapping of BXD samples using publicly available data. Our custom BXD-specific references detected proportionally more eQTLs and alleviated reference bias. Below we will discuss the complexity of assessing the analytical design of RNA-seq and the various strategies to integrate genomic variations in transcriptomic analyses.

Although the analysis of RNA-seq data is often regarded as well established, it remains a complex procedure. A great number of factors are involved, going from experimental design (number of replicates, kit for library preparation, sequencing platform, read length) to softwares, functions, and settings used in the different analytical different steps (quality control, mapping, filtering, normalization). We are not being exhaustive about all these aspects, but nevertheless think our observations and considerations on selected features are useful for the community. One of these observations concerned the unanticipated tissue differences in mappability. We were unable to identify supportive literature for this phenomenon and think it merits a more thorough analysis addressing such tissue effects (e.g. using GTEx).

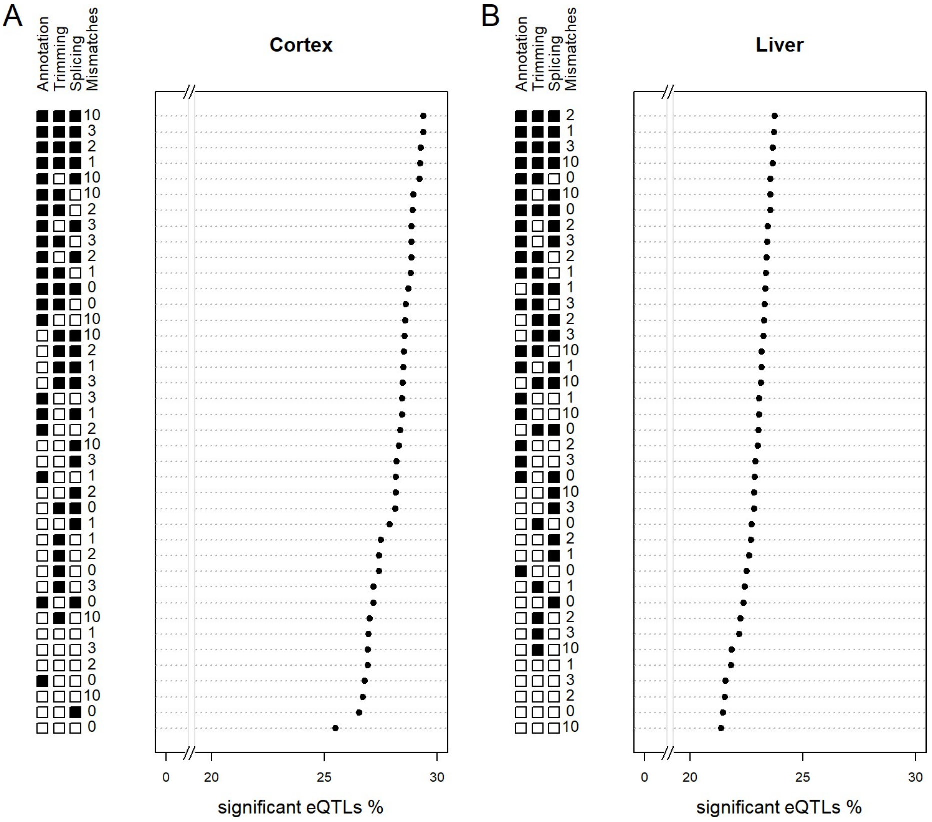


Fig 5. Evaluating mapping parameters. A. The performance on local eQTLs of selected mapping settings on cortex samples (average of the NSD and SD conditions) is measured by the percentage of expressed genes that have a significant local eQTL. The BXD-specific references were used. C. As in A but for liver samples.

https://doi.org/10.1371/journal.pcbi.1010552.g005

Another difficulty is that for mapping of real samples, the true location of reads is unknown. The fraction of uniquely mapped reads is used as a mapping statistic because an RNA molecule can only come from one locus. However, this does not guarantee the correctness of mapping of uniquely mapping reads. Importantly, some reads are expected to be correctly classified as multi-mappers because some regions in the genome are identical or very similar (e.g. repeat elements). Moreover, uniqueness can have slightly different meanings depending on the mappers and parameters used, as reads are not necessarily exactly and fully aligned because of mismatches, indels, and sequencing errors especially at the ends of reads (trimming).

Most RNA-seq studies use standard analytical pipelines with default setting, or with slight modifications such as the number of mismatches allowed. Mismatches have the task to compensate for sequencing errors or small unknown variants to give some flexibility in case an exact match is not found. However, the choice of the number of mismatches allowed is rarely given, even though it has been shown in humans that reducing the number of mismatches allowed increased the difference in uniquely mapped reads when using a general versus an individual-specific reference [10]. We showed that the mapping settings have an effect on mappability and also on the local eQTLs. An interesting benchmark was completed on human RNA-seq data in a differential gene expression (DE) pipeline [22]. They compared filtering methods, along with transcriptome reference sets, normalization and DE detection, alignment and counting software. The authors concluded that the optimal filtering threshold depended on the pipeline parameters. Particularly, the mapping software had the least impact whereas the transcriptome annotation had the higher impact. Also, our study highlighted the importance of transcriptome annotation even if our evaluating measure differed. By focusing on the impact of genetic differences in the reference we could assess different combinations of mapping options. We used the fraction of eQTLs as evaluation measure rather than DE, because we compared B6 versus D2 alleles in a genetic population and not two fixed groups. Indeed, each BXD sample will be considered to be B6 at some loci and D2 at others. A reference bias is likely to influence eQTLs because a variant in a certain gene can modify the mapping of the reads precisely for the samples that have the alternative allele in this region.

An assembly is a global solution because it uses all genomic variants specific to a strain regardless of their size. However, we observed that the technical difference in quality between the two parental assemblies prevents a fair comparison of the genetic difference. In this context it can be noted that we can exclude mix-up of samples as a possible cause of D2 bias in assembly mapping, because it affected all samples: BXD lines, B6 and D2, and because our previous analysis confirmed that this had likely not occurred [18]. Even if transcriptome annotation is likely to be similar in other strains, the different coordinate systems between the assemblies complicate the transfer. Moreover, the actual D2 transcriptome annotation corresponding to the D2 assembly includes manual curation steps that make it hard to update to new releases of the genome assembly or variants. Notably, no study was published using this D2 assembly, except the one from the group that released it [15]. In contrast, our customization of one assembly offers the advantage that the coordinate changes are formalized which allows automatization of transcriptome annotation changes with the same tool used for upgrading versions of an assembly (liftOver). For mapping reads of D2 samples and those of other strains than B6, we currently recommend the use of GRCm38 assembly modified with strain-specific indels and SNVs from dbSNP.

Our custom references combine the specificity of BXD genotypes with the large amount of D2-specific short variants from dbSNP. Importantly, we did not include structural variants (SVs), although many were detected between B6 and D2 strains [23]. SVs can have important phenotypic impacts [24], potentially more than SNVs [25,26]. However, SVs calls will require further efforts in the reporting to ensure the confidence and the format for integration into current workflows. This is due largely to the nature of SVs: their length and large variety implies that the possible number of SVs is greatly superior to that of SNVs, making them less easy to validate and report. Without technologies like long read sequencing and optical genome mapping, those SVs will be very likely inaccessible for mouse models unless an international consortium tackles this issue.

Another limitation is that all murine assemblies are haploid whereas mice are diploid. The diploidy is ignored at the mapping step under the assumption that the genotypes of inbred strains are mostly homozygous. However, the homozygosity and stability of inbred mouse strains is based on a theoretical model that does not consider new mutations [27], although germline mutations are estimated to be between 10 and 30 per generation [28,29]. Unfortunately, the assumption of stability of inbred lines is so strongly anchored in the field that its verification is compromised, because of not searching for heterozygous sites or dismiss them. Indeed, the term heterozygous is sometimes used to call variants uncertain or low quality, and they are always unphased. When mouse assemblies were built, regions with high density of heterozygous sites were used to detect (haploid) assembly errors, ignoring the potential coherence of diploid or polyploid references [15]. A more systematic detection and characterization of heterozygous regions will likely improve the accuracy of transcriptomics studies, particularly for loci with allele-specific expression. However, the read mapping of different possible alleles, which could also be used for not inbred crosses between 2 strains would require a reconciliation step, as implemented for example for human with paternal and maternal allele [9]. Indeed, every read can come from either one of the two alleles but not from both at the same time. We made an effort to improve the D2 parts of reference to map the BXD samples, however the B6 strain itself is also susceptible to mutations, as confirmed by the occurrence of many B6 substrains, even if it affects only a few genes. Therefore, a complete characterization of genetic variants of the BXD by DNA-sequencing could improve the customization of both D2 and B6 parts of BXD, and therefore enhance resolution of downstream analyses.

Conclusion

In current genetic studies using the BXD population, genomic variations are paradoxically ignored at the read mapping step, which as we show here causes a reference bias. The genomic variations need to be explicitly integrated in the reference instead of treated as sequencing errors. Our results show the need for a critical evaluation of the RNA-seq pipeline and the development of more complete genomic variants databases to best approximate the genetics of the samples. Most genetic studies with a transcriptomic component in mice and other model organisms can suffer from reference bias, which could be attenuated by assessing and sequencing those strains. The mouse community could follow the drosophila community (http://dgrp2.gnets.ncsu.edu) and sequence genetic reference populations. Our study can serve as a wake-up call for improving the characterization of genomic variations, and as a concrete guide for analyses in BXD and other genetic populations. As RNA-seq analyses are often a starting point to identify one or a few genes that then are studied in more detail in follow-up experiments, it is worth the extra effort to avoid potential bias by not blindly following traditional pipelines.

Supporting information

S1 Table. Mouse replicates.
(CSV)

S2 Table. Summary statistics of the mouse genome assemblies. Assembly summary statistics calculated with GAAS toolkit (https://github.com/NBISweden/GAAS).
(CSV)

S3 Table. Broad classification of DM genes according to the two types of artefacts identified. S2 Fig provides examples of such artefacts.
(CSV)

S1 Fig. Consequences of mapping reference transcriptome at read mapping level. A. Mappability of all samples on 2 parental assembly transcriptomes using STAR permissive mapping setting. B. Pseudo-mappability of all samples on 2 parental assembly transcriptomes using Kallisto. C. Mappability of parental and F1 samples on 2 parental assembly transcriptomes using STAR restrictive mapping setting. D. Mappability of parental and F1 samples on 2 parental assembly transcriptomes using STAR restrictive mapping setting, but up to 10 mismatches.
(TIFF)

S2 Fig. Examples of artefacts of assemblies sequence and annotation. A. *Nova2* genomic region in Integrative Genomics Viewer (IGV https://software.broadinstitute.org/software/igv/), as a transcriptome annotation artefact. The coverage is very similar between GRCm38 and D2 assemblies, but the annotation differs, which causes the reads to be counted differently. B. *Gm15564* genomic sequence in IGV (https://software.broadinstitute.org/software/igv/), as an

artefact due to difference in completeness of genome assembly. Many reads map to this region on GRCm38 assembly, but not on D2. It appears that in this region of the D2 assembly there are three unknown nucleotides (with label "N"), which supports the interpretation that it is probably due a difference in assembly quality, and not to a genomic variant.
(TIFF)

S3 Fig. Genotype imputation workflow.
(TIFF)

S4 Fig. Consequences of mapping reference at local eQTL level in SD condition. A. For all expressed genes, the best local genetic marker to explain gene expression is selected. The Venn diagrams represent the overlap of this analysis between GRCm38 and BXD-specific references for the three criteria in cortex SD. The marker (in green) indicates changing the reference result in the same genetic marker associated with gene expression. The slope (in blue) is the direction and strength of allele-specific gene expression, it is considered to be overlapping between the references if it varies less than 5%. The qvalue (in pink) is the statistical significance of the marker to gene expression association, it is considered to be overlapping between the references if it varies less than 5%. B. Same than A but in the liver SD.
(TIFF)

Acknowledgments

We thank Mark Ibberson and Brian Stevenson for the helpful discussions and guidance.

Author Contributions

Conceptualization: Nastassia Gobet, Maxime Jan, Paul Franken, Ioannis Xenarios.

Data curation: Nastassia Gobet, Maxime Jan.

Formal analysis: Nastassia Gobet.

Funding acquisition: Paul Franken, Ioannis Xenarios.

Investigation: Nastassia Gobet.

Methodology: Nastassia Gobet, Maxime Jan, Paul Franken, Ioannis Xenarios.

Project administration: Nastassia Gobet, Ioannis Xenarios.

Resources: Paul Franken.

Software: Nastassia Gobet.

Supervision: Maxime Jan, Paul Franken, Ioannis Xenarios.

Validation: Nastassia Gobet, Ioannis Xenarios.

Visualization: Nastassia Gobet.

Writing – original draft: Nastassia Gobet.

Writing – review & editing: Nastassia Gobet, Maxime Jan, Paul Franken, Ioannis Xenarios.

References

1. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing Reference Genome Assemblies. PLOS Biol. 2011 Jul 5; 9(7):e1001091.

2. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin CS, et al. Extending reference assembly models. Genome Biol. 2015 Jan 24; 16(1):13. https://doi.org/10.1186/s13059-015-0587-3 PMID: 25651527

3. Liu X, MacLeod JN, Liu J. iMapSplice: Alleviating reference bias through personalized RNA-seq alignment. PLOS ONE. 2018 Aug 10; 13(8):e0201554. https://doi.org/10.1371/journal.pone.0201554 PMID: 30096157

4. Rivas-Astroza M, Xie D, Cao X, Zhong S. Mapping personal functional data to personal genomes. Bioinformatics. 2011 Dec 15; 27(24):3427–9. https://doi.org/10.1093/bioinformatics/btr578 PMID: 22006915

5. Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. Genome Biol. 2020 May 25; 21(1):124. https://doi.org/10.1186/s13059-020-02038-8 PMID: 32450900

6. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019 Aug 9; 20 (1):159. https://doi.org/10.1186/s13059-019-1774-4 PMID: 31399121

7. Chen NC, Solomon B, Mun T, Iyer S, Langmead B. Reducing reference bias using multiple population reference genomes. bioRxiv. 2020 Mar 7;2020.03.03.975219.

8. Geijn B van de, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015 Nov; 12(11):1061–3. https://doi.org/10.1038/nmeth.3582 PMID: 26366987

9. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011 Jan 1; 7(1):522.

10. Yuan S, Johnston HR, Zhang G, Li Y, Hu YJ, Qin ZS. One Size Doesn't Fit All—RefEditor: Building Personalized Diploid Reference Genome to Improve Read Mapping and Genotype Calling in Next Generation Sequencing Studies. PLOS Comput Biol. 2015 Aug 12; 11(8):e1004448. https://doi.org/10.1371/journal.pcbi.1004448 PMID: 26267278

11. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec; 420(6915):520–62. https://doi.org/10.1038/nature01262 PMID: 12466850

12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1; 29(1):308–11. https://doi.org/10.1093/nar/29.1.308 PMID: 11125122

13. Wang J, Geisert EE, Struebing FL. RNA sequencing profiling of the retina in C57BL/6J and DBA/2J mice: Enhancing the retinal microarray data sets from GeneNetwork. Mol Vis. 2019 Jul 5; 25:345–58. PMID: 31354228

14. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, et al. RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations. Genetics. 2014 Sep 1; 198(1):59–73. https://doi.org/10.1534/genetics.114.165886 PMID: 25236449

15. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet. 2018 Nov; 50(11):1574–83. https://doi.org/10.1038/s41588-018-0223-8 PMID: 30275530

16. Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet. 2004 Apr 29; 5:7. https://doi.org/10.1186/1471-2156-5-7 PMID: 15117419

17. Diessler S, Jan M, Emmenegger Y, Guex N, Middleton B, Skene DJ, et al. A systems genetics resource and analysis of sleep regulation in the mouse. PLOS Biol. 2018 Aug 9; 16(8):e2005750. https://doi.org/10.1371/journal.pbio.2005750 PMID: 30091978

18. Jan M, Gobet N, Diessler S, Franken P, Xenarios I. A multi-omics digital research object for the genetics of sleep regulation. Sci Data. 2019 Oct 31; 6(1):1–15.

19. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1; 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

20. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010 Mar 2; 11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867

21. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet. 2021 Jan; 53(1):120–6. https://doi.org/10.1038/s41588-020-00756-0 PMID: 33414550

22. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf. 2015;2015:6461–4.

23. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 2010 May; 20 (5):623–35. https://doi.org/10.1101/gr.102970.109 PMID: 20308636

24. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019 Nov 20; 20(1):246. https://doi.org/10.1186/s13059-019-1828-7 PMID: 31747936

25. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011 Sep; 477(7364):289–94. https://doi.org/10.1038/nature10413 PMID: 21921910

26. Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. bioRxiv. 2021 Mar 8;2021.03.06.434233. https://doi.org/10.1101/gr.275488.121 PMID: 34544830

27. Casellas J. Inbred mouse strains and genetic stability: a review. animal. 2011 Jan; 5(1):1–7. https://doi.org/10.1017/S1751731110001667 PMID: 22440695

28. Casellas J, Medrano JF. Within-Generation Mutation Variance for Litter Size in Inbred Mice. Genetics. 2008 Aug; 179(4):2147–55. https://doi.org/10.1534/genetics.108.088070 PMID: 18660537

29. Reardon S. Lab mice's ancestral 'Eve' gets her genome sequenced. Nat News. 2017 Nov 16; 551 (7680):281. https://doi.org/10.1038/nature.2017.22974 PMID: 29144484

Chapter 4

Knowledge network

Many methods for different omics are based on independently identifying key biological features, and after trying to find out how they interact. This may manage to capture a strong causal gene, but the risk is to miss smaller impact genes that are working together. Can we use the prior knowledge of the interactions between biological items to retrieve full subnetworks implicated in the regulation of sleep and combine the strength of different molecular layers? The goal of this chapter is to assemble a knowledge graph to consider interactions and to integrate data-driven and knowledge-driven inputs for the identification of multi-gene regulation subnetworks of the sleep phenotypes.

The primary article presenting this BXD sleep resource offers an analytical method to find candidate gene in single locus QTLs [Diessler et al., 2018]. The high heritability of sleep phenotypes is however not fully explained. Another data-driven approach, that I took for example with MOFA, aims at using embeddings to combine features possibly across modalities (omics) to differentiate the biological signals from the experimental noise. The issue with any purely data-driven is that it works independently of previous knowledge acquired and curated over the years. Regulations between genes, interactions of proteins, and reactions of metabolites have been long studied. There are not all discovered but there are good chances that the underlying network of biological known interactions can guide future discoveries to build multi-variate hypotheses explaining complex phenotypes.

4.1 Methods

4.1.1 Overview

A purely data-driven approach is explained with MOFA and embeddings to provide contrast to the main network approach taken with the following workflow (Figure 4.1). **Extraction**: The mouse-specific interactions were retrieved from databases. The only exception is for Rhea where all reactions were retrieved because the reactions do not have an organism specified, but the enzymes retrieved are mouse-specific. **Structuring**: The nodes were structured in the form ”type|identifier” where the identifiers would be ENSEMBL gene identifier for genes and proteins, and ChEBI for metabolites. The concept of metabolite complexes was used for the sides of the reactions. **Harmonization**: The identifiers of the genes and metabolites were mapped to ENSEMBL gene identifiers and ChEBI, respectively. For metabolites, manual processing was required and a pH of 7.3 was chosen to match Rhea metabolites. **Path detection**: Paths up to a certain length are listed for all pairs of genes in the network. **Scoring**: A BXD-expression score was calculated by path and by BXD line. **Ranking by line**: The lines were ordered for each sleep phenotype and for each path. **Correlation**: The Kendall correlation was used to compare BXD lines rankings of sleep phenotypes and paths.

4.1.2 MOFA

To build the MOFA model, 3 datasets of molecular (intermediate) phenotypes were used as input: the top 500 most variable genes in cortex in log2(CPM), the top 500 most variable genes in liver in log2(CPM), and the metabolites concentrations averaged by BXD line (124 metabolites). The samples considered are combinations of BXD line and condition: not sleep deprived (NSD) or sleep deprived (SD). MOFA package (version 1.1.1) was used for the analysis with R (version 3.5.3).

4.1.3 Retrieving datasets from databases

IntAct is a reviewed database of molecular interactions, mainly protein-protein interactions, based on experiments reported in literature. Version 2021-10-13 was downloaded from IntAct website, then only mouse specific interactions were kept by selecting interactions where both interactors have taxid 10090 (Mus musculus), for a total of 30'391 interactions. Notice that some taxid of the interaction may be different depending on the experiment from which the information come from (for example: -1 in vitro). The interactors identifiers were converted to ENSEMBL gene id (ENSG) using the gconvert function from R Bioconductor package gprofiler2, using "mmusculus" as organism and giving maximum one result per identifier (mthreshold=1). Interactions where at least one interactor was not converted were removed.

STRING is a database of functional protein-protein interactions known and predicted, based on other databases, ortholog predictions, and text mining of literature. Due to the diversity of evidence and prediction, there is a confidence score for each type of evidence, and they are all combined in a combined score, going from 0 to 1000 (or 0 to 1), where a higher value indicates a higher confidence in the interaction. However, interaction does not mean physical interaction. There is a R Bioconductor package for the STRING database (STRINGdb) but limited in which information are retrievable. The database for mouse version 11.5 was downloaded ("10090.protein.links.full.v11.5.txt") from STRING website, giving a total of 14'496'358 interactions. The interactors identifiers were converted to ENSEMBL gene id (ENSG) using the gconvert function from R Bioconductor package gprofiler2, using "mmusculus" as organism and giving maximum one result per identifier (mthreshold=1). Interactions where at least one interactor was not converted were removed. One version of the knowledge graph included STRING interactions and one did not.

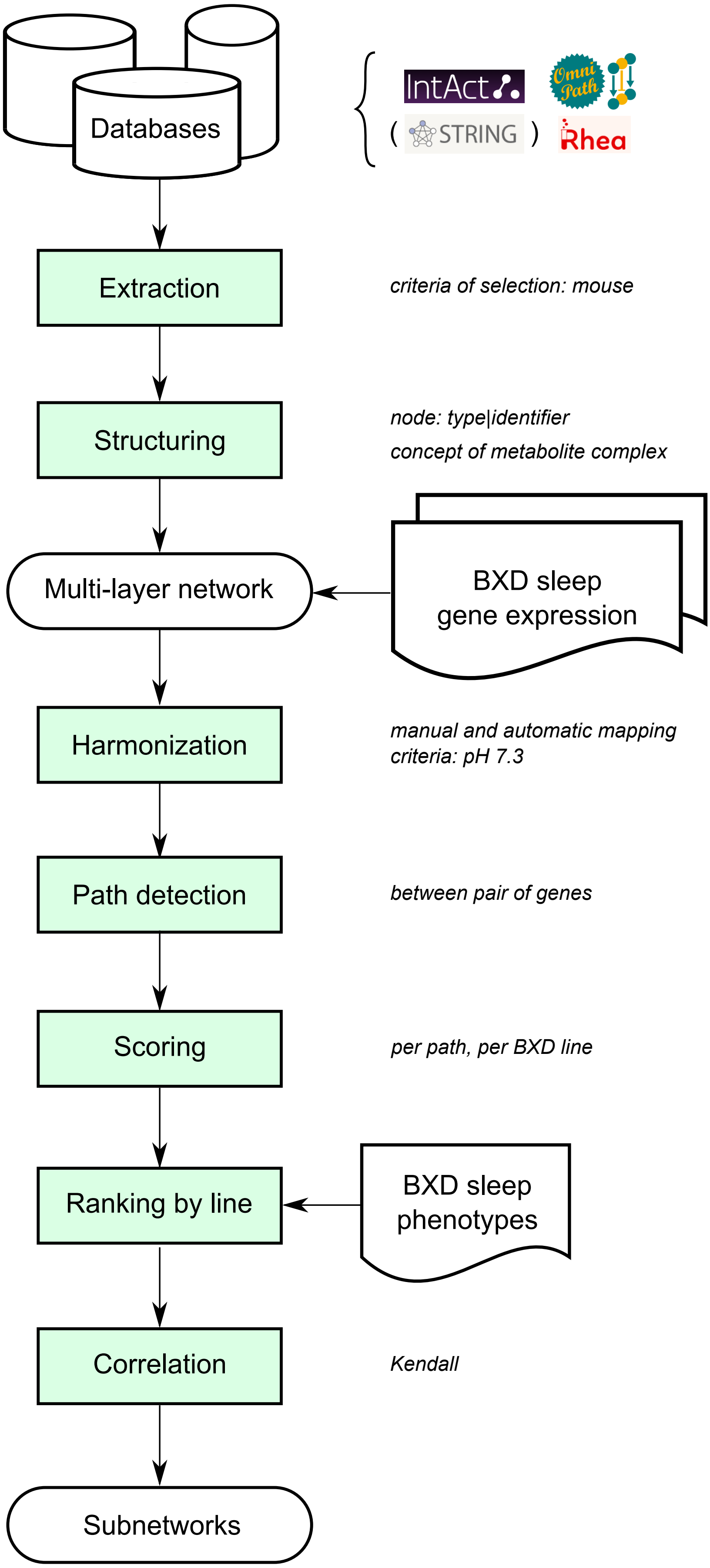


Figure 4.1: Flowchart of the main network approach.

Omnipath is a biological processes database assembled for human but available via homology for mouse and rat [Türei et al., 2021]. Interactions were downloaded using the R bioconductor package: OmnipathR version 3.2.8, with the function import_all_interactions, specifying 10090 for mouse as organism. This gives 144'453 directed protein-protein interactions of 14'643 interactors coming from over 100 resources. The interactors identifiers were converted to ENSEMBL gene id (ENSG) using the gconvert function from R Bioconductor package gprofiler2, using "mmusculus" as organism and giving maximum one result per identifier (mthreshold=1). On the 14'643 interactors, 13'480 were converted to gene id. After removing interactions where at least one interactor was not converted, there were 122'185 interactions, for 13'375 interactors.

Rhea is a metabolic reactions database not specific to a species or tissue [Bansal et al., 2022]. Downloaded release 121 (2022 03 02) from rhea website, which contains 14'281 reactions of 12'385 unique interactors. The sides of reactions were called metabolite complexes and the interactors were called metabolites. The network is built with two types of edges: the reactions linking two metabolite complexes (the two can be the same in the case of transport reaction for example) and the membership relations indicating which metabolite belongs to which metabolite complex(es) where one metabolite complex can have one or more metabolites and a metabolite can be in one or more metabolite complexes. The reactions are undirected, what is the left and right sides of the reactions is arbitrarily determined (since release 90 by the alphabetic order of the metabolites names). Metabolite names were converted to ChEBI identifiers at pH 7.3 using the chebiId_name.tsv mapping file from rhea website. The macromolecules or polymers for examples cannot be converted to ChEBI ids. Only the reactions where all the interactors were converted are kept. The enzymes to rhea information from the rhea2uniprot_sprot.tsv was used to identify mouse specific catalysed reactions. The enzymes were converted from uniport id to ENSG id using the gconvert function from R Bioconductor package gprofiler2, using "mmusculus" as organism and giving maximum one result per identifier (mthreshold=1). For the 1.3% of enzymes that were converted, the metabolite complexes on both sides were considered to be linked to the enzyme. Some metabolites were mapped to SwissLipids by Alan Bridge.

PubTator is a database of scientific literature [Wei et al., 2019]. The genes to citations dataset was downloaded. The number of citations was counted. The NCBI gene identifiers were converted to ENSEMBL gene identifiers using the NCBI key table (gene2ensembl.gz).

4.1.4 Structuration of the knowledge graph

The nodes are called with their type and identifier, for example: gene|ENSMUSG00000032265. For the reactions, the metabolites designate the molecular species, while the term metabolite complex was chosen for side of the reaction, by analogy to protein complex. The datasets were merged and only one occurrence of duplicate edges was kept. Only the maximal component of the network was kept to avoid to have multiple unconnected subgraphs.

4.1.5 Network analysis

All paths up to a maximal length of 2 edges were searched for all gene pairs present in the network. For each path a coverage (or presence) score was given. It is calculated as the number of nodes with a gene expression available in our BXD data, divided by the number of nodes in the path (path length). A BXD-expression score is given to each path, for each BXD line. It is calculated as the sum of the gene expression of the nodes divided by the number of nodes in the path. The BXD-expression scores are used to rank the BXD for each path. The sleep phenotypes are also used to rank the BXD for each path. For each path the Kendall correlation is taken between the BXD-expression rank and the sleep phenotypes rank. For each sleep phenotype, the top 100 paths with the higher absolute value of correlation coefficient are selected. The subgraphs represents the list of these top 100 paths. The subgraphs were aggregated by tissue and condition and the presence or absence of each node and edge was counted.

4.2 Results and Discussion

4.2.1 Data-driven approach with MOFA

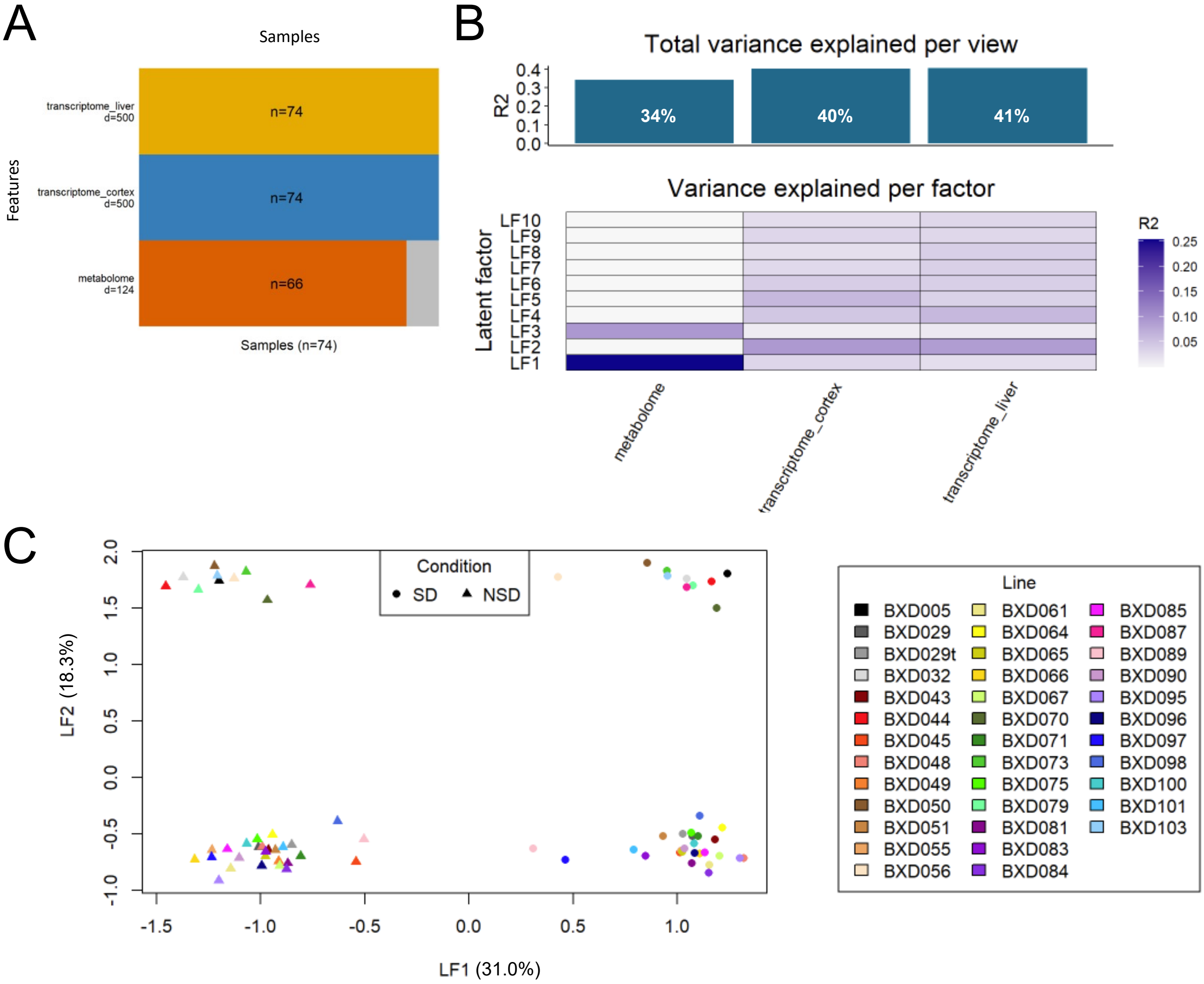


Figure 4.2: MOFA simple use on molecular data

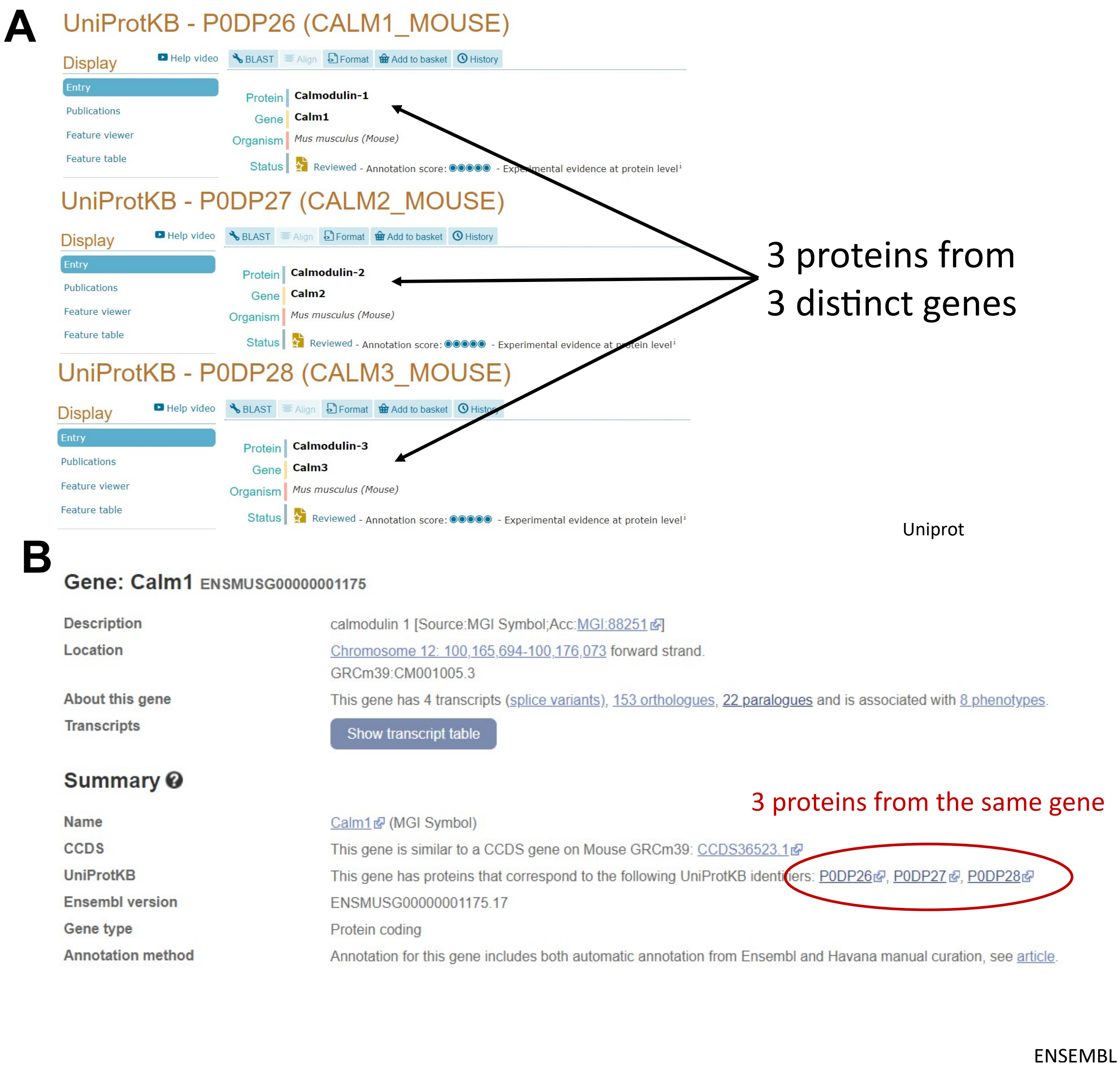
The metabolomics data and the transcriptomics (cortex and liver) data allows to create a model differentiating the sleep-deprived and non-sleep-deprived samples. A) Input data used for the model. B) Variance explained by the model. C) Scatter plots of the samples projected on the two first latent factors of the model, colored by BXD line and shaped by condition.

Combining the three molecular modalities using MOFA allows to split the samples by condition (first factor), and to split the BXD lines into 2 groups (second factor) (Figure 4.2). The lines in the upper group are: BXD005, BXD032, BXD044, BXD050, BXD056, BXD070, BXD073, BXD079, BXD087, and BXD103. This unsupervised method shows both the genetic and environment (external perturbation) have an strong impact at the molecular level. Both the complementarity of data inputs and the ability of the algorithm seem to be relevant. Now as powerful as the data-driven approach we still need then to use the current biological knowledge to interpret it. Now instead of first considering the data-driven way (for example assuming the different metabolites are independent even though we know they are connected), and then adding the knowledge on it, we decided to take another approach where we start from our data and the scientific knowledge (which is actually also data but structured into databases).

4.2.2 Building a multi-layer knowledge graph

During the building of the knowledge graph, the conversion of uniprot identifiers from IntAct database to ENSG gene identifiers lost 16% (using gprofiler2 R package) or 13% (using STRING aliases table) of proteins. The conversion of uniprot identifiers from Omnipath to ENSG gene identifiers lost 8% using gprofiler2 R package, but only less than 1% when starting from gene names. However, since the gene name was present in the initial entry, we can check if the ENSG gene id obtained allows are converted back to the same gene name and in 13% (if starting from uniprot id) or 4% (if starting from gene name) the initial and final gene names are different. The reconversion does not not always give the original result because of inconsistencies in the databases. For example, the gene Calm1 has one ENSEMBL entry linked to 3 proteins, whereas these 3 proteins have separate Uniprot entries each linked to one gene: respectively Calm1, Calm2, and Calm3 (Figure 4.3).

Mapping our metabolome data on the network yielded many losses. Starting from 124 metabolites, 92 lipids had a SwissLipids identifier, in which: 64 had an attached ChEBI identifier on SwissLipids, 85 had an attached ChEBI identifier on MetaNetX, 6 lipids had no ChEBI on both SwissLipids and MetaNetX tools. So 118 metabolites were translated to ChEBI identifiers. Now on those only 30 metabolites are found in the knowledge network. This is due to the fact that there are not many lipids in RHEA. The organisation of the information in the databases, such as which chemical species is more present at pH 7.3 is not always available in



UniProtKB - P0DP26 (CALM1_MOUSE)

Protein: Calmodulin-1
Gene: Calm1
Organism: *Mus musculus (Mouse)*
Status: Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

UniProtKB - P0DP27 (CALM2_MOUSE)

Protein: Calmodulin-2
Gene: Calm2
Organism: *Mus musculus (Mouse)*
Status: Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

UniProtKB - P0DP28 (CALM3_MOUSE)

Protein: Calmodulin-3
Gene: Calm3
Organism: *Mus musculus (Mouse)*
Status: Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

3 proteins from 3 distinct genes

Uniprot

Gene: Calm1 ENSMUSG00000001175

Description	calmodulin 1 [Source:MGI Symbol;Acc:MGI:88251]
Location	Chromosome 12: 100,165,694-100,176,073 forward strand. GRCm39:CM001005.3
About this gene	This gene has 4 transcripts (splice variants), 153 orthologues, 22 paralogues and is associated with 8 phenotypes.
Transcripts	Show transcript table

Summary

Name	Calm1 (MGI Symbol)
CCDS	This gene is similar to a CCDS gene on Mouse GRCm39: CCDS36523.1
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: P0DP26, P0DP27, P0DP28
Ensembl version	ENSMUSG00000001175.17
Gene type	Protein coding
Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.

3 proteins from the same gene

ENSEMBL

Figure 4.3: Example of inconsistency between databases
A) Uniprot entries (arrows and text on the right added for highlight). B) ENSEMBL entry (highlights added in red).

the same place so it makes more manual steps to curate the entries.

4.2.3 Network analysis

The full network without STRING contains 54'259 nodes and 197'436 edges. It has a diameter of 14 (longest shortest path is 14 edges) and a mean distance of 5 (average shortest path is 5 edges long). The network can be described as a small-world (Figure 4.4).

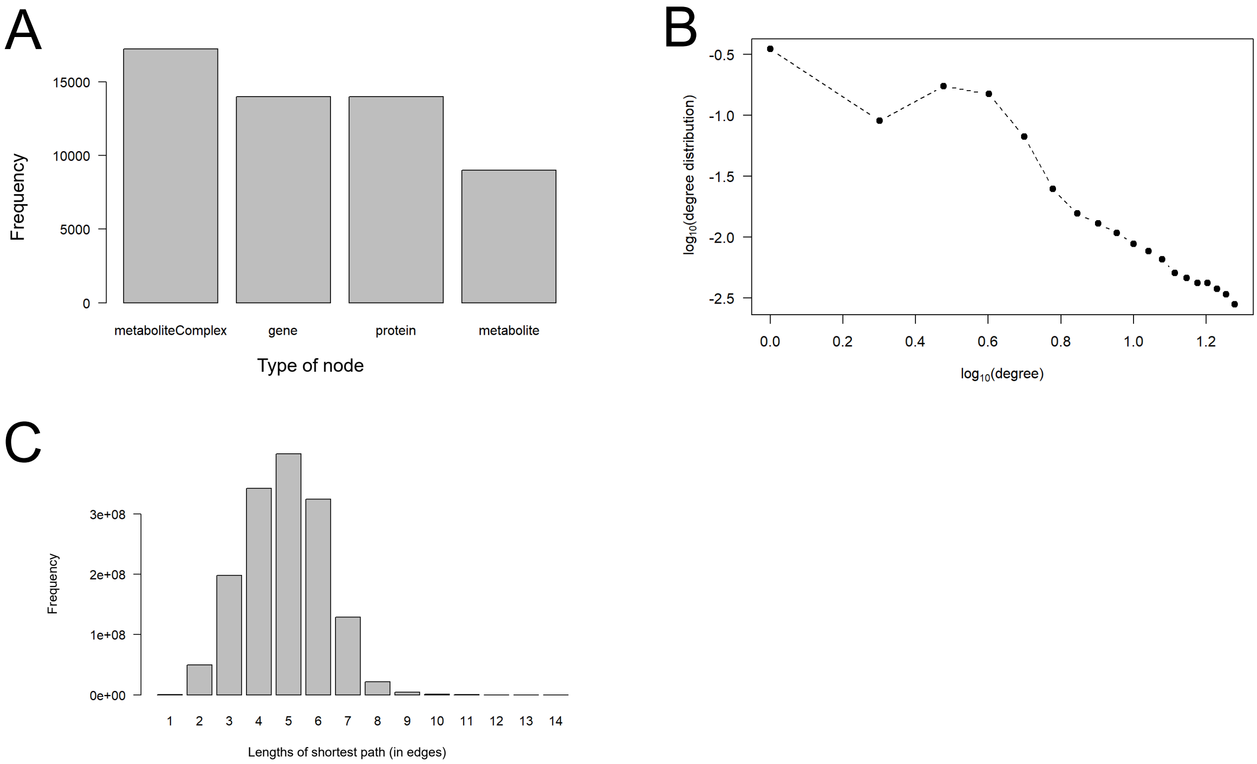


Figure 4.4: Knowledge network description

A) Type of nodes in the network. B) Distribution of the degree of nodes (log and zoom on small degrees). C) Distribution of the length of shortest paths.

As also observed by [Garrido-Rodriguez et al., 2022] in their prior knowledge graph, a literature bias is present in our network (Figure 4.5). It remains hard to know if a biological item is more studied because it has many interactions and a crucial role in biology, or if it is found to interact more because it is more studied.

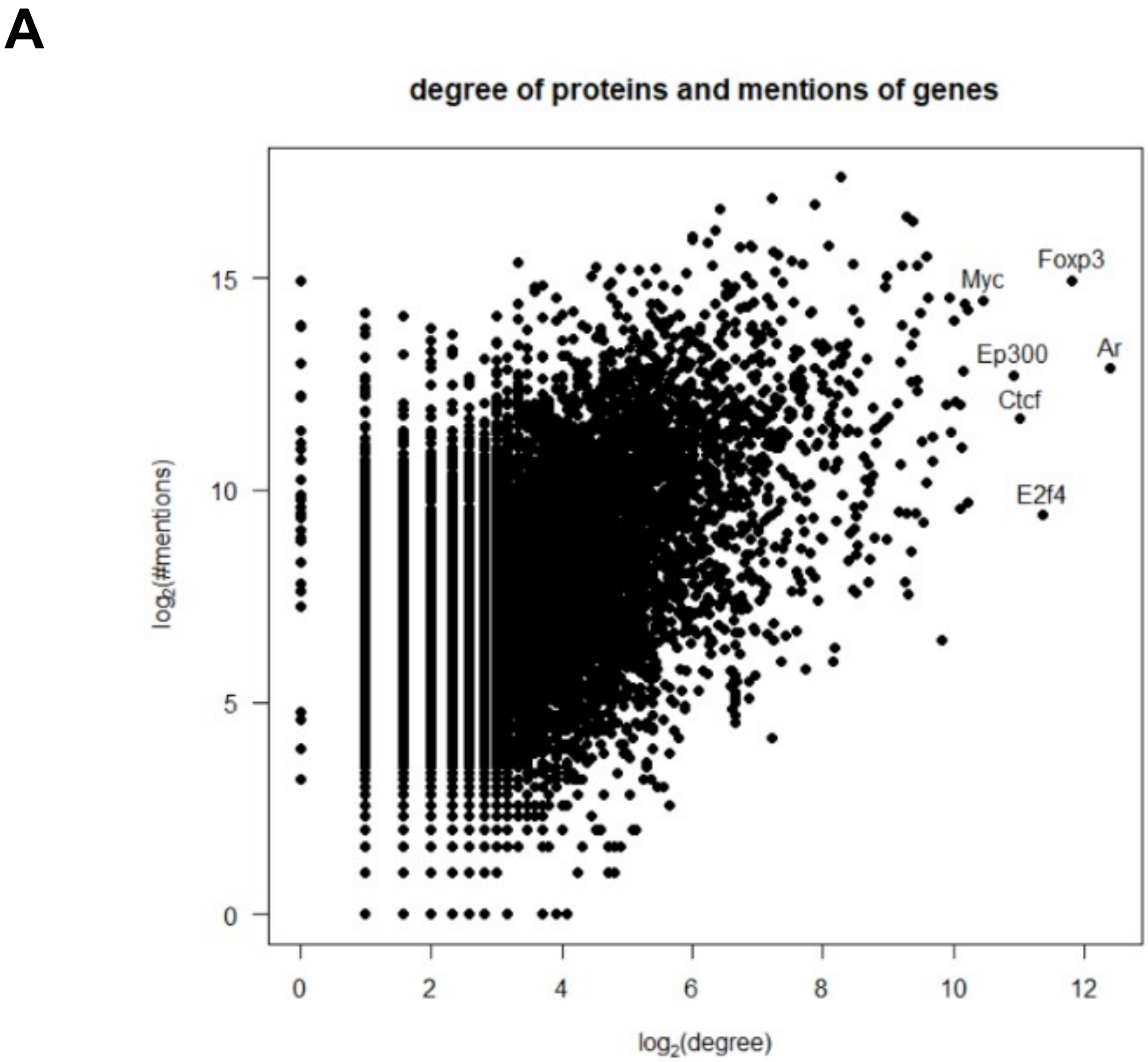
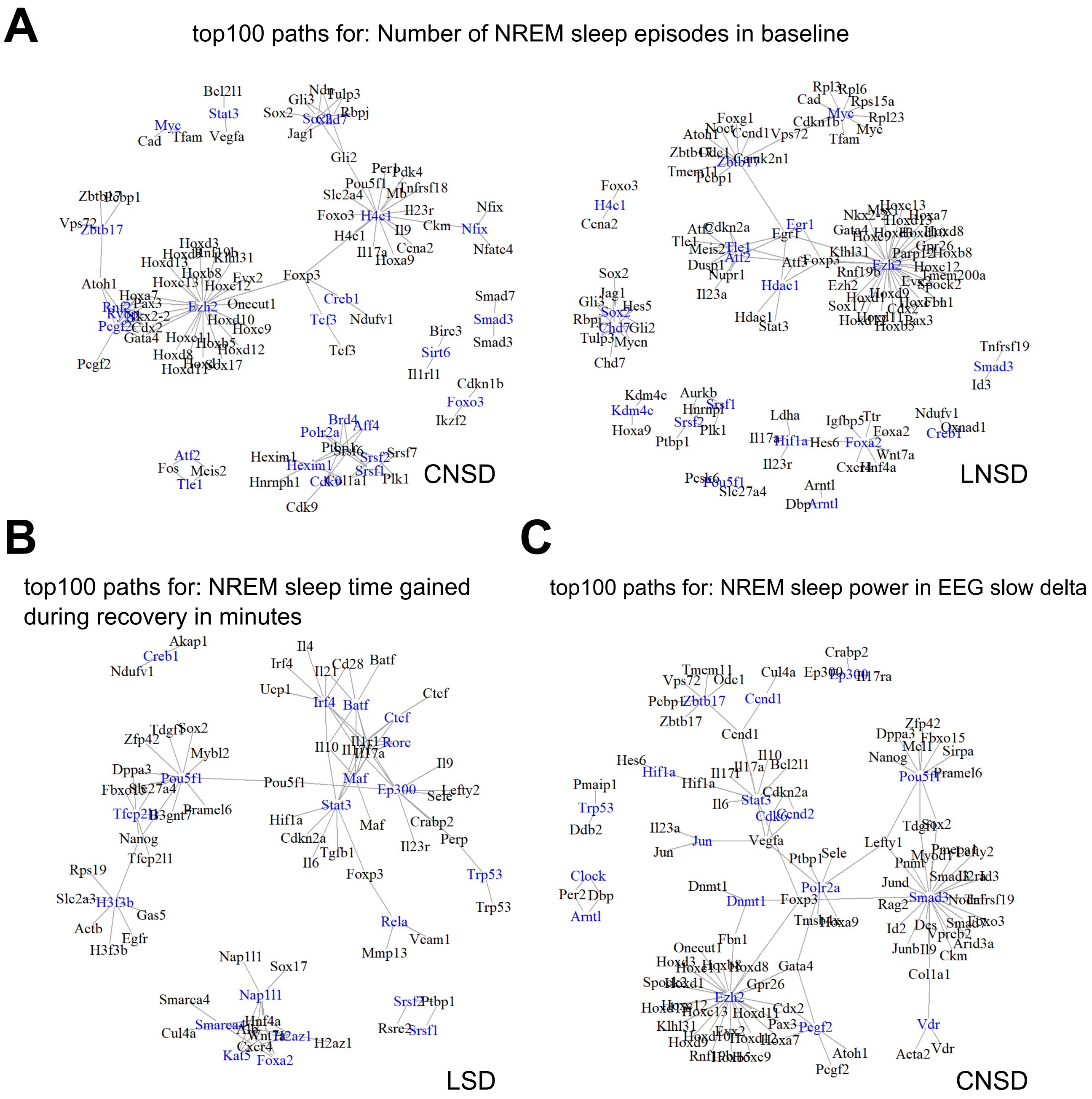


Figure 4.5: Literature Bias in the knowledge network

A) The number of mentions of genes in PubTator in function of the degree of proteins in the network. (Pearson correlation coefficient: 0.57, 95% confidence interval: 0.56 to 0.58, t-test: p-value $<$ 2.2e-16)

The more positively or negatively correlated paths were retrieved for each sleep phenotypes using the BXD expression score. A few examples of subnetworks are shown (Figure 4.6). It has to be noted that the correlation coefficients are overall low (not more than 0.5). Additionally, each path is considered independently of other paths even if they have node(s) in common. We can observe also that the subgraphs tend to have multiple (unconnected) components. This is probably an indication that the path length used for the search was not high enough.



A

top100 paths for: Number of NREM sleep episodes in baseline

B

top100 paths for: NREM sleep time gained during recovery in minutes

C

top100 paths for: NREM sleep power in EEG slow delta

Figure 4.6: Examples of subnetworks from top 100 paths for different sleep phenotypes In all subnetworks, the genes are in black and the proteins in blue. The Fruchterman-Reingold layout is used. A) Subnetwork for the number of NREM sleep episodes in baseline (48h), with scores based on expression in NSD condition in cortex (left) and liver (right) B) Subnetwork for the NREM sleep time (in minutes) gained during recovery compared to baseline, with scores based on expression in SD condition in liver C) Subnetwork for NREM sleep power in EEG slow delta compared to baseline reference power, with scores based on expression in NSD condition in cortex.

To understand which subpaths are specific to some aspects of sleep or common to all, the subgraphs were concatenated. Complete heatmaps for interactions (edges) and interactors (nodes) are in Annexes 2 to 9 while a brief overview is shown in Figure 4.7. The approach requires many fine tunings but is a novelty in the sense that it aims at predicting regulatory networks of most promising candidates for complex traits, instead of trying to retrieve a single gene that is significant on statistical test but even if we do not expect that the behavior is impacted only by one gene.

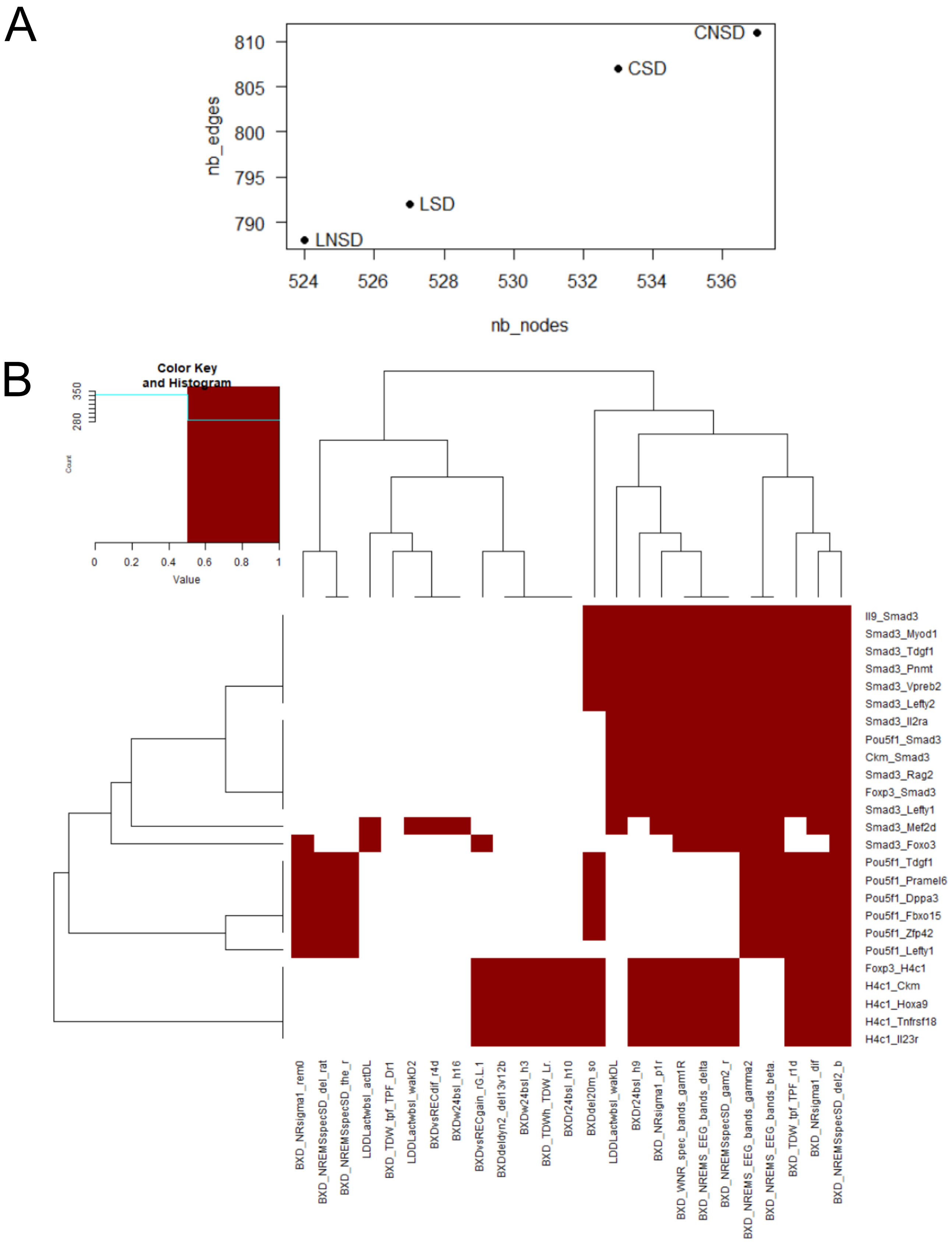


Figure 4.7: Aggregation of top 100 paths by condition and tissues

A) Number of nodes and edges by tissue and condition. B) Heatmap of 25 most frequent interactions and 25 phenotypes with most interactions in the cortex SD.

The data-driven exploration brought evidence that the signal of interest is present in the molecular layers collected. The approach taken was then more generalist to include data-driven and prior knowledge inputs together and to see the sleep phenotypes as multi-gene regulated traits. The databases mining, the mapping of identifiers, and the projection of the sleep specific omics on the knowledge graph are completed, although we are hoping future development of lipids metabolism databases will be able to improve the integration and fill the current gap. The preliminary results of the network analysis are given here for this project still under progress. The more logical extensions of these is first some optimisation on the current settings, including: i) search the full network for paths longer than 2 edges probably with more efficient ways to run the calculations, ii) varying the threshold of top paths taken for each sleep phenotype to assess the robustness, and iii) use of negative controls datasets (not related to sleep) to highlight the specificity of the approach even with the guidance of a general knowledge graph. Additionally, the inclusion of a scoring for multi-locus genetics could help link the different layers and help the identification of the genetic interactors. Another future direction that could be taken would be the inclusion of a time component in the knowledge graph to account for the dynamics observed in the sleep phenotypes (some are the same variable at different periods). As the molecular layers are snapshots of the situation at a single moment in time, it would requires to use time course experiment data [Hor et al., 2019]. That would allow to compare sleep phenotypes to their time-corresponding values of molecular phenotypes.

Chapter 5

Discussion

5.1 Summary of thesis achievements

Sleep research has a lot of unknown, my PhD journey therefore started by establishing solid structure bases with the data and metadata (re)organization. Shortly after followed the assessment of robustness of the results by changing the reference assembly done collaboratively. Then, I adapted a tool made for human research to fit our needs and implemented the D2 blocks imputation of genotypes for the BXD lines and the concept of differential mapping analysis. Finally, I built the prior knowledge graph, discovering a lot on different databases structures, purposes, advantages, and limitations. I progressively developed my network thinking even if I was not able to finish all the analyses that we had conceptualized. I will now discuss points that were not covered in the previous chapters.

5.2 Reproducibility

The FAIR principles are not always easy to apply to a project with multiple types of data. For example, simply choosing where to deposit them is not trivial, because the database more suited for each dataset is different for each omic. On one hand, specialized repositories, such as

GEO for the RNA data, are preferable because they have specific tools to search, explore and retrieve data. However, not all omics have a tailored database. On the other hand, generalist repositories, such as Figshare, allow more freedom on the type of data and formatting, but this makes it harder to explore and compare between datasets because they may capture a similar type of data, without necessarily using the same vocabulary. For genes names, this is relatively well organized, the Gene Ontology can be used and diverse packages allow to transform from gene name to identifiers. However, to standardize the description of our sleep phenotypes, the Human Phenotype Ontology (HPO) was the closest ontology available. However, it was made more to describe disease-linked phenotypes in humans whereas we have healthy phenotypes in the mice. Therefore, the correspondence is not one-on-one and had to be done manually. This situation is not unique to sleep, but can be found in different fields where there is a lack of consortium. More formal languages can help unify different area of biology [Lazebnik, 2004], but there is also a need of more discussion between specialists of different fields and omics [Pinu et al., 2019].

Data and analysis management is crucial but often in the shadow of the scientific results. However, we manage our data whether we want it or not, when talking about data management the big step is often about become aware of that and deciding to actively improve our organisation and communication. And in the end, even with the newest technologies an essential part is actually still carried by humans. Indeed, well documented meta-data and code, multiple research articles in peer-review journals, and data-mining exploratory website are clearly important but the human contact with the people that designed and performed the experiments is what actually brings the most meaning to the data: the anecdotes of highs and lows and interactions are things that words and numbers often miss.

5.3 Mouse model and experimental design

The definition of species and strains are very different across the entirety of living beings. For the mouse, the definition of an inbred strain is that they need to have been inbreed between sibilings for at least 20 generations [Mekada et al., 2009]. However, there are considered substrains if they have known or suspected genetic differences. The issue is that even if we take a low estimate of 10 germline mutations by generation [Reardon, 2017], each inbreeding done to bring closer to the 20 generations necessary for a inbreed strain are actually by definition creating substrains because we should suspect mutations. The assumption that inbred strains are fully fixed and stable across time was anchored in the mentalities which make some researchs being vague about exactly which mice were used and the practice of precising which substrain was used is quite recent [Bryant, 2011].

The best experimental design practices are not always compatible with sleep experiments and mouse breeding. For example the randomization of mice of different lines in different time batches, but the age and breeding success of the mice are sometimes coming in the way. We can also mention the isolation of mouse, since for the gold standard of sleep recording they are one per cage to avoid fights and other mice to climb on other mice cables to go out of the cage. However, mice are social animals and it seems unlikely isolation has no effect on their sleep. Studies using multiple mice per cage have detected social genetic effects (ie. the genotype of another mouse in the same cage has an effect) for a variety of phenotypes including wound healing [Baud et al., 2017]. Interestingly, the social genetic effect is sometimes stronger than the direct genetic effect. So we are using the BXD population but we are studying the genetics of the individuals.

Notice that all the mice for our sleep BXD dataset were male. Presumably to avoid to deal with periodic hormonal changes of female mice and because administratively speaking that represents a burden to justify more mice in the authorization. However, it was shown that sleep regulation in female mice more influenced by the genetic than the hormonal variations associated with the estrous cycle [Koehl et al., 2003]. Studying the sex-difference indeed requires to use more mice in the experiment but is it not important to include a structural variant covering large amount of a chromosome, easy to detect, and known to impact phenotypes? There is a need for a sex perspective in basic sleep and circadian research [Dib et al., 2021], and the inclusion of a sex variable is present in other studies too so it should not be a unstoppable administrative burden [Strefeler et al., 2023]. The lack of female mouse in sleep research is however sadly corresponding to the trend in humans where the women are underrepresented in research studies [Pandi-Perumal et al., 2022].

The characterization of sleep in the mouse is done by measuring objective variables. The large number of them try to capture the full picture, but there has no guarantee that we do not miss some aspects, such as fatigue, or cognitive performance during the following time awake. For example in humans, the objective sleep duration (gold standard) is not necessarily reflecting the subjective sleep duration, which is what matter really for patients and a diagnostic can be based on subjective measures [Benz et al., 2022].

5.4 Genetic-specific RNA mapping

Questioning current practice even in commonly used analyses has its place, even if it means only small incremental progresses it can help to gradually improve things that tend to be inert [Soneson et al., 2016]. In our BXD-specific approach for RNA-seq reference, a key thing that makes the downstream analyses still possible is that the data are aggregated at the gene level. The difficulty for it to be applied to other sequencing methods is the lack of unity of coordinate reference system. For example in ATAC-seq where genomic regions obtained with different references would be harder to compare. An important limitation on the theme of genetic-specificity is that in our analyses we did not controlled for BXD population structure that others have to account for more proximity of some of the BXD lines [Kang et al., 2008, Li et al., 2018].

5.5 Databases and Network

One of the reasons that genes and transcripts databases are more developed than metabolites and lipids ones is that the hierarchy for classification is simpler in the first case. Another reason is that genes are defined and identified by their 1D sequence, whereas for metabolites and lipids the identification is only indirectly linked to their 2D composition or 3D conformation (Figure

5.1). A limitation that we have in our approach to retrieve information from databases is that we did not put in place automatic updates, whereas most databases are updated at a more or less regular frequency. Having fixed data however allows to avoid to constantly reassess the changes that the updates may or not have, the complex dependencies between databases (one meta-resource may have a different update frequency than the individual resources), and because the comparison between different updates is anyway not trivial [Ormond et al., 2021].

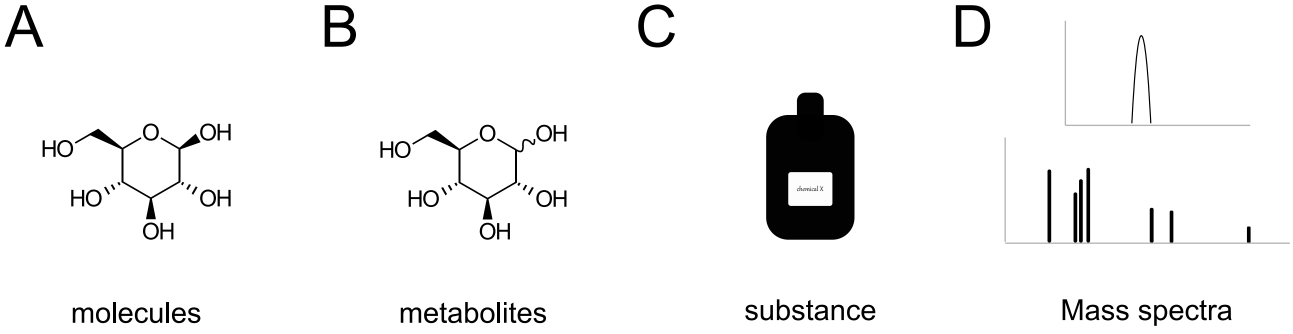


Figure 5.1: Metabolites concepts

A) Molecular structure (2D or 3D) of a metabolite. B) Metabolite as an existing chemical species that can interact, may have more than one possible molecular structure. C) Metabolites as they can be bought or used as reference, may be a mix of different molecules, with possible presence of a solvent. D) Metabolites as they are often identified by Mass Spectrometry (MS). *Figure idea from M. Pagni.*

We do look at gene expression in two key organs for sleep and circadian rhythms, however it remains that these are mix of different cell populations within the tissues and this may blur the signals and limit the insights that we could retrieve on their gene regulation systems [Aguet et al., 2019].

Computationally talented people have developed viable solutions for epistasis eQTLs calculations [Schüpbach et al., 2010, Huang et al., 2013, Trotter et al., 2021]. Certainly larger graphs exists and different algorithms exist to process them efficiently [Sakr, 2013, Slota et al., 2016, Lin et al., 2018]. However, with some hindsight what is most limiting in my approach is not necessarily the length or intensity of the computations but rather that I needed more time to grasp and process the concepts to be able to communicate about them and search for the appropriate tools or resources to put in practice the ideas.

Many tools assume the existence of one or few key genes that drive the entire behavior which makes them inadequate for complex traits [Lee et al., 2023] or are designed for specifically for a clinical or medical outcome [Shu et al., 2016]. Supervised learning works well to find more items of a certain type in large datasets [Libbrecht and Noble, 2015]. However, for sleep which are the core items is not so well defined as genes considered as core circadian genes can be actually in some organs mostly influenced by sleep-wake history, so the highly variable context dependent makes it hard to put reliable labels on genes. Different levels of representation of the data can used are have their pros and cons, with more embedded data being usually faster to process and simplifying the situation [Nelson et al., 2019].

Bibliography

[noa, 2016] (2016). Reality check on reproducibility. *Nature*, 533(7604):437–437. Number: 7604 Publisher: Nature Publishing Group.

[Aguet et al., 2019] Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Parsana, P. E., Flynn, E., Fresard, L., Gaamzon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., Park, Y., Saha, A., Segrć, A. V., Strober, B. J., Wen, X., Wucher, V., Das, S., Garrido-Martín, D., Gay, N. R., Handsaker, R. E., Hoffman, P. J., Kashin, S., Kwong, A., Li, X., MacArthur, D., Rouhana, J. M., Stephens, M., Todres, E., Viñuela, A., Wang, G., Zou, Y., Consortium, T. G., Brown, C. D., Cox, N., Dermitzakis, E., Engelhardt, B. E., Getz, G., Guigo, R., Montgomery, S. B., Stranger, B. E., Im, H. K., Battle, A., Ardlie, K. G., and Lappalainen, T. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, page 787903.

[Argelaguet et al., 2020] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111.

[Argelaguet et al., 2018] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill,

D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

[Atef et al., 2022] Atef, H., Bargiotas, P., and Youness, R. A. (2022). Sleep is a fire with smoke: Time to incorporate sleep as a fundamental component in cancer treatment protocols. *Sleep*, page zsac320.

[Badia-i Mompel et al., 2022] Badia-i Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., and Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1):vbac016.

[Bansal et al., 2022] Bansal, P., Morgat, A., Axelsen, K. B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T. B., Pozzato, M., Blatter, M.-C., Ignatchenko, A., Redaschi, N., and Bridge, A. (2022). Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, 50(D1):D693–D700.

[Baud et al., 2017] Baud, A., Mulligan, M. K., Casale, F. P., Ingels, J. F., Bohl, C. J., Callebert, J., Launay, J.-M., Krohn, J., Legarra, A., Williams, R. W., and Stegle, O. (2017). Genetic Variation in the Social Environment Contributes to Health and Disease. *PLoS Genetics*, 13(1).

[Benz et al., 2022] Benz, F., Riemann, D., Domschke, K., Spiegelhalder, K., Johann, A. F., Marshall, N. S., and Feige, B. (2022). How many hours do you sleep? A comparison of subjective and objective sleep duration measures in a sample of insomnia patients and good sleepers. *Journal of Sleep Research*, n/a(n/a):e13802. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.13802.

[Blake, 2013] Blake, J. A. (2013). Ten Quick Tips for Using the Gene Ontology. *PLOS Computational Biology*, 9(11):e1003343. Publisher: Public Library of Science.

[Borbély, 1982] Borbély, A. A. (1982). A two process model of sleep regulation. *Human Neurobiology*, 1(3):195–204.

[Borbély et al., 2016] Borbély, A. A., Daan, S., Wirz-Justice, A., and Deboer, T. (2016). The two-process model of sleep regulation: a reappraisal. *Journal of Sleep Research*, 25(2):131–143.

[Bryant, 2011] Bryant, C. D. (2011). The blessings and curses of C57BL/6 substrains in mouse genetic studies. *Annals of the New York Academy of Sciences*, 1245:31–33.

[Cantini et al., 2021] Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1):124. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Data integration;Machine learning Subject_term_id: computational-biology-and-bioinformatics;data-integration;machine-learning.

[Chen et al., 2018] Chen, Y., Tan, F., Wei, L., Li, X., Lyu, Z., Feng, X., Wen, Y., Guo, L., He, J., Dai, M., and Li, N. (2018). Sleep duration and the risk of cancer: a systematic review and meta-analysis including dose–response relationship. *BMC Cancer*, 18:1149.

[Cook et al., 2006] Cook, D. N., Whitehead, G. S., Burch, L. H., Berman, K. G., Kapadia, Z., Wohlford-Lenane, C., and Schwartz, D. A. (2006). Spontaneous Mutations in Recombinant Inbred Mice: Mutant Toll-like Receptor 4 (Tlr4) in BXD29 Mice. *Genetics*, 172(3):1751–1755.

[Dib et al., 2021] Dib, R., Gervais, N. J., and Mongrain, V. (2021). A review of the current state of knowledge on sex differences in sleep and circadian phenotypes in rodents. *Neurobiology of Sleep and Circadian Rhythms*, 11:100068.

[Diessler et al., 2018] Diessler, S., Jan, M., Emmenegger, Y., Guex, N., Middleton, B., Skene, D. J., Ibberson, M., Burdet, F., Götz, L., Pagni, M., Sankar, M., Liechti, R., Hor, C. N., Xenarios, I., and Franken, P. (2018). A systems genetics resource and analysis of sleep regulation in the mouse. *PLOS Biology*, 16(8):e2005750.

[Dubin et al., 2016] Dubin, E., Spiteri, M., Dumas, A.-S., Ginet, J., Lees, M., and Rutledge, D. N. (2016). Common components and specific weights analysis: A tool for metabolomic data pre-processing. *Chemometrics and Intelligent Laboratory Systems*, 150:41–50.

[Everson et al., 2014] Everson, C. A., Henchen, C. J., Szabo, A., and Hogg, N. (2014). Cell Injury and Repair Resulting from Sleep Loss and Sleep Recovery in Laboratory Rats. *Sleep*, 37(12):1929–1940.

[Franken, 2013] Franken, P. (2013). A role for clock genes in sleep homeostasis. *Current Opinion in Neurobiology*, 23(5):864–872.

[Franken and Dijk, 2009] Franken, P. and Dijk, D.-J. (2009). Circadian clock genes and sleep homeostasis. *European Journal of Neuroscience*, 29(9):1820–1829. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2009.06723.x.

[Franken et al., 1998] Franken, P., Malafosse, A., and Tafti, M. (1998). Genetic variation in EEG activity during sleep in inbred mice. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 275(4):R1127–R1137. Publisher: American Physiological Society.

[Franken et al., 1999] Franken, P., Malafosse, A., and Tafti, M. (1999). Genetic Determinants of Sleep Regulation in Inbred Mice. *Sleep*, 22(2):155–169. Publisher: Oxford Academic.

[Freedman et al., 2015] Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, 13(6):e1002165. Publisher: Public Library of Science.

[Garrido-Rodriguez et al., 2022] Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S., and Saez-Rodriguez, J. (2022). Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Molecular Systems Biology*, 18(7):e11036. Publisher: John Wiley & Sons, Ltd.

[Gobet et al., 2022] Gobet, N., Jan, M., Franken, P., and Xenarios, I. (2022). Towards mouse genetic-specific RNA-sequencing read mapping. *PLOS Computational Biology*, 18(9):e1010552. Publisher: Public Library of Science.

[Hermjakob et al., 2004] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32(Database issue):D452–D455.

[Hor et al., 2019] Hor, C. N., Yeung, J., Jan, M., Emmenegger, Y., Hubbard, J., Xenarios, I., Naef, F., and Franken, P. (2019). Sleep–wake-driven and circadian contributions to daily rhythms in gene expression and chromatin accessibility in the murine cortex. *Proceedings of the National Academy of Sciences*, 116(51):25773–25783. Publisher: Proceedings of the National Academy of Sciences.

[Huang et al., 2013] Huang, Y., Wuchty, S., and Przytycka, T. M. (2013). eQTL Epistasis – Challenges and Computational Approaches. *Frontiers in Genetics*, 4.

[Hubbard et al., 2020] Hubbard, J., Gent, T. C., Hoekstra, M. M. B., Emmenegger, Y., Mongrain, V., Landolt, H.-P., Adamantidis, A. R., and Franken, P. (2020). Rapid fast-delta decay following prolonged wakefulness marks a phase of wake-inertia in NREM sleep. *Nature Communications*, 11(1):3130.

[Jan et al., 2019] Jan, M., Gobet, N., Diessler, S., Franken, P., and Xenarios, I. (2019). A multi-omics digital research object for the genetics of sleep regulation. *Scientific Data*, 6(1):1–15.

[Jan et al., 2020] Jan, M., O'Hara, B. F., and Franken, P. (2020). Recent advances in understanding the genetics of sleep. *F1000Research*, 9:214.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654. Number: 6804 Publisher: Nature Publishing Group.

[Kang et al., 2008] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3):1709–1723. Publisher: Genetics Section: Investigations.

[Kim et al., 2016] Kim, E., Hwang, S., Kim, H., Shim, H., Kang, B., Yang, S., Shim, J. H., Shin, S. Y., Marcotte, E. M., and Lee, I. (2016). MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. *Nucleic Acids Research*, 44(D1):D848–D854.

[Koehl et al., 2003] Koehl, M., Battle, S. E., and Turek, F. W. (2003). Sleep in female mice: a strain comparison across the estrous cycle. *Sleep*, 26(3):267–272.

[Koutrouli et al., 2020] Koutrouli, M., Karatzas, E., Paez-Espino, D., and Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, 8.

[Krassowski et al., 2020] Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in Genetics*, 11.

[Lazebnik, 2004] Lazebnik, Y. (2004). Can a biologist fix a radio? – Or, what I learned while studying apoptosis, (Cancer Cell. 2002 Sep;2(3):179-82). *Biochemistry. Biokhimiia*, 69(12):1403–1406.

[Lee et al., 2023] Lee, Y. Y., Endale, M., Wu, G., Ruben, M. D., Francey, L. J., Morris, A. R., Choo, N. Y., Anafi, R. C., Smith, D. F., Liu, A. C., and Hogenesch, J. B. (2023). Integration of genome-scale data identifies candidate sleep regulators. *Sleep*, 46(2):zsac279.

[Li et al., 2018] Li, H., Wang, X., Rukina, D., Huang, Q., Lin, T., Sorrentino, V., Zhang, H., Bou Sleiman, M., Arends, D., McDaid, A., Luan, P., Ziari, N., Velázquez-Villegas, L. A., Gariani, K., Kutalik, Z., Schoonjans, K., Radcliffe, R. A., Prins, P., Morgenthaler, S., Williams, R. W., and Auwerx, J. (2018). An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. *Cell Systems*, 6(1):90–102.e4.

[Libbrecht and Noble, 2015] Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

[Lin et al., 2018] Lin, H., Zhu, X., Yu, B., Tang, X., Xue, W., Chen, W., Zhang, L., Hoefler, T., Ma, X., Liu, X., Zheng, W., and Xu, J. (2018). ShenTu: Processing Multi-Trillion

Edge Graphs on Millions of Cores in Seconds. In *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 706–716, Dallas, TX, USA. IEEE.

[Lu et al., 2017] Lu, C., Sun, H., Huang, J., Yin, S., Hou, W., Zhang, J., Wang, Y., Xu, Y., and Xu, H. (2017). Long-Term Sleep Duration as a Risk Factor for Breast Cancer: Evidence from a Systematic Review and Dose-Response Meta-Analysis. *BioMed Research International*, 2017:4845059.

[Mahmoud et al., 2019] Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20(1):246.

[Mang and Franken, 2012] Mang, G. M. and Franken, P. (2012). Sleep and EEG Phenotyping in Mice. *Current Protocols in Mouse Biology*, 2(1):55–74.

[Manouchehri et al., 2021] Manouchehri, E., Taghipour, A., Ghavami, V., Ebadi, A., Homaei, F., and Latifnejad Roudsari, R. (2021). Night-shift work duration and breast cancer risk: an updated systematic review and meta-analysis. *BMC women's health*, 21(1):89.

[Marai et al., 2019] Marai, G. E., Pinaud, B., Bühler, K., Lex, A., and Morris, J. H. (2019). Ten simple rules to create biological network figures for communication. *PLOS Computational Biology*, 15(9):e1007244. Publisher: Public Library of Science.

[Mekada et al., 2009] Mekada, K., Abe, K., Murakami, A., Nakamura, S., Nakata, H., Moriwaki, K., Obata, Y., and Yoshiki, A. (2009). Genetic Differences among C57BL/6 Substrains. *Experimental Animals*, 58(2):141–149.

[Mongrain et al., 2010] Mongrain, V., Hernandez, S. A., Pradervand, S., Dorsaz, S., Curie, T., Hagiwara, G., Gip, P., Heller, H. C., and Franken, P. (2010). Separating the Contribution of Glucocorticoids and Wakefulness to the Molecular and Electrophysiological Correlates of Sleep Homeostasis. *Sleep*, 33(9):1147–1157.

[Mongrain et al., 2011] Mongrain, V., La Spada, F., Curie, T., and Franken, P. (2011). Sleep Loss Reduces the DNA-Binding of BMAL1, CLOCK, and NPAS2 to Specific Clock Genes in the Mouse Cerebral Cortex. *PLoS ONE*, 6(10).

[Nath et al., 2017] Nath, R. D., Bedbrook, C. N., Abrams, M. J., Basinger, T., Bois, J. S., Prober, D. A., Sternberg, P. W., Gradinaru, V., and Goentoro, L. (2017). The jellyfish Cassiopea exhibits a sleep-like state. *Current biology : CB*, 27(19):2984–2990.e3.

[Nelson et al., 2019] Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., and Sharan, R. (2019). To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in Genetics*, 10.

[Orchard et al., 2013] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363.

[Ormond et al., 2021] Ormond, C., Ryan, N. M., Corvin, A., and Heron, E. A. (2021). Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Briefings in Bioinformatics*, (bbab069).

[O'Callaghan et al., 2019] O'Callaghan, E. K., Green, E. W., Franken, P., and Mongrain, V. (2019). Omics Approaches in Sleep-Wake Regulation. In Landolt, H.-P. and Dijk, D.-J., editors, *Sleep-Wake Neurobiology and Pharmacology*, Handbook of Experimental Pharmacology, pages 59–81. Springer International Publishing, Cham.

[Pandi-Perumal et al., 2022] Pandi-Perumal, S. R., Cardinali, D. P., Zaki, N. F. W., Karthikeyan, R., Warren Spence, D., Reiter, R. J., and Brown, G. M. (2022). Timing is everything: circadian rhythms and their role in the control of sleep. *Frontiers in Neuroendocrinology*, page 100978.

[Pinu et al., 2019] Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., and Wishart, D. (2019). Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites*, 9(4):76.

[Reardon, 2017] Reardon, S. (2017). Lab mice's ancestral 'Eve' gets her genome sequenced. *Nature News*, 551(7680):281.

[Rey et al., 2011] Rey, G., Cesbron, F., Rougemont, J., Reinke, H., Brunner, M., and Naef, F. (2011). Genome-Wide and Phase-Specific DNA-Binding Rhythms of BMAL1 Control Circadian Output Functions in Mouse Liver. *PLOS Biology*, 9(2):e1000595.

[Sakr, 2013] Sakr, S. (2013). Processing large-scale graph data: A guide to current technology.

[Sanchez-Lengeling et al., 2021] Sanchez-Lengeling, B., Reif, E., Pearce, A., and Wiltschko, A. B. (2021). A Gentle Introduction to Graph Neural Networks. *Distill*, 6(9):e33.

[Schulz, 2022] Schulz, H. (2022). The history of sleep research and sleep medicine in Europe. *Journal of Sleep Research*, 31(4):e13602. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.13602.

[Schüpbach et al., 2010] Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K. (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics (Oxford, England)*, 26(11):1468–1469.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504.

[Shu et al., 2016] Shu, L., Zhao, Y., Kurt, Z., Byars, S. G., Tukiainen, T., Kettunen, J., Orozco, L. D., Pellegrini, M., Lusis, A. J., Ripatti, S., Zhang, B., Inouye, M., Mäkinen, V.-P., and

Yang, X. (2016). Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC genomics*, 17(1):874.

[Slota et al., 2016] Slota, G. M., Rajamanickam, S., Devine, K., and Madduri, K. (2016). Partitioning Trillion-edge Graphs in Minutes. arXiv:1610.07220 [cs].

[Soneson et al., 2016] Soneson, C., Love, M. I., and Robinson, M. D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521.

[Spiegel et al., 2005] Spiegel, K., Knutson, K., Leproult, R., Tasali, E., and Cauter, E. V. (2005). Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes. *Journal of Applied Physiology*, 99(5):2008–2019. Publisher: American Physiological Society.

[Spiegel et al., 2009] Spiegel, K., Tasali, E., Leproult, R., and Van Cauter, E. (2009). Effects of poor and short sleep on glucose metabolism and obesity risk. *Nature reviews. Endocrinology*, 5(5):253–261.

[Strefeler et al., 2023] Strefeler, A., Jan, M., Quadroni, M., Teav, T., Rosenberg, N., Chatton, J.-Y., Guex, N., Gallart-Ayala, H., and Ivanisevic, J. (2023). Molecular insights into sex-specific metabolic alterations in Alzheimer's mouse brain using multi-omics approach. *Alzheimer's Research & Therapy*, 15:8.

[Szkiela et al., 2020] Szkiela, M., Kusidel, E., Makowiec-Dabrowska, T., and Kaleta, D. (2020). Night Shift Work—A Risk Factor for Breast Cancer. *International Journal of Environmental Research and Public Health*, 17(2):659.

[Szkiela et al., 2021] Szkiela, M., Kusidel, E., Makowiec-Dabrowska, T., and Kaleta, D. (2021). How the Intensity of Night Shift Work Affects Breast Cancer Risk. *International Journal of Environmental Research and Public Health*, 18(9):4570.

[Szklarczyk et al., 2019] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Mering,

C. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(Database issue):D607–D613.

[Sügis et al., 2019] Sügis, E., Dauvillier, J., Leontjeva, A., Adler, P., Hindie, V., Moncion, T., Collura, V., Daudin, R., Loe-Mie, Y., Herault, Y., Lambert, J.-C., Hermjakob, H., Pupko, T., Rain, J.-C., Xenarios, I., Vilo, J., Simonneau, M., and Peterson, H. (2019). HENA, heterogeneous network-based data set for Alzheimer's disease. *Scientific Data*, 6(1):151.

[Tariq et al., 2020] Tariq, H., Weng, Q. D., Garavan, T. N., Obaid, A., and Hassan, W. (2020). Another sleepless night: Does a leader's poor sleep lead to subordinate's poor sleep? A spillover/crossover perspective. *Journal of Sleep Research*, 29(1):e12904.

[Trotter et al., 2021] Trotter, C., Kim, H., Farage, G., Prins, P., Williams, R. W., Broman, K. W., and Sen, S. (2021). Speeding up eQTL scans in the BXD population using GPUs. *G3 (Bethesda, Md.)*, 11(12):jkab254.

[Türei et al., 2021] Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T., and Saez-Rodriguez, J. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3).

[von Mering et al., 2005] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(suppl_1):D433–D437.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442. Number: 6684 Publisher: Nature Publishing Group.

[Wei et al., 2019] Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

[Xenarios and Eisenberg, 2001] Xenarios, I. and Eisenberg, D. (2001). Protein interaction databases. *Current Opinion in Biotechnology*, 12(4):334–339.

[Xu* et al., 2018] Xu*, K., Hu*, W., Leskovec, J., and Jegelka, S. (2018). How Powerful are Graph Neural Networks?

[Ying et al., 2019] Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. arXiv:1903.03894 [cs, stat].

Softwares

R and Rmarkdown were used for analyses and data visualization.

LaTeX via Overleaf for writing the main thesis.

Zotero was used for collection and organisation of the references.

Inkscape was use for the generation of diagrammatic figures and combining of images.

Acronyms

ATAC-seq assay for transposase-accessible chromatin using sequencing. 73

B6 C57BL/6J (mouse strain). 36

BMAL1 Brain and Muscle Arnt-Like 1. 10, 18

Bmal1 Brain and Muscle Arnt-Like 1. 10, 11

BXD recombinant inbred derived from a cross between B6 and D2 strains. v, vi, 12, 13, 36, 37, 55, 56, 73

CCSWA Common Components and Specific Weights Analysis. 15

ChIP chromatin immunoprecipitation. 18

CLOCK Circadian Locomotor Output Cycles Kaput. 10

Clock Circadian Locomotor Output Cycles Kaput. 10, 11

Cry1 Cryptochrome Circadian Regulator 1. 11

Cry2 Cryptochrome Circadian Regulator 2. 11

D2 DBA/2J (mouse strain). 36

DOI Digital Object Identifier. 14

EEG electroencephalography. 11, 12

EMG electromyography. 11, 12

FAIR findable, accessible, interoperable in order to be reusable. 13

GEO gene expression omnibus. 71

GNN Graph Neural Networks. 16

GO Gene Ontology. 15

Homer1a homer scaffolding protein 1a. 11

HPO Human Phenotype Ontology. 71

LF Latent Factor. 15

MOFA multi-omics factor analysis. vii, 14, 15, 55, 56, 61

MS Mass Spectrometry. 74

Npas2 Neuronal PAS domain-containing protein. 11

NREM non-REM. 11, 13

NSD not sleep deprived. 56, 107, 109, 115, 117

PCA Principal Component Analysis. 14, 15

Per1 Period Circadian Regulator 1. 11

PER2 Period Circadian Regulator 2. 10

Per2 Period Circadian Regulator 2. 10, 11

REM rapid-eye-movement sleep. 11, 13

SCN suprachiasmatic nucleus. 10

SD sleep deprived. 56, 68, 111, 113, 119, 121

SD sleep deprivation. 11

ZT Zeitgeber time. 13

The pages were the acronym is used are indicated after each definition.

Annexes

1. Review on Structural Variant calling

Mahmoud *et al. Genome Biology* (2019) 20:246
https://doi.org/10.1186/s13059-019-1828-7

Genome Biology

REVIEW

Open Access

Structural variant calling: the long and the short of it

Medhat Mahmoud[1†], Nastassia Gobet[2,3†], Diana Ivette Cruz-Dávalos[3,4], Ninon Mounier[3,5], Christophe Dessimoz[2,3,4,6,7*] and Fritz J. Sedlazeck[1*]

Abstract

Recent research into structural variants (SVs) has established their importance to medicine and molecular biology, elucidating their role in various diseases, regulation of gene expression, ethnic diversity, and large-scale chromosome evolution—giving rise to the differences within populations and among species. Nevertheless, characterizing SVs and determining the optimal approach for a given experimental design remains a computational and scientific challenge. Multiple approaches have emerged to target various SV classes, zygosities, and size ranges. Here, we review these approaches with respect to their ability to infer SVs across the full spectrum of large, complex variations and present computational methods for each approach.

Keywords: Structural variant (SV) detection, De novo assembly, Short-read, Long-read, Mapping, Hybrid, RNA-Seq, Gene fusion

Introduction

Structural variants (SVs) are large genomic alterations, where large is typically (and somewhat arbitrarily) defined as encompassing at least 50 bp. These genomic variants are typically classified as deletions, duplications, insertions, inversions, and translocations describing different combinations of DNA gains, losses, or rearrangements [1–3]. Copy number variations (CNVs) are a particular subtype of SVs mainly represented by deletions and duplications (reviewed in Carvalho and Lupski [4]). SVs are typically described as single events, although more complex scenarios involving combinations of SV types exist [5, 6]. Chromothripsis, which is a large and complex combination of rearrangements reported in cancer [7], is an example. While the average genomic variation between two humans is 0.1% in terms of single nucleotide variants (SNVs), when taking SVs into account, this increases to 1.5% [8]. In particular, telomeric regions are affected by a higher rate of SVs [9].

SVs can have a pronounced phenotypic impact—disrupting gene function and regulation or modifying gene dosage. Multiple studies have highlighted their role in functional changes across populations [1, 10, 11] and species [12]. Their importance in medicine and molecular biology has been highlighted by multiple recent studies. For instance, in neurological diseases, SVs have been often discussed based on ATTCC repeat extensions in Parkinson [13] or CAG expansions in Huntington disease [14]. Furthermore, a retrotransposon insertion in an intron of the TAF1 gene has been associated with early stages of linked dystonia-parkinsonism disease [15]. In cancer, different types of SVs have been highlighted as causing various types of dysfunction: (i) deletions or rearrangements truncating genes [16]; (ii) amplification of genes leading to overexpression, for example, due to homologous recombination (HR) that leads to an inactivation of BRCA1 and BRCA2 [17, 18]; (iii) gene fusions, such as Her2-positive SKBR3 breast cancer that combines multiple genes across chromosomes [19]; and (iv) alteration of the location of gene regulatory elements, causing changes in the gene expression [4, 20]. In Mendelian studies, multiple diseases have been associated with deletions or duplications of genic regions. For example, three complex SVs affecting ARID1B (Coffin-Siris syndrome), HNRNPU (hypotonia), and CDKL5

* Correspondence: Christophe.Dessimoz@unil.ch; fritz.sedlazeck@bcm.edu
†Medhat Mahmoud and Nastassia Gobet contributed equally to this work.
[2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, USA
Full list of author information is available at the end of the article

(early infantile epileptic encephalopathy is a severe intellectual disability and Rett-like features) have been reported [21]. Another more recent study showed the complexity of these CNVs and an increase in mutation rates for Potocki-Lupski and Smith-Magenis syndrome [22].

SVs are also playing an essential role in plants including having a direct phenotypic impact [23]. For example, SVs play important roles in tolerance for multiple plants: (i) in maize, a tandem triplication over the AMTE1 genes is reported to be associated with aluminum resistance [24]; (ii) an amplification of Bot1 plays an important role in boron toxicity in barley [25]; and (iii) for weeds, a tolerance against the herbicide glyphosate based on amplification of EPSPS has been reported in response to extensive use of glyphosate [26]. Other SVs have a positive impact on fruit yield and quality. For example, a transposon insertion near Ruby, a MYB transcriptional activator, leads to the increase of anthocyanin concentration in blood orange compared to pumelo and mandarin [27]. In tomatoes, a transposon insertion in JOINTLESS2 (J2) results in undesirable branching of flower-bearing shoots (inflorescences) in genetic backgrounds that also carry a cryptic variant for the close homolog enhancer of J2. This combination results in excessive flower production. However, an additional tandem duplication in fresh-market breeding lines across this region leads to a threshold of correctly spliced product and thus to a healthy phenotype with higher fruit yield [28].

Despite all these evidences of the importance of SVs, they have been largely understudied, compared to SNVs, because they are much more difficult to identify. In principle, taken individually, each type of SV induces a distinctive pattern in mapping reads that can be used to infer the underlying mutation. For example, a deletion forms a lack of a sequence and thus a gap in the alignment of the sample relative to a reference (Fig. 1). However, in practice, it is much more complicated. First, sequencing and mapping errors blur the patterns. Indeed, in contrast to SNVs and smaller insertions and deletions, SVs can cover a large portion of a read or even be larger than the read length—which complicates mapping [5]. Second, the patterns induced by the different SV types can be very similar. For example, it is often hard to distinguish tandem duplications from novel insertions for genomic alignments (Fig. 1). Finally, multiple SVs can overlap or be nested, giving rise to much more complex mapping patterns than when considered individually [5, 20]. Such complex patterns may preclude mapping altogether, forcing researchers to assemble each genomic sample de novo—a difficult and more costly task with conventional sequencing.

However, great strides have recently been made, thanks to technological and methodological developments. The advent of long-read sequencing technology, in particular, Pacific Biosciences (PacBio) and Oxford Nanopore technologies (ONT), makes it possible to produce reads of several thousand base pairs, even reaching up to 2 Mbp for Oxford Nanopore [29]. Furthermore, as we shall review in more detail below, technologies such as linked reads (e.g., 10x Genomics), optical mapping, and Strand-Seq have also been developed to improve the quality of assemblies and/or SV calling. Long reads help to increase the detection of SVs as they considerably ease de novo genome assembly and mapping. Nevertheless, the increased length and the higher error rate of emerging long-read technologies can pose new methodological challenges. Complementary to long reads, another noteworthy development has been the repurposing of transcriptomics (RNA-Seq) to detect SVs—in particular, rearrangements. Indeed, by identifying apparent RNA fusions, which are thus inherently transcribed, it is possible to focus on SVs with potential functional implications. Lastly, recent progress in benchmarking is greatly improving our understanding of the strengths and weaknesses of each approach. Current efforts such as Genome in the Bottle [30] and the FDA-led initiative SEQC2 (https://www.accessdata.fda.gov/scripts/fdatrack/view/track_project.cfm?program=nctr&id=NCTR-DBB-PM-SEQC2-Phase-II) aim at better characterizing false positives and false negatives in SV calling.

In this review, we give an overview of methods to detect SVs utilizing DNA and RNA-Seq from both short and long reads (Fig. 2). We provide a snapshot of the main methods currently available for detecting SVs (Table 1), with practical guidance as to which approach is suitable for which type of study. We conclude the review with a discussion of open challenges and future directions.

De novo assembly-based approach

De novo genome assembly has traditionally been used to generate reference genomes. Multiple strategies have been proposed, utilizing long and short reads or leveraging both. We refer the interested reader to the review of Nagarajan and Pop [72], which provides a critical overview of de novo assembly methods.

To detect SVs, such de novo-assembled sequences can be aligned to a reference or other assembly (Fig. 1), and the alterations between the two can be systematically identified: the comparison of each position in one genome to its corresponding position in the other genome should allow the identification of all forms of variations [3, 73]. Discontinuities that arise from certain types of SVs during a whole-genome alignment result in different patterns (Fig. 1). However, although conceptually simple, genome alignment is computationally far from trivial [74].

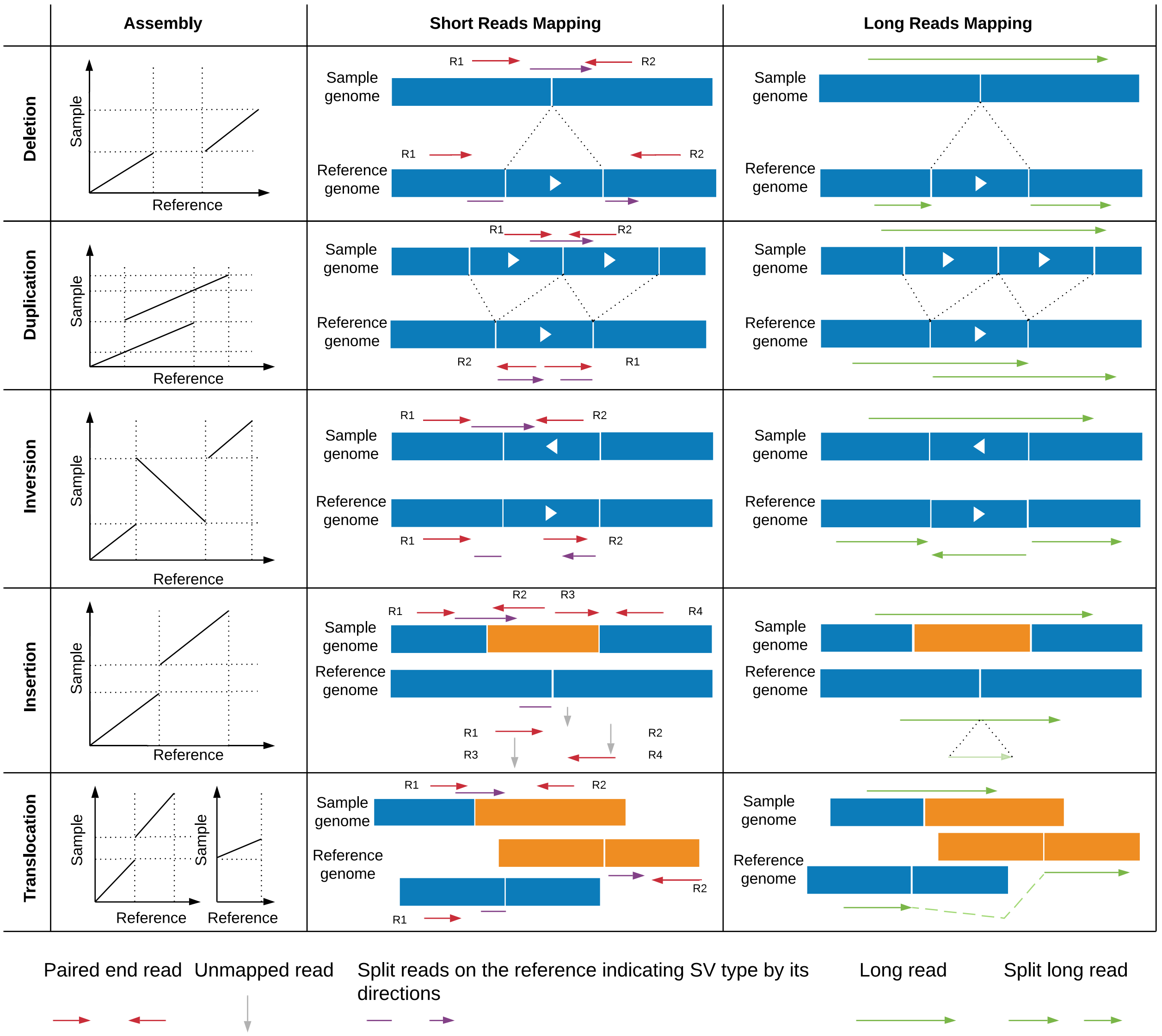


Fig. 1 Comparison between de novo assembly, short-read and long-read mapping approaches to identify structural variants. For de novo assembly approaches, the relative positions of the segments in the dot plot indicate the type and size of the SV. For short-read-based mapping approaches, paired-end (red) and split reads (purple) are typically used to decipher the type size and location. In addition, the coverage can be used to improve the detection of deletions and duplications. Long-read-based mapping approaches typically leverage the alignment patterns of long reads (green) to detect the different types of SVs

Multiple methods have been proposed to identify SVs based on a genomic alignment. These can be distinguished by whether they construct an assembly graph or operate directly on the already assembled sequences. Methods that construct the assembly graph are typically slower, but can provide more insights, as they are leveraging the read information directly. Cortex is one of these methods that use short-read sequencing data and can simultaneously assemble several genomes. Cortex uses a colored de Bruijn graph (see Table 2 for definition) to simultaneously infer SVs and complex combinations of SNVs, indels, and rearrangements [31]. SGVar [32] is a more recent string graph-based (see Table 2 for definition) de novo assembly pipeline based on the SGA assembler [75] that also uses short-read sequencing data. SGVar uses a stringent read preprocessing based on the read length and read quality. It requires a perfect match to merge reads or sequences, which improves the assembly quality. Using both simulated and real data (chromosome six of the human genome), SGVar has been shown to outperform other methods, such as Cortex, for insertion and deletion identification [32].

The other group of methods operate based on previously assembled contigs or scaffolds and are thus

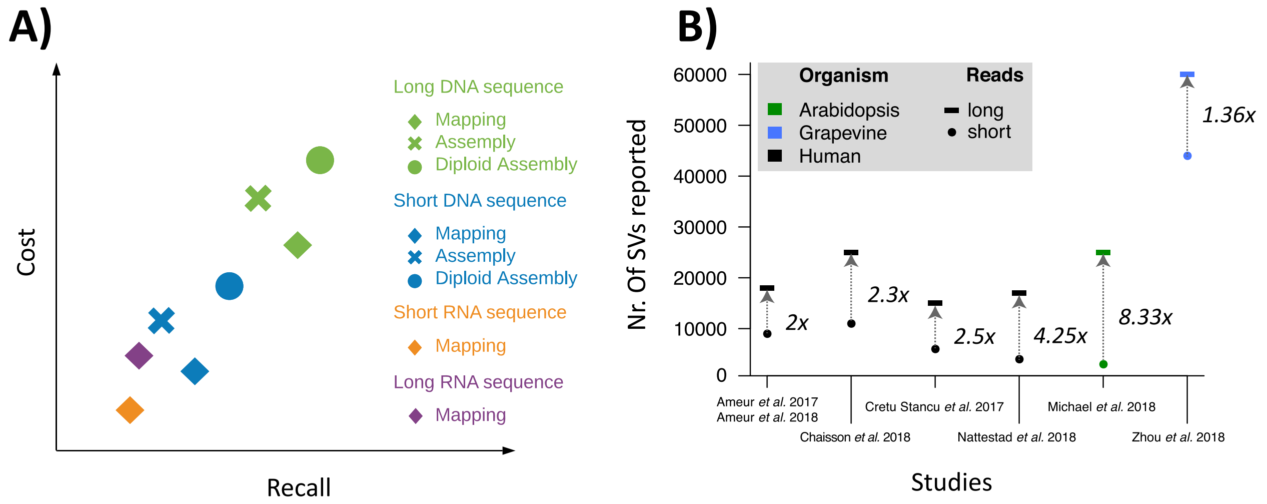


Fig. 2 Qualitative overview of structural variant calling methodology using short reads and long reads and their associated costs. **a**, A qualitative comparison of the different SV methodologies ranging across technologies (whole genome and RNA-Seq using short and long reads) to different approaches (mapping vs. assembly) with respect to their costs and recall. **b**, The ratio of improvement in the number of SVs detected from using long reads across four human and two non-human studies. Overall, each study shows a clear improvement of using the longer reads. Additional file 1: Table S1 shows the details of each study

independent of the sequencing technology (see Table 2 for definition). Basically, they rely on alignments between an assembly and a reference assembly, computed with aligners such as BlasR [76], MUMmmer [77], or Minimap2 [35]. Assemblytics [34] is a web application that relies on MUMmer and identifies insertions and deletions up to 10 kbp. It distinguishes between contractions and expansions of repetitive elements in contrast to insertions and deletions in a unique sequence. This can be an important distinction since it already annotates the type of event to provide further insight. Another method paftools.js [35] uses Minimap2 alignments, which are typically many fold faster than MUMmmer-based approaches. Similar to Assemblytics, it calls insertions and deletions but only runs on the command line. SMARTie-SV was recently introduced to detect insertions, deletions, and inversions, using BlasR. It has been applied to study SVs across great apes (gorillas, chimpanzees, orangutans) and humans [12].

Theoretically, all forms of structural variants can be identified given a fully contiguous and complete de novo assembly. The main strength of de novo assembly-based approaches compared to other approaches lies in detecting larger insertions (3+ kbp) [34, 32]. One major challenge is the lack of haplotype representation. Thus, heterozygous SVs are often missed simply by the fact that a standard de novo assembly only represents one haplotype. Nevertheless, there are de novo assembly methods to account for this such as trio-sga [78], Falcon-Unzip [79], or Trio-Canu [80] that often require additional coverage and/or parental information. They can provide diploid information of the genome and thus enable a better representation of heterozygous SVs. However, some challenges remain even for a haplotype representation, such as the de novo assembly quality and improving the genomic alignments by taking a larger genomic context sequence into account. Therefore, the de novo assembly-based approach is often used for a small number of challenging samples or for studying organisms that do not have a genome of reference.

Short-read alignment approach

Short paired-end sequencing data dominates most of the publicly available datasets. Typically, these paired-end reads are mapping in the opposite orientation and within a certain distance of each other (e.g., 500 bp). In the presence of SVs, these pairs are abnormally oriented and or spaced (Fig. 1). In addition, split reads can be used to improve the breakpoint resolution (Fig. 1). SV calling using paired-end reads is currently the standard approach and has been applied to single samples up to large cohorts (e.g., 1000 genomes).

In this section, we first focus on DNA-Seq-based methods then on RNA-Seq-based ones.

Short-read DNA-Seq mapping

Over the past decade, more than 100 short read-based mappers have been introduced, yet read mapping is still not entirely solved—for example, when it comes to reliably aligning reads to highly polymorphic regions [81]. Once the reads are mapped, their insertion size, orientation, and alignment length can be used to identify SVs candidates. Figure 1 gives a detailed overview of the

Table 1 Overview of multiple methods representative for the different SV methodologies currently used. Input types indicate the required data at start being either: De novo assembly (a), Oxford Nanopore (o), PacBio (p), 10X Genomics (x), Hi-C (h), Strand-Seq (t), Optical mapping (c) or Short reads (s)

Category	Name	Input types (a, c, h, o, p, s, tx)	Description	Link	Paper
De novo assembly	Cortex	s	Insertions, deletions, combinations of SNVs—inversions and deletions—rearrangements	http://cortexassembler.sourceforge.net/	[31]
	SGVar	s	Large insertions and deletions, complex SV		[32]
	HySA	p, s	Small (11 to 50 bp) to large (> 50 bp) insertions and deletions, complex SV	https://bitbucket.org/xianfan/hybridassemblysv/overview	[33]
	Assemblytics	a	Insertions and deletions (1 bp to 10 kb), repeat expansions/contractions	https://github.com/MariaNattestad/Assemblytics	[34]
	Paftools	a	Insertions, deletions	https://github.com/lh3/minimap2/tree/master/misc	[35]
	Smartie-sv	a	Insertions, deletions, inversions	https://github.com/zeeev/smartie-sv	[12]
	BreaKmer	s	Insertions, deletions, translocations, inversions, duplications	https://github.com/ccgd-profile/BreaKmer	[36]
	novoBreak	s	Deletions, duplications, inversions, translocations	https://sourceforge.net/projects/novobreak/	[37]
Short-read mapping	BreakDancer	s	Deletions, insertions, inversions, intra-chromosomal and inter-chromosomal translocations	https://github.com/genome/breakdancer	[38]
	BreakSeq		Insertions, deletions, translocations, inversions, duplications	http://sv.gersteinlab.org/breakseq/	[39]
	CREST	s	Insertions, deletions, translocations, inversions, duplications	https://www.stjuderesearch.org/site/lab/zhang	[40]
	DELLY	s	Deletions, inversions, duplications, inter-chromosomal translocations	https://github.com/dellytools/delly	[41]
	EricScript	s	Gene fusion	https://sourceforge.net/projects/ericscript/	[42]
	FusionCatcher	s	Gene fusion	https://github.com/ndaniel/fusioncatcher	[43]
	GRIDSS	s	Insertions, deletions, translocations, inversions, duplications	https://github.com/PapenfussLab/gridss	[44]
	Gustaf	s	Deletions, inversions, duplications, translocation	http://www.seqan.de/apps/gustaf/	[45]
	IDP-fusion	p, s	Gene fusion	https://www.healthcare.uiowa.edu/labs/au/IDP-fusion/	[46]
	JAFFA	p, s	Gene fusion	https://github.com/Oshlack/JAFFA/wiki	[47]
	LUMPY	s	Deletions, duplications, inversions, translocations	https://github.com/arq5x/lumpy-sv	[48]
	Manta	s	Insertions, deletions, translocations, inversions, duplications	https://github.com/Illumina/manta	[49]
	Meerkat	s	Insertions, deletions, translocations, inversions, duplications	http://compbio.med.harvard.edu/Meerkat/	[50]
	Pindel	s	Insertions, deletions, translocations, inversions, duplications	https://github.com/genome/pindel	[51]
	STAR-Fusion	s	Gene fusion	https://github.com/STAR-Fusion/STAR-Fusion/wiki	[52]
	SQUID	s	Gene fusion	https://github.com/	[53]

Table 1 Overview of multiple methods representative for the different SV methodologies currently used. Input types indicate the required data at start being either: De novo assembly (a), Oxford Nanopore (o), PacBio (p), 10X Genomics (x), Hi-C (h), Strand-Seq (t), Optical mapping (c) or Short reads (s) *(Continued)*

Category	Name	Input types (a, c, h, o, p, s, tx)	Description	Link	Paper
				Kingsford-Group/squid	
	TARDIS	s	Discovery of tandem and interspersed segmental duplications	https://github.com/BilkentCompGen/tardis	[54]
	TIGRA	s	Insertions, deletions	https://bitbucket.org/xianfan/tigra	[55]
	Tophat-Fusion	s	Gene fusion	http://ccb.jhu.edu/software/tophat/fusion_index.shtml	[56]
	Ulysses	s	Insertions, deletions, translocations, inversions, duplications	https://github.com/gillet/ulysses	[57]
	SvABA	s	Insertion, deletions, somatic rearrangments	https://github.com/walaj/svaba	[58]
Long-read mapping	NanoSV	o	Local SV (LSV): duplications, deletions, inversions; insertions (transposons, intra-chromosomal (> 1 Mb away) and inter-chromosomal insertions)	https://github.com/mroosmalen/nanosv	[59]
	PBHoney	p	Insertions, deletions, duplications, inversions, translocations	https://sourceforge.net/projects/pb-jelly/	[60]
	PBSV	p	Insertions (20 bp to 5 kb), deletions (20 bp to 100 kb), inversions (200 bp to 5 kb), intra-chromosomal (> 100 kb away) and inter-chromosomal translocations, complex SV	https://github.com/PacificBiosciences/pbsv	
	SMRT-SV	p	Insertions, deletions, duplications, inversions, translocations	https://github.com/EichlerLab/pacbio_variant_caller	[61]
	Sniffles	o, p	Insertions, deletions, translocations, inversions, duplications, complex SV (nested SV)	https://github.com/fritzsedlazeck/Sniffles	[62]
Multimethods SV caller	FusorSV	s	Combining LUMPY, DELLY, and GenomeSTRiP	https://github.com/TheJacksonLaboratory/SVE	[63]
	MetaSV	s	Combining BreakSeq, Breakdancer, Pindel, CNVnator	http://bioinform.github.io/metasv/	[64]
	Parliament2	s	Combining LUMPY, DELLY, Manta, BreakSeq, CNVnator	https://github.com/dnanexus/parliament2	[65]
	SURVIVOR	a, o, p, s	Can combine/compare any SVs VCF	https://github.com/fritzsedlazeck/SURVIVOR	[10]
Hi-C technology	Hic_breakfinder	h	Detects SVs based on optical mapping, Hi-C, short reads	https://github.com/dixonlab/hic_breakfinder	[66]
	HiCnv	h	Pipeline to identify CNVs from Hi-C data	https://github.com/ay-lab/HiCnv	[67]
	HiCtrans	h	Identify potential translocations using change-point statistics	https://github.com/ay-lab/HiCtrans	[67]
Optical mapping		c	Commercial tools; visualization and analysis of Bionano data	https://bionanogenomics.com/support-page/bionano-access-software/	
Strand-Seq technology	Strandseq-InvertR	t	R package to locate putative inversions	https://sourceforge.net/projects/strandseq-invertr/	[68]
10x Genomics	Gemtools	x	Downstream and in-depth analysis of SVs from linked-read data	https://github.com/sgreer77/gemtools	[69]
	GROC-SVs	x	Identify large-scale SVs based on barcode information	https://github.com/grocsvs/grocsvs	[70]
	LongRanger	x	Align reads, call and phase SNPs, indels, identify SVs	https://support.10xgenomics.com/genome-exome/software/	[16]

Table 1 Overview of multiple methods representative for the different SV methodologies currently used. Input types indicate the required data at start being either: De novo assembly (a), Oxford Nanopore (o), PacBio (p), 10X Genomics (x), Hi-C (h), Strand-Seq (t), Optical mapping (c) or Short reads (s) *(Continued)*

Category	Name	Input types (a, c, h, o, p, s, tx)	Description	Link	Paper
				downloads/latest	
	NAIBR	x	Identifies novel adjacencies created by SVs events	https://github.com/raphael-group/NAIBR	[71]

patterns of abnormally mapped paired reads and how they relate to SVs types. For example, a deletion in a sequenced sample leads to a larger insert size (the distance of the pairs). In addition, the coverage in the allele is half (heterozygous) or zero (homozygous) compared to the surrounding regions. For duplications, the coverage is increased, and for rearrangements, the pairs are abnormally spaced or oriented while the coverage is not affected. This signal is often filtered by coverage, mappability, or other measurements, such as an increase in substitutions.

The methods for detecting SVs from short reads vary in the type of information they exploit. Early methods relied exclusively on the distance and orientation of paired-end reads (Fig. 1). For example, BreakDancer [38] classifies each read into normal or SV depending on the mapping distance and orientation between the read and its mate. Regions with an excess of reads fitting into an SV category are then identified and assigned a confidence score. This can lead to missed variations, e.g., smaller deletions, for which the length is within the variability of the paired-end distribution. To increase the resolution, split reads can also be used. DELLY [41] integrates the analysis of split reads into its search of abnormal distances and orientations among pairs of reads. Although this increases the accuracy of breakpoint prediction and enables the detection of smaller deletions (20+ bp), the larger events remain hard to distinguish from mapping artifacts. To overcome this, some methods have integrated coverage information as a third kind of input signal. For example, LUMPY [48] does a joint analysis of the read depth, paired-end read discordance, and split-reads. Another tool that leverages all three types of information is Manta [49], which includes a highly parallel strategy that can be used on an individual sample or on a small set of samples including

Table 2 Glossary**.** Here, positive (P) or negative (N) describes the SV detection (or SV calling), and true (T) or false (F) describes if the calling was correct. Thus, SVs are true positive (TP) if they are called or false negatives (FN) if they are not called but present in the sample. Conversely, SVs that are not in the sample are true negatives (TN) if they are not called or false positives (FP) if they are called

Word	Definition
Accuracy	Proportion of correctly identified events (T) to the overall events: (TP + TN)/(TP + TN + FP + FN).
Breakpoints	Positions on the genome denoting the start and end of SVs relative to the reference genome.
Contigs	Contiguous sequence stretches assembled from reads.
De Bruijn graph	Directed graph consisting of nodes with exactly *n* incoming and *n* outgoing edges. In genome assemblies, a de Bruijn graph is built where the nodes are *k*-mers (sequences of length *k*) and the edges correspond to the overlap on *k* – 1 bases between nodes.
String graph-based assembly	Similar method to De Bruijn graph-based assembly, but in this case, the overlaps between all read pairs (instead of *k*-mers) are computed to construct a string graph based on the overlaps.
Insert size	The distance between the two paired-end reads.
Overhang	Portion of a mapped read that cannot be aligned and thus could indicate a structural variation.
Phasing	The identification of two or more heterozygous variations are co-occurring on the same or different DNA molecule.
Precision (or positive predictive value)	Proportion of predictions (FP + TP) that are correct (TP).
Recall (or sensitivity or true-positive rate)	Proportion of the total positives (FN + TP) that were correctly identified (TP).
Scaffold	Connected contiguous sequence stretches, with unresolved sequence stretches in between.
Split reads	Reads containing parts that map in different loci on the reference genome. They are found by splitting the read in sub-segments, align individually each sub-segment, and then grouping sub-fragments from one read.
Tandem sequence	A specific type of repetitive region that was repeated directly adjacent to each other.

tumor-normal pairs. This is achieved by parallelly building graphs across regions of the genome and testing for a specific variant hypothesis. The nodes of such graphs are regions that may contain one or more breakpoints, and the edges represent the evidence (i.e., reads) of breakpoints between the regions (see Table 2 for definition). The evidence accumulated around every pair of genomic regions is then evaluated for specific SVs hypotheses. GRIDSS [44], on the other hand, retains only the reads that provide evidence for SVs and then assembles them via a positional *de Bruijn* graph. The alignment of the subset of reads enhances the accurate identification of SVs, thus achieving an increased recall. Regarding precision (the proportion of inferred SVs that are correct), GRIDSS's authors show similar performance to LUMPY, with an estimated precision rate of 90% (evaluated from 1000 previously validated deletions) [44]. In the same study, BreakDancer, Pindel, DELLY, and Manta exhibited lower precision rates, ranging from 70 to 85%. However, GRIDSS has the disadvantage of reporting any type of SV event as a simple breakpoint (i.e., BND), and this makes the interpretation of the underlying SV type difficult. More recently, to detect more complex events such as a tandem duplication where the second copy is inverted, methods such as TARDIS have been proposed [54].

The aforementioned methods specialize in the detection of specific types of variants, but none of them is able to reliably identify all SV types and size regimes [5, 10, 82, 83]. Meta-methods seek to fill in this gap by combining calls from different tools and selecting the variants identified by multiple methods. Ideally, meta-methods can combine the strengths of multiple methods while overcoming their individual weaknesses. In practice, this works up to a certain point, but these methods can also serve to adjust the precision-recall trade-off more flexibly. MetaSV [64], Parliament2 [65], and SURVIVOR [10] have been reported to yield higher recall than a single caller, at the cost of moderately reduced precision. Using different parameters, SURVIVOR can also be used to increase precision, at the cost of a moderately reduced recall [10, 19]. Furthermore, SURVIVOR can also incorporate the information from short and long reads to further improve precision and recall.

Overall, short-read-based methods are well established and widely used. Nevertheless, the recall is often reported to be between 10 [61] and 70% [1, 5, 10] and the false-positive rates are very high (up to 89%) [60, 73, 84, 85] depending on the size and type of SVs. While rearrangements or certain larger (500+ bp) deletions are robustly identified, mid to larger size insertions remain a major challenge. These insertions are often disturbing the accurate alignment of reads and thus can lead to misinterpretations [5]. These cases might be resolved by using a localized assembly approach, for example using SvABA [58]. In addition, these methods are often blind to certain regions (e.g., low complexity, highly repetitive, highly mutated) of the genome. To sum up, while we can control the precision of these short-read-based methods, the recall can only reach a certain point and certain complex types of SVs will remain hidden [1, 5, 19, 82]. Thus, we may be reaching the limits of DNA mapping approaches based on short reads. Indeed, the emergence of meta-methods may well be indicative of diminishing returns in a maturing field.

RNA-Seq mapping

In contrast to the genome approaches, RNA-Seq-based approaches focus only on expressed regions. Here, the challenges are different, and thus, specialized methods have been proposed. In general, RNA-Seq methods aim to identify gene fusions, which are connections between parts or full lengths of two or more genes. Using RNA-Seq, we can detect if the variant observed is expressed and measure the amount of expression in comparison with other genes.

Multiple methods have been developed to detect gene fusions. These methods work based on mapping of short RNA-Seq paired-end reads to the reference genome and or transcriptome. Subsequently, the abnormal spaced paired and split reads (see Table 2 for definition) between different genes are identified, summarized, and filtered. Recent benchmarks highlighted the impact of the read quality and length to detect gene fusions but disagreed about their recommendation [46, 86–88].

For gene fusion detection, the methods mainly differ in how strictly they use existing gene annotations. Reliance on gene and exon annotations can increase precision by disregarding or correcting mapping errors. For instance, methods such as FusionCatcher [43] and EricScript [42] inherently focus on the annotated parts of the genome. FusionCatcher is designed to identify somatic fusion genes, by aligning reads to a transcriptome using Bowtie [89] guided by Ensembl annotation. It removes the reads that align to rRNA and tRNA or trim them if they have a low base quality to improve the prediction of gene fusions. EricScript follows a novel approach mapping first the paired-end reads and performing a localized assembly across fusion candidates to obtain better exon junction candidates. The reads are then mapped back to the fusion catalog, and annotation candidates are subsequently scored and filtered.

On the other hand, methods that do not strictly rely on the annotation of a genome can have a higher sensitivity. Indeed, annotations are typically incomplete, even for well-characterized organisms such as humans [90], let alone for non-model organisms. A loose reliance on annotations is further relevant when dealing with cancer samples [19], which can contain complex non-canonical gene fusion patterns. One of the earliest fusion detection methods was TopHat-Fusion [56], which used a specialized version of TopHat [91]. Of note, TopHat is outdated, and its authors recommend to use HISAT2 [92] instead. STAR-Fusion [52] is leveraging the speed and accuracy of the STAR RNA-Seq aligner [93] by selecting parameters optimized for gene fusion detection (e.g., allowing chimeric alignments, setting a low minimum overhang for a chimeric junction) (see Table 2 for definition). STAR-Fusion uses single or paired-end reads mapped to a reference and annotation index. SQUID [53] constructs a graph based on the regions with discordant reads. The graph represents candidates of gene fusions and the reference where the individual neighboring regions (nodes) are connected. The connections are subsequently weighted by the number of supportive reads. Linear programming is then used to traverse the graph and report gene fusions.

The last group of RNA-Seq fusion detection methods has been conceived to also take advantage of long reads—in particular, those obtained from the PacBio Isoform Sequence protocol. IDP-fusion [46] and Jaffa [47] are gene fusion identification tools that consolidate long-read with short-read RNA sequencing data. IDP-fusion requires both long and short reads while it is optional for Jaffa. The long reads are used primarily to identify fusion candidates. Subsequently, short reads are used to improve the breakpoint accuracy and precision.

Overall, RNA-Seq-based SV detection has the advantage of determining if an allele is expressed or not. Although this is no guarantee that this variant has an impact on the phenotype (the protein might not be translated or stable), RNA-Seq helps with prioritizing fusions that affect gene structure. However, there are multiple disadvantages. First, the underlying SV type can be uncertain for the gene fusion. This might complicate the interpretation, as well as the validation. Second, the coverage levels vary with the expression of the gene. Thus, lower expressed genes and their variations are likely to be missed. Third, SVs that impact promoter regions, introns, or non-transcribed regions are not detectable. This is especially the case for some of the methods penalizing read mapping outside of annotated regions. And fourth, previous benchmarks have shown that gene fusion studies often suffer from high false-positive rates, for example, due to chimeric regions [94].

Long-read mapping-based approach

Long reads are advantageous for SV calling because they can span repetitive or other problematic regions. Thus, these longer reads (5+ kbp) have the potential to improve the mapping and also to capture larger SVs better compared to short reads alone [5, 60, 76, 82, 83]. Both PacBio and Oxford Nanopore methods can generate reads of thousands of base pairs but present two major disadvantages. First, the costs for sequencing are higher to obtain the same coverage compared to short-read sequencing. Second, the high sequencing error rate (~ 8–20%) [95] has to be considered for both alignment and SV calling steps. Thus, specialized methods to align long reads such as BLASR [76], Minimap2 [35], and NGMLR [5] were recently developed. The identification of SVs is still at an early stage with only a few methods available.

With long reads, the SV detection methods are often tailored to the underlying technology—mainly PacBio or Oxford Nanopore. One exception is Sniffles [5], which employs a parameter estimation in the beginning and thus adjusts itself to the underlying error model. Sniffles operates on a per read base, also capable of reporting very low-frequency SVs in the sample. This is particularly useful in cancer or in mosaic variation. Furthermore, Sniffles allows the detection of more complex or adjacent SVs such as inversions flanked by deletions or inverted tandem duplications. It implements a statistical framework to reduce the number of false-positive calls.

For PacBio, three main specialized methods have been proposed. PBHoney [60] uses coverage and split read information relying on BLASR alignments. PacBio structural variant calling and analysis tools (PBSV) is a method developed by PacBio to detect SVs within the range of 20+ bp (https://github.com/PacificBiosciences/pbsv). Reads supporting a putative SV are used to generate a consensus, which is then re-aligned to the reference genome. SMRT-SV [61] includes de novo assembly and a specialized genotyping module. Reads are first aligned to the reference and, subsequently, a local assembly is performed for each multiple kbp window across the entire genome. The resulting assemblies are then aligned back to the reference, and structural variants (insertion, deletions, and inversions) are identified.

For Oxford Nanopore, NanoSV was one of the first methods developed [59]. NanoSV preferentially uses as input an alignment from LAST [96], which uses adaptive seed rather than fixed-length seed for speed optimization [96]. Of note, NanoSV reports only breakpoints (BND) which again makes the interpretation of the SVs type difficult.

Overall, long-read mapping-based methods for SV calling often show a better performance than short-read ones (Fig. 2). Indeed, longer continuous reads can be aligned more accurately, even after accounting for the higher sequencing error rate. Furthermore, the enhanced length enables a full capture of most of the alleles for SVs—in contrast to short reads where multiple pieces of information have to be put together to infer single SVs. However, there are still some performance deficiencies for larger (5+ kbp) insertions compared to de novo assemblies. This is because, as with short reads, the allele is getting longer than the read itself. Current efforts perform a localized assembly to improve, but do not fully solve, this issue when looking at very large insertions or inversions that are flanked with large low-complexity repeats (e.g., 5 kbp). Nevertheless, multiple papers have reported a significant improvement in precision and recall for SV calling using long reads compared to short-read mapping approaches [2, 5, 19, 82, 97–99].

Alternative approaches for the identification of structural variants

While this review focuses on SV calling methods utilizing short and long reads, there are other technologies that have recently improved our ability to call SVs. In this section, we provide a brief overview of these technologies and the associated software packages and refer the interested reader to other reviews for more details [95, 100–102].

Linked reads produced by 10x Genomics enable to pair reads over distances of up to 150 kb, and multiple methods have been developed to detect SVs from the linked reads. The challenge here is to identify an SV based on sparse coverage of the molecule with paired-end Illumina reads. These methods typically have a specific target SV size resolution because the barcode identifying the paired-end reads per molecule is not unique and the distance between the individual paired-end reads is undefined. Prominent methods for this technology include LongRanger [16] (50+ bp for deletions, 30+ kbp for rearrangements), GROC-SVs [70] (min 10 kbp) utilizing a localized assembly, and NAIBR (1+ kbp) [71], which uses a probabilistic model that combines multiple signals in barcoded reads.

Another technology relying on short-read sequencing is Hi-C, which is used to identify regions that are in close proximity in 3D space, which provides longer-range information than standard short read. An alteration of these pairs is likely caused by an SV allele at the location. Several methods have been devised to directly detect SVs based on Hi-C data. While some methods, such as Hic_breakfinder (1+ Mbp), can potentially identify all types of SVs [66], others, such as HiCnv (> 1 Mbp) and HiCtrans [67], only aim to detect CNVs and translocations, respectively.

Strand-Seq is a new sequencing method that preserves strand directionalities. Thus, when the reads are aligned to the reference genome, the individual homologs for each chromosome can be distinguished [101]. This helps in identifying inversions, for example, using Strandseq-InvertR [68] (min ~ 1 kbp), and can also be applied at a single-cell level.

Optical mapping, e.g., provided by BioNano, uses a different approach based on restriction enzyme maps which labels 7-bp markers. Optical mapping is a highly cost-efficient method to detect SVs but is often limited in terms of breakpoint accuracy and in terms of distinguishing SVs that are close to one another. Furthermore, BioNano cannot provide the sequence of an allele (e.g., insertions). SV calling from BioNano data can be performed using the vendor's software, called BioNano Access (https://bionanogenomics.com/support-page/bionano-access/).

Discussion

SVs are increasingly being recognized as an important class of variants, which need to be considered in evolutionary, population, and clinical genomics. In this review, we delved into different available algorithms to call SVs, highlighting their advantages and disadvantages. It transpires that SV calling methods based on short-read mapping offer a cost-efficient way to search for most known SV alleles (genotyping) [103], but they struggle to detect novel SVs, especially insertions [5, 82, 83]. On the other hand, SV calling approaches from de novo assembly require a contiguous, haplotype-resolved and complete representation of the sample, something which can only be achieved through costly high-coverage sequencing. This makes them currently impractical when dealing with multiple samples (e.g., > 20)—which, for instance, is needed for population-scale studies. However, they are necessary to reliably detect and resolve complex SVs alleles. As for the long-read-based SV mapping approaches, they are at the "bleeding edge". Long-read sequencing is currently more expensive and less widespread than short-read sequencing. However, this is currently changing with continuous reductions from both Oxford Nanopore and PacBio cost per base. It is already apparent that SV calling from long-read mapping can be more effective than from short-read mapping approaches. In addition, mapping approaches are often less expensive than de novo assemblies. For applications requiring the elucidation of very long or very complex SVs, it is still possible to perform a localized long-read de novo assembly. Phasing SVs can further improve the overall quality by identifying which SVs violate the diploid genome assumption. Clearly, this needs to be adopted, given copy number alterations or genomes with higher ploidy. Due to the complexity, only few studies were able to do this so far with a success of 78.7%, even though parental genomic data was used [59].

Regardless of the sequencing technology and SV calling algorithm, a challenge that remains is the comparison and interpretation of SVs. For example, a tandem duplication will result in having the second paired read or part of the read mapped before the first (Fig. 1). Interspersed duplications induce very different mapped read patterns, which can easily be confounded with an inversion or deletion (if the duplication is on the same chromosome) or with a translocation (duplication on a different chromosome). This is caused by molecules that have recombined between different regions, an event which can occur in cancer. In such cases, the reads of these regions will map back to their original locations along the genomes, forming larger gaps in their alignments. These gaps are then misinterpreted sometimes as different SV types flanking the duplicated regions, depending on their distance to each other (Fig. 1). As for insertions, while a novel sequence will indeed be identified as an insertion, a sequence that is similar to a region in the genome (e.g., 80% identity or more) can be called depending on the location of the region as a translocation, inversion, or deletion event. Lastly, when comparing de novo assembly-based calls and mapping-based calls, duplications and insertions can be hard to distinguish: while a genomic alignment may indicate a novel sequence between two genomes, mapping-based approaches might highlight the same event as a tandem duplication if the inserted sequence shares similarity to the neighboring region. As these examples illustrate, comparing different SV call sets and reconciling them can add a whole new layer of difficulty to the problem.

For methods to progress, benchmarking is critical. Currently, the performance of each method remains hard to assess, because precision and recall are typically estimated on different datasets, each presenting different challenges, often using inconsistent operational definitions (e.g., a minimum length of 20 vs. 50 bp to be considered a SV). Furthermore, most benchmarks to date are limited to simulated datasets: this is advantageous in that the truth is known with certainty, but it is often unclear how such results generalize to real datasets. To establish gold standards and facilitate the comparison of different methods, several efforts are underway, such as Genome in a Bottle (led by the US National Institute of Standards and Technology) and SEQC2 (lead by the US Food and Drug Administration). Both seek to obtain a better gold standard and understanding of the underlying bias. This is achieved by sequencing trios very deeply with multiple technologies (Genome in a Bottle) or sequencing a sample multiple times by different laboratories and different sequencing machines (SEQC2). The results of these studies will further highlight the advantages of certain approaches over others.

Ultimately, for SVs to be routinely considered in evolutionary and medical studies, standard methods and reference databases will be required. An improved differentiation between germline and somatic SVs would be desirable, similar to that of SNVs, to improve the categorization of SVs. Currently, only few methods exist that offer an initial assessment (e.g., Manta [49]). Databases of allele frequencies such as gnomAD [104] are available for SNVs, but we completely lack them for SVs. The annotation of SVs is often more difficult because their length needs to be taken into account, and the underlying sequence itself needs to have a reliable allele frequency assessment. Furthermore, although SVs can be reported using the standard Variant Call Format (VCF), there are inconsistencies in the way different methods report SVs. Some methods fail to report sufficient information to determine the exact type of SV or report valuable extra information in an ad hoc format. Standardization would greatly facilitate SV calling across multiple samples. One possible solution would be to extend the format in a similar way as with the Genomic VCF format (gVCF) for SNVs. In that format, for SNVs and smaller insertion and deletions, the reference information is also included to enable subsequent genotyping of variants that might not have been called in the initial assessment. Such an approach greatly speeds up the assessment and often increases the accuracy.

Likewise, before SV calling becomes routine in clinical settings, several challenges will need to be overcome. Besides the challenges in detection and correct genotyping, we are lacking an assessment and annotation of SVs. One of the best indicators if a variant is a candidate for pathogenicity is if this variant occurs at a low frequency (e.g. < 0.5%) in the population. While it is possible to assess the frequency of SNVs using reference datasets such as gnomAD/ExAC [104], this is much more difficult for SVs [103]. Indeed, while there is only a small number of possible SNVs at each site (typically one or two alleles, but only up to four given the nature of DNA), the number of possible SVs affecting each site is much larger, due to their size and type differences. This also complicates our ability to compare SVs with each other. Finally, because of the need for certification and quality assurance in a clinical setting, the aforementioned lack of format standardization and metadata information is even more acute in clinical applications than in research.

In conclusion, the current state of SV calling is akin to that of SNV calling about 10 years ago: its value is unquestionable, but the technology and methods are still evolving very rapidly, and the lack of standard protocols, benchmarks, and reference databases means that SV calls require careful interpretation. Considering the intense competition among long-read sequencing providers and the need for SV characterization for clinical applications—in particular for cancer diagnostic and treatment—it will not be long before SV analysis becomes routine.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1828-7.

Additional file 1: Table S1.

Additional file 2: Review history.

Acknowledgements

We thank Michael M. Khayat and Adam W. Hansen for the helpful discussions. We also thank Monique Zahn for the helpful editing. This review started as an assignment in the graduate course *Reviews in Quantitative Biology* at the University of Lausanne, Switzerland. DICD thanks Anna-Sapfo Malaspinas for the supervision and advice on this project.

Review history

The review history is available as Additional file 2.

Authors' contributions

All authors wrote, read, and approved the final manuscript.

Funding

MM and FJS were supported by the National Institutes of Health award (UM1 HG008898). NG was funded by a grant of the Swiss National Science Foundation (31003A_173182) to Paul Franken (Center for Integrative Genomics, University of Lausanne). DICD was funded by a European Research Council starting grant (CAMERA). NM was funded by the Swiss National Science Foundation (31003A-143914). This work was supported by the Swiss National Science Foundation grant 150654 (to CD).

Competing interests

FJS obtained a Pacbio SMRT grant in 2018 and had multiple travels sponsored by Pacific Biosciences, Inc. and Oxford Nanopore Technologies Ltd. CD has been providing consulting services for Pacific Biosciences, Inc. All other authors declare that they have no competing interests.

Author details

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, USA. [2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [4]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [5]University Center for Primary Care and Public Health, Lausanne, Switzerland. [6]Centre for Life's Origins and Evolution, Department of Genetics, Evolution & Environment, University College London, London, UK. [7]Department of Computer Science, University College London, London, UK.

Received: 11 April 2019 Accepted: 19 September 2019
Published online: 20 November 2019

References

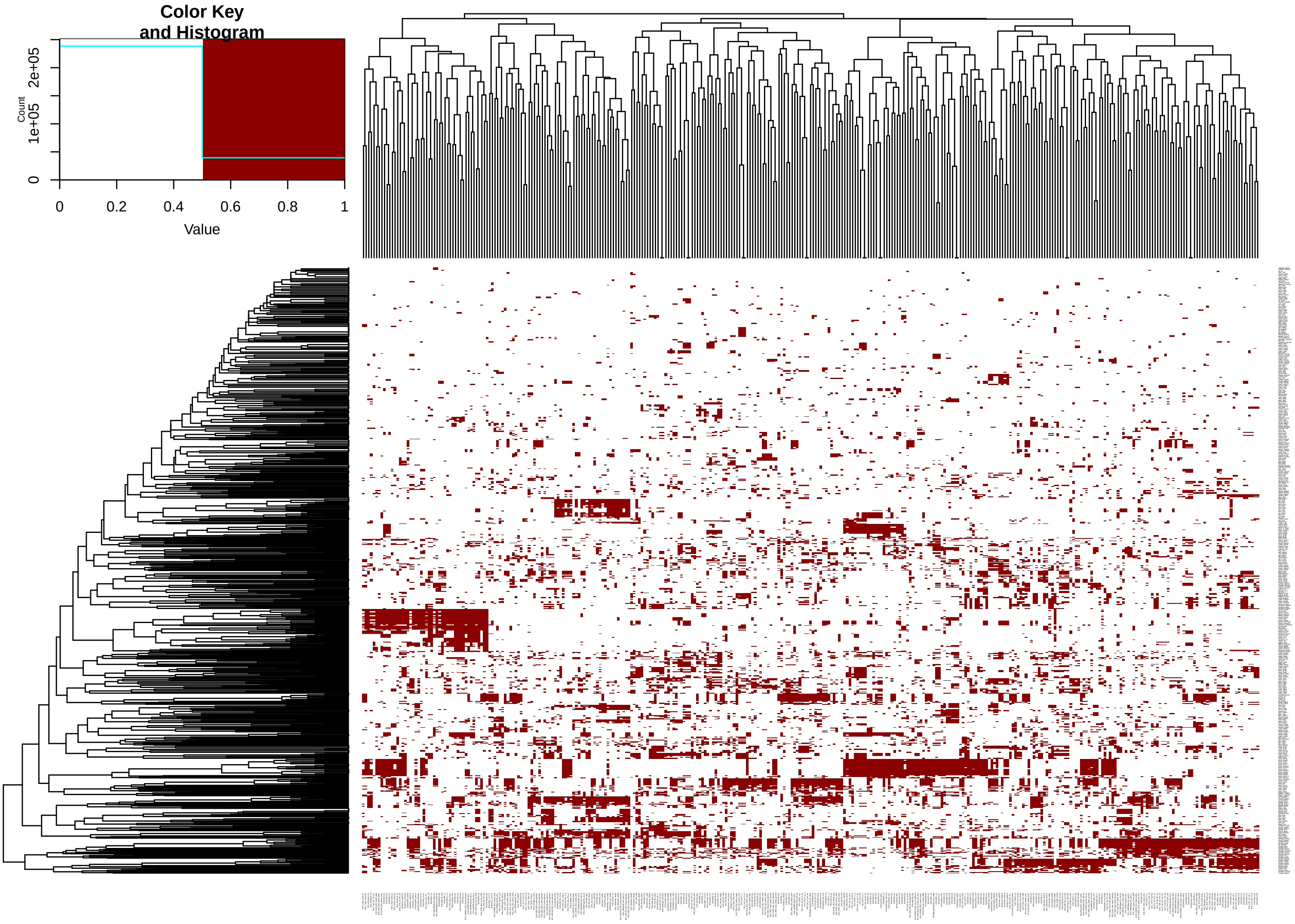
1. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
2. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.
3. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.
4. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38.
5. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
6. Tian C, Li D, Liu P, Jiao L, Gao X, Qiao J. A de novo complex chromosome rearrangement associated with multisystematic abnormalities, a case report. Mol Cytogenet. 2017;10:32.
7. Meyerson M, Pellman D. Cancer genomes evolve by pulverizing single chromosomes. Cell. 2011;144:9–10.
8. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11:R52.
9. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun. 2019;10:1025.
10. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck FJ. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8:14061.
11. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256–60.
12. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. High-resolution comparative analysis of great ape genomes. Science. 2018; 360(6393):eaar6343.
13. Schule B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, Liu M, Byrne C, Gopi R, Huang N, et al. Parkinson's disease associated with pure ATXN10 repeat expansion. NPJ Parkinsons Dis. 2017;3:27.
14. McColgan P, Tabrizi SJ. Huntington's disease: a clinical review. Eur J Neurol. 2018;25:24–34.
15. Bragg DC, Mangkalaphiban K, Vaine CA, Kulkarni NJ, Shin D, Yadav R, Dhakal J, Ton ML, Cheng A, Russo CT, et al. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. Proc Natl Acad Sci U S A. 2017;114:E11020–8.
16. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nat Commun. 2014;5:4846.
17. Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE, King MC. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. Nat Genet. 1994;8:399–404.
18. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995;378:789–92.
19. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res. 2018;28:1126–35.
20. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. Exp Mol Med. 2018;50:98.
21. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Megy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018;10:95.
22. Beck CR, CMB C, Akdemir ZC, Sedlazeck FJ, Song X, Meng Q, Hu J, Doddapaneni H, Chong Z, Chen ES, et al. Megabase length hypermutation accompanies human structural variation at 17p11.2. Cell. 2019;176(6):1310–24.
23. Gabur I, Chawla HS, Snowdon RJ, Parkin IAP. Connecting genome structural variation with complex traits in crop plants. Theor Appl Genet. 2019;132:733–50.
24. Maron LG, Guimaraes CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, et al. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A. 2013;110:5241–6.
25. Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P. Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science. 2007;318:1446–9.
26. Gaines TA, Zhang W, Wang D, Bukun B, Chisholm ST, Shaner DL, Nissen SJ, Patzoldt WL, Tranel PJ, Culpepper AS, et al. Gene amplification confers glyphosate resistance in Amaranthus palmeri. Proc Natl Acad Sci U S A. 2010;107:1029–34.
27. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24:1242–55.
28. Soyk S, Lemmon ZH, Sedlazeck FJ, Jimenez-Gomez JM, Alonge M, Hutton SF, Van Eck J, Schatz MC, Lippman ZB. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. Nat Plants. 2019;5:471–9.
29. Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: a graphical viewer for Oxford Nanopore bulk fast5 files. bioRxiv. 2018;35:312256.
30. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025.
31. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012;44:226–32.
32. Tian S, Yan H, Klee EW, Kalmbach M, Slager SL. Comparative analysis of de novo assemblers for variation discovery in personal genomes. Brief Bioinform. 2018;19:893–904.
33. Fan X, Chaisson M, Nakhleh L, Chen K. HySA: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. Genome Res. 2017;27:793–800.

34. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32:3021–3.
35. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
36. Abo RP, Ducar M, Garcia EP, Thorner AR, Rojas-Rudilla V, Lin L, Sholl LM, Hahn WC, Meyerson M, Lindeman NI, et al. BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. Nucleic Acids Res. 2015;43:e19.
37. Chong Z, Ruan J, Gao M, Zhou W, Chen T, Fan X, Ding L, Lee AY, Boutros P, Chen J, Chen K. novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods. 2017;14:65–7.
38. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81.
39. Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28:47–55.
40. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011;8:652–4.
41. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
42. Benelli M, Pescucci C, Marseglia G, Severgnini M, Torricelli F, Magi A. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics. 2012;28:3232–9.
43. Nicorici D, Şatalan M, Edgren H, Kangaspeska S, Murumägi A, Kallioniemi O, Virtanen S, Kilkku O. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv. 2014:011650.
44. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27:2050–60.
45. Trappe K, Emde AK, Ehrlich HC, Reinert K. Gustaf: detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics. 2014;30:3484–90.
46. Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, Zabner J, Korlach J, Wong WH, Au KF. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res. 2015;43:e116.
47. Davidson NM, Majewski IJ, Oshlack A. JAFFA: high sensitivity transcriptome-focused fusion gene detection. Genome Med. 2015;7:43.
48. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
49. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
50. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell. 2013;153:919–29.
51. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71.
52. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, Bankapur A, Ganote C, Doak T, Pochet N, et al. STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. bioRxiv. 2017:120295.
53. Ma C, Shao M, Kingsford C. SQUID: transcriptomic structural variation detection from RNA-seq. Genome Biol. 2018;19:52.
54. Soylev A, Le T, Amini H, Alkan C, Hormozdiari F. Discovery of tandem and interspersed segmental duplications using high throughput sequencing. Bioinformatics. 2019;35(20):3923–30.
55. Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res. 2014;24:310–7.
56. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12:R72.
57. Gillet-Markowska A, Richard H, Fischer G, Lafontaine I. Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. Bioinformatics. 2015;31:801–8.
58. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28:581–91.
59. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8:1326.
60. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014;15:180.
61. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27:677–85.
62. Sedlazeck FJ, Lemmon Z, Soyk S, Salerno WJ, Lippman Z, Schatz MC. SVCollector: optimized sample selection for validating and long-read resequencing of structural variants. bioRxiv. 2018:342386.
63. Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, Sargent J, Shanker K, Mil-Homens A, Cerveira E, et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 2018;19:38.
64. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HY. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31:2741–4.
65. Zarate S, Carroll A, Krasheninina O, Sedlazeck FJ, Jun G, Salerno W, Boerwinkle E, Gibbs R. Parliament2: fast structural variant calling using optimized combinations of callers. bioRxiv. 2018:424267.
66. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardimci GG, Chakraborty A, Bann DV, Wang Y, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018;50:1388–98.
67. Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. 2017;34(2):338–45.
68. Sanders AD, Hills M, Porubsky D, Guryev V, Falconer E, Lansdorp PM. Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Res. 2016;26:1575–87.
69. Greer SU, Ji HP. Structural variant analysis for linked-read sequencing data with gemtools. Bioinformatics. 2019;35(21):4397–99.
70. Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods. 2017;14:915–20.
71. Elyanow R, Wu HT, Raphael BJ. Identifying structural variants using linked-read sequencing data. Bioinformatics. 2017;34(2):353–60.
72. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14:157–67.
73. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol. 2015;3:92.
74. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I. Strategies and tools for whole-genome alignments. Genome Res. 2003;13:73–80.
75. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22:549–56.
76. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13:238.
77. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res. 1999;27:2369–76.
78. Malinsky M, Simpson JT, Durbin R. trio-sga: facilitating de novo assembly of highly heterozygous genomes with parent-child trios. bioRxiv. 2016:051516.
79. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13:1050–4.
80. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy A. Complete assembly of parental haplotypes with trio binning. bioRxiv. 2018;36:271486.
81. Smolka M, Rescheneder P, Schatz MC, von Haeseler A, Sedlazeck FJ. Teaser: individualized benchmarking and optimization of read mapping results for NGS data. Genome Biol. 2015;16:235.
82. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176:663–75 e619.
83. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. Resolving the

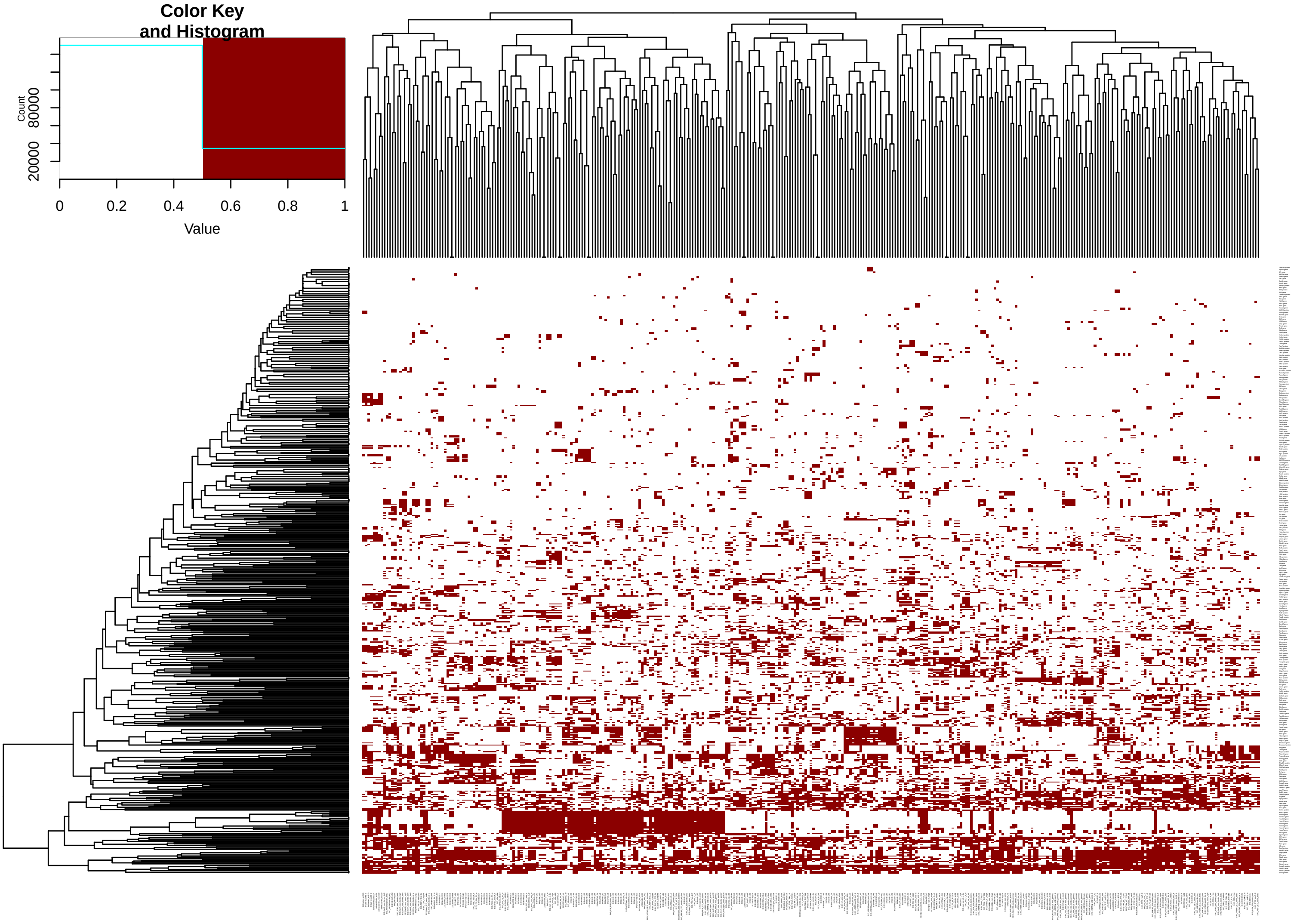
complexity of the human genome using single-molecule sequencing. Nature. 2015;517:608–11.
84. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470:59–65.
85. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28:2711–8.
86. Kumar S, Razzaq SK, Vo AD, Gautam M, Li H. Identifying fusion transcripts using next generation sequencing. Wiley Interdiscip Rev RNA. 2016;7:811–23.
87. Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, Kim S, Ma T, Chang TY, Priedigkeit NM, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res. 2016;44:e47.
88. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep. 2016;6:21597.
89. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
90. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
91. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
92. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
93. 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;29:15–21.
94. Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, Calogero RA. State-of-the-art fusion-finder algorithms sensitivity and specificity. Biomed Res Int. 2013;2013:340620.
95. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
96. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–93.
97. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784.
98. Leija-Salazar M, Sedlazeck FJ, Toffoli M, Mullin S, Mokretar K, Athanasopoulou M, Donald A, Sharma R, Hughes D, Schapira AHV, Proukakis C. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. Mol Genet Genomic Med. 2019;7:e564.
99. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
100. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58:268–76.
101. Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat Protoc. 2017;12:1151–76.
102. Cao H, Hastie AR, Cao D, Lam ET, Sun Y, Huang H, Liu X, Lin L, Andrews W, Chan S, et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience. 2014;3:34.
103. Chander V, Gibbs RA, Sedlazeck FJ. Evaluation of computational genotyping of structural variation for clinical diagnoses. Gigascience. 2019;8(9):giz110.
104. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

Publisher's Note

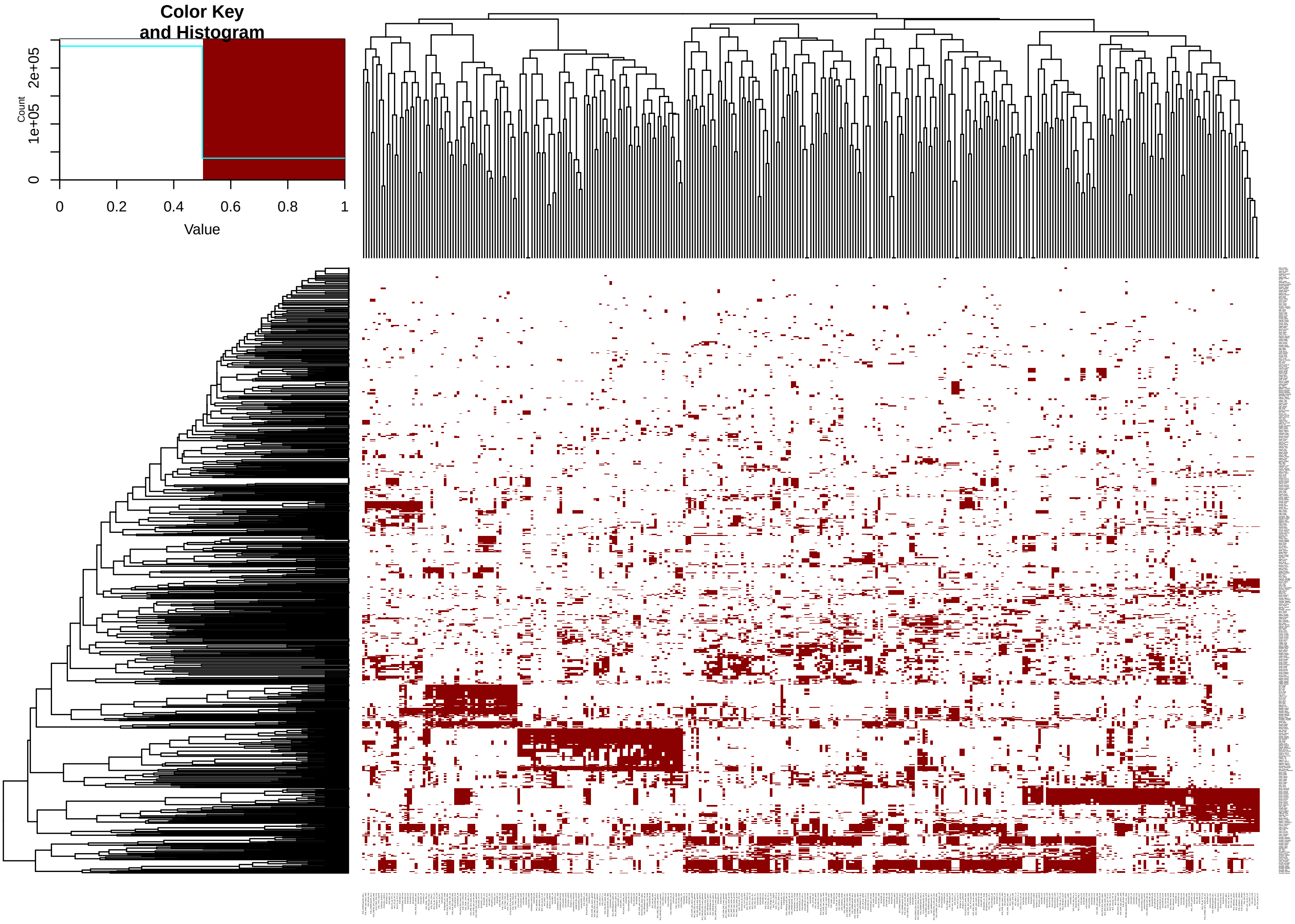
2. Aggregated top 100 paths for Cortex NSD: interactions


Color Key
and Histogram
Count
0
1e+05
2e+05
0
0.2
0.4
0.6
0.8
1
Value

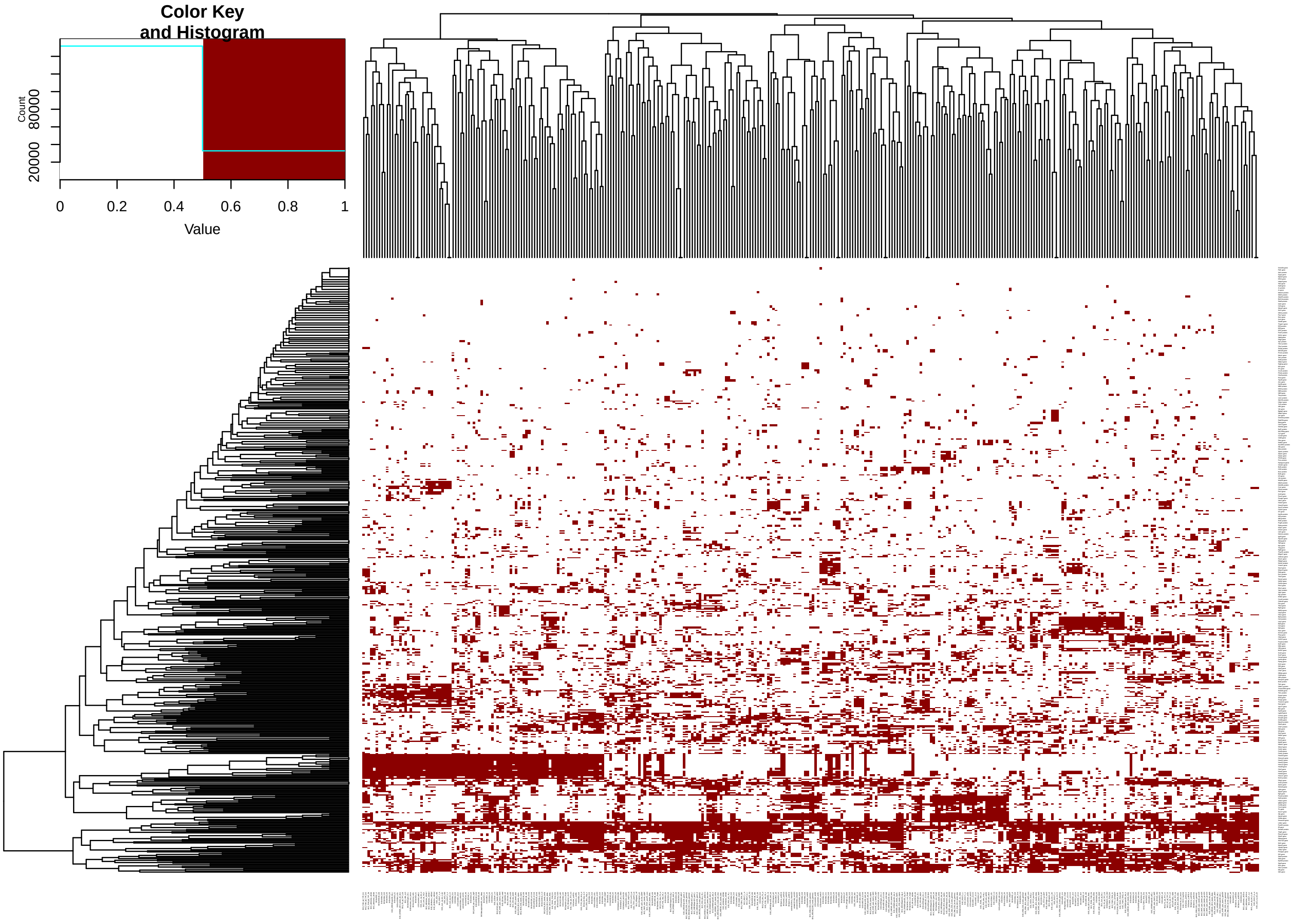
3. Aggregated top 100 paths for Cortex NSD: interactors


Color Key
and Histogram
Count
20000
80000
0
0.2
0.4
0.6
0.8
1
Value

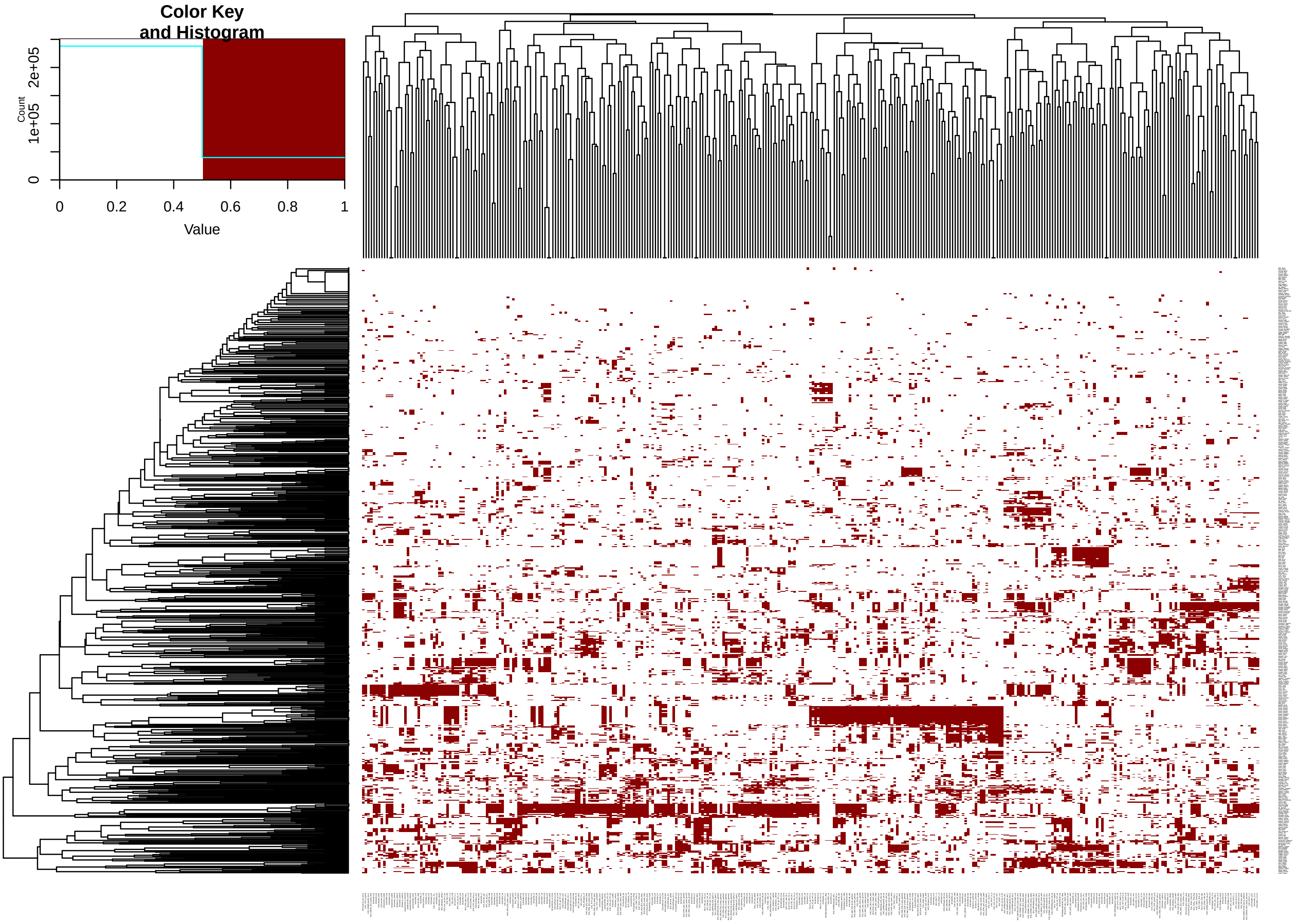
4. Aggregated top 100 paths for Cortex SD: interactions


Color Key
and Histogram
Count
0
1e+05
2e+05
0
0.2
0.4
0.6
0.8
1
Value

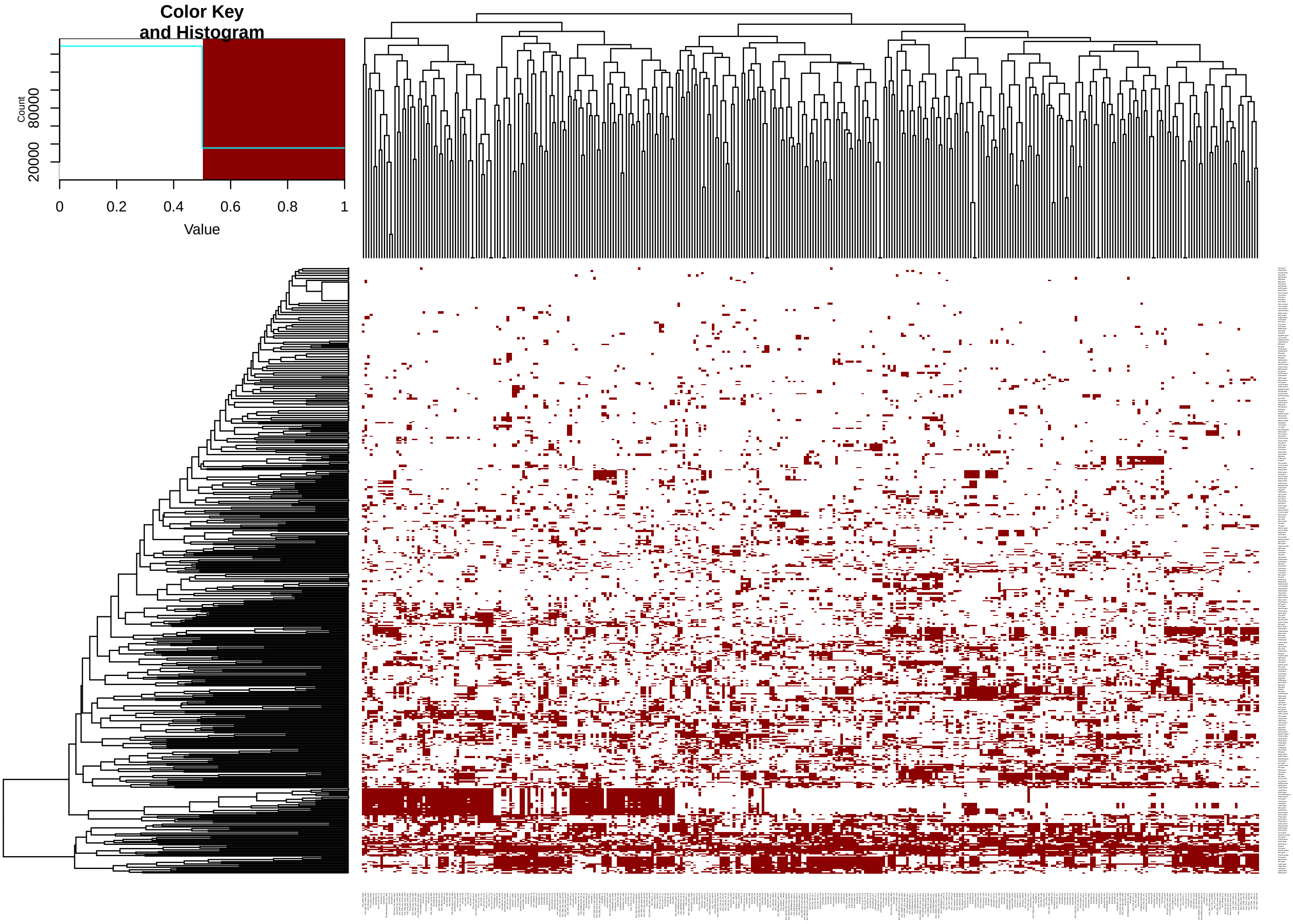
5. Aggregated top 100 paths for Cortex SD: interactors


Color Key
and Histogram
Count
20000
80000
0
0.2
0.4
0.6
0.8
1
Value

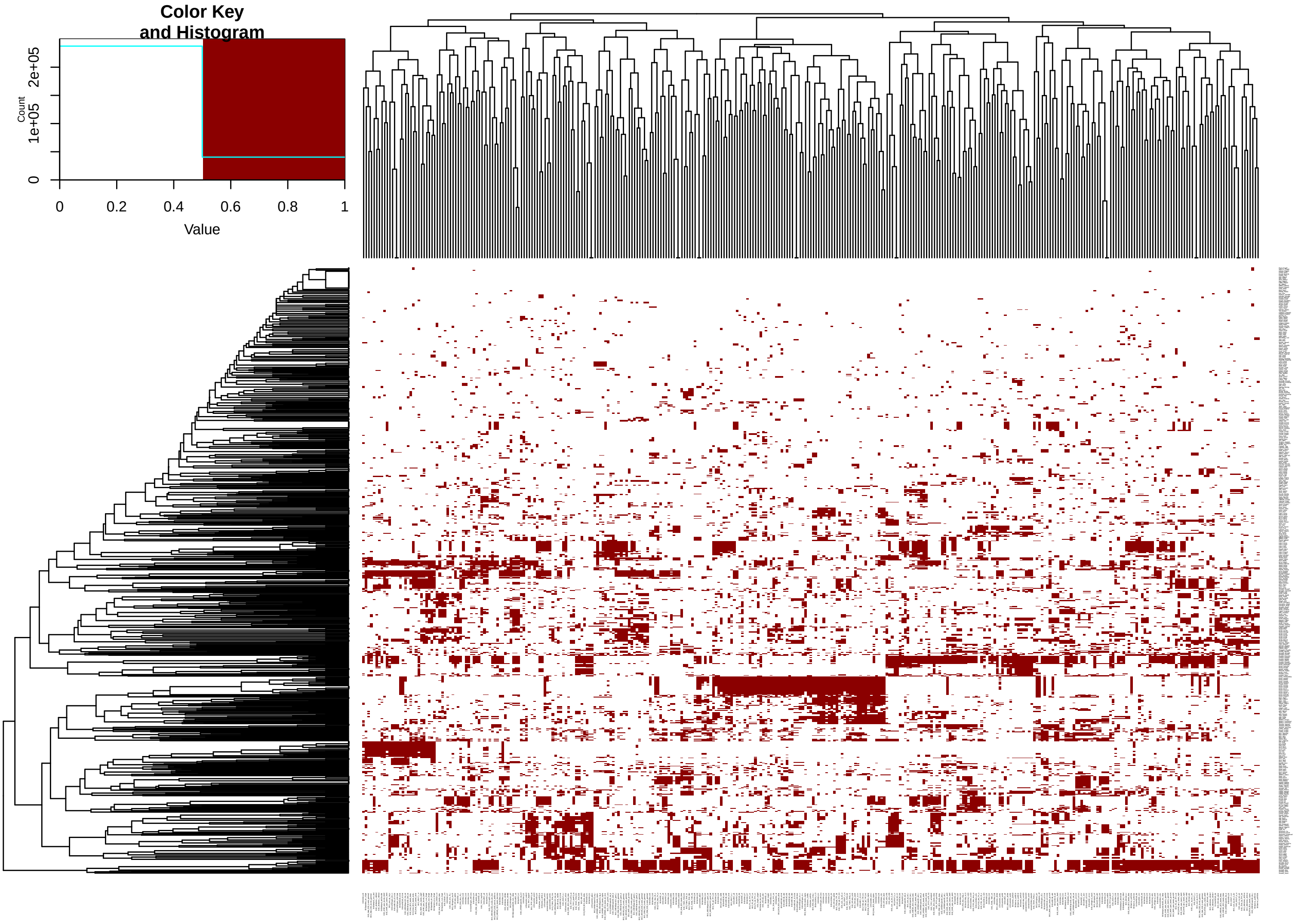
6. Aggregated top 100 paths for Liver NSD: interactions


Color Key
and Histogram
Count
0
1e+05
2e+05
0
0.2
0.4
0.6
0.8
1
Value

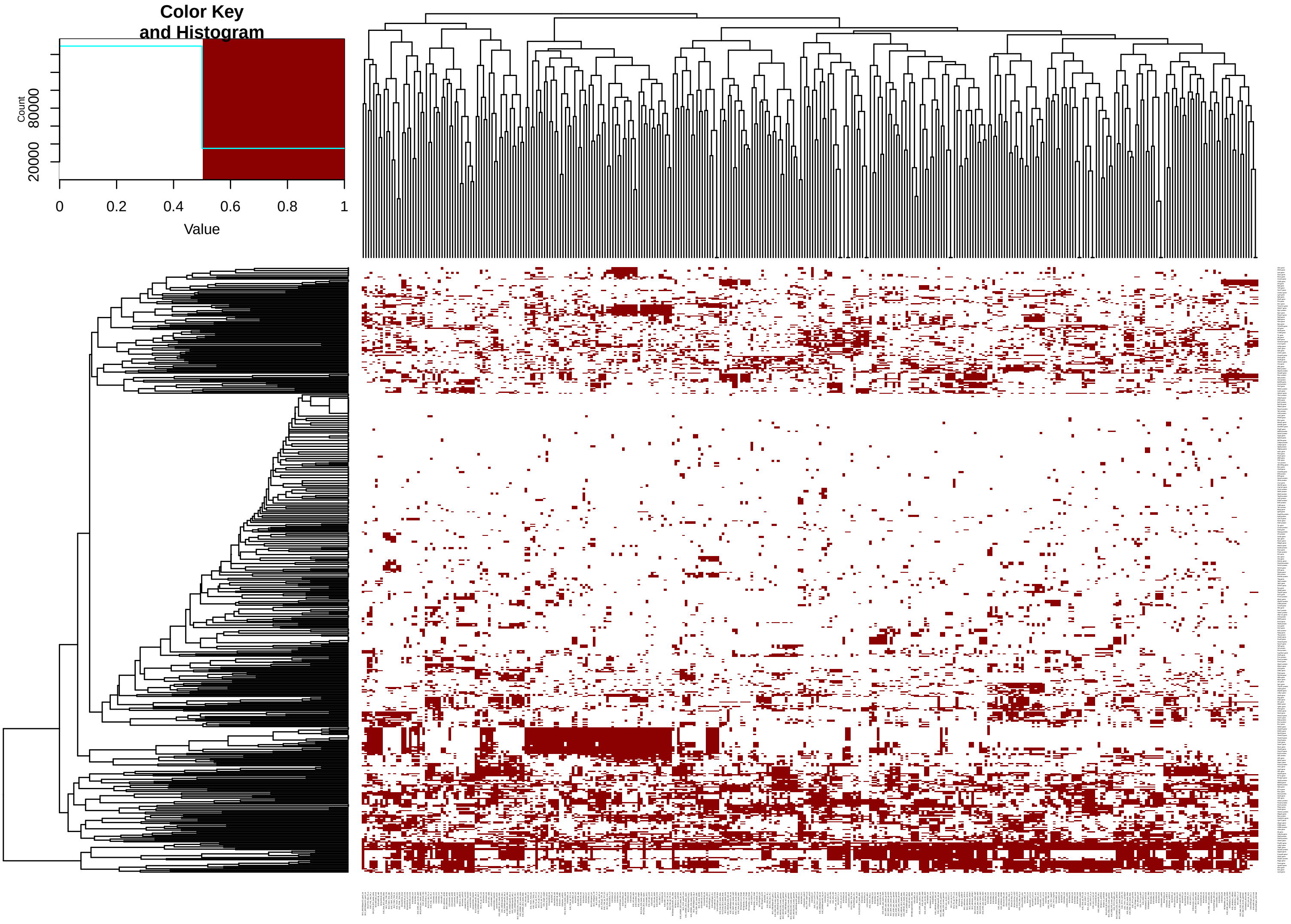
7. Aggregated top 100 paths for Liver NSD: interactors


Color Key
and Histogram
Count
20000
80000
0
0.2
0.4
0.6
0.8
1
Value

8. Aggregated top 100 paths for Liver SD: interactions


Color Key
and Histogram
Count
0
1e+05
2e+05
0
0.2
0.4
0.6
0.8
1
Value

9. Aggregated top 100 paths for Liver SD: interactors


Color Key
and Histogram
Count
20000
80000
0
0.2
0.4
0.6
0.8
1
Value