Year : 2018

AN INTEGRATIVE SYSTEMS BIOINFORMATICS APPROACH OF THE ENVIRONMENTAL, GENETIC AND MOLECULAR FACTORS REGULATING SLEEP

Jan Maxime

UNIL | Université de Lausanne

Faculté de biologie et de médecine

Centre intégratif de génomique & Swiss Institute of Bioinformatics

AN INTEGRATIVE SYSTEMS BIOINFORMATICS APPROACH OF THE ENVIRONMENTAL, GENETIC AND MOLECULAR FACTORS REGULATING SLEEP

Thèse de doctorat ès sciences de la vie (PhD)

Présentée à la

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

Par

Maxime JAN

Biologiste diplômé de l'Université de Lausanne, Suisse

Jury

Prof. Marc Robinson-Rechavi, Président

Prof. Ioannis Xenarios, Directeur de thèse

Prof. Paul Franken, Co-directeur de thèse

Prof. Alexandre Reymond, expert

Prof. Emmanouil Dermitzakis, expert

Lausanne 2018

IECB doctoral program

UNIL | Université de Lausanne

SIB
Swiss Institute of
Bioinformatics

Imprimatur

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

Président·e	Monsieur	Prof.	Marc **Robinson-Rechavi**
Directeur·trice de thèse	Monsieur	Prof.	Ioannis **Xenarios**
Co-directeur·trice	Monsieur	Prof.	Paul **Franken**
Expert·e·s	Monsieur	Prof.	Alexandre **Reymond**
	Monsieur	Prof.	Emmanouil **Dermitzakis**

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

Monsieur Maxime Jan

Master en sciences moléculaires du vivant Université de Lausanne

intitulée

An integrative systems bioinformatics approach of the environmental, genetic and molecular factors regulating sleep

Lausanne, le 18 mai 2018

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Marc Robinson-Rechavi

Content

Acknowledgments

First of all, I want to deeply thank my two thesis directors, Prof Ioannis Xenarios & Prof. Paul Franken for their guidance during the four years of my PhD, and moreover for giving me their trust and the opportunity to realize my thesis on an impressive, state-of-the-art dataset merging computer science, statistics, bioinformatics, genetics, sleep and metabolism.

I thank also all the members of my thesis jury, Prof Alexandre Reymond & Prof. Emmanouil Dermitzakis, to give their time for the evaluation of my work and to come at my thesis defense, as well as the president of my jury, Prof. Marc Robinson-Rechavi.

My work would not have been possible without the excellent contribution of all the persons that were involved in this project. I would particularly thank Dr. Shanaz Diessler for the colossal work she did on hundreds of mice to create this dataset. I also would thank Yann Emmennegger and Dr. Nicolas Guex for the hours spent to generate high quality EEG annotation. Concerning the bioinformatics analyses, I greatly thank Dr. Mark Ibberson, Frédéric Burdet and Dr. Marco Pagni for their precious inputs and advices. And I am also highly grateful to Dr. Robin Liechti, Lou Gotz and Dr. Martial Sankar for their work on the BXD web application. I would like to mention also the contribution of Dr. Charlotte Hor during this project, for her input and help with article writing.

Thanks to all the people I worked with or spent time with in Vital-IT and Franken's group, in the office, in the lab, at lunch, on the edge of the lake, in a snowstorm snowshoeing or on a paddle.

Finally, I would like to thank my family and friends for all the merry moments around a table (with food or dice) or on skis. And more than anything, I thank my parents that supported and encouraged me during these years.

Abstract

Environmental changes and genetic variations are two important drivers of biological diversity. In complex traits, a multitude of genetic and environmental factors interact and combine in cryptic ways to direct the phenotypic variation. Sleep is a classic illustration of a complex trait that is vital and heritable but still poorly understood. Many aspects of sleep like the timing, duration and quality are regulated by the interaction of two processes: the circadian oscillations and the sleep homeostasis.

In the context of a study that aimed at uncovering more clearly the molecular pathways regulating the sleep homeostat through the ambiguous relationship that exists between sleep-wake cycle and metabolism, we built, assembled, analyzed an extensive multi-scaled dataset using the systems genetics design. Machine learning algorithms and novel high-throughput sequencing technology permit to appraise more precisely and broadly the plethora of physiological and molecular phenotypes that contribute to sleep under disparate circumstances and genetic background, in order to build novel hypotheses based on data-driven discoveries. This dataset is composed of 33 recombinant inbred lines (RIL) from the BXD panel that were interrogated under sleep deprivation and undisturbed conditions for 341 sleep-wake related physiological phenotypes, 124 blood plasma metabolites, and cortical and liver transcriptomics. First analyses pointed out the pervasive effects of sleep deprivation and genetics both at the molecular and behavioral level and the complex interaction between genetic and environmental factors at all phenotypic layers. Then, two novel integrative methods were developed, the first to prioritize candidate genes within large associated genomic regions for physiological or metabolic phenotypes and the second to visualize the meta-dimensionality of the molecular network using the deterministic structure of hiveplots. Our findings led to the discovery of a bidirectional relationship between fatty acid turnover and sleep homeostasis but also between brain slow-waves activity and ionotropic glutamate receptor transport. Using markup language and cloud-based technologies, we aimed at transforming this resourceful, multidisciplinary dataset into an exploitable digital research object. The generation of dynamic analysis reports and workflow metadata promoted the reproducibility this data-object. In addition, tools were developed for the exploration and mining of integrated data. The resulting database and associated web interface ensures the reusability of this dataset and associated methodologies.

Résumé

La diversité biologique est dirigée par deux opérateurs importants, les changements environnementaux ainsi que les variations génétiques. Pour les traits dits complexe, leur variation est le fruit de nombreux facteurs génétiques et environnementaux qui vont interagir et se combiner, souvent de manière cryptique. Le sommeil est un exemple-type de trait complexe, il est vital et héritable mais fondamentalement méconnu. La régulation de nombreux aspects du sommeil comme sa durée, timing ou qualité fait intervenir deux processus : les oscillations circadiennes et l'homéostasie du sommeil.

Afin de mieux cerner les voies qui régulent le mécanisme d'homéostasie du sommeil, en particulier celle mêlant le métabolisme, nous avons créé, assemblé et analysé un grand set de données en utilisant une approche dite de génétique des systèmes. Avec l'aide d'algorithmes d'apprentissage automatique et de nouvelles technologies de séquençage à haut-débit, nous avons pu mesurer dans des conditions et contextes génétiques différents de nombreux phénotypes moléculaires ou physiologiques qui contribuent à la régulation du sommeil. Notre approche étant ainsi principalement axée sur la construction d'hypothèse guidée par les données. Ce set est composé de 33 lignées de souris consanguines recombinantes (BXD) dont on a examiné, dans des conditions de privation de sommeil et de contrôle : 341 phénotypes physiologiques liés au sommeil et à l'éveil, 124 métabolites du plasma sanguin, ainsi que leur transcriptome du cortex et du foie. Les premières analyses ont pointé l'effet aigu de la privation de sommeil, de la génétique ainsi que leur interaction sur tous les niveaux de phénotypes. Ensuite, deux nouvelles méthodes d'intégration ont été développées, la première pour prioritiser les gènes opérateurs du sommeil et du métabolisme à l'intérieur de grande région génomique, la deuxième pour visualiser la méta-dimensionalité des données moléculaires via une structure de 'hiveplot'. Nous avons mis en avant une relation bidirectionnelle entre les modifications d'acides gras et l'homéostasie du sommeil, ainsi que l'activité des ondes lentes du cerveau et le transport de récepteur au glutamate ionotropique. En utilisant le langage de balisage ainsi que des technologies basées sur le cloud, nous avons cherché à transformer ce jeu de données en un objet de recherche numérique. La reproductibilité de cet objet a été améliorée par la génération de rapports d'analyse dynamiques ainsi que de métadonnées. De plus, des outils ont été développés pour l'exploration et l'extraction de données via une interface web et assurent ainsi la réutilisation de ce set et de ces méthodologies associées.

Chapter 1. Introduction

1.1 Complex Systems

1.1.1 Genetic & environmental contributions

The origin of biological variability has been studied for centuries. Scientists have explored the causes for phenotypes to differ from a predefined normality that in many cases fit the standard gaussian distribution (Büttner 1998), with still unresolved and debated questions about the exact origin of this observed variation, nature vs nurture or most likely both. These questions are immediately relevant for issues concerning health and sensitivity or resistance to develop disease (Theodoratou, Timofeeva et al. 2017) but also for some of our innate characteristics like physical ability (Issurin 2017), aging (Merkwirth, Jovaisaite et al. 2016, Ryu, Mouchiroud et al. 2016) and propensity to sleep (Viola, Archer et al. 2007).

In biology, many major breakthroughs helped to apprehend and confine the genetics and environmental factors driving phenotypic variability. On the genetics side, the discoveries of Gregor Mendel's laws of inheritance, the uncovering of the DNA structure together with the development of high-throughput sequencing technologies are but a few examples that have greatly improved our understanding of how and what constitutes some of the predetermined characteristics and that are transmitted to next generations. The DNA variations that are encountered between individuals, like Single Nucleotide Polymorphism (SNP), Insertion-Deletion (Indels) or Structural Variation (SV) can alter gene expression or protein structure and are components of the Mendelian or polygenic inheritance of traits. On the environmental side, the interaction between the organism and its environment directly influences its biology. Already in the XI century in China, it was observed that humans that were intentionally inoculated by smallpox became resistance to this disease before spreading to western world (Riedel 2005). Organisms can undergo molecular adaptation upon a challenge thereby gaining novel characteristics that can be either short- or long-lived through diverse mechanisms like epigenetics (Hunter 2012).

Several approaches have been used to investigate environmental and genetic variables independently. The environmental factors contributing to phenotypic variance are generally

symbolized by *E* and the genetic factors by *G*. Organisms with identical genomes are used worldwide to study environmental driven hypotheses. The representative organism are clonal cell lines or animal models that are maintained and distributed by organization such as the co-isogenic C57BL/6J mouse, an inbred strain maintained by the Jackson Laboratory that is fully sequenced. On the human side, monozygotic twins are commonly used to remove the genotypic factors diversity from the experiment as they share common DNA. They are also compared with dizygotic twins to calculate the contribution of genetic factors and estimate the heritability of a trait like sleep as they often share common environment (Linkowski 1999). Two distinct types of heritability can be estimated, the narrow-sense (h^2) and the broad-sense (H^2). Narrow-sense heritability represents the additive contribution of independent genetic factors and broad-sense represents the contribution of all genetic factors and their combined effects, including recessive/dominant relationship and epistasis (*G*x*G*).

Two general approaches were used in genetics to uncover the relationship between phenotypic variation of interest and the causative genetic factor(s): *forward* and *reverse* genetics. The *forward* approach aimed at determining the causative genetic variation that underlie an observed phenotype (from phenotype-to-gene). Many designs exist like using natural occurring variation within a population or random mutagenesis screening. On the other hand, *reverse* genetics uses a gene-to-phenotype approach by targeting specific genes and observing the consequences of a gain or loss of function. Again, many designs exist using a gene partial or complete knock-out, or more recently, enhanced and precise approaches were developed like gene editing using CRISPR/Cas9 (Shah, Davey et al. 2015, Gurumurthy, Grati et al. 2016).

These two concepts of genetic and environmental factors have been widely investigated independently using reductionist approaches, where their respective roles on biological phenotypes were investigated separately. However, a huge number of interactions underlie biological systems, which led to e.g. the now well-known issue of the missing heritability in Genome Wide Association Studies (GWAS). GWAS (a *forward* genetics approach) associated many genetic markers with a given traits like height (Maher 2008) or diabetes (Stančáková and Laakso 2016). Even so the heritability was high (80%), the collection of markers could only explain for a very small amount of the variance (5%). In the same way, biological interaction can greatly affect single marker effect as observed within our own data, where an association with a biogenic amine was found with genotypes on chromosome 2 in mice (Figure 1 left), but

measures on other lines from the same panel could display highly divergent concentration of this metabolite given the first hypothesis (Figure 1 right). A simple SNP-phenotype association is often not enough for a biological systems characterization that is consistent with genetics and environmental background that were not yet observed.
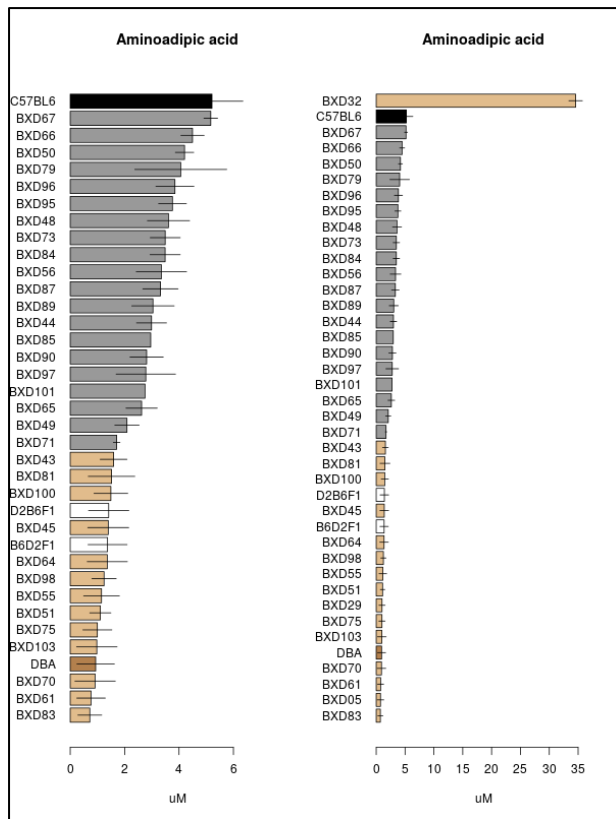
Figure 1 : Alpha-amino adipic acid (α-AAA) level in BXD panel.

Left: Plasma blood concentration of the α-AAA within the mouse BXD/RwwJ panel, ordered by concentration [µM]. Genotype-Phenotype linkage analysis revealed a QTL on chromosome 2 that explains 72.5% of the α-AAA variance within the investigated mouse population. The QTL includes *Dhtkd1*, encoding an enzyme subunit involved in lysine degradation known to controls α-AAA levels in BXD lines (Wu, Williams et al. 2014). Color indicates parental strains and the allele carried by the BXD on the associated region, grey = C57BL/6J allele, beige = DBA2/J allele. **Right**: Identical values used on the left panel, now including to 3 older BXD/TyJ strains (BXD32, BXD05, BXD029). BXD032/TyJ is a DBA2/J allele-carrier but the α-AAA blood concentration is 10 times higher than other BXD. This excessive concentration is replicable in BXD032 and unique in our 124 different metabolites population. This difference could be explained by higher-order interaction (GxG) or differences within generated BXD population (see BXD section and discussion).

1.1.2 Toward systems approaches:

Novel experimental designs and associative methods were proposed to improve the classical reductionist approaches and examine in more detail the results of the *E* and *G* factors interactions. A few simplified examples related to health, sleep and life style are illustrated here to apprehend the general complexity and high connectivity that can be found behind polygenic phenotypes. In section 1.3, the sleep regulatory systems will be addressed more profoundly. As mentioned previously, the large effect of *E* and *G* factors on their own were highlighted using epidemiological studies or GWAS. Currently, our life style under high stress (*E*) have negative

repercussions on our health (Kopp, Stauder et al. 2008), furthermore artificial light generated by recent technologies tends to disrupt our sleep (Falbe, Davison et al. 2015) which reduces its restorative function. The combination of diverse risk factors as mentioned (E x E) can have multiplicative negative repercussion on our health, as emphasized in (Ding, Rogers et al. 2015) where long/short sleep duration, prolonged sitting, and physical inactivity increased significantly the risk of mortality. We often associate these risk factors with other side effects like with weight gain (Bayon, Leger et al. 2014) and type-2 diabetes (Shan, Ma et al. 2015). The negative consequences of stress and sleep perturbation can be viewed as a singular event or be considered as a causal chain, where chronic disorder and the inability of biological systems to return to a balanced state can induce other complications at longer term ($E \rightarrow E$), highlighted using temporal disease trajectories (Jensen, Moseley et al. 2014). Therefore, a central question is to know which pathways are perturbed, how they are balanced, and which genetic factors may increase or reduce the chances of complications. The accumulation of visceral fat during weight gain for example is associated with increased type-2 diabetes risk (Sabag, Way et al. 2017) and is under genetic control (G) (Fox, Liu et al. 2012).

At a mechanistic level, the underlying molecular phenotypes (transcript or protein) responsible for this phenotype could be driven by cis-regulation or trans-regulation ($G \rightarrow G$), but the complete regulatory system involved is more complex. These interactions are also highly likely to be epistatic (G x G) given the millions of molecular interactions that exist only at the protein level (Szklarczyk, Franceschini et al. 2015). Even minor genetic differences between individuals may highly reshape molecular networks and phenotypic outcome, like the genetically closed mouse sub-strains: C57BL6/J and C57BL6/N that only distinguish for 34 SNPs, 2 indels and 15 SVs (Simon, Greenaway et al. 2013), but these few variations led to different observations for levels of a hepatotoxic biomarker and results interpretation (Bourdi, Davies et al. 2011). At the molecular level, both environmental and genetic factors (G x E) influence the organism transcriptional environment. In diploid organism, the gene Allelic Specific Expression (ASE) depend on both GxE and cis-/trans- regulatory element (Buil, Brown et al. 2015).

This highly interactive architecture limits the development of preventive medicine (Narimatsu 2017) such as in coronary artery disease (CAD) (Lanktree and Hegele 2009) using reductionist approaches or current genetic-based predictive models (Schrodi, Mukherjee et al. 2014). It was also proposed to be the cause of the GWAS missing heritability problem, that

could be solved by integrating intermediate phenotypes (Blanco-Gómez, Castillo-Lluva et al. 2016), epigenetics (Trerotola, Relli et al. 2015), epistasis (G x G) (Zuk, Hechter et al. 2012) and host-microbiome interaction (G x E) (Sandoval-Motta, Aldana et al. 2017). It is hypothesized that biological complexity arises from the non-additive interactions among the factors driving biological variability resulting in a system that is far more complex than the sum of its individual constituent parts. This phenomenon, referred to as *emergence* (Macklem 2008), necessitates to explore life's complexity as an ensemble/systems rather than a collection of isolated entities.

Therefore, the understanding of the underlying *E* and *G* factors interactions within a molecular system context was intended to be a major step toward better precision medicine and insight on disease co-occurrence (Carlsten, Brauer et al. 2014, Hu, Thomas et al. 2016), but also for an enhanced comprehension of normal traits. Thus, to "*deciphering the dark matter*" (Crawford 2016) surrounding these complex traits caused by the vast number of molecular interactions ubiquitous in regulatory networks, systems genetics was proposed as a novel multidimensional, population based, experimental model (Civelek and Lusis 2014). In this design, intermediate phenotypes such as transcriptomics, proteomics, metabolomics or microbiome are integrated in parallel of the end-phenotypes of interest. These phenotypes are measured under the different genetic backgrounds from the population observed (*G*) and different condition (*E*). The systems genetics approach was successfully used in human for epilepsy, neuronal development, obesity or CAD (Johnson, Behmoaras et al. 2015, Johnson, Shkura et al. 2015, Talukdar, Foroughi Asl et al. 2016, He, Sun et al. 2017), F2 and Recombinant Inbred (RI) lines of mouse for stress and fear behavior (Park, Gale et al. 2011, Jiang, Scarpa et al. 2015), drosophila (Ayroles, Carbone et al. 2009), plants (Basnet, Del Carpio et al. 2016), and virus (Kollmus, Wilk et al. 2014) in host-pathogen interaction. The population genetics and different environmental conditions are the motor of phenotypic variation that is then used for associative statistical methods, like Quantitative Trait Locus (QTL), gene co-expression (Plaisier, Horvath et al. 2009, Park, Gale et al. 2011, Langfelder, Castellani et al. 2012) or Bayesian network inference (Jiang, Scarpa et al. 2015). The phenotyping approach demands precise and broad measurement, both of molecular and of physiological phenotypes, to uncover the mechanisms underlying a trait of interest. The efforts in next generation sequencing and quantitative mass-spectrometry favored sensitivity and specificity of intermediate molecular phenotype measurement in genomics, transcriptomics, proteomics and

metabolomics. But ultimately disease/physiological responses also demand efforts on the phenotyping strategy (Großkinsky, Svensgaard et al. 2015) to improve the accuracy and our understanding on genotypic-phenotypic contribution (Ala-Korpela, Kangas et al. 2011). These approachs require novel protocols and computational approaches like image-analysis or Machine Learning (ML) (Singh, Ganapathysubramanian et al. 2016) for large dataset assimilation and classification.

Finally, the quality of acquired data is a major step toward data integration between the different actors of the systems genetics community. Multiple phenotypes can be recorded by a single group and/or experiments, but the combination and integration of multiple independent dataset is a key step toward reproducible and improved molecular pathways discovery. Approaches using a single QTL association were already shown to be successful in parental behavioral (Bendesky, Kwon et al. 2017), hypoglycemia (Picard, Soyer et al. 2016) or glaucoma (Chintalapudi, Maria et al. 2017) traits but the need of intermediate phenotype is often required for candidate gene identification and further validation. Moreover, these types of approaches require significant association of sufficient effect size which are expected to be less present in complex traits because of the underlying polygenic structure, whereas large integrative initiative would better to detect low effect size associations. The current problems and solutions for data generation, integration, reusability and generally the bioinformatic data-management strategy of systems genetics design are further described in section 1.4 and chapter 3.

A proper animal model must be selected to privilege data-sharing, the use of F2 mice designed for a single experiment like (Jiang, Scarpa et al. 2015) impose some limitation for the data integration by others as these mouse are not maintained and/or widely distributed. Furthermore, they have a limited amount of recombination event which led to large associated region that can contains sometimes hundreds of genes.

1.2 The BXD mouse panel:

Animal models have been used for a long time to study complex traits as their use offers many advantages compared to humans with respect to e.g. controlling both environmental and genetic factors (Schughart, Libert et al. 2013) and tissue sampling for quantification of the intermediate phenotypes driving complex traits that could not be properly controlled or measured in humans. Despite some controversial use for translating mouse mechanisms to human mechanisms in traits like inflammation (Seok, Warren et al. 2013, Takao and Miyakawa 2015), mice were proposed to be a good model for mechanism discovery related to sleep disorder (Toth and Bhargava 2013) and sleep physiology (Franken, Malafosse et al. 1999). On the genetic side, the use of a Genetic Reference Population (GRP) seems to be the most appropriate choice to build a systems genetics resource because of their known and fixed genotypes for reusable and reproducible mapping results.

The BXD genetic reference population is one of the most advanced GRP in mouse that was specifically bred for complex traits investigation and QTL analysis. Other mouse populations with higher genetic diversity exist and are further described in the Discussion. Multiple batches of BXD have been produced since 1971 and are still generated with up to date 198 available lines. This panel contains Recombinant Inbred (RI; BXD/*TyJ* lines 1-42) and Advanced Recombinant Inbred (ARI; BXD/*RwwJ* lines 43-220) lines that were generated from two co-isogenic strains: C57BL6/J (B6 or B) and DBA2/J (D2 or D). The ARI lines were published in 2004 (Peirce, Lu et al. 2004) and contained twice more recombination events compared to conventional RI. The older lines should be considered separately from the new lines (Figure 2) as they show behavioral differences and significant different mapping results (Philip, Duvvuru et al. 2010). One explanation is that BXD were generated with two different DBA sub-strains: DBA2/Rj and DBA2/J, due to genetic drift that have accumulated over the years (Taylor, Wnek et al. 1999, Shin, Pandey et al. 2014). Each BXD mouse represents a unique combinatorial mixture of C57BL/6J and DBA2/J genotypes (with mitochondrial and Y chromosome exception) and by consequence with unique transcriptional, metabolomic and behavioral signatures.
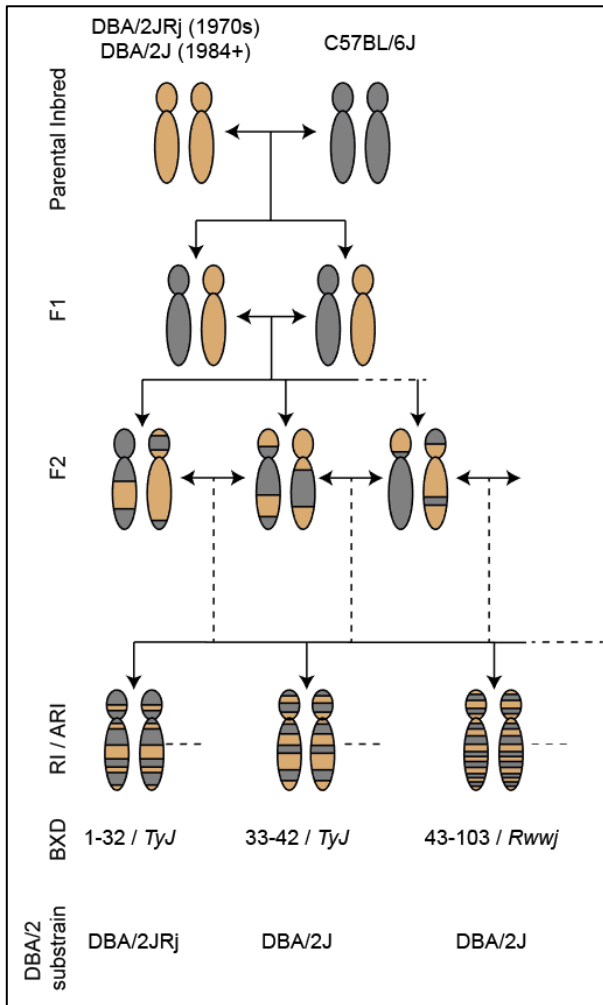
The BXD mouse panel are Recombinant Inbred (RI) and Advanced Recombinant Inbred (ARI) lines developed by crossing DBA2/J and C57BL6/J co-isogenic strains. These two strains differ for around 5 million sequence variants (Wang, Pandey et al. 2016). Some variations can explain part of observed phenotypic divergence, for example *Nnt* functional allele in DBA2/J enhance insulin secretion (Yeadon 2014). BXD/TyJ 1-42 (RI) were developed using conventional F2 sib-mating for at least 20 generation. BXD/RwwJ 43-103 (ARI) were generated using advanced intercross lines (9-14 intercross generations) followed by 14-18 generations of sib-mating for genetic stabilization. ARI have approximately twice as many recombination events compared to RI lines. Before 1984, BXD lines were generated using the DBA/2JRJ sub-strain as parental strain which contains other variations, like functional *Klrd1* (CD94). DBA2/J sub-strain used after 1984 have a deletion on the last exon of *Klrd1*. The BXD family contains mainly DBA2/J maternal mitochondrial chromosome and C57BL6/J paternal Y chromosome, with exceptions for 10 lines that should be considered as DXB (see discussion).

1.2.1 Phenotype-Genotype association

For genome-wide association, statistical power is strongly affected by sample number, allelic frequency and effect size (Manolio, Collins et al. 2009, Arnar, Andersen et al. 2016), therefore QTL mapping in mouse compared to GWAS offers some advantages (Flint and Eskin 2012). Compared to GWAS which require thousands of samples in order to each the 5e-8 p-value threshold for significant association (Fadista, Manning et al. 2016), the BXD panel can use less than 40 samples to reach an acceptable statistical power (Andreux, Williams et al. 2012). To diminish rare variants number, GWAS can focus on isolated population (Zeggini 2011, Jeong, Alkorta-Aranburu et al. 2014) which limit studies to specific phenotypes of interest or use more

genetically homogeneous population, whereas in BXD only a few regions have a low allelic frequency in the complete panel.

However, the associated region of the BXD are large linkage disequilibrium (LD) blocks of a few mega base-pairs (Mbps) that may contain lots of genes (Figure 3) using interval mapping methods, therefore a lot of effort was done on increasing recombination rate in the BXD panel and decreasing associated region size. On the other side, current GWAS methods can use direct marker association which is more precise for candidate gene identification but can be problematic in case of allelic heterogeneity, however novel techniques using machine learning methods for pattern mining may change this last issue (Mieth, Kloft et al. 2016, Llinares-López, Papaxanthos et al. 2017). Another analytical simplification of the BXD panel is the genomic homozygosity that is found in a majority of loci, with a few exceptions of heterozygosity. Homozygosity reduces complexity of the model, by removing the possibility for allelic specific expression that is found in non-inbred diploid organism, and the dominant-recessive effect can be inspect using F1 lines.
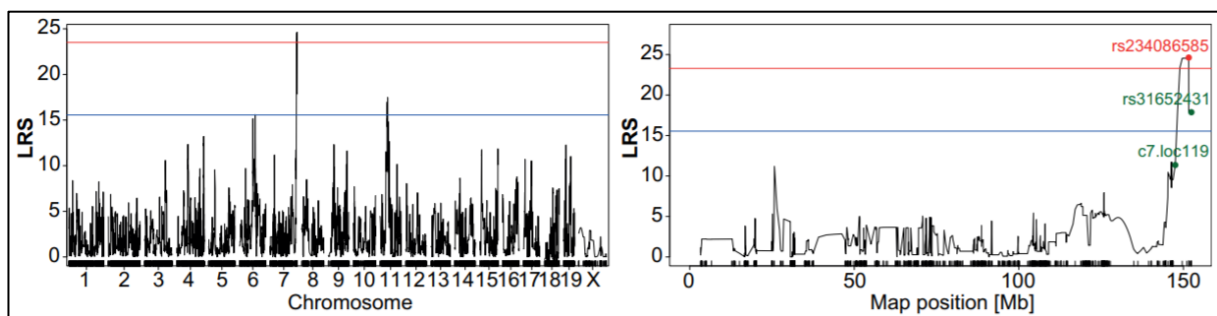
Figure 3 QTL analysis for glucagon ratio in BXD.

Example of interval QTL mapping for genotype-phenotype association in the BXD panel. **Left**: The association test is performed genome-wide by likelihood ratio test given gaussian distribution parameters estimation by Expected-Maximization (EM) algorithm. False Discovery Rate (FDR) is estimated using 1000 permutations, red line indicates significant threshold (FDR<0.05), blue line indicates suggestive threshold (FDR<0.63). **Right**: Confident interval (green markers) for the true QTL location is appraised using 1.5 log odd (LOD) difference from the peak (red marker). Adapted from (Picard, Soyer et al. 2016)

1.3 Sleep regulation

1.3.1 Fundamentals of sleep regulation:

In biology, one of the most enigmatic phenomena is sleep, while this behavior seems to be universally conserved among species and quite easy to master: "*one sleeps when one is tired — mostly at night—and awakens the next day usually feeling rested and refreshed*" (Spiro 2005), its function (Krueger, Frank et al. 2016), regulation and the molecular processes involved still remain elusive (Mignot 2008). In mammals, we observed a large variety of sleep parameters like duration (Ollila, Kettunen et al. 2014), brain activity (Franken, Malafosse et al. 1998) or restorative benefit (Wilkinson and Shapiro 2012) that can be quite difficult to fully capture and comprehend due to the multi-factorial processes involved. Time is an essential factor, in human, for most of us a third of our lifespan will be occupied sleeping, more during development in new-born thus with age-related changes (Dijk, Duffy et al. 2000) but also with seasonal effects (Kume, Makabe et al. 2017) and long perturbation after sleep loss (Fifel, Meijer et al. 2018). In addition, the genetics and environmental factors such as jet-lag, night-shift work or technology can modify important sleep properties, which can lead to chronic sleep disorders (Nunn, Samson et al. 2016) with consequences on our health (Altevogt and Colten 2006).

In mammals, some sleep attributes can be defined by cortical activity recording and the identification of vigilance states over prolonged periods. Two sleep states should be distinguished: (i) Rapid-Eye-Movement sleep (REM, or paradoxical sleep) characterized in the mouse by the prevalence of theta oscillation (6-9 Hz) in the EEG and muscle atonia and (ii) non-REM (NREM) sleep, the predominant sleep state, with a high contribution of slow oscillations (0.75-4.5 Hz) in the EEG. The third state is wakefulness (Wake) with a higher variability in EEG and EMG. The transformation from a time domain signal into a frequency domain signal to identify the prevailing frequencies to a complex waveform such as the EEG is performed using Fast Fourier Transform (FFT) (Figure 4). This transformation allows for signal transformation to perform power spectrum analysis but also computational automation. The annotation of spectral data can be done with visual inspection by an expert where the time to perform this task will mostly depend on the length of the epoch (time-window or annotation resolution) and period of recording. Several classification machine learning algorithms were
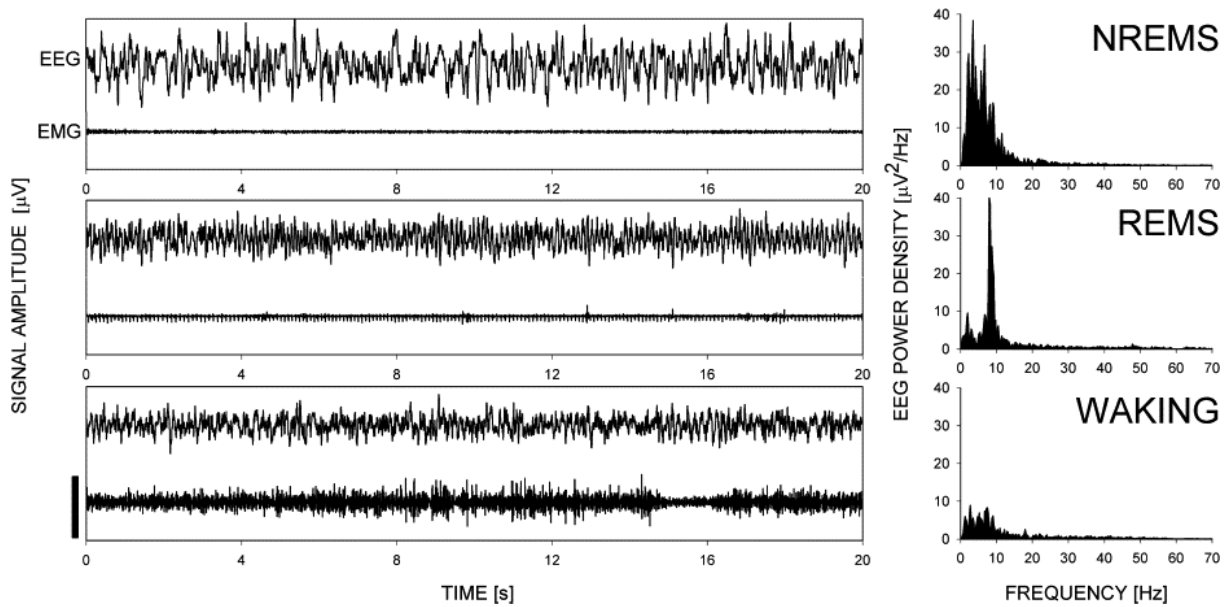
Figure 4 : EEG/EMG signal during Wake, NREM sleep and REM sleep

Examples EEG/EMG signal in the time domain and EEG spectral power transformation into frequency domain using FFT during wakefulness, NREM sleep and REM sleep. Adapted from (Xie, Dumas et al. 2005)

developed to help the annotator (Lampert, Plano et al. 2015) or automatically annotate signals (Sunagawa, Séi et al. 2013). The main difficulties of classification is signal normalization (Katsageorgiou, Lassi et al. 2015), misclassification by single classifier approach (Gao, Turek et al. 2016) and the great diversity of EEG/EMG power bands when different mouse strains are used (Franken, Malafosse et al. 1998).

A model for sleep regulation was proposed 35 years ago (Borbély 1982), known as the two-process model (Figure 5) where sleep-wake cycle is controlled by two distinct mechanisms, circadian rhythm (C) and sleep homeostasis (S). Process C will dictate the optimal time in the day for a resting phase and process S will compensate for the sleep loss through the accumulation of sleep pressure during wakefulness and its decrease during sleep. Sleep homeostasis will attempt to maintain the level of required sleep close to a set-point and compensate by negative feedback loop for deviation due to the everyday activity or disturbance like sleep deprivation. From a mechanistic point of view, the circadian process (C) was shown to be mainly driven by the SupraChiasmatic Nucleus (SCN) of the hypothalamus that acts as a master circadian pacemaker (Dibner, Schibler et al. 2010) via clock genes (Franken and Dijk
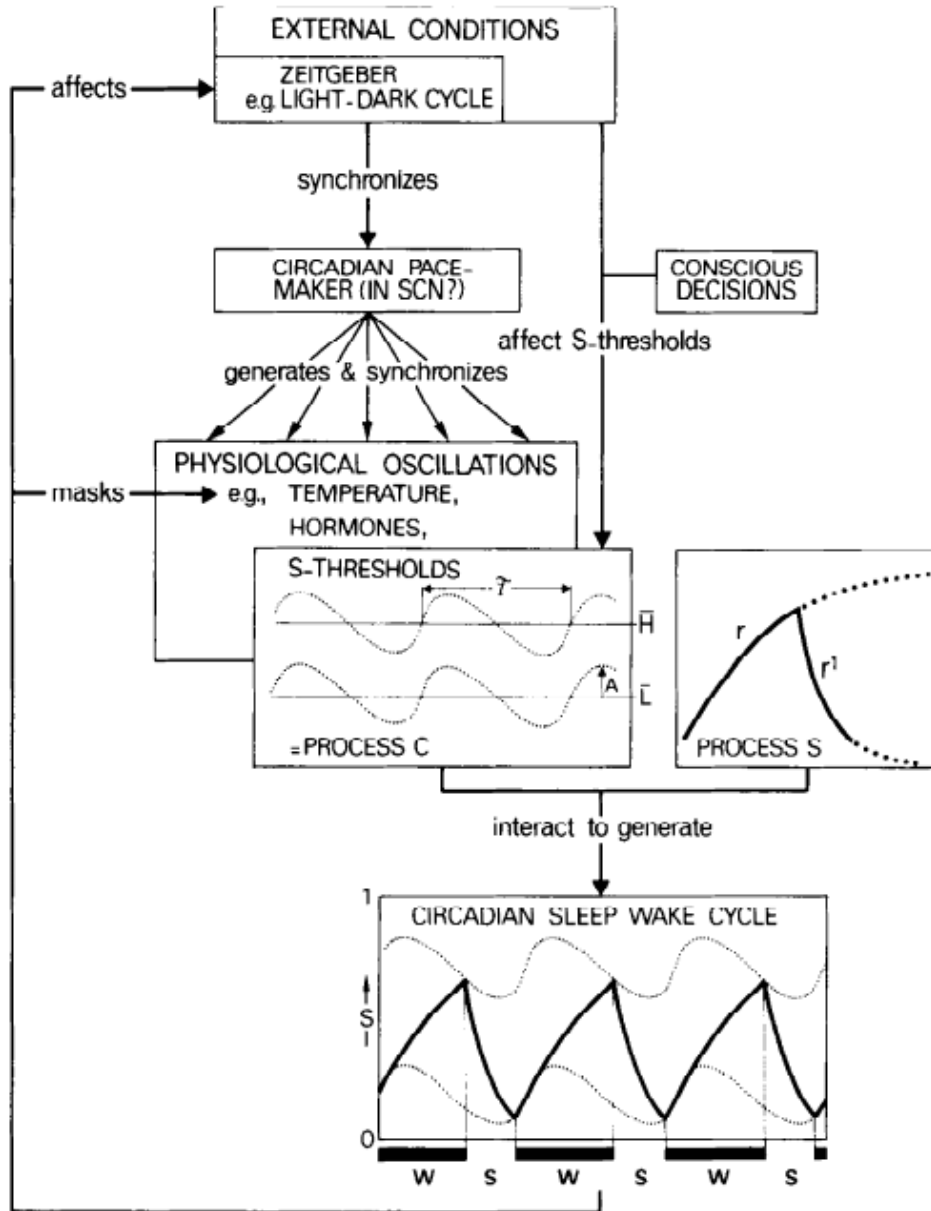
Figure 5 : The two-process model

The two-process model for sleep-wake regulation as proposed by Borbély. Circadian rhythm oscillations (C) are entrained by external factors (Zeitgebers) to favor the optimal time for propensity to sleep or be awake during the day (in rodent, light will induce sleep). Sleep homeostasis (S) process maintains the adequate quantity of sleep that is needed by the organism through the accumulation of a *sleep pressure* by time spent awake, and a decrease during sleep. As a homeostatic regulatory process, 3 essentials component are expected to regulate sleep need in a feedback loop: (i) a sensor for sleep pressure, (ii) a controller for a mediated response compared to an optimal set-point and (iii) an effector that modify the sleep behavior (Mang and Franken 2013). Adapted from (Daan, Beersma et al. 1984)

2009), that set internal and peripherical rhythms and is synchronized daily by external factors (Zeitgebers) like the 24h light-dark cycle of our planet. The increase need of sleep (Process S) can be measured by the increase of the prevalence and amplitude of EEG slow-wave (delta

waves: 0.75-4.5 Hz) during NREM sleep that follow sleep-wake distribution (Tobler and and clinical 1986) or the increase of time spent asleep after challenging the sleep homeostat by sleep deprivation (Mang and Franken 2015). Unlike the SCN driving circadian rhythms, studies focusing on brain structure or core mechanisms regulating the sleep homeostasis process did not reveal similar 'master' regulators. The two-process model was described as two interactive processes (Borbély 1982, Daan, Beersma et al. 1984) that allows the organism at the physiological level to stay awake during its waking phase despite an increasing sleep propensity and sleep during its resting phase despite a decreasing need for sleep (Dijk and Franken 2005). And at the molecular level some clock genes like *Per2* were shown to participate at both processes using *forward* and *reverse* genetics approach, reviewed in (Mang and Franken 2015). Furthermore the desynchrony of sleep wake-cycle and circadian rhythm in human strongly affect molecular balance (Archer, Laing et al. 2014) and is suspected to be involved in depression that could be treated with the help of the two-process model (Borbély, Daan et al. 2016). The complete underlying mechanisms and relations between these two processes are still unclear.

1.3.2 Genetics of sleep:

The contribution of genetic factors to many aspects of sleep phenotypes was revealed to be particularly important. In twins, the estimated heritability of NREM sleep power spectra between 8-16 Hz (De Gennaro, Marzano et al. 2008) was found to be 96%, making EEG spectral content one of the most heritable traits in human. Other sleep-wake aspects were found to be determined by genetic factors such as sleep disorder (Kimura and Winkelmann 2007,), neurobiology of vigilance state via neurotransmitters (Crocker and Sehgal 2010,) or clock genes (Franken 2013). Therefore, *forward* and *reverse* genetic approaches were often used to disentangle the genetics behind the different sleep parameters (Franken and Tafti 2003, Andretic, Franken et al. 2008) and identify single variation effects. For instance, with the help of the BXD1-32/*TyJ* mouse panel, QTLs were associated with delta power increase after sleep deprivation (Dps1) (Franken, Chollet et al. 2001) and *Homer1a* was proposed as a probable candidate gene (Maret, Dorsaz et al. 2007, Mackiewicz, Paigen et al. 2008). Homer1a, a metabotropic glutamate receptor regulator (Kammermeier and Worley 2007), was suggested to link sleep homeostasis and synaptic down-scaling (Diering, Nirujogi et al. 2017), which is a

proposed function of NREM sleep for the Synaptic Homeostasis Hypothesis (SHY) involved in synaptic plasticity and memory (Tononi and Cirelli 2003, Tononi and Cirelli 2006, Tononi and Cirelli 2014). Among the proposed molecular pathways for sleep homeostasis and synaptic plasticity regulation, the product of AMPA-type glutamate receptors cascade (Shepherd 2012) are being explored. Other genetic approaches have been used, e.g. random mutagenesis screening (Funato, Miyoshi et al. 2016) identified *Sik3* and *Nalcn* with large effect over reciprocally sleep need and REM sleep amount.

1.3.3 Sleep and metabolism interplay:

Extensive research was performed on brain circuitry and brain molecular machinery during sleep and wakefulness using mainly EEG signals (Brown, Basheer et al. 2012). However alternative mechanisms were also highlighted, to be part of, or a co-factor of the sleep regulatory system (Frank 2012, Frank 2013). Other physiological systems like temperature (Franken, Dijk et al. 1991), glucocorticoid (Gronfier, Luthringer et al. 1997, Mongrain, Hernandez et al. 2010) or immune systems and cancer disease (Berger and Redline 2014) were connected to sleep characteristics.

Obesity and type-2 diabetes on their side, were associated with sleep disorders and prolonged bad sleep behavior (Spiegel, Knutson et al. 2005, Spiegel, Tasali et al. 2009) leading to the notion of possible interconnected physiological processes, between sleep regulation and metabolism, with shared circuits (Adamantidis and de Lecea 2008). This is also supported by *forward* genetics approaches that identified common loci between sleep disturbance and metabolic trait by GWAS (Lane, Liang et al. 2017).

On one side, sleep restriction shows that the human metabolome was sensitive to the increase of sleep pressure (Davies, Ang et al. 2014), which would hint that the organism's metabolism is driven by sleep-wake cycle. On the other hand, other examples suggested another direction, the sleep regulation driven by metabolic disruption. In BXD mouse, a short acyl-CoA dehydrogenase enzyme, *Acads,* was selected as a candidate gene within a QTL for REM sleep theta peak frequency (Tafti, Petit et al. 2003), important for memory consolidation (Boyce, Glasgow et al. 2016). In drosophila, lipid metabolism was shown to modify sleep homeostasis and learning using a *reverse* genetics approach involving the *bmm* and *Lsd2* genes, homologues

of lipid regulatory genes in human (Thimgan, Suzuki et al. 2010). Muscle metabolism was also suggested to directly be connected to sleep regulation, by *Bmal1* rescuing expression specifically in the muscle, restoring mouse rhythmicity, while *Bmal1* rescuing in the brain did not rescue mouse rhythmicity (Ehlen, Brager et al. 2017). The bidirectional relationship found between sleep-wake cycle and metabolism homeostasis was largely reviewed through the interplay other factors such as clock genes (Bass and Takahashi 2010), stress (Hirotsu, Tufik et al. 2015), brain circuitry (Huang, Ramsey et al. 2011).

With the systems genetics approach, we aimed at deciphering the close relationship that exists between sleep-wake cycle and metabolism. The quantification of liver transcriptomics and plasma blood targeted metabolomics in addition to EEG/behavioral phenotyping and cortex transcriptomics permit a broad and precise capture of the mouse metabolic environment to associate with sleep aspects.

1.4: Data life cycle management:

With the development of computational power, communication protocols and high-throughput technologies, biology has now acceded to the area of big-data. It is expected that around 2025, genomics data only will generate near 2-40 Exabytes per year (10^{18} bytes) representing 1 zetta-basepairs (10^{21}) sequenced per year (Stephens, Lee et al. 2015). This data explosion requires many innovations and novel bioinformatics paradigms for better information manipulation and extraction, including computational infrastructures with cloud-based data distribution or in an algorithmic context with ingenious, reproducible analysis pipelines and integrative methods (Muir, Li et al. 2016, He, Ge et al. 2017). The systems genetics designs as presented in previous sections generate typical features of big-data dataset, with large and heterogeneous amount of data characterized by an important variability as dint of high-throughput omics use and EEG/EMG signals recording but also by the fundamental complexity of this regulatory system. While the predictive capacity of big-data approach is questioned, in particular using blind data collection without accurate mathematical model of dynamic systems (Coveney, Dougherty et al. 2016), the power of these datasets lies through exploration, discovery and mechanistic hypothesis building, which is ameliorated by a data-driven and (semi-)hypothesis free research (McCue and McCoy 2017) in combination with data integration (Zierer, Menni et al. 2015).

1.4.1 Data integration:

Data integration is the combination of multiple sources from a same system to improve knowledge. It is now seen as one of the most challenging approaches of big-data that can be viewed from two different but ultimately linked perspectives, with distinct problems to solve for the functions of "*data-exploitation*" and "*data-source-discovery*" (Gomez-Cabrero, Abugessaisa et al. 2014). The objective function of *data-exploitation* is the detection of relevant interactions between the integrated resources to gain new insight in a biological system. The methodological challenges surrounding this function of data integration are the development and implementation of novel mathematical, statistical, computational (Gligorijević and Pržulj 2015) or visualization (Nekrutenko and Taylor 2012) strategies. The function of *data-source-discovery* is more community related, with the improvement of the communication and of the

knowledge sharing systems surrounding a resource, meaning the possibility for the public and collaborative projects to find and maximize the use an already existing relevant resource and thereby increasing the value of this resources for the benefit of the scientific community.

The many steps required for a correct data manipulation, integration and sharing urged the need for initiatives and guidelines promoting better data-handling. For example, complete meta-data documentation about the organisms and environmental conditions at the origins of the resources in combination with standardization of methods and common file formats diminish errors, allowing correction for stochastic noise and unwanted heterogeneity that can be misleading (Ward, Schmieder et al. 2013, Alyass, Turcotte et al. 2015, Lapatas, Stefanidakis et al. 2015). These procedures for a planned and active supervision of the digital data, from their collection to their processing, analysis, preservation, publication, reproducibility and reusability are gathered under the term of Data Life-Cycle Management (DLCM).

Initiatives like the FAIR principles developed by FORCE11 (Wilkinson, Dumontier et al. 2016) focused on human or computer knowledge sharing by proposing data objects to be Findable, Accessible, Interoperable and Reusable through many guidelines, notably by extensive metadata annotation. Other initiatives have integrated the FAIR data principles as a core component of their development strategy, like the *Big Data To Knowledge* (Margolis, Derr et al. 2014, Jagodnik, Koplev et al. 2017) framework of the National Institute of Health (NIH) or the *ELIXIR* program in Europe (Jiménez, Kuzak et al. 2017) for the maximization of biomedical dataset use through software and infrastructure development. Conjointly, open and reproducible science have also shown many possibilities for improvement through entire data life cycle (Munafò, Nosek et al. 2017). Only a few biomedical journals currently have an explicit data-sharing requirement (Vasilevsky, Minnier et al. 2017) or reproducibility policy (Iqbal, Wallach et al. 2016). However, the current demand for reproducible research (eLife Jul 14, 2017) articles have encouraged journals to support actively open technologies for publication (eLife Sep 7, 2017). The implementation of these initiatives can be quite time consuming in the working process, simply moving to a reproducible workflow can take up to many years (Lowndes, Best et al. 2017), but data-sharing and open analytical pipelines are essential for the reusability and reproducibility of the datasets.

PhD objectives & goals

The research initiated by the group of Prof. Paul Franken on genetics of sleep and the tight relationship highlighted between sleep processes and peripheral factors such as corticosterone or cellular metabolism serve as foundation for <u>decrypting the regulatory pathways linking metabolism and sleep homeostasis</u> using the systems genetics approach with in the BXD panel. The amount of data to treat, analyze and manage from the plasma metabolomic, cortical + liver transcriptome and EEG/EMG signal recording necessitated the collaboration with the group of Prof. Ioannis Xenarios for the many bioinformatics elements that needed to be incorporated and developed in this project. My work was at the interface of this collaborative association between sleep research, systems biology and bioinformatics with a main component toward computational and statistical methods. My objectives for this interdisciplinary project were:

- <u>The apprehension, implementation and enhancement of a supervised deep machine learning algorithm</u> designed by Dr. Nicolas Guex for the semi automation of EEG/EMG signal annotation for over 262 BXD mice and 22.5 million epochs.

- <u>The processing and analysis of multiomics datasets</u> for gene quantification and construction of genotype maps.

- <u>Identify and implement the most adapted statistical methods</u> for genetics association and appraise the sleep deprivation consequences on all data-types.

- <u>Develop a framework for data integration and visualization</u> in the interest of the identification of key driver genes and pathways for hypothesis building and experimental validation.

- <u>The generation of a reproducible & reusable dataset</u> with the implementation of an explicit interactive report and the development/supervision of a web resource interface for data-mining.

- <u>Participation to the bioinformatics and sleep research life</u> through group meetings, corresponding topic-oriented conferences and sleep deprivation experiments.

- <u>Annex projects</u>: I participated in annex projects that helped for the generation and follow-up of this dataset: (i) Evaluation of variant callers performance in the context of DREAM challenges, (ii) help with QTL mapping for diabetic related traits in BXD and (iii) set-up an ATAC-seq peak calling pipeline.

Chapter 2 presents this resource with the experimental design performed for *omics and physiological data extraction, followed by some analysis and first results interpretation. The principal focus of the resulting paper was the sleep deprivation consequences coupled with genetics interaction (GxE) and candidate genes that link sleep regulation and fatty acid metabolism.

Chapter 3 presents in more detail the bioinformatics strategies and data-management methodologies implemented during the project to maximize the resource reusability, reassessment and value.

Finally, a short annex presents a first view on the epigenetics modification that arose from sleep deprivation in C57BL/6J using ATAC-sequencing and the consequences on the sleep-wake driven or circadian genes. This project is still under preparation for publication and will direct follow-up experiments for epigenetics modification within the BXD panel, to complement our resource for better mechanistic characterization of the sleep loss and genetics interactive effect on transcriptome.

Chapter 2. Explore the sleep regulatory pathways through the prism of systems genetics.

2.1 Results summary

In this paper, we present the strong genetic contribution and the pervasive effect of short sleep loss, in addition to their interaction for each the BXD mice intermediate phenotype and their implication on physiological and behavioral end-phenotype. The estimated narrow-sense heritability for physiological, behavioral and metabolic phenotypes was overall high with a median value above 0.5. An association mapping revealed 61 significant QTL for behavioral/EEG and metabolic phenotypes, as well as more than 5000 genes driven by a cis-eQTL. The short 6h disruption of sleep extensively reshaped the BXD phenotypic landscape by altering 60-78% of the liver and cortical transcriptomes, 60% of the plasma blood metabolome, and many physiological phenotypes such as NREM sleep amount or EEG delta power. We showed that genetic affects both the magnitude and direction of change induced by sleep deprivation (G x E) on genes expression, for example on enzyme promoting mitochondrial fatty-acid oxidation (*Mlycd*), but also on metabolite level such as the acylcarnitine C18:1. Two integrative methods were developed in this project, the first to identify candidate gene driving the behavioral/EEG and metabolic phenotypes, the second to visualize the meta-dimensional network underlying behavioral/EEG phenotype. The combination of both methods first identified the implication of cortical AMPA-receptor trafficking and DNA-helicase amplitude change for the physiological phenotypes of the EEG delta-power bands during the recovery phase, a correlate of sleep pressure. In a second time, we identified an intriguing relationship between branched-amino acid and arachidonic-acid (ARA) metabolism for the shift of theta-peak frequency in REM sleep, an important aspect of memory consolidation. It leaved open question about the potential relationship between synthesis and derivates of ARA, like prostaglandins and endocannabinoids on the synaptic plasticity. Finally, we identified a large fatty acid turn-over after sleep deprivation, with effect on the NREM sleep recovery during the active phase of mice (dark) where most of NREM sleep loss was catched up. The implication of the associated acyl-CoA thioesterase enzyme (*Acot11*) was tested using knock-out and showed a deficiency in NREM sleep gain compared to their wild-type controls in the dark.

2.2 Contribution

Beside my contributions mentioned in *PhD objectives & goals* (i.e. participation to EEG/EMG annotation automation, bioinformatics analyses and methods development), I participated to the results interpretation and paper writing with the help of all co-authors, I generated and formatted all figures (with exception of Fig5A & 6A) with comments/inputs of Prof.Paul Franken, Dr. Shanaz Diessler and Dr. Charlotte Hor and I participated to sleep deprivation experiment for *Acot11* validation

2.3 Publication:

Shanaz Diessler, **Maxime Jan**, Yann Emmenegger, Nicolas Guex, Benita Middleton, Debra J. Skene, Mark Ibberson, Frederic Burdet, Lou Götz, Marco Pagni, Martial Sankar, Robin Liechti, Charlotte N. Hor, Ioannis Xenarios, Paul Franken. *An integrative systems genetics analysis of sleep regulation in the mouse.* **In Review**

A systems genetics analysis of
sleep regulation in the mouse

Shanaz Diessler[1*], Maxime Jan[1, 2*], Yann Emmenegger[1], Nicolas Guex[2], Benita Middleton[3],
Debra J. Skene[3], Mark Ibberson[2], Frederic Burdet[2], Lou Gotz[2], Marco Pagni[2], Martial Sankar[2],
Robin Liechti[2], Charlotte Hor[1], Ioannis Xenarios[2†], Paul Franken[1,5†]

[1]Center for Integrative Genomics; University of Lausanne, Switzerland; [2] Vital-IT Systems
Biology Division; Swiss Institute of Bioinformatics; Lausanne; Switzerland; [3]Faculty of Health
and Medical Sciences, University of Surrey, Guildford, United Kingdom

*share first authorship

†share last authorship

[5]Lead Contact

Running title:

Systems genetics of sleep loss

SUMMARY

Sleep is essential for optimal brain functioning and health, but the biological substrates through which sleep delivers these beneficial effects remain largely unknown. We used a systems genetics approach in the BXD genetic reference population of mice and assembled a comprehensive experimental knowledge base comprising a deep 'sleep-wake' phenome, central and peripheral transcriptomes, and plasma metabolome data, collected under undisturbed baseline conditions and after sleep deprivation. We present analytical tools to interactively interrogate the database, visualize the molecular networks altered by sleep loss, and identify driver genes. We found that a one-time, short disruption of sleep already extensively reshaped the systems genetics landscape by altering 60-78% of the transcriptomes and the metabolome with numerous genetic loci affecting the magnitude and direction of change. Systems genetics integrative analyses drawing on all levels of organization imply AMPA-receptor trafficking and fatty acid turn-over as substrates of the negative effects of insufficient sleep. Our analyses demonstrate that genetic heterogeneity and the effects of insufficient sleep itself on the transcriptome and metabolome are far more widespread than previously reported.

INTRODUCTION

Insufficient or disrupted sleep are widespread in our 24/7 society and represent a serious public health concern, as it is associated with increased risk for e.g. obesity, diabetes and high blood pressure, and impairs cognitive performance, which in turn increases the likelihood of accidents, medical errors, and loss of productivity (Liu et al, 2016; Schmid et al, 2015). Several hypotheses concerning sleep's still elusive function converge on the notion that staying awake imposes a burden that can only be efficiently alleviated during sleep (Benington & Heller, 1995; Krueger et al, 2008; Maquet, 1995; Tononi & Cirelli, 2014; Xie et al, 2013). This concept of a need for sleep accumulating during wakefulness and recovering while asleep is central in sleep research and is referred to as sleep homeostasis. Insight into the molecular substrates of the sleep homeostatic process is instrumental in advancing our basic understanding of sleep need under both physiological and pathological conditions.

The impact of acute sleep deprivation on recovery sleep and cognitive performance is under strong genetic control (Dissel et al, 2015; Franken et al, 2001; Kuna et al, 2012; Lo et al, 2012; Mang & Franken, 2015; Urry & Landolt, 2015), and genetic approaches therefore seem promising in uncovering the molecular pathways important in sleep homeostasis. Reductionist studies in mice and flies deleting genes through gene targeting [for review see (Mang & Franken, 2015)] or in mutagenesis screens (Cirelli et al, 2005a; Funato et al, 2016; Koh et al, 2008) have demonstrated that single genes can have large effects on various aspects of sleep, including its homeostatic regulation. Such large single gene (Mendelian) effects, often assessed on one genetic background only, are however likely to be the exception. Indeed, susceptibility to sleep loss in the general population is assumed to be determined by the interactions of many genes, their natural allelic variants, and their interaction with the environment (lifestyle), a complexity that only recently has begun to be appreciated. Such complexity can best be assessed in so-called Genetic Reference Populations (GRPs), which are designed for the study of complex traits inherited in a non-Mendelian fashion. The BXD panel of advanced recombinant inbred lines (ARIL) is the largest and best characterized GRP to date, consisting of well over 100 lines in which two parental [C57BL/6J (B6) and DBA/2J (D2)] now fully sequenced genomes are segregating [www.genenetwork.org; (Peirce et al, 2004)]. As each line represents a reproducible clone of animals, many mutually reinforcing datasets can be collected and compared at multiple levels across many biological systems. This approach has been termed 'systems genetics', which in essence allows for making inferences about biological phenomena

by assessing the flow of information from DNA-to-phenotype at the level of a population, and how this flow is perturbed by environmental challenges. Because systems genetics generalizes results to a population level, it is considered critical towards predicting disease susceptibility (Civelek & Lusis, 2014). Systems genetics has been applied with great success in the BXD set for e.g. mitochondrial function and metabolic- and aging-related phenotypes (Andreux et al, 2012; Merkwirth et al, 2016; Williams et al, 2016).

Systems genetics approaches for sleep have been pioneered in the fly and mouse (Harbison et al, 2009; Jiang et al, 2015), but neither study reported on effects of sleep loss and its impact on intermediate phenotypes such as the metabolome and transcriptome. Here we present an extensive and comprehensive dataset interrogating the BXD set at the levels of the genome, the brain and liver transcriptomes, the plasma metabolome, and finally, the phenome including sleep-wake state, EEG-, and locomotor activity-related phenotypes, both under undisturbed baseline conditions and after an acute sleep deprivation challenging the sleep homeostatic process. We observed that sleep deprivation profoundly impacted all three phenotypic levels, and that genetic background not only determined the magnitude, but also the direction of the sleep-deprivation evoked changes. The molecular pathways underlying some of these effects will be illustrated here to introduce our integrated data resource. The molecular signaling circuitry underlying the equally profound phenotypic differences observed under baseline conditions will be reported in subsequent molecular driven validations.

Systems genetics is an emerging field, and innovative ways to improve data access, portability, and reproducibility, tools to display and mine these data, as well as statistical models to extract the multi-dimensional relationships across datasets are an area of intense research (Baliga et al, 2017). The size and complexity of our current dataset necessitated the development of new analytical tools and data sharing strategies such as (i) a supervised machine learning based algorithm to annotate sleep-wake states on EEG/EMG tracks, (ii) a gene prioritization strategy that draws on all levels of the experimental dataset to assist the search for candidate genes within QTL intervals, (iii) the implementation and integration of a recently developed systems genetics visualization tool (Krzywinski et al, 2012) in a dynamic web-based interface which, in addition, provides access to the data presented and enables interactive data-mining (https://bxd.vital-it.ch).

RESULTS

Study design

We subjected mice from 33 BXD/RwwJ lines (see https://bxd.vital-it.ch; Downloads, General_Information.xlsx for a listing), the two parental strains (B6 and D2), as well as F1 individuals from reciprocal crosses between the parental lines, to a deep behavioral and molecular phenotyping across four levels of organization. The BXD lines were randomly chosen from the then available, newly generated ARIL panel (Peirce et al, 2004), although lines with documented poor breeding performance were not considered. In one set of mice, we recorded sleep-wake behavior, brain activity (by electroencephalography or EEG), and locomotor activity (LMA) during 4 days (Figure 1, Experiment 1). On day 3, mice underwent a sleep deprivation challenge (SD) during the first 6 hours of the light period when mice normally sleep most of the time. We refer collectively to EEG and LMA derived measures as 'EEG/behavioral' phenotypes. A second set of mice, representing the same lines, was processed in parallel for collection of brain, liver, and plasma (Figure 1, Experiment 2). Half these animals underwent a sleep deprivation alongside animals of Experiment 1, while the other half were left undisturbed and used as controls (Ctr). We measured gene expression in cortex and liver, and metabolites in plasma. These transcriptomic and metabolomic data are collectively referred to as (intermediate) molecular phenotypes. The phenotypic variability in EEG/behavioral and molecular phenotypes was assessed in relation to the genetic variation present in the BXD panel using Quantitative Trait Locus (QTL) analysis. The resulting QTLs were divided into ph-, m-, and eQTLs, for phenotype (EEG/behavioral), metabolite, and expression QTLs, respectively. The entirety of the multi-level dataset was integrated in a systems genetics analysis to chart molecular pathways underlying the many facets of sleep and the EEG, using newly developed computational tools to (i) interactively display results and pathways, and (ii) to identify driver genes.

EEG/behavioral phenotyping

From the EEG and LMA recordings we could extract a comprehensive set of 341 EEG/behavioral phenotypes (see https://bxd.vital-it.ch; Downloads, General_Information.xlsx) in 227 mice (~6 mice/line, see Methods). These phenotypes were separated into 3 main biological categories related to: (i) LMA, (ii) EEG signal features, and (iii) the prevalence and

time structure of sleep-wake state [i.e., wakefulness and its substate theta-dominated wakefulness (TDW), NREM sleep, and REM sleep], collectively referred to as 'LMA', 'EEG', and 'State', respectively. Sleep-wake states were semi-automatically derived from the EEG/EMG signals using a specifically developed supervised machine learning based algorithm (see Methods and Appendix Figure S1). The 3 phenotypic categories were divided further into sub-categories (Figure 2A) and by experimental condition (baseline, sleep deprivation, and recovery). Because some phenotypes were tightly linked (e.g., the time spent in NREM sleep and wakefulness), we estimated the total number of distinct phenotypic classes to be 148 (Appendix Figure S2 and Methods). However, we still used all available phenotypes in our analyses to detect potential regulatory differences among even closely related phenotypes and to avoid analysis bias arising from selecting a 'representative' phenotype.

Molecular phenotyping

We quantified 124 metabolites (see https://bxd.vital-it.ch; Downloads, General_Information.xlsx) in 249 blood plasma samples (~3 samples/line/condition, see Methods) using targeted metabolomics. Although targeted metabolomics does not provide a comprehensive assessment of the entire metabolome, it does cover important metabolite classes (i.e., amino acids, biogenic amines, acylcarnitines, sphingolipids, and glycerophospholipids) and the results obtained using this method are highly reproducible and readily interpretable (Siskos et al, 2017).

Cortex and liver transcript levels were measured using RNA-sequencing (RNA-seq). Gene boundaries and names used were extracted from the Refseq database to ensure reusability of the data. We detected 14.9K expressed genes in the cortex and 14.1K genes in the liver after filtering and normalization. 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 (Appendix Figure S3A), and we observed a somewhat higher coefficient of variation in the raw gene read count in liver than in cortex (Appendix Figure S3B).

Genotyping

We used the RNA-seq alignments also to genotype the lines to verify that no mix-up occurred during the breeding and data collection phase, and to increase mapping resolution. The GATK variant-calling pipeline (McKenna et al, 2010) was used for genotyping, and we compared the 500K detected genotypes with the publicly available 3.5K genotype set for the same BXD lines from GeneNetwork (2005 release; see Methods). We observed only a ~1% discrepancy and merged both genotype sets, resulting in a set of 11K tag variations which increased the number of haplotype blocks from 551 (GeneNetwork) to 1071 (RNA-seq + GeneNetwork). All analyses we report here were based on our merged map (see https://bxd.vital-it.ch; Downloads, Genotypes.GeneNetwork2005AndRNAseq.geno). Of note, by the completion of this publication, an updated set of BXD genotypes was released with an estimated haplotype block number of 816 for the specific lines we used (GeneNetwork, 2017 release http://genenetwork.org). Of the 61 significant phQTLs we detected (see below), 54 were also detected using either GeneNetwork genotypes (the 2005 or 2017 release), while the remaining 7 significant phQTLs were unique to our merged genotype map.

Although overall, a close to 50/50 balance between B6 and D2 genotypes was observed across the genome, a minority of sites displayed a strong imbalance towards either genotype (Appendix Figure S4a). We also confirmed a minor but general trend towards more D2 than B6 genotypes per strain (Appendix Figure S4Bb), which was also found in the GeneNetwork genotypes for the BXD strains used in our study.

Heritability and QTLs

To obtain a first sense of the contribution of genetic factors to the phenotypic variability contained within our BXD set, we examined the heritability of the EEG/behavioral and metabolic phenotypes. The estimated narrow sense heritability (Hegmann & Possidente, 1981) among the EEG/behavioral phenotypes was high overall (median h2=0.68, Figure 2A), consistent with what has been reported in previous human and mouse studies (Andretic et al, 2008). We also confirm that various aspects of the EEG signal are among the most heritable traits with, in our dataset, theta-peak frequency (TPF) in REM sleep ranking highest (h2=0.89). The heritability for differential EEG/behavioral phenotypes (i.e., recovery vs. baseline; green symbols in Figure 2A) were consistently lower by ca. 0.2 points compared to the heritabilities

39

obtained for recovery or the baseline values per se. We found that this effect did not simply reflect increased variability due to combining recovery and baseline values, and might thus suggest a smaller genetic contribution to the response to sleep loss. The overall heritability of plasma metabolite levels was somewhat lower than for EEG/behavioral phenotypes (median h2=0.50) with alpha-aminoadipic acid (α-AAA) displaying the highest heritability (h2=0.88; Figure 2A).

Average-to-high heritabilities are a requirement to attribute phenotypic variation to gene loci, but even then there is no guarantee to find genome-wide significant QTL(s); e.g., for the TPF in REM sleep phenotype mentioned above, only 4 suggestive phQTLs of small effect size were identified (see https://bxd.vital-it.ch; Downloads, QTL_Mapping.xlsx) that together could nevertheless account for 58% of the variance, suggesting that perhaps higher order loci interactions (e.g. epistasis), which cannot be captured using the single marker linkage analysis we used here, underlie differences in this EEG trait. Genome scans revealed a total of 61 'significant' (FDR≤0.05), 65 'highly suggestive' (0.05<FDR≤0.10), and 923 'suggestive' (0.10<FDR≤0.63) (Burgess-Herbert et al, 2008; Lander & Kruglyak, 1995) phQTLs and 21 significant, 40 highly suggestive, and 528 suggestive mQTLs (Figure 2B). Given the large number of EEG/behavioral and metabolic traits of average-to-high heritability, the number of significant QTLs might seem a modest yield. We assume, however, that most of the traits are complex/polygenic and thus influenced by many smaller effect-size QTLs. The high number of suggestive QTLs might reflect this. Because for each genome scan one suggestive QTL is expected to be found by chance, the 1556 suggestive ph- and mQTLs obtained are 2.2 times higher than the number expected by chance alone, given the 713 EEG/behavioral and metabolic phenotypes analyzed. These suggestive QTLs are therefore expected to be biologically meaningful, each having a small contribution to the variance of polygenic traits. The large portion of the variance explained by the 4 suggestive phQTLs associated with REM sleep TPF (see above) attests to this.

Several phenotypes from distinct phenotypic categories were associated with overlapping genomic regions. For example, differences in baseline wake consolidation, gain in REM sleep time after sleep deprivation, EEG delta power (1.0-4.0 Hz) in REM sleep, baseline levels of serotonin and PC-aa-C34:4, and levels of PC-aa-C34:4 and PC-aa-C36:6 after sleep deprivation, all mapped to one 30Mb region on chromosome 10 (50-80 Mb), each with a

significant or highly-suggestive QTL (Figure 2B). These overlapping QTLs may point to pleiotropic effects of one underlying gene or close but distinct underlying QTLs.

We also performed QTL analysis for gene expression, but because many more linkage tests were required for transcriptome mapping, we used a more suitable method than for ph- and mQTL mapping. The format for reporting eQTLs will therefore differ from that used for ph- and mQTLs (see Methods). The expression of individual genes was mapped separately for cis-eQTLs with genetic markers within a 2 Mb window, and trans-eQTLs with markers positioned throughout the genome (see Methods). The transcriptome of BXD mice showed strong linkage with genotypic variation. For example, in the cortex, the expression of 5704 genes (i.e., 38% of all expressed genes) was significantly driven by a cis-variation (Figure 2C and https://bxd.vital-it.ch; Downloads, cis_eQTL.xlsx). Moreover, 2465 (34%) of all genes under cis-eQTL effect in both tissues passed the 0.05 FDR cutoff in a single condition and tissue. Factors contributing to this tissue/condition specificity are the absence of gene expression in one of the two tissues, or a different gene regulatory environment on which sleep deprivation had pervasive effects (see below). This important tissue/condition specificity also applied to trans-eQTLs with 5537 (53% of 10450) being under trans-eQTL effect only in one specific tissue or condition. Although the observation that a large portion of eQTLs reached significance in one tissue and condition only does suggest wide-spread gene x environment interactions regulating gene expression, reaching the 0.05 FDR threshold or not, does not prove this. We therefore compared linkage strength of significant cis-eQTLs that were specific for one tissue and condition with that in the 3 other RNAseq sets. Among the 870 genes with a significant cis-eQTL effect in sleep deprived cortex only (Figure 2C), 175 (20%) showed a significant difference in linkage signal (FDR<0.05). This proportion was similar in the control cortex and liver (19 and 21%, respectively) and somewhat higher in sleep deprived livers (32%).

Systems genetics visualization

The complexity of multi-level networks can only be appreciated through visual aids. Because the widely used 'hairball' representation, in which biological factors are represented as 'nodes' and their interconnections as 'edges', is hardly interpretable due to its non-deterministic structure (Figure 3A), we opted for a structured representation more suitable for the visualization of complex systems, namely 'hiveplots' (Krzywinski et al, 2012). The hiveplots

were laid out as follows (Figure 3B, see Methods for details): each plot represents one EEG/behavioral phenotype and its underlying associated molecular network, i.e. only the genes and metabolites strongly associated with a given phenotype are displayed. Each hiveplot is composed of 3 radial axes containing the molecular data with nodes assigned to the 2 bottom axes for genes expressed in the cortex (Figure 3B Left in blue) and liver (Figure 3B Right in red), while nodes on the vertical axis (Figure 3B Top in yellow) represent metabolites. On top, we added a separate 'genetic' axis (Figure 3B Top white) containing the genotypes. The node position on the 3 (molecular) radial axes was determined by the response to sleep deprivation, i.e. molecules positioned closer to the center were down-regulated more strongly, while more up-regulated genes/metabolites can be found closer to the axes' perimeter. Edges connecting nodes represent positive/negative correlations (red/blue, respectively) between measurements of expression/metabolite levels. Genetic markers linked to genes by eQTLs connect the genetic and molecular space.

The hiveplot representation allows investigation of the underlying network of an EEG/behavioral phenotype in a structured manner, and comparison of phenotypes using all intermediate phenotypic layers available in the dataset. The difference in presence or absence of nodes/edges between 2 phenotypes indicates which association was gained or lost. Furthermore, the importance of the sleep deprivation effect on these nodes can be visually estimated by their position along the axis (Figure 3C). Although the inter-phenotype connectivity present in the hairball representation is lost in the printed format of these hiveplots, this aspect can be easily accessed through our web interface (https://bxd.vital-it.ch) by highlighting common edges. The web interface also allows for an in-depth exploration of the data by displaying node details such as gene and metabolite name, and variation identifiers. It also lets the user modify the parameter settings used to generate the hiveplots (see Appendix Figure S5 and the tutorial on https://bxd.vital-it.ch; Help).

Systems genetics prioritization

We developed an unbiased, data-driven approach to identify candidate genes underlying our EEG/behavioral and metabolic phenotypes. We focused mainly on genes located in the associated genomic regions found by QTL analyses (see Figure 2B). To investigate these often quite large regions (mean=9.8 Mb, range=0.7-34.7 Mb for significant and highly suggestive

phQTLs), we implemented a scoring strategy inspired by the 'similarity profiling prioritization strategy' (Moreau & Tranchevent, 2012), which combines multiple sources to prioritize a gene. For each gene, we computed an integrated score comprised of (i) the genomic position of the gene with respect to the ph-/mQTL peak, (ii) a detected cis-eQTL driving the expression of the gene, (iii) a protein-damaging annotation of a variant, (iv) differential expression after sleep deprivation, and (v) correlation between expression and phenotype of interest (Figure 3D, see Methods for details). Our prioritization strategy thus aimed at identifying genes which are sensitive to sleep loss, correlated with the phenotype being evaluated, and associated to a cis-eQTL, and/or carrying a protein-damaging variant that could contribute to trait variance. A Henikoff weighting algorithm was applied to correct for intrinsic correlations among the 5 analysis scores. The integrated score for each gene was computed with the given formula (Figure 3D) and a FDR was computed by performing 10K permutations (Appendix Figure S6 and Methods). For each QTL, we kept the gene with the highest significant integrated score. This scoring strategy was applied to cortex and liver data separately.

To illustrate our prioritization algorithm, we applied it to the metabolite with the highest heritability, α-AAA (see above), and for which we obtained a highly significant mQTL on chromosome 2 [LOD=9.25, 1-11Mb]. We readily identified Dhtkd1 as the top-ranked significant candidate gene in liver within the chromosome 2 mQTL (Figure 3E), because of (i) the strong correlation of Dhtkd1 expression with α-AAA levels, (ii) Dhtkd1 is under a cis-eQTL effect (rs222492362, chr2: 5.8 Mb, q=1.5e-17), (iii) the marker of the cis-eQTL is located within the peak of the mQTL, and (iv) both α-AAA and Dhtkd1 levels are affected similarly by sleep deprivation. This result can be taken as a first validation of our scoring strategy, because Dhtkd1 encodes an enzyme subunit involved in lysine degradation known to control α-AAA levels in BXD lines (Wu et al, 2014). Although with the examples that will be highlighted below we primarily focus on regions associated with significant QTLs, our prioritization strategy also allows identification of candidate genes within suggestive QTL intervals as shown for the 'δ1 power gain' phenotype (see below and Figure 5).

Pervasive effects of sleep deprivation at all levels

The EEG/behavioral and molecular phenotypes were assessed both under undisturbed baseline conditions and after 6h sleep deprivation. Sleep deprivation profoundly and

significantly impacted a majority of measurements at all levels. We observed the well-known increase in EEG delta power (1.0-4.0 Hz) during NREM sleep as well as the increase in the time spent asleep (Figure 4A), both reflective of an accumulated homeostatic sleep pressure during sleep deprivation. The gain in time spent in NREM sleep was strongest during the initial 12h following the sleep deprivation, with an average gain of +23 minutes (compared to values reached during corresponding baseline hours) during the first 6h after the SD [ZT6-12] and +32 minutes during the first 6h of the following dark period [ZT12-18]. The most strongly affected sleep phenotype concerned time spent in REM sleep, which displayed a 3.3-fold gain during the first 6 hours of darkness [ZT12-18] after sleep deprivation (Figure 4A). Sleep deprivation thereby doubled the proportion of REM sleep to NREM sleep in this interval. Locomotor activity and waking phenotypes were generally decreased during the light period immediately following the sleep deprivation (ZT6-12).

In addition, the plasma metabolome was profoundly altered by sleep deprivation. Of the 124 measured metabolites, 75 (60%) were significantly up- or down-regulated. The levels of all amino acids were significantly altered after sleep deprivation, the majority being down-regulated with the exception of glutamine, glutamate, and tryptophan, which were up-regulated (Figure 4B). A recent publication reported similar effects on amino acid levels in brain dialysates of sleep deprived rats (Marini et al, 2017) suggesting that plasma can report on central changes in amino acid levels. By contrast, tryptophan was the only amino acid that was found significantly changed during sleep deprivation in humans using the same methodology (Davies et al, 2014). The two acylcarnitines present in our dataset (C18:1 and C18:2) were both strongly up-regulated with a greater than two-fold change. Similar results were found in humans, with acylcarnitines levels increased in blood and carnitines increased in urine after sleep loss (Davies et al, 2014; Giskeødegård et al, 2015).

The transcriptome was especially sensitive to sleep deprivation, with 78% of all expressed genes being differentially expressed in cortex and 60% in liver. In cortex, the most strongly differentially expressed genes were Arc, Egr2, and Plin4, with an almost 8-fold increase in expression after sleep deprivation (see Appendix Table S1). Arc (Activity regulated cytoskeletal-associated protein) is an immediate early gene crucial for long-term synaptic plasticity and memory formation (Korb & Finkbeiner, 2011). Arc is among the most consistently up-regulated transcripts after sleep deprivation (Wang et al, 2010a) and features in a short-list of 78 genes the expression of which we found reliably and significantly changed by

extended wakefulness under a number of experimental conditions (Mongrain et al, 2010). Forty-nine other genes in this short-list also featured among the top 5% most affected transcripts of the current experiment (Appendix Table S1 and blue symbols in Figure 4C Left). Similarly, Egr2 (Early growth response 2) is one of three Egr genes that are rapidly induced by sleep deprivation in several species (Wang et al, 2010a). Egr1 and -3 appear on our short-list and all 3 Egrs are among the top 100 differentially expressed cortical genes in the current study (Appendix Table S1). The Egr family are immediate early genes encoding transcription factors important in neuronal plasticity (O'Donovan et al, 1999). Plin4 (Perilipin 4), which encodes a lipid-droplet associated protein involved in lipid storage (Itabe et al, 2017), has not been reported previously as part of the sleep deprivation response. Ttll8 (tubulin tyrosine ligase-like family, 8) encoding a ligase which glycylates microtubules (Rocha et al, 2014), and Fam107a (family with sequence similarity 107, A) a stress- and glucocorticoid-regulated gene (Masana et al, 2015; Schmidt et al, 2011), were the top differentially expressed genes in liver (Appendix Table S2). Although the short-list was based on forebrain samples, 17 genes were also present in the top 5% differentially expressed genes in the liver (blue dots Figure 4C Right). Moreover, 13 genes were common to the top 5% list in cortex, liver, and the 78 genes of the short-list (Hspa1a/b, Cirbp, Fos, P4ha1, Chordc1, Dusp1, Slc5a3, Hsph1, Creld2, Tra2a, Zbtb40, and Pfkfb3). These genes might be interesting candidates for tissue-independent biomarkers of sleep pressure.

Genetics of the effects of sleep deprivation

In the context of our project, a key question to be addressed is whether genetic background modifies these pervasive effects of sleep deprivation. We found evidence for this at all 3 levels of organization and detected genomic loci predicting differences not only in the magnitude of the response to sleep deprivation, but also in the direction of the response (illustrated in Figure 4D-F). In the analyses, we included both the levels reached after the sleep deprivation, and these levels contrasted to their baseline levels. These contrasts will be referred to as 'change', 'increase', 'gain', 'decrease', or 'differential expression' (DE).

For 7 EEG/behavioral 'gain' phenotypes we discovered a significant QTL (https://bxd.vital-it.ch; Downloads, QTL_Mapping.xlsx). Illustrated in Figure 4D is the gain in time spent in REM sleep which mapped significantly to chromosome 18 (LOD=3.9; 57-62 Mb) with B6-

allele carriers gaining more REM sleep than D2-carriers (genotype-SD interaction: p=2.0e-5). Three more 'gain' phenotypes will be discussed in detail below (see Figures 5-7). Also illustrated in Figure 4D is an EEG/behavioral gain phenotype with a pronounced genotype effect on the direction of change. The sleep-deprivation induced changes in EEG activity in the fast gamma band (55-80 Hz) in NREM sleep mapped suggestively to chromosome 6 (LOD=2.83; 77-89 Mb) with a majority of B6-allele carriers at the QTL peak position having a significant decrease in fast gamma while several D2-allele carriers showed a significant increase (genotype x SD interaction: p=1.0e-5). Examples of two genetically driven metabolic responses to sleep deprivation are illustrated in Figure 4E. The change in phosphatidylcholine acyl-alkyl 32:2 (PC-ae-C32:2) after sleep deprivation mapped significantly to chromosome 5 (LOD=3.6; 58-69 Mb; genotype x SD interaction: p=2.0e-3). The change in acylcarnitine C18:1, the strongest among all metabolites assayed (Figure 4B), mapped suggestively to chromosome 18 (LOD=3.6; 73-75 Mb; genotype x SD interaction: p=2.0e-3). For an additional 79 metabolites, a significant genotype x SD interaction was obtained that mapped at the suggestive level (see https://bxd.vital-it.ch; Downloads, Genotype_SD_Interaction.xlsx). Finally, significant cis-eQTLs were detected for the differential expression (DE, i.e. recovery vs. control) of 195 genes after sleep deprivation (see https://bxd.vital-it.ch; Downloads, Genotype_SD_Interaction.xlsx). The strongest cis-allele was found for the DE of cortical Pla2g4e (rs47077493, chr2: 118.3 Mb, q=1.2e-9) with a down-regulation that was 2-fold larger in B6- than in D2-allele carriers (genotype x SD interaction: p=1.0e-9; Figure 4F). Also illustrated are the effects of sleep deprivation on Mlycd expression for which a cis-eQTL was identified (rs33610973, chr8: 120.8 Mb, q=1.9e-5). In BXD lines carrying a B6-allele at the cis-eQTL position, a down-regulation of Mlycd was observed, while the opposite was true for D2-allele carriers (genotype x SD interaction: p=2.0e-4; Figure 4F). Pla2g4e (phospholipase A2, group IVE) encodes a phospholipase promoting the formation of free-fatty acids, while Mlycd (malonyl-CoA decarboxylase) encodes an enzyme promoting mitochondrial fatty-acid oxidation. One last example of a significant differential cis-eQTL, for Wrn, will be discussed in detail below (see Figure 5).

Systems genetics of the effects of sleep deprivation

In the following 4 sections, we highlight 4 phenotypes during recovery from sleep deprivation that emerged from our systems genetics analyses because of the presence of strong

genetic evidence at all levels of organization, allowing causal/meaningful relational networks to be built linking the flow of information from DNA to phenotype. Two concern the levels of EEG delta power reached after sleep deprivation, one concerns the gain in time spent in NREM sleep during recovery, and, as a last example, the changes in TPF during REM sleep in recovery. While for the first 3 phenotypes abundant evidence exists documenting their change with sleep deprivation and their relevance in optimal day-time functioning and health, the latter phenotype, which has not been reported on previously, illustrates that depending on genotype, a phenotype can either increase or decrease after sleep loss. Moreover, this example shows that phenotypes considered strictly 'central' (i.e., the frequency of hippocampal theta oscillations during REM sleep) are strongly associated with genomic loci affecting gene expression in the periphery and not in the brain. Finally, it is important to point out that the genomic loci identified for these 4 recovery phenotypes appear after sleep deprivation only and not (even at the suggestive level) under baseline conditions.

Example 1: **Genetic heterogeneity in the gain of slow and fast EEG delta power after sleep deprivation.**

The prevalence and amplitude of EEG oscillations in the delta frequency range (1.0-4.0 Hz) during NREM sleep can be quantified as EEG delta power. The sleep-wake dependent changes in EEG delta power have been widely used as a marker of the sleep homeostatic process and form the basis of leading hypotheses on sleep-wake regulation and function (Daan et al, 1984; Krueger et al, 2008; Tononi & Cirelli, 2014). In the older BXD panel (BXD1-42/TyJ), we previously mapped a significant QTL on chromosome 13 (Dsp1; MGI: 2135996) that could explain a large portion of the variance in the increase (or gain) in EEG delta power (compared to the lowest values reached in baseline between ZT8-12) during NREM sleep immediately following a sleep deprivation (Franken et al, 2001). Subsequent studies identified Homer1 as a credible candidate gene (Mackiewicz et al, 2008; Maret et al, 2007), and specifically its short isoform Homer1a might mechanistically link sleep homeostatic drive to one proposed function of NREM sleep, namely synaptic down-scaling (Diering et al, 2017). In the current BXD set (BXD43-161/RwwJ), we could not replicate the Dps1 QTL. Among the many possible explanations for the differences between the two studies are (i) the strong bias in allele

frequency toward the B6 genotype at the Dps1 region in the 33 BXD lines we used in the current study (28/33 lines, Figure S3A), a situation which decreases statistical power for QTL detection and is also thought to contribute to replication issues among QTL studies (Gatti et al, 2009), (ii) the two BXD sets were derived from two different DBA/2 sub-strains [i.e., DBA/2Rj vs. DBA/2J; new vs. old (Shin et al, 2014)], (iii) genetic drift (Reardon, 2017; Shifman et al, 2006), (iv) interaction with other loci that are differently represented in the two BXD sets, and (v) differences in phenotyping. In an attempt to address some of these issues, we re-phenotyped two of the older BXD lines, namely BXD5/TyJ and BXD32/TyJ, which ca. 17 years earlier gave the highest and lowest EEG delta power increase after sleep deprivation, respectively (Franken et al, 2001). While for BXD5, which carries a B6-allele at Dps1 as well as for the B6 parental strain, we obtained a close to perfect match (B6: 221.2 vs. 223.1%, BXD5: 223.3 vs. 223.6% over baseline, current vs. previous), for BXD32, a D2-allele carrier at the Dsp1 locus, and the D2 parental line we observed notable discrepancies (D2: 168.9 vs. 179.5%; BXD32: 159.2 vs. 133.0%, current vs. previous). Thus each BXD set must be regarded as a GRP in its own right and QTL analyses can lead to different sets of equally valid genetic associations due to genetic drift and other above-mentioned reasons.

In the current dataset, we did not find loci with strong linkage for the gain in EEG delta power after sleep deprivation when analyzed for the entire delta frequency range (1.0-4.0 Hz). We previously noticed that this increase in EEG power after sleep deprivation is, however, not homogeneous across the delta band (Franken et al, 2006; Vassalli & Franken, 2017) and therefore divided it into a slow ($\delta1$: 1.0-2.25 Hz) and a fast ($\delta2$: 2.5-4.25 Hz) delta band. As a neurophysiological correlate of increased EEG activity in $\delta1$ after sleep deprivation, increased noradrenergic tone in the cortex has been proposed (Cirelli et al, 2005b). In addition, the acceleration of the clock-like delta oscillations generated by thalamocortical neurons at increasing levels of hyperpolarization that accompany deep NREM sleep could contribute to increases in $\delta2$ activity (Amzica & Steriade, 1998; Dossi et al, 1992). In the current dataset, we confirmed that EEG activity in the two bands respond differently to sleep deprivation. E.g., lines that showed the lowest/highest gain in $\delta1$ power (i.e., BXD81 and -67, respectively; Figure 5A Bottom) only ranked 12th and 23rd (out of 33) for the gain in $\delta2$ power. Moreover, we found different genetic loci contributing to their respective increases over baseline. The separation of delta in a $\delta1$ and a $\delta2$ band increased the number of QTLs detected: while one suggestive QTL (LOD = 2.58; chr1: 165-176 Mb) was found using the full, 1.0-4.0 Hz band, we detected one

suggestive QTL on chromosome 8 (LOD = 2.86; 18-37 Mb) for the gain in δ1 power explaining 33% of the phenotypic variance across the BXD lines, and 5 suggestive QTLs for δ2 power gain, none of which overlapped with the δ1 and 'full' delta power gain QTLs. Although each of these 5 QTLs explained only <5% of the total variance in δ2 power gain, together they explained no less than 75% of the variance. These genetic findings extend our previous observations that δ1 and δ2 power gain are regulated through distinct signaling pathways.

Gene prioritization significantly scored the DNA-helicase Wrn (Werner syndrome RecQ like helicase) as a candidate for δ1 power gain, while no significant candidates were found for the full delta gain and for the δ2 power gain, probably due to the low effect size for each of the 5 suggestive QTLs. Wrn is located within the suggestive QTL on chromosome 8 (Figure 5B Bottom) and its expression was strongly associated with a long phosphatidylcholine (PC-ae-C38:5; Figure 5C Bottom). We found that Wrn expression in the cortex specifically was driven by a cis-eQTL (rs51740715, chr8:35.2 Mb, q=1.9e-7) with D2-allele carriers having higher expression levels than B6 carriers under control conditions. Moreover, this same cis-eQTL region determined the magnitude of the sleep-deprivation induced decrease in Wrn expression, such that after sleep deprivation D2-allele carriers now displayed lower levels than B6-allele carriers (genotype x SD interaction: p=5.2e-10; Figure 5D Right). Moreover, a higher gain in δ1 power was associated with a stronger down-regulation of Wrn after sleep deprivation (Figure 5E Right).

Wrn encodes a DNA-repair protein involved in several aging-related diseases (Muftuoglu et al, 2008) and is regulated by Sirt1 (Lee et al, 2015) which, in turn, is involved in redox homeostasis, senescence, and wakefulness (Mouchiroud et al, 2013; Panossian et al, 2011). Down-regulation of Wrn alters redox homeostasis through a metabolic shift, impacts glucose metabolism and increases oxidative stress (Li et al, 2014; Massip et al, 2006). Wrn helicase mutants also showed up-regulation of long phosphatidylcholines (Aumailley et al, 2015) relevant for the significant association between Wrn expression and PC-ae-C38:5 we reported above. The down-regulation of Wrn after sleep deprivation and its association with the sleep-wake dependent changes in EEG delta power raise questions concerning its involvement in the known sleep-loss related increases in oxidative stress (Everson et al, 2014; Villafuerte et al, 2015) and the age-related reduction in EEG delta power (Dijk, 2010; Hasan et al, 2012).

***Example 2:* The level of fast delta activity in the NREM sleep EEG after sleep deprivation**

Apart from the sleep-wake driven gain in EEG delta power discussed above, the prevalence and magnitude of the delta oscillations per se are under strong genetic control both in human and mouse (Landolt, 2011; Maret et al, 2005). These two aspects of EEG delta power represent two unrelated EEG phenotypes governed by different genetic factors (Franken, 2012). Accordingly, the QTLs associated with the gain in δ1 power and in δ2 power presented above, did not associate with levels of δ1 power and δ2 power reached after sleep deprivation. Moreover, as for the δ1 and δ2 gain phenotypes, the levels in delta power measured after sleep deprivation differed between the δ1 and δ2 frequency bands. For example, the lowest/highest delta power values for the δ2 band were found in BXD75 and -44, respectively (Figure 5A Top), while these two lines ranked 2nd and 14th (out of 33) for δ1 power. For δ2 power after sleep deprivation, a significant QTL was identified on chromosome 2 (LOD = 3.87; 136-144 Mb) that explained 42% of the variance. This QTL was specific for recovery sleep and did not associate with δ1 power for which no QTL was found. We did however find a suggestive QTL at the same locus for the power in the full delta band explaining a mere 2% of the variance in the full delta band.

Our prioritization strategy revealed Kif16b (Kinesin family member 16B) as the top significant candidate gene for δ2 power after sleep deprivation (Figure 5B Top). The high prioritization score was based on the strong cis-eQTL associated with Kif16b expression in both cortex and liver (Figure 5C highlight; novel marker, chr2:142.4 Mb, q=1.3e-15 in cortex, q=7.13e-5 in liver), the pronounced down-regulation of Kif16b expression in cortex after sleep deprivation (Figure 5D Left), and the positive correlation between δ2 power after sleep deprivation and Kif16b expression (Figure 5E Left). Lines carrying a B6-allele at the chromosome 2-associated region displayed higher δ2 power after sleep deprivation and a significantly higher Kif16b expression compared to D2-allele carriers (Figure 5E Left).

Kif16b encodes a kinesin involved in early endosome and receptor transport, including of receptors that play a role in sleep regulation such as fibroblast growth factor (FGF) (Ueno et al, 2011), nerve growth factor (NGF) (Yasuda et al, 2007), and ionotropic glutamate (AMPA) (Farkhondeh et al, 2015) receptors. AMPA receptor (AMPA-R) levels are sleep-wake driven, associated with changes in EEG delta power, and have been explored as therapeutic targets to counter the deleterious effects of sleep deprivation on cognition (Boyle et al, 2012; Del Cid-Pellitero et al, 2017; Lanté et al, 2011; Porrino et al, 2005; Vyazovskiy et al, 2008). Our results

thus corroborate a link between fast delta EEG activity after sleep deprivation and AMPA-R trafficking, and implicate Kif16b as a candidate molecular go-between. Of interest, given the large changes in Arc expression after sleep deprivation reported above, is that increased Arc expression reduces the number of AMPA-Rs through its direct interaction with components of the endocytic pathway, thereby contributing to homeostatic synaptic scaling (Korb & Finkbeiner, 2011; Rial Verde et al, 2006). Whether Arc-dependent AMPA-R trafficking through the endocytic pathways involves Kif16b's role in the localization of early endosomes requires further study.

Example 3: **Sleep deprivation shifts theta peak frequency (TPF) in the REM sleep EEG**

The EEG during REM sleep in the mouse is dominated by an almost single-frequency theta oscillation in the 5 to 9 Hz range of hippocampal origin (Buzsáki, 2002), the main frequency of which can be easily determined with a Fourier transformation (Figure 6A). Theta activity during REM sleep is important for memory consolidation (Boyce et al, 2016). Our current data (see h2 analysis above) confirm our previous observations that most of the variance in theta-peak frequency (TPF) among inbred strains of mice can be explained by additive genetic factors (Franken et al, 1998; Tafti et al, 2003). Here we discovered that increased sleep pressure shifts REM sleep TPF (compared to REM sleep TPF in corresponding baseline hours, i.e. ZT6-12) and that the direction of this shift strongly depends on genetic background (Figure 6A).

For this phenotype, we found a significant QTL on chromosome 4 (LOD = 4.94, 104-123 Mb; 50% variance explained) and a suggestive QTL chromosome 8 (LOD=2.73, 0-15 Mb; 32% variance explained; Figure 6C Top). The prioritization strategy identified Cyp4a32 (Cytochrome P450, family 4, subfamily a, polypeptide 32) as the top candidate gene in the liver (Figure 6C Bottom). Cyp4a32, which was not expressed in cortex, is located within the associated chromosome 4 phQTL region, is under strong cis-eQTL effect (rs27480007, chr4: 115.2 Mb, q=1.0e-12) greatly increasing its expression in D2-allele carriers, and contains a non-synonymous protein-damaging variation in the coding region of the D2 allele (V314E, PolyPhen2 score = 1.0, see Methods). Sleep deprivation causes TPF to accelerate in carriers of the D2-allele at the Cyp4a32 cis-eQTL locus and to slow down in B6 carriers. Cyp4a32 expression in D2-allele carriers is high in baseline and increases further after sleep deprivation, while remaining low and stable under both conditions in B6-allele carriers (Figure 6D). The

two F1 hybrids both have a positive TPF-shift, suggesting a dominance of the D2 allele, although some D2-allele carrying lines did show a negative TPF-shift (Figure 6F), indicating that this variation is not sufficient and possibly interacts with other loci, such as the suggestive QTL on chromosome 8, and with metabolites. Hiveplot visualization revealed that the TPF-shift was associated with several amino acids (Figure 6B), which were all significantly down-regulated after sleep deprivation (see above and Figure 4B). The 3 top-ranked associated amino acids were the branched-chain amino acids (BCAAs) leucine, isoleucine, and valine. Plasma levels of valine, in turn were significantly linked to Cyp4a32 expression (Figure 6E highlight) although no common mQTL was found. Cyp4a32, and its human ortholog CYP4A11, are part of the Cyp4a gene family encoding cytochrome 450 liver enzymes that can ω-hydroxylate fatty acids and which are induced by starvation and diabetes (Kroetz et al, 1998). Cyp4a32 encodes a peptide targeting the degradation of arachidonic acid (ARA) specifically. ARA is abundant in the brain, but its levels largely depend on supply by blood (Bazinet & Layé, 2014). ARA and its metabolites, such as prostaglandins and endocannabinoids, are involved in many processes in the brain including signaling, synaptic plasticity, long-term potentiation, and neurogenesis, and have been associated in cognitive performance, mood, and neurodegenerative disease (Bazinet & Layé, 2014; DeCostanzo et al, 2010; Williams et al, 1989). The relation between TPF and fatty acid metabolism has already been suggested with the identification of Acads, an acyl-CoA dehydrogenase involved not only in fatty-acid β-oxidation but also in BCAA degradation (KEGG: mmu00280), as the causative gene explaining REM sleep TPF differences between two inbred strains (but not the sleep deprivation induced shift in TPF reported here) (Tafti et al, 2003). BCAAs, in turn, are involved in fatty acid biosynthesis (Crown et al, 2015; van der Hoeven & Steffens, 2000) and are also implicated in insulin resistance (Newgard, 2012). These results suggest a pathway relating the sleep deprivation effects on BCAA, and possibly arachidonic acid, through fatty acid metabolism in the periphery, with the marked sleep-deprivation induced changes in TPF during REM sleep.

Example 4: **Compensation for NREM sleep time lost**

During recovery sleep, mice compensate for the sleep lost during the preceding sleep deprivation not only by sleeping deeper (quantified as the increase in EEG delta power discussed above), but also by sleeping more (Franken et al, 1999). We quantified the gain in NREM and REM sleep time over the 24h recovery period following the sleep deprivation by

contrasting these recovery values to time-matched baseline values within individual mice. We found that the gain for both NREM and REM sleep was largest in the first 6h of the recovery dark period (ZT12-18; Figure 7A), consistent with our earlier observations (Franken et al, 1999). However, only for the NREM sleep gain during that period did we identify a significant QTL on chromosome 4 [LOD=4.38; 103-110 Mb] explaining 45% of the variance in this trait (Figure 7C). A second, suggestive QTL was found on chromosome 1 [LOD=3.14; 169-173 Mb]. Together the two loci explained 55% of the variance in NREM sleep gain. Neither QTL was associated with the gain in REM sleep during this period (not even at the suggestive level), further underscoring the different regulation, both genetic and physiological, of these two sleep states.

Hiveplot visualization of NREM sleep gain over the 24h recovery period readily revealed the contrasting systems genetics 'landscapes' for the four consecutive 6h recovery intervals with the ZT12-18 interval yielding far more connections at all 4 levels of analysis (Figure 7B). For instance, during this interval, 36 metabolites were highly correlated to NREM sleep gain, while in the second half of the recovery dark period (ZT18-24) only 2 remained correlated. Of these 36 metabolites, 32 were phosphatidylcholines and for 2 among those [PC-ae-C38:2 (LOD=3.27; chr4: 101-110 Mb; 37% variance explained) and PC-ae-C42:5 (LOD=3.02; chr4: 101-110 Mb; 31% variance explained)] suggestive mQTLs were identified both mapping to the chromosome 4 phQTL for NREM sleep gain (Figure 7C). Gene prioritization identified Acot11 as the top candidate gene independently for the gain in NREM sleep and for PC-ae-C38:2 levels (Figure 7C), while for PC-ae-C42:5 no gene passed the prioritization FDR threshold. Nevertheless, both metabolites were significantly linked to Acot11 expression, as can be seen in the hiveplot highlight for NREM sleep gain at ZT12-18, along with 13 other metabolites (Figure 7D).

A significant cis-eQTL was found that explained the differences in Acot11 (acyl-CoA thioesterase 11) expression levels among BXD lines in liver after sleep deprivation (rs28135130, chr4:106.3Mb, q=1e-13), but not in cortex. Liver Acott11 expression in mice carrying the D2-allele at the cis-eQTL region was close to zero (Figure 7E,F). This near-zero expression in D2-allele carriers was even more pronounced for the shorter Acot11 isoform (NM_025590), which was the more abundant isoform in the liver of B6-allele carriers. By contrast, the D2-allele did not alter the expression of the short isoform in the cortex, and in both genotypes its expression was higher than that of the longer isoform (Appendix Figure S7).

Moreover, expression of the less prevalent, longer isoform (NM_001347159) was not affected by genotype. Besides the tissue- and isoform-specific regulation of Acot11 expression, sleep deprivation differentially modified Acot11 expression in cortex and liver. The strong cis-eQTL effect associated with Acot11 expression in the liver after sleep deprivation was not present in the cortex for the control condition (q=0.5) and only marginal after sleep deprivation (q=7e-4). Acot11 was down-regulated in liver after sleep deprivation, but up-regulated in the cortex (Figure 7F).

D2-allele carriers display lower plasma PC-ae-C38:2 levels and have a larger NREM sleep gain during ZT12-18 (Figure 7E). While the majority of the BXD lines compensated by sleeping significantly more than baseline during ZT12-18 (+33.0 minutes on average), only BXD83 showed a negative gain (-1.3 minutes, Figure 7A). BXD83 is also the line with the highest PC-ae-C38:2 plasma levels and the 3rd highest Acot11 expression in liver after sleep deprivation (Figure 7E). It is intriguing that the NREM sleep gain and PC-ae-C38:2 levels measured in the parental strains are closer to that in BXD lines carrying the opposite allele (Figure 7A,E). This reinforces the idea that these phenotypes are due to multiple gene x gene interactions. It should be kept in mind that in these analyses the metabolome and transcriptome data were obtained in tissues collected immediately after the sleep deprivation (ZT6), while the gain in NREM sleep time was quantified in the ensuing recovery. Thus, changes in Acot11 expression and/or PC-ae-C38:2 levels seem to predispose to differences in NREM sleep recovery occurring later.

Acot11 is an acyl-CoA thioesterase which catalyzes the hydrolysis of long fatty acyl-CoAs to form free fatty acids and therefore important in the homeostatic regulation and turn-over of free fatty acids (Cohen, 2013). Acot11 knockout mice show increased energy expenditure and are resistant to diet-induced obesity and its metabolic consequences (Zhang et al, 2012). We used these Acot11 knockout mice to verify the causal involvement of Acot11 in NREM sleep gain in mice. In the line used, the knock-out allele was brought onto a B6 background through repeated (>20) backcrossing. Both heterozygous and homozygous null allele carriers were deficient in NREM sleep gain compared to their wild-type littermate controls (Figure 7G), confirming that Acot11 is causally implicated in NREM sleep recovery. The difference in NREM sleep gain occurred, however, in the second half (ZT18-24) and not, as was the case in the BXD panel, in the first half (ZT12-18) of the recovery dark period.

In humans, sleep deprivation induces an increase in circulating free fatty acids (FFA) (Broussard et al, 2015). Because both elevated plasma FFA levels (Boden, 2003; DeFronzo, 2004) and sleep restriction (Buxton et al, 2010; Spiegel et al, 2005) can lead to insulin resistance and predispose to metabolic disease including type 2 diabetes, Broussard and colleagues proposed that the effects of sleep restriction on FFA levels might present a mechanism by which sleep restriction causes insulin resistance and increased type 2 diabetes risk (Broussard et al, 2015). Our data implicate Acot11 as a molecular player in this mechanistic link between sleep restriction and its adverse effects on fatty acid metabolism.

DISCUSSION

We have generated a rich multi-dimensional experimentally determined knowledge base drawing on 4 levels of organization from the DNA level, to steady-state RNA levels in brain and liver, circulating metabolites, and a deep phenome of sleep-wake related phenotypes, all under two experimental conditions. At the core of this knowledge base is the BXD ARIL resource. This mouse GRP provides a 'population model' with a controlled and stable degree of genetic variation, each line carrying a fixed and unique pattern of recombination of the two parental chromosomes (Peirce et al, 2004). The panel segregates for ~5.2 million sequence variants corresponding to about half of all common genetic variation among classic laboratory mouse strains (Wang et al, 2016). This level of genetic complexity exceeds that in many human populations, such as the Icelandic and Finnish populations that have been so useful in genetics of disease (Arnar et al, 2016; Milani et al, 2015; Peltonen et al, 2000). Our results underscore the power of the BXD panel in discovering the genetic and molecular underpinnings of clinically relevant traits already demonstrated in other research fields (Andreux et al, 2012; Merkwirth et al, 2016; Williams et al, 2016).

We extracted 341 sleep-wake related phenotypes belonging to 148 distinct phenotypic classes from each individual mouse. Half of these phenotypes had a higher than 0.68 heritability, indicating that they are amenable to genetic dissection even when using only 33 ARI lines. Although numerous knockout studies have shown that (lack of) single genes impact many of the phenotypes we quantified [for review see (Franken & Tafti, 2003; Mang & Franken, 2015)], we demonstrate here that even highly heritable traits are determined by the interaction of several small effect loci. Two striking examples of such traits are TPF during REM sleep and the gain in δ2 power after sleep deprivation for which we identified 4 and 5

suggestive QTLs, respectively, that together explained 58 and 75% of the overall variance in these 2 traits. Thus, while reductionist approaches have been successful at identifying genes affecting sleep in a Mendelian fashion, when studied at a more natural, population level most of these phenotypes represent complex traits, and Mendelian (null) alleles are likely to play a lesser role. To systematically explore these non-additive, multi-loci interactions at the level of the whole genome, innovative algorithms in the area of machine learning are needed. Currently, more than 2-way epistatic interactions are computationally challenging. We are therefore now exploring novel multi-loci epistatic approaches to extract this type of information [see e.g. (He & Parida, 2016; Llinares-Lopez et al, 2015)].

With the 4 examples described, we could only illustrate a fraction of all the novel information contained in our experimentally derived knowledge base. Here we focused on the effects of sleep loss exclusively, because systems genetics resources in this research domain are lacking and because of the immediate clinical relevance of these effects. Importantly, the pathways we identified were unique to the sleep-deprivation condition and did not explain phenotypic variance of the respective traits under undisturbed baseline conditions. This illustrates that already a relatively mild sleep disruption (preventing sleep during half of the rest phase) extensively reshapes the systems genetics landscape.

The power of systems genetics lies in generating hypotheses. In the current dataset, several observations imply sleep deprivation to challenge fatty acid turnover. Besides Acot11, which regulates the levels of free fatty acids and, as we show here, the recovery of NREM sleep, also Cyp4a32, which contributes to the sleep-deprivation induced shift in the frequency of theta oscillation in REM sleep, encodes an enzyme regulating fatty acid levels. This frequency shift was strongly correlated with levels of the branched amino acids leucine, isoleucine, and valine, which, in turn, are part of a fatty acid biosynthesis pathway. The link between Cyp4a32 and the dominant frequency of theta oscillatory activity also illustrates the importance of a peripheral molecular pathway in regulating brain activity, as Cyp4a32 was not expressed in brain. This finding is of relevance because although many studies have emphasized the deleterious effect of sleep loss on peripheral systems, research on the substrate of sleep need largely remains brain centric. In addition, Pla2g4e and Mlycd, the two genes with the strongest cis-eQTL effect for their differential expression after sleep deprivation, both encode enzymes affecting fatty acid metabolism. Acot11, the Cyp4a gene family, free fatty acid levels, and sleep restriction have all been linked to obesity and insulin resistance (Boden, 2003; Buxton et al, 2010; DeFronzo,

2004; Kroetz et al, 1998; Spiegel et al, 2005; Zhang et al, 2012). Another pathway of importance in mediating the effects of sleep loss concerns AMPA-R trafficking supported by the 8-fold increase in cortical Arc expression and Kif16b's role in shaping δ2 power after sleep deprivation. Both genes encode proteins involved in the endosomal trafficking of AMPA-Rs (see Results) that have already been explored as therapeutic targets to counter the deleterious effects of sleep deprivation on cognition (Boyle et al, 2012; Porrino et al, 2005). Finally, Wrn's association with EEG slow waves during NREM sleep offers a model system to mechanistically study the molecular pathways underlying the characteristic age-related decrease in the prevalence of EEG slow waves and sleep quality.

Hypotheses concerning the involvement of the pathways in the sleep homeostatic process we discovered need to be further tested experimentally. With a reverse genetics approach, we could already confirm Acot11's role in the recovery of sleep time lost. This approach is, however, not always informative or possible because a lack of protein on a given genetic background is unlikely to mimic the impact of an allelic variant in a genetically diverse population, or the knock-out might be lethal as is the case for Kif16b (Ueno et al, 2011). Efforts to comprehensively phenotype (including sleep), knockouts for all known and predicted mouse genes by the International Mouse Phenotyping Consortium (IMPC; www.mousephenotype.org) are ongoing, but unfortunately, no knockouts for the 4 genes we highlight here have been submitted for phenotyping. Another important community resource is the mostly mouse oriented database GeneNetwork (www.genenetwork.org), which hosts a massive amount of phenotypic and molecular information collected by the many researchers using the same BXD resource. We are in the process of structuring our database to enable sharing the integrated data in GeneNetwork according to the FAIR data management concepts (Wilkinson et al, 2016). Furthermore, cross-species validation in, e.g., humans, flies, and C. elegans and GWAS and biobank database searches are important additional ways of validating and extending our mouse observations. According to the human GWAS databases grasp.nhlbi.nih.gov and www.ebi.ac.uk/gwas/, SNP variants in Acot11 are significantly associated with (among others) the rate of cognitive decline in Alzheimer disease, behavioral disinhibition, cardiovascular disease, and triglyceride levels. Variants in Wrn are associated with aging and time-to-death, cardiovascular disease, cholesterol, and daytime rest. Finally, variants in the human ortholog of Cyp4a32, CYP4A11, are associated with blood metabolite levels including amino acids and acyl carnitines, and Kif16b variants with intelligence.

A first evaluation of the systems genetics field has highlighted a clear need for better communication, 'open science', and collaboration among groups (Baliga, Björkegren et al. 2017). Towards this aim, we have shared our results and analyses through an easily accessible and reproducibility-oriented web interface that accompanies this publication. We hope that the interactivity of the web interface will encourage the reader to further mine our data, thereby reproducing our conclusions and, hopefully, discovering other key regulators and pathways. In our analyses, we have also strived to follow the concepts of the FAIR data management approach (Wilkinson et al, 2016), resulting in a data life-cycle management plan, open access provided by the web interface for datamining, and, importantly, interoperability. The implementation of the FAIR approach will be illustrated in an accompanying publication.

In summary, we have applied a systems genetics approach to uncover new genes and pathways associated with the effects of sleep loss, an approach thought critical towards predicting disease susceptibility (Civelek & Lusis, 2014). This integrative, multi-level approach allowed us to follow the flow of information from DNA variants, to molecular intermediate phenotypes, to behavioral and electrophysiological end phenotypes, and to assess how this network of multi-scale effects is perturbed by an environmental challenge. The information gained could not have been achieved through other genetic approaches that are based on the '1-gene-to-1-phenotype' approach. Moreover, with the tools and web interface we developed, our open access knowledge base provides a unique resource which goes well beyond merely cataloguing and ranking ph-, m-, and eQTLs. Furthermore, owing to the use of a GRP, the database and its content are easily scalable. A first challenge will be to complement the dataset with females of the same lines. In addition, we are expanding the database with an additional intermediate phenotype namely the sleep-deprivation induced changes in chromatin accessibility, aiming to identify the variants in non-coding regulatory elements that could predict the varying molecular and phenotypic response to sleep loss. Adding proteome, microbiome, and inflammasome data are obvious other intermediate phenotypes that will further strengthen this knowledge base and increase its value to e.g. assist with identifying biomarkers gauging sleep pressure and potential therapeutic targets for sleep-wake related disorders.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTONS

SD supervised and coordinated all aspects of animal breeding and the experiments, sleep deprived mice, collected tissues, performed transcriptome and metabolome analyses, interpreted the results, and wrote the manuscript. MJ performed all systems genetics analyses, interpreted the results, and wrote the manuscript. YE performed all EEG surgeries and visually annotated sleep-wake states, sleep deprived mice, and collected tissues. NG developed the semi-automatic annotation of sleep-wake state. BM and DJS were responsible for the metabolomics analyses and DJS commented on the manuscript. MI and FB designed the RNA-seq analyses pipe-line and guided the systems genetics analyses. LG, MS, and RL built the web interface and implemented the interactive hiveplots and data mining tools. MP helped with developing the prioritization strategy. CNH helped guide the analyses, interpreted the results, and wrote the manuscript. IX guided and supervised all aspects of the systems genetics analyses and commented on the manuscript. PF designed the experiments, supervised the project, analyzed all sleep-wake related phenotypes, interpreted the results, and wrote the manuscript. All authors agreed with the final version of the submitted manuscript.

CONFLICT of INTEREST

The authors have no conflicts of interest to declare.

MATERIALS & METHODS

Animals, breeding, and housing conditions

Experiments were approved by the local veterinary authorities (SCAV authorization #2534). We phenotyped 33 BXD RI strains originating from the University of Tennessee Health Science Center (Memphis, TN, USA). 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/2J (D2) and C57B/6J (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 12h light/12h dark cycle [LD12:12, fluorescent lights, intensity 6.6cds/m2, with Zeitgeber time ZT0 and -12 designating light and dark onset, respectively]. Male mice aged 11-14 weeks 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 Figure 1) [mean = 6.2/line; 5≤n≤7] and for molecular phenotyping (Experiment 2 in Figure 1) [mean = 6.8/line; 6≤n≤9]. Additionally, to assess the stability of outcome variables over time, parental lines and reciprocal F1 offspring were phenotyped twice; i.e., at the start (labelled B6-1 and D2-1) and end (labelled B6-2 and D2-2) 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 256 mice for tissue collection for transcriptome and metabolome analyses (Experiment 2), the latter being divided into sleep deprived (SD) and controls (Ctr; see Experimental design section below). We strived to randomize the lines across the experimental cohorts so that biological replicates of one line were collected/recorded on more than one 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).

Experimental design

Figure 1 summarizes the experimental design which consisted of two experiments; i.e., Experiment 1 and 2. Animals of both experiments were maintained under the same housing conditions. Animals in Experiment 1 underwent surgery and, after a >10-day recovery period, EEG and locomotor activity (LMA) were recorded continuously for a 4-day period starting at ZT0. The first two days were considered baseline (B1 and B2). The first 6h of Day 3 (ZT0-6), animals were sleep deprived in their home cage by "gentle handling" (Mang & Franken, 2012). The remaining 18h of Day 3 and Day 4 were considered recovery (R1 and R2). Half of the animals included in Experiment 2 were sleep deprived (SD) alongside the animals of Experiment 1. The other half was left undisturbed in another room (i.e., control or Ctr). Both SD and Ctr mice of Experiment 2 were sacrificed at ZT6, i.e., immediately after the end of the sleep deprivation, for sampling of liver and cerebral cortex tissue, as well as trunk blood. All mice were left undisturbed for at least two days prior to sleep deprivation.

Experiment 1: EEG/EMG and LMA recording and analysis: EEG/EMG surgery was performed under deep anesthesia according to our standard methods (Mang & Franken, 2012). EEG and EMG signals were amplified, filtered, digitized, and stored using EMBLA (Medcare Flaga, Thornton, CO, USA) hardware (A10 recorder) and software (Somnologica). Locomotor activity (LMA) was recorded by passive infrared (PIR) sensors (Visonic Ltd, Tel Aviv, Israel) at 1-min resolution for the duration of the 4-day experiment using ClockLab (ActiMetrics, IL, USA).

Offline, the sleep-wake states wakefulness (W), rapid-eye movement (REM) sleep, and non-REM (NREM) sleep were annotated on consecutive 4-s epochs, based on the EEG and EMG patterns. To assist the annotation of this extensive dataset (ca. 20 million 4-s epochs), we developed a semi-automated 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 (Mang & Franken, 2012). 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 one 'reference' mouse, even per line, to reliably annotate sleep-wake states for the others (Appendix Figure S1A). To overcome this problem, one day out of 4 (i.e., Day 3 or R1, which includes the sleep deprivation) was visually annotated for each mouse. These 4-s sleep-wake scores were used to train the semi-

61

automatic scoring algorithm, which took as input 82 numerical variables derived from the analyses of EEG and EMG signals using frequency- (Discrete Fourier Transform or DFT) and time-domain analyses performed at 1-s resolution. We then used these data to train a series of Support Vector Machines (SVM) (Meyer et al, 2014) 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 one 4-s 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. A representative example of prediction is shown in Appendix Figure S1B. To prevent over-fitting and assess the expected performance of the predictor, only 50% of the R1 manually annotated data from each mouse was used for training. The classification performance was assessed by comparing the automatic and visual scoring of the fully manually annotated 4-day recordings of 43 mice. The global accuracy was computed using a confusion matrix (Kuhn et al, 2014) of the completely predicted days (B1, B2, and R2; Appendix Figure S1C). 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.

We quantified 341 phenotypes based on the sleep-wake states, LMA, and the EEG signal, constituting three broad phenotypic categories. The 96h sleep-wake sequence of each animal was used to directly assess traits in three '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 vs NREM, (ii) aspects of their distribution over the 24h cycle, e.g., time course of hourly values, midpoint of the 12h 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, overall activity counts per day as well as per unit of time-spent-awake, and the distribution of activity over the 24h cycle were extracted from the LMA data. 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 4s epochs

matching the sleep-wake states using DFT (0.25Hz 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; 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 inter-individual variability in absolute power. Moreover, the frequency of dominant EEG rhythms was extracted as phenotypes. In particular, a sub-state of wakefulness, referred to as TDW, was quantified according to the prevalence of theta-activity (6.0-10.0 Hz) in the EEG during waking (Buzsáki, 2002; Welsh et al, 1985), according to the algorithm described in (Vassalli & Franken, 2017). We assessed the time spent in this state, the fraction of total wakefulness it represents, and its distribution over 24h. Finally, discrete, paroxysmal events were counted, such as sporadic spontaneous seizures and neocortical spindling, which are known features of D2 mice (Ryan, 1984), which we also found in some BXD lines.

All phenotypes were quantified in baseline and recovery separately, and the effect of sleep deprivation on all variables was computed as recovery vs. baseline differences or ratios. The recovery-to-baseline contrasts are the focus of this paper. Obviously, some of the 341 phenotypes are strongly correlated (e.g. the time-spent-awake and –asleep in a given recording interval), resulting in identical QTLs (albeit with different association strengths). To estimate the number of unique phenotypes, we clustered highly correlated phenotypes into modules. We then counted the number of phenotype categories and subcategories within each module (Appendix Figure S2). We estimated 148 separable groups of phenotypes. Please see the 'Swiss-BXD' web interface (https://bxd.vital-it.ch; Downloads, General_Information.xlsx) for a full listing of all phenotypes quantified.

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 minutes per mouse. Blood was collected at the decapitation site into tubes containing 10 ml heparin (2U/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.

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 256 mice initially sacrificed for tissue collection yielded 222 cortex and 222 liver samples of good quality.

Equal amounts of RNA from biological replicates (same strain and experimental condition) were pooled, yielding 78 samples for library preparation. RNA-seq libraries were prepared from 500ng 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). 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 41M 100-bp single-end reads were obtained [29M $\leq$ n $\leq$ 63M].

Targeted metabolomics analysis was performed using flow injection (FIA) and liquid chromatography/mass spectrometry (LC/MS) as described in (Davies et al, 2014; Isherwood et al, 2017). 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 (ACT), amino acids (AA), biogenic amines (BA), hexoses, glycerophospholipids (GPL), and sphingolipids (SM). Plasma samples were prepared according to the manufacturer's instructions. Sample order was randomized and 3 levels of quality controls (QCs), run on each 96-well plate. Data were normalized between batches using the results of QC 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 (Isherwood et al, 2017). Out of the 188 metabolites assayed, 124 passed these criteria across samples and were used in subsequent analyses. 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, -85, and -101 with respectively 2, 1 and 1 animal/condition used; see above under 'Animals'). Note that in contrast to the RNAseq experiment, samples were not pooled, but analyzed individually.

In the same plasma samples, we determined corticosterone levels using an enzyme immunoassay (corticosterone EIA kit; Enzo Life Sciences Inc, Lausen, Switzerland) according to the manufacturer's instructions. All samples were diluted 40 times in the provided buffer, kept on ice during the 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 five C57BL/6 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.

RNA-sequencing analyses

RNA sequencing (RNA-seq) data were processed using the Illumina Pipeline Software version 1.82. All RNA-seq samples passed FastQC quality thresholds (version 0.10.1) and could thus be used in subsequent analysis. For gene expression quantification, we used a standard pipeline that was already applied in a previous study (Picard et al, 2016). Reads were mapped to MGSCv37/mm9 using the STAR splice aligner with the 2pass pipeline (Dobin et al, 2013). Count data was generated using htseq-count from the HTseq package using parameters 'stranded=reverse' and 'mode=union' (Anders et al, 2015). Gene boundaries were extracted from the mm9/refseq/reflat dataset of the UCSC table browser. EdgeR was then used to normalize read counts by library size. Genes with a mean raw read count below 10 were excluded from the analysis, and the raw read counts were normalized using the TMM normalization (Robinson & Oshlack, 2010) and converted to log counts per million (CPM). To assess the differential expression between the sleep deprived and control conditions, we used the R package limma (Ritchie et al, 2015) with the voom weighting function followed by the limma empirical Bayes method (Law et al, 2014). RNA-seq data are deposited in NCBI GEO (accession code GSE1xxxxx).

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 (McKenna et al, 2010). To improve coverage depth, two additional RNA-seq datasets from other projects using the same BXD lines were added (Picard et al, 2016). In total, six BXD datasets from four different tissues (cortex, hypothalamus, brainstem, and liver) were used. A hard filtering procedure was applied as suggested by the GATK pipeline (DePristo et al, 2011; McKenna et al, 2010; Van der Auwera et al, 2013). Furthermore, genotypes with more than 10% missing information, low quality (<5000), and redundant information were removed. GeneNetwork genotypes, which were discrepant with our RNAseq experiment, were tagged as 'unknown' (mean of 1% of the GeneNetwork genotypes/strain [0.05% ≤n≤ 8%]). Finally, GeneNetwork and our RNA-seq genotypes were merged into a unique set of 11'000 genotypes, which was used for all subsequent analyses. This set of genotypes was already used successfully in a previous study of BXD lines (Picard et al, 2016), and is available through our 'Swiss-BXD' web interface (https://bxd.vital-it.ch; Downloads, Genotypes.GeneNetwork2005AndRNAseq.geno).

QTL mapping

The R package qtl/r (Broman et al, 2003) was used for interval mapping of behavioral/EEG phenotypes (phQTLs) and metabolites (mQTLs). Pseudo-markers were imputed every cM and genome-wide associations were calculated using the Expected-Maximization (EM) algorithm. p-values were corrected for false discovery rate (FDR) using permutations 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 (Burgess-Herbert et al, 2008; Lander & Kruglyak, 1995). 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. We used the following calculation to obtain the variance explained for a single QTL effect: $1 - 10^{\wedge}(-(2/n)*LOD)$. To compute the additive effect of multiple suggestive, highly suggestive, and significant QTLs, we used the fitqtl function of the qtl/r package (Broman & Sen, 2009). For detection of expression QTLs (eQTLs), cis-eQTLs were mapped using FastQTL (Ongen et al, 2016) with a 2Mb window for which adjusted p-values were computed with 1000 permutations and beta distribution fitting. The R package qvalue (Bass JDSwcfAJ, 2015) was then used for multiple-testing correction as proposed by (Ongen

et al, 2016). Only the q-values are reported for each cis-eQTL in the text. Trans-eQTL detection was performed using a modified version of FastEpistasis (Schupbach et al, 2010), on several million associations (15K genes x 11K markers), applying a global, hard p-value threshold of 1E-4.

Protein damage prediction

Variants detected by our RNA-seq variant calling were annotated using Annovar (Wang et al, 2010b) with the RefSeq annotation dataset. Non-synonymous variations were further investigated for protein disruption using Polyphen-2 version 2.2.2 (Adzhubei et al, 2010), which was adapted for use in the mouse according to recommended configuration.

Hiveplot visualization

Hiveplots were constructed with the R package HiveR (Krzywinski et al, 2012) for each phenotype. Gene expression and metabolite levels represented in the hiveplots come from either the Ctr (control) or SD (sleep deprivation) 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 threshold of 0.5 for genes and 0.4 for metabolites. Cross-associations between genes 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 [see Results and Figure 4B,C for details].

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, Figure 2C), (ii) correlation analysis, (iii) expression QTL (eQTL, Figure 2B), (iv) protein damaging variation prediction, and (v) differential expression (DE, Figure 3A). Each result was transformed into an "analysis score" using a min/max normalization, where the contribution of extreme values was reduced by a winsorization of the results (Appendix Figure S6A). 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 one integrated score in cortex and one 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 simply attributed to the corresponding gene. To associate a gene with the ph-/mQTL analysis score (which is associated to markers), we used the central position of the gene to infer the associated ph-/mQTL 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 (Appendix Figure S6B). 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 (Appendix Figure S6C). The integrated score (formula in Figure 4D) was calculated separately for cortex and liver. We performed a 10'000 permutations procedure to compute a FDR for the integrated scores. For each permutation procedure, all 5 analyses 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 quantified with one or more suggestive/significant QTL after sleep deprivation, 40 had at least 1 gene significantly prioritized with Henikoff weighting, against 32 without.

Data accessibility and the 'Swiss-BXD' web interface

Published data and interactive hiveplot figures can be visualized and data-mined through our web interface: https://bxd.vital-it.ch. See Appendix Figure S5 for an outline of the web interface. The help section in the interface provides a first-step guide.

Figure 1

Figure 2

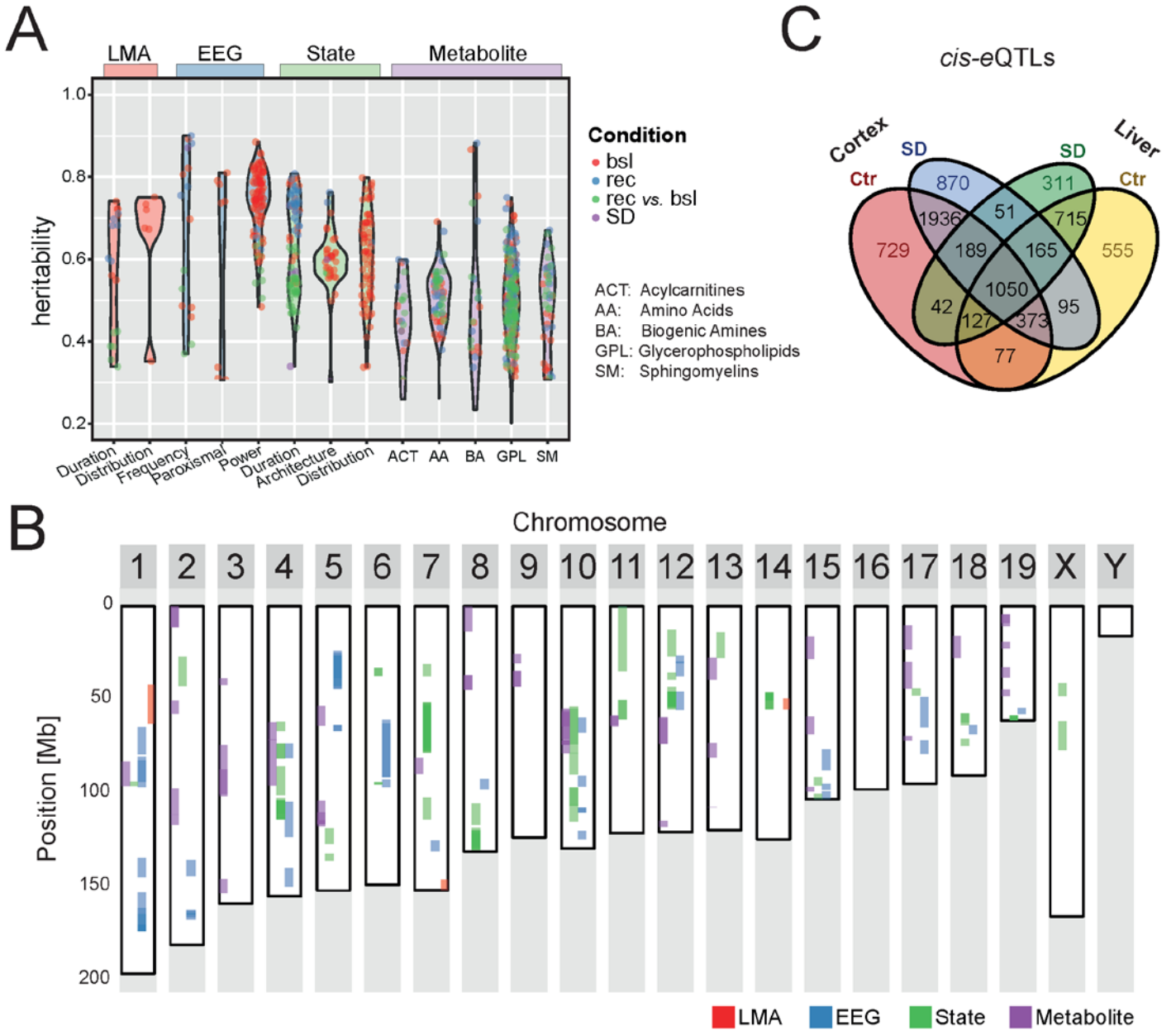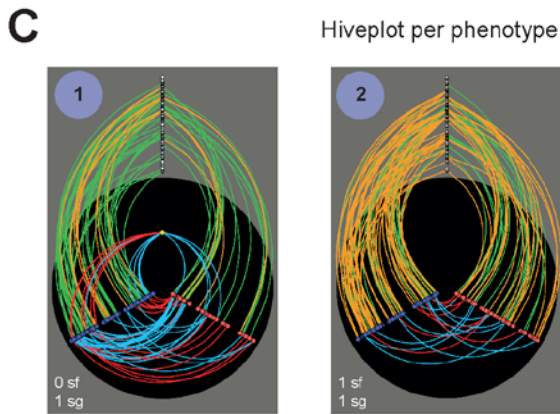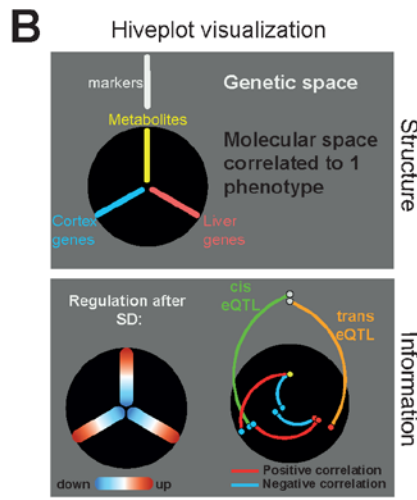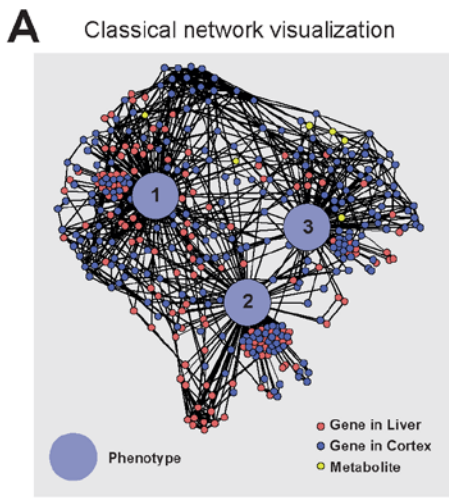**Systems genetics visualization**

A Classical network visualization

- Gene in Liver
- Gene in Cortex
- Metabolite
- Phenotype

B Hiveplot visualization

Structure

markers
Genetic space
Metabolites
Molecular space correlated to 1 phenotype
Cortex genes
Liver genes

Information

Regulation after SD:
down — up

cis eQTL
trans eQTL
Positive correlation
Negative correlation

C Hiveplot per phenotype

1 — 0 sf 1 sg
2 — 1 sf 1 sg
3 — 1 sf 2 sg

sf: significant; sg suggestive QTL

Systems genetics prioritization strategy

D Scoring for a single phenotype

Genes (i) on genome (I)	eQTL	QTL	DE	Correlation	Variations	Ranking of score (S)

High Scores
Low Scores

Henikoff weighting (W): weight weight weight weight weight

Ranking: 1, 2, 3, 4, 5, 6

Scoring formula: $\forall i \in I. \quad S_i = \sum_{k=1}^{5} V_k^i W_k$

E Gene prioritization for the α-AAA QTL

Dhtkd1

QTL LOD score / −log FDR
Markers
Chr 2 [Mb]: 0 4 8 12 16 20 24 28

Liver genes

Gene	Score	log FDR
Dhtkd1	0.662	2.95
Vipr2	0.532	0.95
Camk1d	0.511	0.72

Figure 3

Figure 4

Figure 5

A EEG θ spectra in REM sleep

baseline BXD61
recovery
baseline BXD101
recovery

[% of total power; 1/f corrected]

EEG frequency [Hz]

B TPF-shift

1 sf
1 sg

BCAA

C TPF-shift QTL

lod

Chromosome

TPF-shift prioritization, liver

lod   - log FDR

Cyp4a32

Chr 4 [Mb]

D Liver *Cyp4a32*

DE *p* = 1.9e-4

[CPM]

D2 allele
B6 allele

Ctr   SD   Ctr   SD

E *Cyp4a32* highlight

Valine

F TPF-shift   Liver *Cyp4a32*

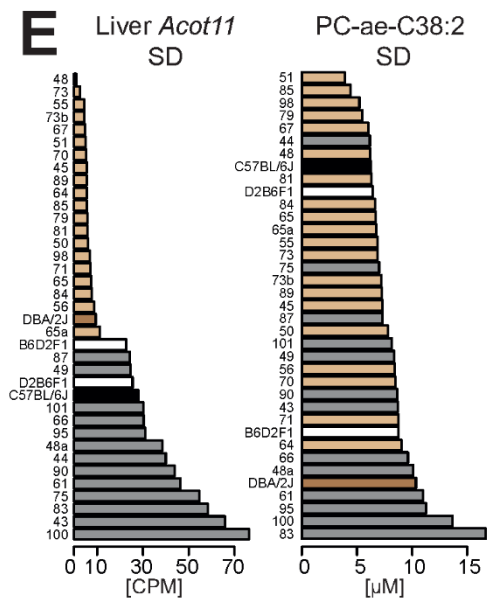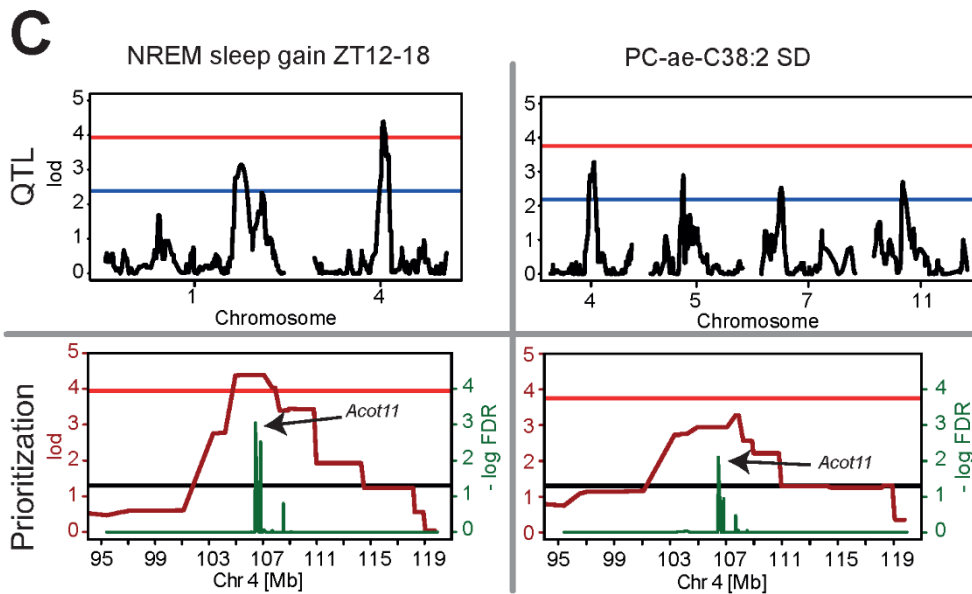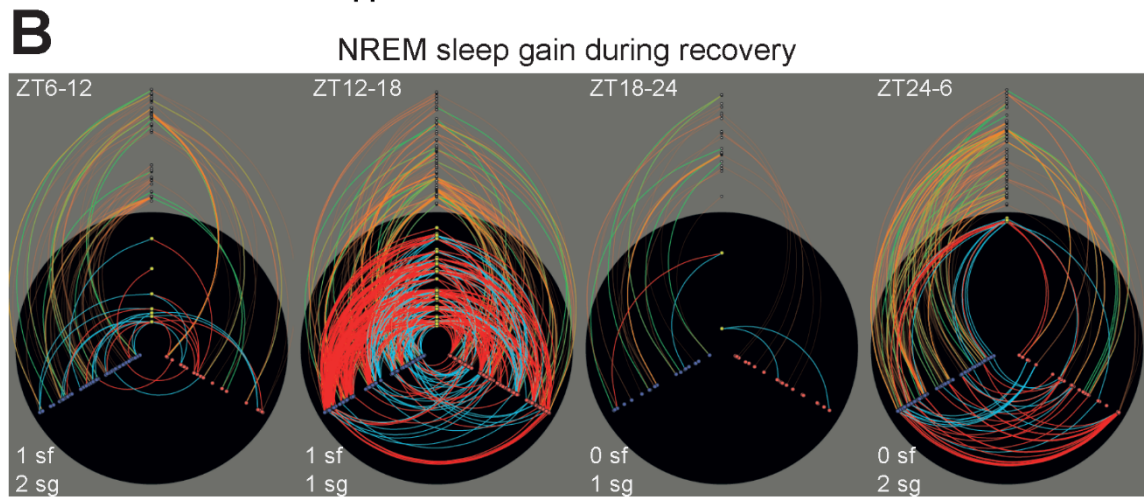[ΔHz]   [CPM]
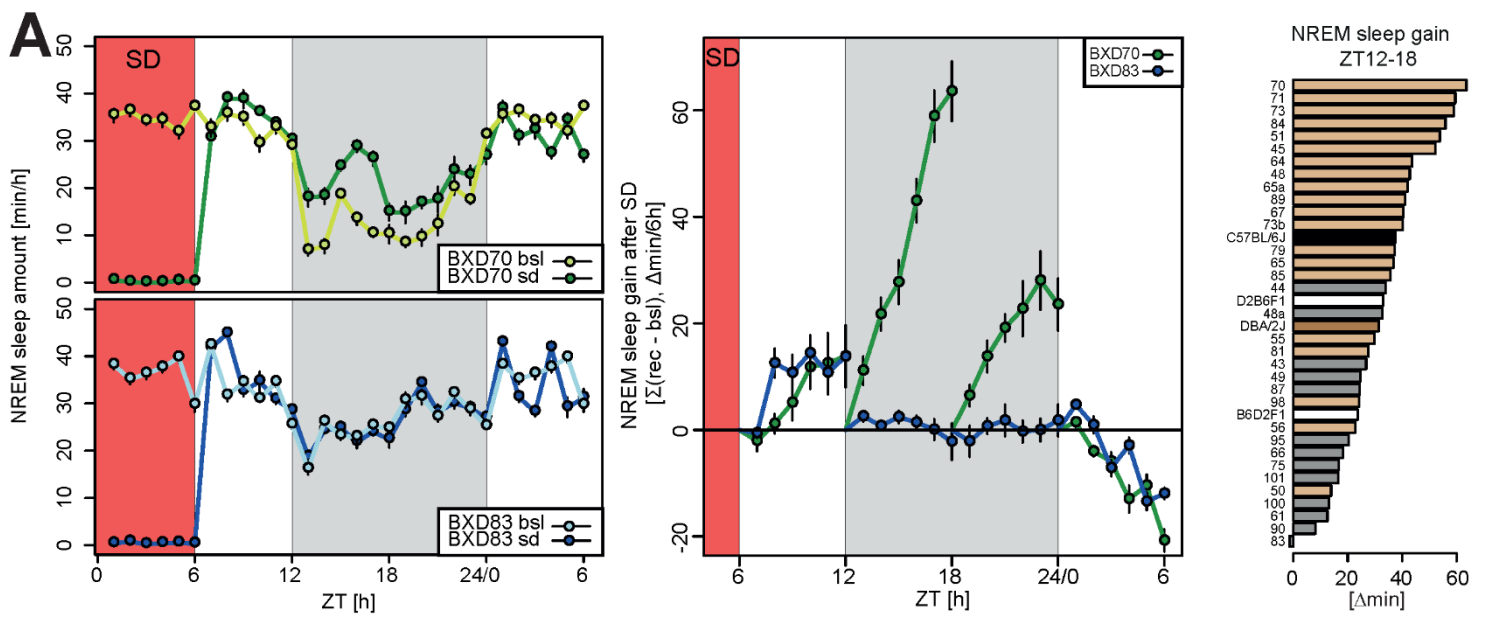
eQTL *p* = 1e-15
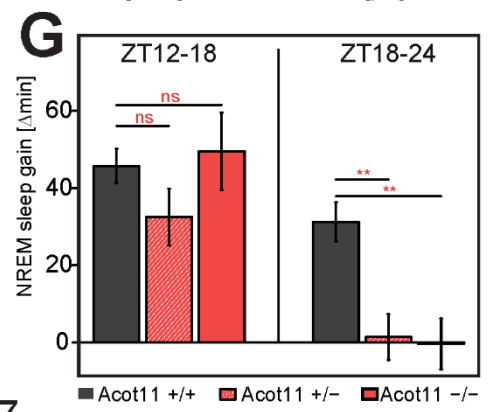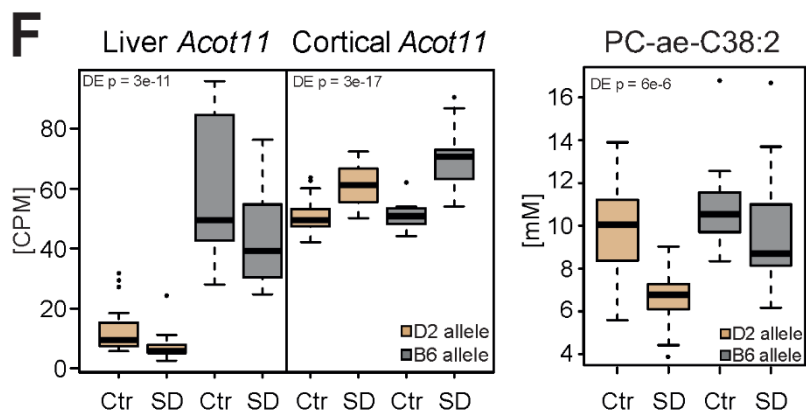
Figure 6

Figure 7

FIGURE LEGENDS

Figure 1. Experimental design. Thirty-three BXD lines plus the 2 parental strains and their reciprocal F1 progeny were phenotyped. Mice were submitted to either one of two experiments. In Experiment 1 (Left) EEG/EMG signals and locomotor activity (LMA) were recorded under standard, 12:12h light-dark conditions (white and black bars under Top-Left panel) for two baseline days (B1, B2), a 6h sleep deprivation (SD; red bar) from ZT0-6 (ZT0 = light onset), followed by two recovery days (R1, R2). The deep sleep-wake phenome consists of 341 sleep-wake state-, LMA-, and EEG-related phenotypes quantified in each mouse, among which time spent in NREM sleep (grey area spans mean max and min NREM sleep time among BXD lines, respectively, for consecutive 90min intervals). Mice in Experiment 2 (Right) were used to collect cortex, liver, and blood samples at ZT6. Half of the mice were challenged with a SD as in Experiment 1, the other half was left undisturbed and served as controls (Ctr). Cortex and liver samples were used to quantify gene expression by RNA-sequencing (RNA-seq), blood samples for a targeted analysis of 124 metabolites by liquid chromatography–mass spectrometry (LC-MS) or with flow injection analysis-mass spectrometry (FIA-MS). For mapping of phenotypic, metabolic, and expression QTLs (i.e., ph-, m-, and eQTLs, respectively), a high-density genotype dataset (Genome; 11K SNPs) was created merging identified RNA-seq variants with a publicly available database (www.genenetwork.org). Systems genetics analyses draws on the four -omics datasets and the 3 types of QTLs.

Figure 2. Genetic diversity in the BXD panel greatly impacts behavioral, metabolic, and molecular traits. The phenome was divided into 3 phenotypic categories: (i) locomotor activity (LMA), (ii) EEG features (EEG), and (iii) sleep-wake state characteristics (state), which were subdivided further (see Methods). The five classes of metabolites and the gene expression represent intermediate molecular phenotypic categories. (A) Heritability for EEG/behavioral and metabolite phenotypes. Dots represent single phenotypes within each category and sub-category indicated along the x-axis. Red dots represent phenotypes recorded in baseline (bsl; B1 and B2), blue in recovery (rec; R1 and R2), purple during sleep deprivation (SD), and green dots refer to the recovery-to-baseline contrasts. Values represent narrow-sense heritability (B) Overview of significant and highly suggestive (FDR<0.1) QTLs obtained for all 341 EEG/behavioral phenotypes (phQTLs: LMA in red, EEG in blue, and sleep-wake state in green) and 124 blood metabolite levels in baseline and recovery (mQTLs; purple). Note that overlap of neighboring QTLs render color shading darker. (C) Venn diagram of genes under significant

cis-eQTL effect in liver and cortex for the two experimental conditions [sleep deprived (SD) and controls (Ctr)].

Figure 3. How to illustrate multi-dimensional networks and prioritize candidate genes? (A) Classical network visualization methods strongly depend on the layout algorithm used for positioning nodes, making structure interpretation and reproducibility difficult. (B) Hiveplot network visualization and structure strategy. See text for details. (C) The classical network visualization for the 3 phenotypes (blue nodes 1-3) in panel A can be represented with our method with 1 hiveplot per phenotype. Phenotype 1 showed more cortex-liver correlations than the two other phenotypes through one metabolite, connecting up- and down-regulated genes in cortex after SD and down-regulated genes in liver. Phenotype 2 shows genomic regions with strong allelic effect over multiple genes in liver and cortex through a high number of trans-eQTLs. Phenotype 3 was mostly connected to cortically expressed genes correlating strongly with up-regulated metabolites; most cis/trans-eQTLs affected only cortical genes. The number of significant (sf) and suggestive (sg) phQTLs detected for each phenotype are indicated on Bottom Left. The 3 phenotypes were related to active wake behaviors during recovery (Phenotype 1 and 2: LMA per h awake and time in TDW, respectively, both during ZT12-24; Phenotype 3: Gain in time-spent-awake during ZT24-6). (D) Gene prioritization strategy to identify driver genes underlying phenotype/metabolite variation, illustrated for 6 genes. Five types of analyses were integrated into a single score for each gene reflecting its strength as candidate driver gene, namely from Left to Right (i) and (ii) QTL mapping for gene expression (eQTLs) and phenotypes/metabolites (ph- or mQTLs), respectively, (iii) differential expression (DE) after sleep deprivation, (iv) gene expression/phenotype correlations, and (v) analysis of protein damaging genetic variations relating genes to an allelic effect. See text for further details. (E) To illustrate and validate our scoring strategy, genes in liver were prioritized for levels of alpha-aminoadipic acid (αAAA) after sleep deprivation. Dhtkd1 was identified as top-ranked driver gene. Results from QTL mapping (red line) and prioritization analysis (green line); red and black horizontal lines indicate significant thresholds for the QTL and prioritization, respectively.

Figure 4. Profound effects of sleep deprivation on transcriptome, metabolome, and phenome. EEG/behavioral phenotypes, metabolites, and transcripts are organized into 3 'columns' (from Left to Right). Top 3 panels show the sleep-deprivation (SD) response (recovery/baseline fold-change). Bottom 3 panels depict examples of allelic effects on the SD responses with color-coding indicating the presence of a C57BL/6J or DBA/2J haplotype under the mapped QTL peaks (B6: grey for BXD and black for parental; D2: light brown for BXD and dark-brown for parental). White bars mark the F1s and hatched bars strain in which haplotype could not be unambiguously determined. (A) Phenotypic changes after SD. The top significantly changed phenotype was the increase in NREM sleep EEG delta power (1-4 Hz) after SD (far left blue data point). The most up-regulated phenotype was time spent in REM sleep during the first 6h of darkness (ZT12-18) after SD (highest green data point). (B) Metabolite changes after SD. Most amino acids (blue) were down-regulated and most sphingolipids (brown) up-regulated after SD. The acylcarnitines C18:1 and C18:2 (highest red dots) increased the most. Vertical red line: significant threshold (FDR adjusted p-value=0.05). (C) Differential gene expression (DE) analysis (SD/Ctr) for cortex (Left) and liver (Right). Genes were sorted according to their ranked p-value along the x-axis. Significantly affected transcripts in red (FDR adjusted p-value <0.05), non-significant results in black. Blue dots indicate 78 genes considered core molecular components of the sleep homeostatic response in the cortex (Mongrain et al, 2010). Note that no low fold-change threshold was applied. (D-F) Examples of genetically driven EEG/behavioral, metabolic, and transcriptional responses to SD, respectively. See text for details.

Figure 5. EEG delta power in NREM sleep after sleep deprivation (SD) is associated with Kif16b and Wrn. (A) NREM sleep EEG spectra in the first 3h after SD (ZT6-9) for the 2 BXD lines that displayed the lowest and highest EEG activity in the fast delta frequency band (2.5-4.25 Hz, δ2; Top, see panel E) and for the 2 BXD lines that displayed the smallest and largest increase (or gain) in EEG power in the slow delta band (1.0-2.25 Hz, δ1; Bottom, see panel E). Spectra were '1/f-corrected' (and therefore not directly comparable to the values in panel E) for better visualization of activity in higher frequency bands [theta (5-9 Hz, θ), sigma (11-16 Hz, σ), beta (18-30 Hz, β), and slow (32-55 Hz, γ1) and fast gamma (55-80 Hz, γ2)]. Subsequent analyses were performed without this correction. (B) QTL mapping and prioritization for δ2-power identified a significant association on chromosome 2 and Kif16b in

cortex as top-ranked gene (Top). For the δ1-increase after SD, we obtained a suggestive QTL on chromosome 8 and a significant prioritization score for the DNA-helicase Wrn. (C) Hiveplot visualization of network connections for the δ1- and δ2-power after SD (Top Left panels) and the SD-induced increase in δ1 and δ2 power over baseline (Bottom Left panels). Note the marked differences in the networks and QTLs regulating the expression of these two delta bands. Right hiveplots highlight Kif16b in the δ2 power associated network (Top), and Wrn in the network associated with the δ1-increase (Bottom). Only Kif16b expression in the cortex was linked to the chromosome 2 cis-eQTL and was not associated with any metabolite. Wrn expression was significantly linked to the chromosome 8 cis-eQTL and to the long phosphatidylcholine, PC-ae-C38:5. (D) Kif16b is highly significantly down-regulated in cortex (Left), while it remains unchanged in liver after SD (p=0.15; not shown). Also Wrn expression was strongly down-regulated by SD in cortex (Right) and only marginally so, albeit significantly, in liver (p=0.02; not shown). (E) Strain distribution patterns. BXD lines carrying a B6-allele on the chromosome 2 associated region showed higher δ2 power after SD (Left) and a significantly higher Kif16b expression (p=1.3e-15; 2nd-to-Left) than D2-allele carriers. D2-allele carriers of the chromosome 8-associated region showed a larger δ1-increase after SD (2nd-to-Right) as well as a significantly larger decrease in Wrn expression after SD (Right) than B6-allele carriers. For color-coding of genotypes see Figure 4.

Figure 6. Changes in the frequency of theta oscillation during REM sleep after sleep deprivation are associated with Cyp4a32. (A) Spectral profiles of the REM sleep EEG for two strains displaying an opposite shift in the frequency of theta oscillations after SD relative to baseline. This shift was quantified by the decrease and increase in theta-peak frequency (TPF) for BXD61 and -101, respectively (see panel F). (B) Hiveplot for the SD-induced shift in TPF. (C) One significant QTL for the TPF-shift was detected on chromosome 4 and one suggestive QTL on chromosome 8. Prioritization yielded Cyp4a32 as the top-ranked significant gene, based on the significant cis-eQTL modifying its expression in liver and a predicted damaging variation (V314E). (D) Effects of SD and genotype on liver Cyp4a32 expression. Carrying a B6-allele at the Cyp4a32 cis-eQTL associated marker greatly decreased its expression. (E) Hiveplot for the SD-induced shift in TPF highlighting Cyp4a32's links to the amino acid Valine and the chromosome 4 eQTL marker. (F) Strain distribution patterns for TPF differences and liver Cyp4a32 expression after SD. B6-allele carriers at the chromosome 4

associated region had lower Cyp4a32 liver expression and a decrease in TPF after SD, while D2-carriers increase TPF and have higher Cyp4a32 expression.

Figure 7. NREM sleep gain in the first 6h of the dark period after sleep deprivation is associated with Acot11. (A) Time course of hourly values of time spent in NREM sleep in baseline, sleep deprivation (SD, red area) and recovery for the two BXD lines showing the largest (BXD70; green) and lowest (BXD83; blue) NREM sleep gain during ZT12-18 (Left). NREM sleep gain during 4 consecutive 6h intervals during recovery compared to corresponding baseline intervals shows that in the recovery dark period (grey area) BXD83 mice did not accumulate extra NREM sleep while BXD70 mice gain 88min (Middle). Strain distribution of ZT12-18 NREM sleep gain (Right). B6-allele carriers compensated less for NREM sleep lost during SD than D2-allele carriers. For color-coding see Figure 4. **(B)** Hiveplots for NREM sleep gain in four 6h recovery intervals after the end of SD at ZT6. Compared to the other 3 intervals, NREM sleep gain was strongly associated with a number of metabolites during the second 6h interval; i.e. ZT12-18. **(C)** NREM sleep gain during ZT12-18 mapped to a significant QTL on chromosome 4, explaining 45% of the total phenotypic variance (Top Left). The phosphatidylcholine C38:2 (PC-ae-C38:2) mapped suggestively to the same region (Top Right). Prioritization of liver transcripts for both phenotypes yielded Acot11 as top-ranked, significant gene (Bottom). **(D)** Hiveplot for the ZT12-18 NREM sleep gain, highlighting Acot11. Acot11 was positively correlated with several phosphatidylcholines and to Ovgp1 expression in the cortex. **(E)** Allelic effect of the chromosome 4-associated region on Acot11 expression and PC-ae-C38:2 levels in the BXDs. Acot11 expression in liver after SD was under a strong eQTL effect ($p=1.6e-13$) with B6-allele carriers showing a higher Acot11 expression than D2-allele carriers. B6-allele

carriers also showed higher PC-ae-C38:2 levels after SD. **(F)** Both Acot11 and PC-ae-C38:2 levels changed after SD. Acot11 in liver and PC-ae-C38:2 in blood were significantly down-regulated. In the cortex, Acot11 was, however, significantly up-regulated and the chromosome 4-associated region did not modulate cortical Acot11 expression. **(G)** Mice carrying one or two KO alleles for Acot11 displayed less extra NREM sleep during recovery. In contrast to the BXD panel, this difference was present in the second (ZT18-24, Right), and not during the first (ZT12-18, Left panel) 6h of the recovery dark period.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7: 248-249

Amzica F, Steriade M (1998) Electrophysiological correlates of sleep delta waves. *Electroencephalogr Clin Neurophysiol* 107: 69-83

Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166-169

Andretic R, Franken P, Tafti M (2008) Genetics of sleep. *Annual review of genetics* 42: 361-388

Andreux PAA, Williams EG, Koutnikova H, Houtkooper RH, Champy M-FF, Henry H, Schoonjans K, Williams RW, Auwerx J (2012) Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. *Cell* 150: 1287-1299

Arnar DO, Andersen K, Thorgeirsson G (2016) Genetics of cardiovascular diseases: lessons learned from a decade of genomics research in Iceland. *Scandinavian cardiovascular journal : SCJ* 50: 260-265

Aumailley L, Garand C, Dubois MJ, Johnson FB, Marette A, Lebel M (2015) Metabolic and Phenotypic Differences between Mice Producing a Werner Syndrome Helicase Mutant Protein and Wrn Null Mice. *PloS one* 10

Baliga NS, Björkegren JL, Boeke JD, Boutros M, Crawford NP, Dudley AMM, Farber CR, Jones A, Levey AI, Lusis AJ, Mak HC, Nadeau JH, Noyes MB, Petretto E, Seyfried NT, Steinmetz LM, Vonesch SC (2017) The State of Systems Genetics in 2017. *Cell systems* 4: 7-15

Bass JDSwcfAJ DA, and, D, R. (2015) qvalue: Q-value estimation for false discovery rate control. *R package version 230*

Bazinet RP, Layé S (2014) Polyunsaturated fatty acids and their metabolites in brain function and disease. *Nature reviews Neuroscience* 15: 771-785

Benington JH, Heller HC (1995) Restoration of brain energy metabolism as the function of sleep. *Progress in neurobiology* 45: 347-360

Boden G (2003) Effects of free fatty acids (FFA) on glucose metabolism: significance for insulin resistance and type 2 diabetes. *Experimental and clinical endocrinology & diabetes : official journal, German Society of Endocrinology [and] German Diabetes Association* 111: 121-124

Boyce R, Glasgow SD, Williams S, Adamantidis A (2016) Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science (New York, NY)* 352: 812-816

Boyle J, Stanley N, James LM, Wright N, Johnsen S, Arbon EL, Dijk DJ (2012) Acute sleep deprivation: the effects of the AMPAKINE compound CX717 on human cognitive performance, alertness and recovery sleep. *Journal of psychopharmacology* 26: 1047-1057

Broman KW, Sen S (2009) *A Guide to QTL Mapping with R/qtl*, Vol. 46: Springer.

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19: 889-890

Broussard JL, Chapotot F, Abraham V, Day A, Delebecque F, Whitmore HR, Tasali E (2015) Sleep restriction increases free fatty acids in healthy men. *Diabetologia* 58: 791-798

Burgess-Herbert SL, Cox A, Tsaih S-WW, Paigen B (2008) Practical applications of the bioinformatics toolbox for narrowing quantitative trait loci. *Genetics* 180: 2227-2235

Buxton OM, Pavlova M, Reid EW, Wang W, Simonson DC, Adler GK (2010) Sleep restriction for 1 week reduces insulin sensitivity in healthy men. *Diabetes* 59: 2126-2133

Buzsáki G (2002) Theta oscillations in the hippocampus. *Neuron* 33: 325-340

Cirelli C, Bushey D, Hill S, Huber R, Kreber R, Ganetzky B, Tononi G (2005a) Reduced sleep in Drosophila Shaker mutants. *Nature* 434: 1087-1092

Cirelli C, Huber R, Gopalakrishnan A, Southard TL, Tononi G (2005b) Locus ceruleus control of slow-wave homeostasis. *J Neurosci* 25: 4503-4511

Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. *Nature reviews Genetics* 15: 34-48

Cohen DE (2013) New players on the metabolic stage: How do you like Them Acots? *Adipocyte* 2: 3-6

Crown SB, Marze N, Antoniewicz MR (2015) Catabolism of Branched Chain Amino Acids Contributes Significantly to Synthesis of Odd-Chain and Even-Chain Fatty Acids in 3T3-L1 Adipocytes. *PloS one* 10

Daan S, Beersma DG, Borbély AA (1984) Timing of human sleep: recovery process gated by a circadian pacemaker. *The American journal of physiology* 246: 83

Davies SK, Ang JE, Revell VL, Holmes B, Mann A, Robertson FP, Cui N, Middleton B, Ackermann K, Kayser M, Thumser AE, Raynaud FI, Skene DJ (2014) Effect of sleep deprivation on the human metabolome. *Proceedings of the National Academy of Sciences of the United States of America* 111: 10761-10766

DeCostanzo AJ, Voloshyna I, Rosen ZB, Feinmark SJ, Siegelbaum SA (2010) 12-Lipoxygenase regulates hippocampal long-term potentiation by modulating L-type Ca2+ channels. *J Neurosci* 30: 1822-1831

DeFronzo RA (2004) Dysfunctional fat cells, lipotoxicity and type 2 diabetes. *International journal of clinical practice Supplement*: 9-21

Del Cid-Pellitero E, Plavski A, Mainville L, Jones BE (2017) Homeostatic Changes in GABA and Glutamate Receptors on Excitatory Cortical Neurons during Sleep Deprivation and Recovery. *Frontiers in systems neuroscience* 11: 17

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491-498

Diering GH, Nirujogi RS, Roth RH, Worley PF, Pandey A, Huganir RL (2017) Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. *Science (New York, NY)* 355: 511-515

Dijk D-JJ (2010) Slow-wave sleep deficiency and enhancement: implications for insomnia and its management. *The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry* 11 Suppl 1: 22-28

Dissel S, Melnattur K, Shaw PJ (2015) Sleep, Performance, and Memory in Flies. *Curr Sleep Med Rep* 1: 47-54

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21

Dossi RC, Nunez A, Steriade M (1992) Electrophysiology of a slow (0.5-4 Hz) intrinsic oscillation of cat thalamocortical neurones in vivo. *J Physiol* 447: 215-234

Everson CA, Henchen CJ, Szabo A, Hogg N (2014) Cell injury and repair resulting from sleep loss and sleep recovery in laboratory rats. *Sleep* 37: 1929-1940

Farkhondeh A, Niwa S, Takei Y, Hirokawa N (2015) Characterizing KIF16B in neurons reveals a novel intramolecular "stalk inhibition" mechanism that regulates its capacity to potentiate the selective somatodendritic localization of early endosomes. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 35: 5067-5086

Franken P (2012) Chapter 4: Genetic mechanisms underlying rhythmic EEG activity during sleep *Sleep and Brain Activity, Ed Frank; Academic Press pp 59-89, ISBN: 0123849950*

Franken P, Chollet D, Tafti M (2001) 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

Franken P, Dudley CA, Estill SJ, Barakat M, Thomason R, O'Hara BF, McKnight SL (2006) NPAS2 as a transcriptional regulator of non-rapid eye movement sleep: genotype and sex interactions. *Proceedings of the National Academy of Sciences of the United States of America* 103: 7118-7123

Franken P, Malafosse A, Tafti M (1998) Genetic variation in EEG activity during sleep in inbred mice. *The American journal of physiology* 275: 37

Franken P, Malafosse A, Tafti M (1999) Genetic determinants of sleep regulation in inbred mice. *Sleep* 22: 155-169

Franken P, Tafti M (2003) Genetics of sleep and sleep disorders. *Frontiers in bioscience : a journal and virtual library* 8: 97

Funato H, Miyoshi C, Fujiyama T, Kanda T, Sato M, Wang Z, Ma J, Nakane S, Tomita J, Ikkyu A, Kakizaki M, Hotta-Hirashima N, Kanno S, Komiya H, Asano F, Honda T, Kim SJ, Harano K, Muramoto H, Yonezawa T et al (2016) Forward-genetics analysis of sleep in randomly mutagenized mice. *Nature* 539: 378-383

Gatti DM, Harrill AH, Wright FA, Threadgill DW, Rusyn I (2009) Replication and narrowing of gene expression quantitative trait loci using inbred mice. *Mammalian Genome* 20: 437-446

Giskeødegård GF, Davies SK, Revell VL, Keun H, Skene DJ (2015) Diurnal rhythms in the human urine metabolome during sleep and total sleep deprivation. *Scientific reports* 5: 14843

Harbison ST, Carbone MA, Ayroles JF, Stone EA, Lyman RF, Mackay TF (2009) Co-regulated transcriptional networks contribute to natural genetic variation in Drosophila sleep. *Nature genetics* 41: 371-375

Hasan S, Dauvilliers Y, Mongrain V, Franken P, Tafti M (2012) Age-related changes in sleep in inbred mice are genotype dependent. *Neurobiology of aging* 33

He D, Parida L (2016) Muse: A Multi-Locus Sampling-Based Epistasis Algorithm for Quantitative Genetic Trait Prediction. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 22: 426-437

Hegmann JP, Possidente B (1981) Estimating genetic correlations from inbred strains. *Behavior genetics* 11: 103-114

Isherwood CM, Van der Veen DR, Johnston JD, Skene DJ (2017) Twenty-four-hour rhythmicity of circulating metabolites: effect of body mass and type 2 diabetes. *FASEB J*

Itabe H, Yamaguchi T, Nimura S, Sasabe N (2017) Perilipins: a diversity of intracellular lipid droplet proteins. *Lipids in health and disease* 16: 83

Jiang P, Scarpa JR, Fitzpatrick K, Losic B, Gao VD, Hao K, Summa KC, Yang HS, Zhang B, Allada R, Vitaterna MH, Turek FW, Kasarskis A (2015) A systems approach identifies networks and genes linking sleep and stress: implications for neuropsychiatric disorders. *Cell reports* 11: 835-848

Koh K, Joiner WJ, Wu MN, Yue Z, Smith CJ, Sehgal A (2008) Identification of SLEEPLESS, a sleep-promoting factor. *Science (New York, NY)* 321: 372-376

Korb E, Finkbeiner S (2011) Arc in synaptic plasticity: from gene to behavior. *Trends in neurosciences* 34: 591-598

Kroetz DL, Yook P, Costet P, Bianchi P, Pineau T (1998) Peroxisome proliferator-activated receptor alpha controls the hepatic CYP4A induction adaptive response to starvation and diabetes. *The Journal of biological chemistry* 273: 31581-31589

Krueger JM, Rector DM, Roy S, Van Dongen HP, Belenky G, Panksepp J (2008) Sleep as a fundamental property of neuronal assemblies. *Nature reviews Neuroscience* 9: 910-919

Krzywinski M, Birol I, Jones SJ, Marra MA (2012) Hive plots--rational approach to visualizing networks. *Briefings in bioinformatics* 13: 627-644

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Team tRC (2014) caret: Classification and Regression Training. *http://CRANR-projectorg/package=caret*

Kuna ST, Maislin G, Pack FM, Staley B, Hachadoorian R, Coccaro EF, Pack AI (2012) Heritability of performance deficit accumulation during acute sleep deprivation in twins. *Sleep* 35: 1223-1233

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics* 11: 241-247

Landolt H-PP (2011) Genetic determination of sleep EEG profiles in healthy humans. *Progress in brain research* 193: 51-61

Lanté F, Toledo-Salas J-CC, Ondrejcak T, Rowan MJ, Ulrich D (2011) Removal of synaptic Ca^{2+}-permeable AMPA receptors during sleep. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31: 3953-3961

Law CW, Chen JC, Shi W, Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts.

Lee S-YY, Lee H, Kim E-SS, Park S, Lee J, Ahn B (2015) WRN translocation from nucleolus to nucleoplasm is regulated by SIRT1 and required for DNA repair and the development of chemoresistance. *Mutation research* 774: 40-48

Li B, Iglesias-Pedraz JM, Chen L-YY, Yin F, Cadenas E, Reddy S, Comai L (2014) Downregulation of the Werner syndrome protein induces a metabolic shift that compromises redox homeostasis and limits proliferation of cancer cells. *Aging cell* 13: 367-378

Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB (2016) Prevalence of Healthy Sleep Duration among Adults--United States, 2014. *MMWR Morb Mortal Wkly Rep* 65: 137-141

Llinares-Lopez F, Grimm DG, Bodenham DA, Gieraths U, Sugiyama M, Rowan B, Borgwardt K (2015) Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics* 31: i240-249

Lo JC, Groeger JA, Santhi N, Arbon EL, Lazar AS, Hasan S, von Schantz M, Archer SN, Dijk DJ (2012) Effects of partial and acute total sleep deprivation on performance across cognitive domains, individuals and circadian phase. *PLoS One* 7: e45987

Mackiewicz M, Paigen B, Naidoo N, Pack AI (2008) Analysis of the QTL for sleep homeostasis in mice: Homer1a is a likely candidate. *Physiological genomics* 33: 91-99

Mang GM, Franken P (2012) Sleep and EEG Phenotyping in Mice. *Current protocols in mouse biology* 2: 55-74

Mang GM, Franken P (2015) Genetic dissection of sleep homeostasis. *Current topics in behavioral neurosciences* 25: 25-63

Maquet P (1995) Sleep function(s) and cerebral metabolism. *Behavioural brain research* 69: 75-83

Maret S, Dorsaz S, Gurcel L, Pradervand S, Petit B, Pfister C, Hagenbuchle O, O'Hara BF, Franken P, Tafti M

(2007) Homer1a is a core brain molecular correlate of sleep loss. *Proceedings of the National Academy of Sciences of the United States of America* 104: 20090-20095

Maret S, Franken P, Dauvilliers Y, Ghyselinck NB, Chambon P, Tafti M (2005) Retinoic acid signaling affects cortical synchrony during sleep. *Science (New York, NY)* 310: 111-113

Marini S, Santangeli O, Saarelainen P, Middleton B, Chowdhury N, Skene DJ, Costa R, Porkka-Heiskanen T, Montagnese S (2017) Abnormalities in the Polysomnographic, Adenosine and Metabolic Response to Sleep Deprivation in an Animal Model of Hyperammonemia. *Front Physiol* 8: 636

Masana M, Jukic MM, Kretzschmar A, Wagner KV, Westerholz S, Schmidt MV, Rein T, Brodski C, Müller MB (2015) Deciphering the spatio-temporal expression and stress regulation of Fam107B, the paralog of the resilience-promoting protein DRR1 in the mouse brain. *Neuroscience* 290: 147-158

Massip L, Garand C, Turaga RV, Deschênes F, Thorin E, Lebel M (2006) Increased insulin, triglycerides, reactive oxygen species, and cardiac fibrosis in mice with a mutation in the helicase domain of the Werner syndrome gene homologue. *Experimental gerontology* 41: 157-168

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-1303

Merkwirth C, Jovaisaite V, Durieux J, Matilainen O, Jordan SD, Quiros PM, Steffen KK, Williams EG, Mouchiroud L, Tronnes SU, Murillo V, Wolff SC, Shaw RJ, Auwerx J, Dillin A (2016) Two Conserved Histone Demethylases Regulate Mitochondrial Stress-Induced Longevity. *Cell* 165: 1209-1223

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2014) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. *http://CRANR-projectorg/package=e1071*

Milani L, Leitsalu L, Metspalu A (2015) An epidemiological perspective of personalized medicine: the Estonian experience. *Journal of internal medicine* 277: 188-200

Mongrain V, Hernandez SA, Pradervand S, Dorsaz S, Curie T, Hagiwara G, Gip P, Heller HC, Franken P (2010) Separating the contribution of glucocorticoids and wakefulness to the molecular and electrophysiological correlates of sleep homeostasis. *Sleep* 33: 1147-1157

Moreau Y, Tranchevent L-CC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews Genetics* 13: 523-536

Mouchiroud L, Houtkooper RH, Auwerx J (2013) NAD^+ metabolism: a therapeutic target for age-related metabolic disease. *Critical reviews in biochemistry and molecular biology* 48: 397-408

Muftuoglu M, Oshima J, von Kobbe C, Cheng W-HH, Leistritz DF, Bohr VA (2008) The clinical characteristics of Werner syndrome: molecular and biochemical diagnosis. *Human genetics* 124: 369-377

Newgard CB (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance. *Cell metabolism* 15: 606-614

O'Donovan KJ, Tourtellotte WG, Millbrandt J, Baraban JM (1999) The EGR family of transcription-regulatory factors: progress at the interface of molecular and systems neuroscience. *Trends Neurosci* 22: 167-173

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32: 1479-1485

Panossian L, Fenik P, Zhu Y, Zhan G, McBurney MW, Veasey S (2011) SIRT1 regulation of wakefulness and senescence-like phenotype in wake neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31: 4025-4036

Peirce JL, Lu L, Gu J, Silver LM, Williams RW (2004) A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC genetics* 5: 7

Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. *Nat Rev Genet* 1: 182-190

Picard A, Soyer J, Berney X, Tarussio D, Quenneville S, Jan M, Grouzmann E, Burdet F, Ibberson M, Thorens B (2016) A Genetic Screen Identifies Hypothalamic Fgf15 as a Regulator of Glucagon Secretion. *Cell reports* 17: 1795-1806

Porrino LJ, Daunais JB, Rogers GA, Hampson RE, Deadwyler SA (2005) Facilitation of task performance and removal of the effects of sleep deprivation by an ampakine (CX717) in nonhuman primates. *PLoS biology* 3

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

Rial Verde EM, Lee-Osbourne J, Worley PF, Malinow R, Cline HT (2006) Increased expression of the immediate-early gene arc/arg3.1 reduces AMPA receptor-mediated synaptic transmission. *Neuron* 52: 461-474

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47

Robinson M, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data.
Rocha C, Papon L, Cacheux W, Marques Sousa P, Lascano V, Tort O, Giordano T, Vacher S, Lemmers B, Mariani P, Meseure D, Medema JP, Bièche I, Hahne M, Janke C (2014) Tubulin glycylases are required for primary cilia, control of cell proliferation and tumor development in colon. *The EMBO journal* 33: 2247-2260

Ryan LJ (1984) Characterization of cortical spindles in DBA/2 and C57BL/6 inbred mice. *Brain research bulletin* 13: 549-558

Schmid SM, Hallschmid M, Schultes B (2015) The metabolic burden of sleep loss. *The lancet Diabetes & endocrinology* 3: 52-62

Schmidt MV, Schülke J-PP, Liebl C, Stiess M, Avrabos C, Bock J, Wochnik GM, Davies HA, Zimmermann N, Scharf SH, Trümbach D, Wurst W, Zieglgänsberger W, Turck C, Holsboer F, Stewart MG, Bradke F, Eder M, Müller MB, Rein T (2011) Tumor suppressor down-regulated in renal cell carcinoma 1 (DRR1) is a stress-induced actin bundling factor that modulates synaptic efficacy and cognition. *Proceedings of the National Academy of Sciences of the United States of America* 108: 17213-17218

Schupbach T, Xenarios I, Bergmann S, Kapur K (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 26: 1468-1469

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J (2006) A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS biology* 4

Shin D-LL, Pandey AK, Ziebarth JD, Mulligan MK, Williams RW, Geffers R, Hatesuer B, Schughart K, Wilk E (2014) Segregation of a spontaneous Klrd1 (CD94) mutation in DBA/2 mouse substrains. *G3 (Bethesda, Md)* 5: 235-239

Siskos AP, Jain P, Römisch-Margl W, Bennett M, Achaintre D, Asad Y, Marney L, Richardson L, Koulman A, Griffin JL, Raynaud F, Scalbert A, Adamski J, Prehn C, Keun HC (2017) Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Analytical chemistry* 89: 656-665

Spiegel K, Knutson K, Leproult R, Tasali E, Van Cauter E (2005) Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes. *Journal of applied physiology (Bethesda, Md : 1985)* 99: 2008-2019

Tafti M, Petit B, Chollet D, Neidhart E, de Bilbao F, Kiss JZ, Wood PA, Franken P (2003) Deficiency in short-chain fatty acid beta-oxidation affects theta oscillations during sleep. *Nature genetics* 34: 320-325

Tononi G, Cirelli C (2014) Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* 81: 12-34

Ueno H, Huang X, Tanaka Y, Hirokawa N (2011) KIF16B/Rab14 molecular motor complex is critical for early embryonic development by transporting FGF receptor. *Developmental cell* 20: 60-71

Urry E, Landolt HP (2015) Adenosine, caffeine, and performance: from cognitive neuroscience of sleep to sleep pharmacogenetics. *Curr Top Behav Neurosci* 25: 331-366

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11 10 11-33

van der Hoeven RS, Steffens JC (2000) Biosynthesis and elongation of short- and medium-chain-length fatty acids. *Plant physiology* 122: 275-282

Vassalli A, Franken P (2017) Hypocretin (orexin) is critical in sustaining theta/gamma-rich waking behaviors that drive sleep need. *Proceedings of the National Academy of Sciences of the United States of America* 114

Villafuerte G, Miguel-Puga A, Rodríguez EM, Machado S, Manjarrez E, Arias-Carrión O (2015) Sleep deprivation and oxidative stress in animal models: a systematic review. *Oxidative medicine and cellular longevity* 2015: 234952

Vyazovskiy VV, Cirelli C, Pfister-Genskow M, Faraguna U, Tononi G (2008) Molecular and electrophysiological evidence for net synaptic potentiation in wake and depression in sleep. *Nature neuroscience* 11: 200-208

Wang H, Liu Y, Briesemann M, Yan J (2010a) Computational analysis of gene regulation in animal sleep deprivation. *Physiological genomics* 42: 427-436

Wang K, Li M, Hakonarson H (2010b) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164

Wang X, Pandey AK, Mulligan MK, Williams EG, Mozhui K, Li Z, Jovaisaite V, Quarles LD, Xiao Z, Huang J, Capra JA, Chen Z, Taylor WL, Bastarache L, Niu X, Pollard KS, Ciobanu DC, Reznik AO, Tishkov AV, Zhulin IB et al (2016)

Joint mouse-human phenome-wide association to test gene function and disease risk. *Nature communications* 7: 10464

Welsh DK, Richardson GS, Dement WC (1985) A circadian rhythm of hippocampal theta activity in the mouse. *Physiology & behavior* 35: 533-538

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3: 160018

Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, Houten SM, Amariuta T, Wolski W, Zamboni N, Aebersold R, Auwerx J (2016) Systems proteomics of liver mitochondria function. *Science (New York, NY)* 352

Williams JH, Errington ML, Lynch MA, Bliss TV (1989) Arachidonic acid induces a long-term activity-dependent enhancement of synaptic transmission in the hippocampus. *Nature* 341: 739-742

Wu Y, Williams EG, Dubuis S, Mottis A, Jovaisaite V, Houten SM, Argmann CA, Faridi P, Wolski W, Kutalik Z, Zamboni N, Auwerx J, Aebersold R (2014) Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* 158: 1415-1430

Xie L, Kang H, Xu Q, Chen MJ, Liao Y, Thiyagarajan M, O'Donnell J, Christensen DJ, Nicholson C, Iliff JJ, Takano T, Deane R, Nedergaard M (2013) Sleep drives metabolite clearance from the adult brain. *Science (New York, NY)* 342: 373-377

Yasuda K, Churchill L, Yasuda T, Blindheim K, Falter M, Krueger JM (2007) Unilateral cortical application of interleukin-1beta (IL1beta) induces asymmetry in fos, IL1beta and nerve growth factor immunoreactivity: implications for sleep regulation. *Brain research* 1131: 44-59

Zhang Y, Li Y, Niepel MW, Kawano Y, Han S, Liu S, Marsili A, Larsen PR, Lee C-HH, Cohen DE (2012) Targeted deletion of thioesterase superfamily member 1 promotes energy expenditure and protects against obesity and insulin resistance. *Proceedings of the National Academy of Sciences of the United States of America* 109: 5417-5422

Appendix to:

A systems genetics analysis of sleep regulation in the mouse

Shanaz Diessler[1], Maxime Jan[1,2], Yann Emmenegger[1], Nicolas Guex[2], Benita Middleton[3], Debra J. Skene[3], Mark Ibberson[2], Frederic Burdet[2], Lou Götz[2], Marco Pagni[2], Martial Sankar[2], Robin Liechti[2], Charlotte N. Hor[1], Ioannis Xenarios[2†], Paul Franken[1,5†]*

Contents:

Appendix Figure S1: EEG/EMG semi-automatic scoring. (**A**) Comparison of the normalized signal for two individual mice (Top and Bottom rows) of two BXD lines (Left and Right) and one parental line (DBA/2J; Middle), visually annotated by an expert scorer. Plotted are the peak-to-peak EMG amplitude (y-axis) against EEG delta (1.0-4.0 Hz) power (x-axis). (**B**) Example of predicted sleep-wake states of a representative 28-minute section (420 4-s epochs) of mouse BXD045-1. Top row: state manually assigned by the expert. Second row: consensus of the automated prediction. Third row: results obtained for 11 distinct SVM (support vector machine) predictors from which the consensus prediction is derived. (**C**) Accuracy values of the prediction for the 43 mice for which the 4-day recording were fully annotated by the expert. The SVM were trained on the R1 recording and then used to predict sleep-wake state for days B1, B2, and R2. Predicted sleep-wake states were compared to manual annotation using a confusion matrix (see Methods).
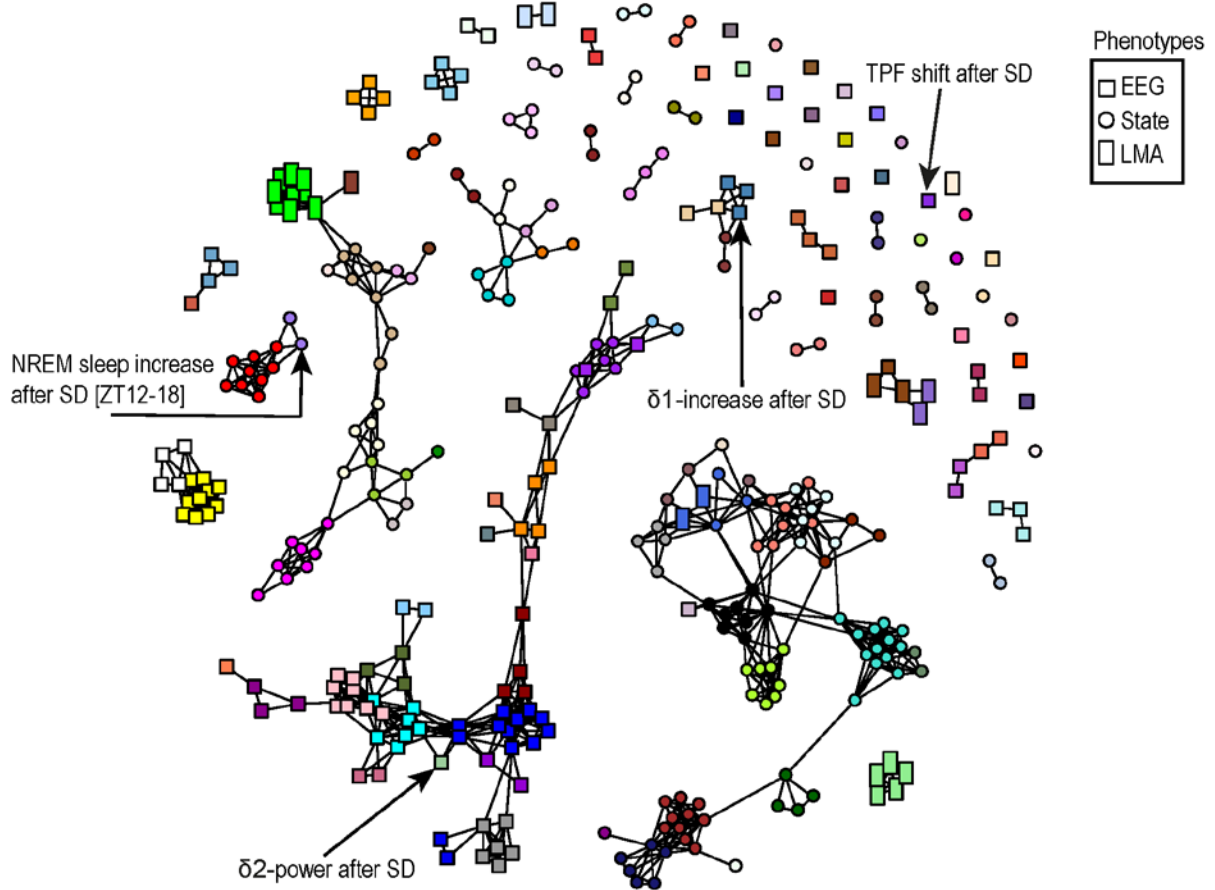
Appendix Figure S2: Relatedness among EEG and behavioral phenotypes. To quantify the relationship among phenotypes and to identify unique phenotype modules, we cross-correlated all 341 phenotypes using Spearman correlations followed by hierarchical clustering (average linkage). The resulting dendrogram was cut at a height of 0.3, thereby defining 121 modules of mean size 2.8 [range: 1 to 10]. Phenotypes belonging to the same module but not to the same (sub-) category were counted separately, yielding 148 distinct phenotypes. The modules are represented by node color, and phenotype categories by node shape (see Figure 2A and Methods). Edges were filtered for top correlation (|s| >= 0.7).

A

log2 Raw Counts Cortex Ctr

log2 Raw Counts Cortex SD

log2 Raw Counts Liver Ctr

log2 Raw Counts Liver SD

B6-2 B6-1 BXD048a BXD043 BXD045 BXD048 BXD065a BXD049 BXD100 BXD065 BXD051 BXD098 BXD081 BXD070 BXD061 BXD085 BXD101 BXD055 BXD066 BXD095 BXD075 BXD071 BXD056 BXD044 BXD090 BXD084 BXD073 BXD050 BXD083 BXD087 BXD073b BXD089 BXD064 BXD079 BXD067 D2-1 D2-2 DB-F1 BD-F1

B

Cortex | Liver

Coefficient of variation

Ctr SD Ctr SD

Appendix Figure S3: RNA-seq raw gene count. (**A**) Distribution of raw gene read counts using HTSeq (see Methods) in cortex and liver samples for both the control (Ctr) and sleep deprived (SD) conditions. Parental strains B6 and D2 are filled with black and brown, respectively. (**B**) Coefficients of variation in the 4 datasets after normalization. Genes in the liver display a slightly higher coefficient of variation than in cortex.
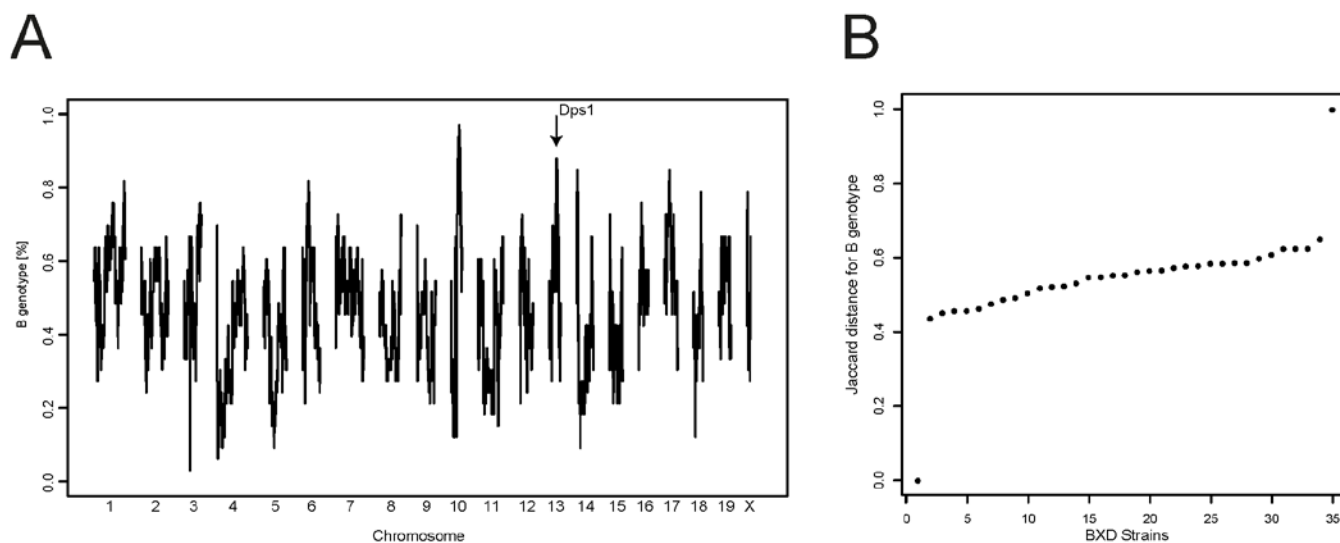
Appendix Figure S4: Allelic distribution in the BXD set. (A) Allelic ratios in the 33 BXD lines at all markers. Several genomic regions display a higher genetic imbalance (either toward the D2- or B6-genotype), among which a region on chromosome 13 containing the QTL Dps1 (MGI:2135996; see main text). Such imbalance decreases statistical power, making it less likely to map QTLs in these regions. (B) To measure the similarity of the BXD set with C57BL6, we used the Jaccard distance metric with our 11K genotypes. We found that a majority of BXD lines have slightly more D2 alleles than B6 alleles.
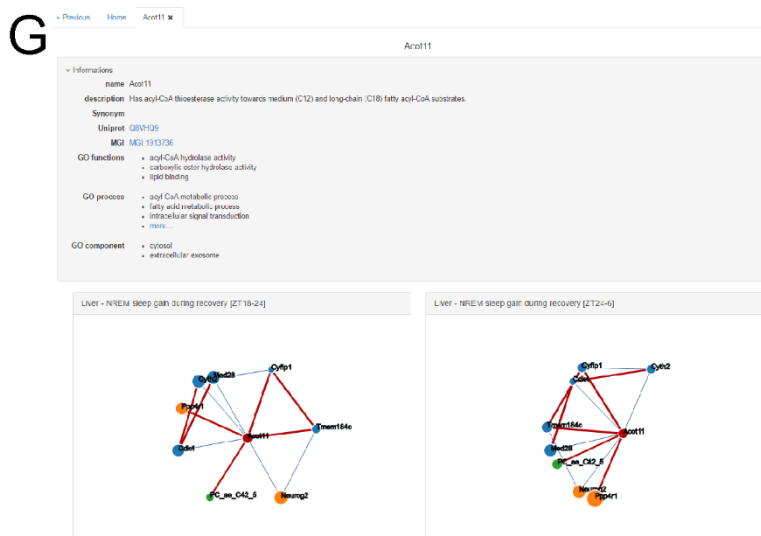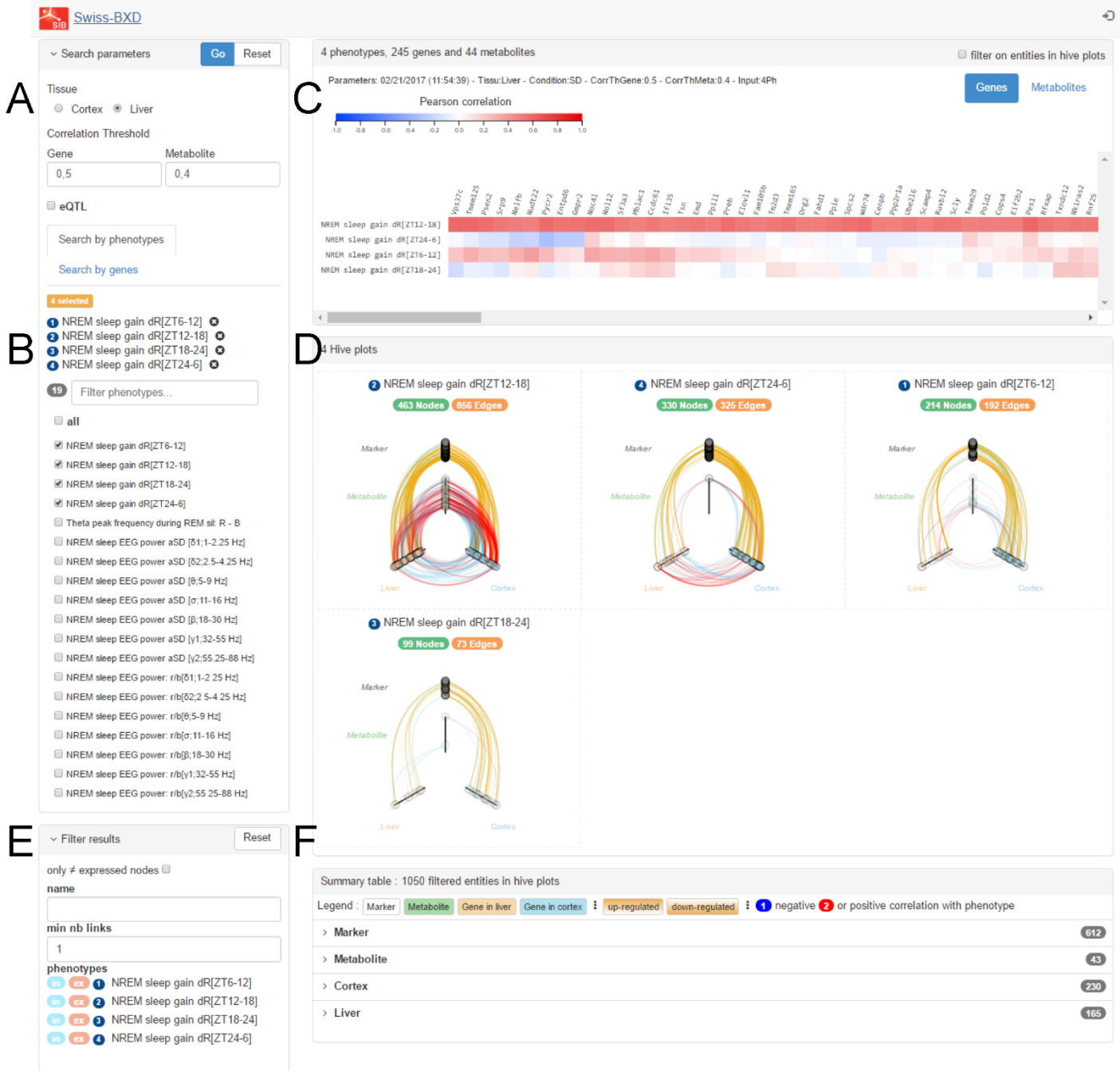
Appendix Figure S5: BXD Web application. All data presented are available in our web application: https://bxd.vital-it.ch. Examples and a tutorial can be found on the website. (**A**) Options to search genes and metabolites in either cortex or liver, with Pearson correlation thresholds selection. (**B**) Search can be initiated by phenotypes or by genes. A search by phenotype(s) will output genes correlated (>= threshold set in A) to the submitted phenotype(s) and vice versa. (**C**) Output is displayed as a heatmap. (**D**) The related hiveplot of each phenotype present in the heatmap is displayed. (**E**) Filtering options specific for the hiveplots. (**F**) Tables containing all genes, markers, and metabolites in the hiveplots, and their relation. (**G**) Gene details: known functions, link to other databases, and strongest relations in the BXD dataset with other genes and metabolites. For details, see the on-line tutorial.

5
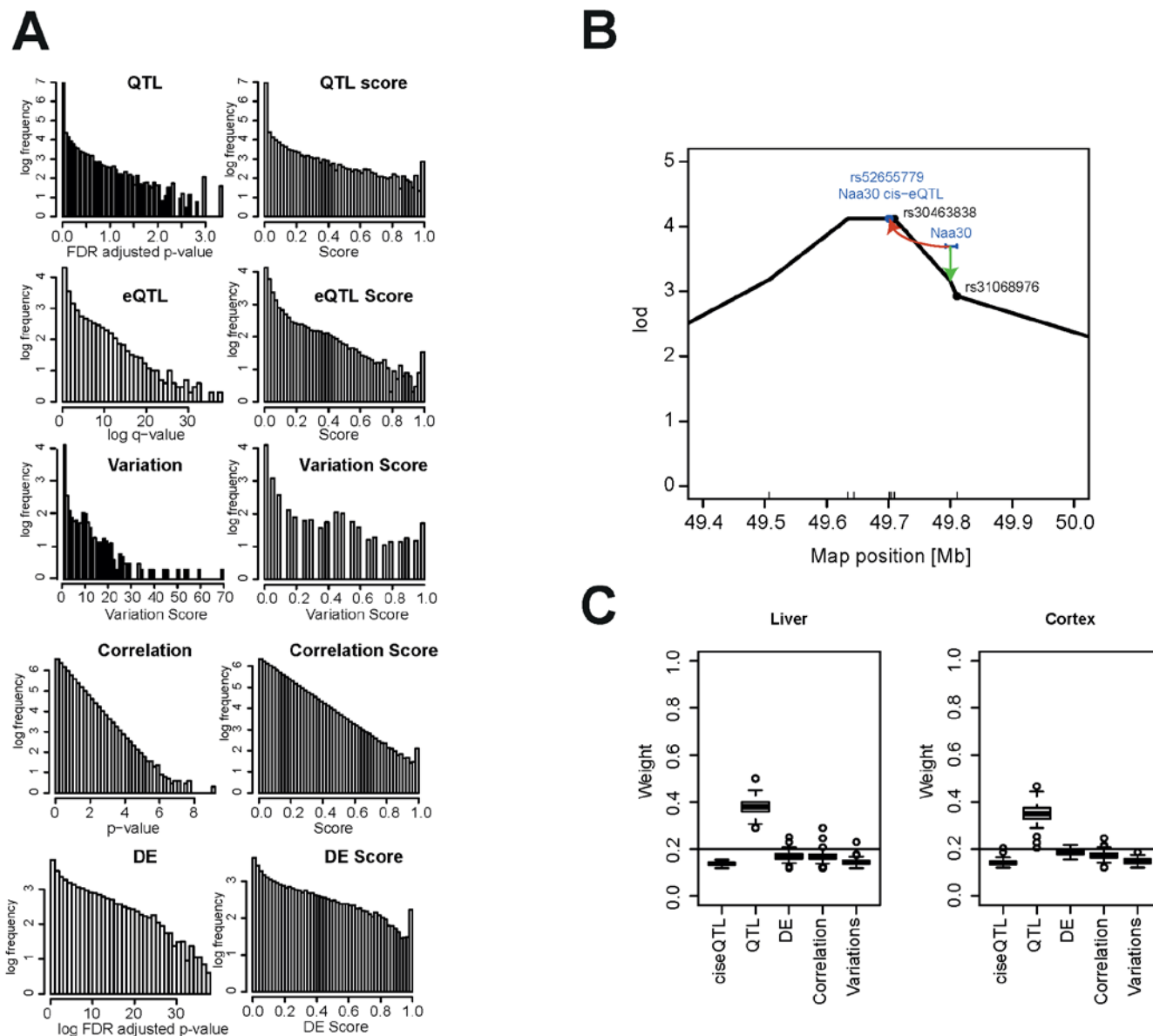
Appendix Figure S6: Gene prioritization strategy. (**A**) Five analysis scores (Right; see Figureure 3, main text, and Methods) are derived from the actual statistics (Left) for i) ph-/mQTL FDR adjusted p-value, ii) eQTL q-value, iii) genetic variant annotation, iv) Pearson correlation p-value, and v) differential expression (DE) FDR adjusted p-value (from Top to Bottom). To compute a single gene variant score, we sum the following values for each gene and for each detected variant: splicing = 10; stop-gain = 10; stop-loss = 10; frameshift indel = 10; nonsynonymous = 10 * polyphen2-probability value. (**B**) We used the central position of the gene to infer the associated ph-/mQTL analysis score at that position. However, in cases where the associated cis-eQTL score or the damaging gene variant score gave a higher value than the ph-/mQTL score, the position of the relevant associated marker was used instead. A case of the former is illustrated with the gene Naa30. This gene is located near a recombinant region with the central gene position (green arrow) located in a low ph-/mQTL associated region, while the cis-eQTL associated marker (red arrow) is located in a highly associated ph-/mQTL region. In the case of Naa30 its associated cis-eQTL score was used. (**C**) Henikoff weighted scores computed for each phenotype after sleep deprivation. The black line at 0.2 represents the line of equality among the 5 scores (summed weight = 1.0). The ph-/mQTL scores generally have higher weights than the other 4 scores because it is the only non-transcript derived score. The other scores are based in part on the RNA-seq data, and the Henikoff lowers their respective weights because of this dependency.
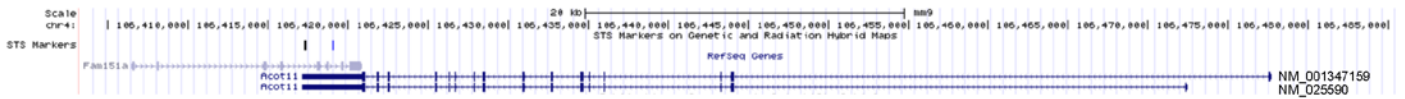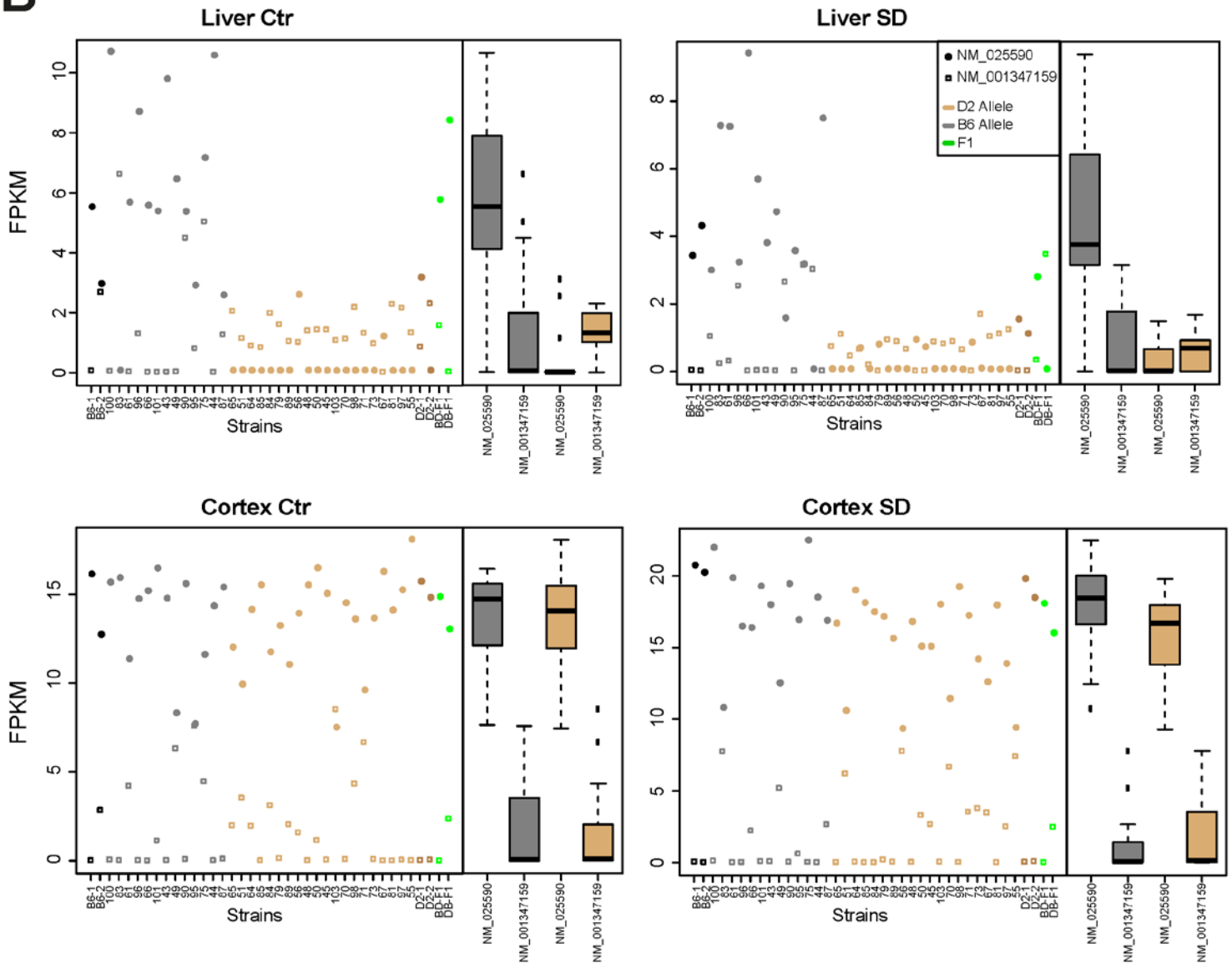
Appendix Figure S7: Acot11 isoforms. (**A**) Structure of the two Acot11 isoforms: NM_001347159 and NM_025590. The two isoforms differ by a single exon at the start of the transcript. (**B**) Estimated expression of the two Acot11 isoforms [FPKM] for cortex and liver samples, under the control (Ctr) and sleep deprivation (SD) conditions. In cortex, isoform NM_025590 is highly expressed compared to NM_001347159, independent of condition and genotype. Note that for 22 out of the 39 lines, NM_001347159 expression is near 0. In liver, NM_025590 is only highly expressed for carriers of the B6 allele for the chromosome 4 associated region for Acot11 expression, while D2 carriers have close to 0 levels. As in cortex, liver expression of the long isoform is low. Expression was estimated using Cufflinks with option -G for the Acot11 refseq file.

Appendix Table S1: Top 100 differentially expressed cortical genes after sleep deprivation. Genes are sorted according to fold-change. Down-regulated genes are highlighted in gray. Of the 78 genes we considered core molecular components of the sleep homeostatic response in the cortex (Mongrain et al, 2010), 13 also made it to this top 100 list (*) and 36 more are among the top 5% most significantly affected genes in the current experiment.

Gene	Log2 FC	Adjusted p-value		Gene	Log2 FC	Adjusted p-value		Gene	Log2 FC	Adjusted p-value		Gene	Log2 FC	Adjusted p-value
Plin4	2.95	4.66E-24		Nr4a1*	1.52	4.78E-28		Zglp1	1.36	3.09E-24		8430408G22Rik	1.22	1.76E-14
Arc*	2.90	1.64E-32		Serinc2	1.49	2.10E-27		Slc23a3	1.35	6.24E-21		Hist2h3c1	1.22	1.66E-25
Egr2	2.88	3.74E-25		Dusp4	1.48	4.29E-29		Stac3	1.35	6.56E-29		Derl3	1.22	1.02E-20
Tekt4	2.23	2.14E-16		Tmem252	1.48	1.72E-16		Cstad	-1.35	2.47E-24		BC021891	-1.22	6.32E-30
Fos*	2.20	3.56E-24		Sik1	1.47	1.29E-27		Hspa1b*	1.34	1.72E-25		A730020M07Rik	1.21	6.27E-16
Plekhg4	2.12	7.81E-34		Npas4*	1.47	2.86E-26		Htr5b	1.33	4.46E-19		Junb	1.20	6.88E-25
Slco1b2	2.08	3.40E-24		Samd3	-1.46	3.29E-22		Hspa5*	1.32	6.58E-38		Arhgef33	-1.20	1.36E-07
Gh	-1.98	4.21E-05		Trib1	1.45	4.99E-34		Fam150b	1.31	2.20E-22		Pcsk9	-1.19	1.64E-24
Tnfrsf25	1.96	3.39E-34		Cbln4*	1.45	1.06E-34		Cyr61	1.29	1.09E-21		Mir6982	1.19	8.40E-17
Misp	1.90	7.87E-19		0610040B10Rik	-1.45	1.11E-20		Map3k6	1.28	1.65E-21		Npbwr1	-1.19	1.30E-19
Rtp1	1.87	2.27E-15		1700001L05Rik	-1.45	4.28E-33		Dusp5	1.28	7.41E-27		Pla2g4e	-1.19	1.52E-18
1700102P08Rik	1.80	9.32E-29		Hspb1	1.44	2.34E-20		Gck	-1.28	6.61E-18		Nr4a3*	1.19	2.42E-22
Fosb	1.75	7.68E-29		Fosl2	1.44	9.67E-29		Itga10	1.27	3.75E-34		Insl5	1.18	2.14E-17
Maff	1.72	2.22E-16		Mybpc3	-1.43	1.04E-11		Alox12	-1.27	9.03E-15		Vwa3a	-1.17	8.23E-32
Tm6sf2	1.68	3.63E-27		Rasl11a	1.43	2.85E-28		Cirbp*	-1.27	6.52E-34		AW549542	-1.17	5.98E-22
Sdf2l1*	1.66	5.32E-37		Dok3	1.43	6.29E-34		Ptgs2	1.27	4.21E-24		Map3k19	1.17	2.95E-27
Bace2	-1.63	3.10E-28		Gkn3	-1.42	3.71E-20		Mdga1	1.26	3.87E-31		Spry4	1.17	2.47E-32
Cdkn1a	1.62	1.29E-24		Parpbp	-1.42	3.56E-20		Cd28	-1.26	8.20E-11		Espnl	1.17	2.36E-15
Klhdc9	-1.60	3.33E-24		Dio2*	1.40	1.47E-33		Neurog2	-1.25	3.64E-19		Tmem82	1.16	1.22E-11
Noxred1	1.59	5.36E-28		Egr1*	1.40	5.36E-28		Hist1h2be	1.25	6.13E-27		Gli1	-1.16	1.55E-22
Rasd1	1.59	4.83E-27		Cml5	1.39	2.16E-22		Fat2	-1.25	1.72E-02		Tmod4	1.16	2.35E-15
Lrrc29	-1.55	4.07E-25		Gpr3	1.37	9.92E-29		Pglyrp1	1.25	3.95E-29		Hfe2	-1.15	1.65E-22
Fam83d	1.55	5.22E-22		Acr	-1.37	3.20E-26		Gm16062	1.24	2.79E-21		4930426L09Rik	-1.15	9.27E-16
Egr3*	1.53	2.59E-34		Gm1141	1.37	8.80E-17		Hif3a	1.24	2.56E-23		Nlrc4	-1.15	1.02E-13
Timp1	1.52	1.14E-08		B430319G15Rik	-1.37	8.64E-19		Arl4d	1.22	7.45E-32		Rho	1.14	6.10E-14

8

Appendix Table S2: Top 100 differentially expressed liver genes after sleep deprivation. Genes are sorted according to fold-change. Down-regulated genes are highlighted in gray.

Gene	Log2 FC	Adjusted p-value	Gene	Log2 FC	Adjusted p-value	Gene	Log2 FC	Adjusted p-value	Gene	Log2 FC	Adjusted p-value
Ttll8	2.94	2.82E-21	Fmo3	1.59	1.24E-09	Hsph1	1.33	1.20E-32	Lefty1	-1.15	7.73E-13
Fam107a	2.62	2.07E-13	Pnpla5	-1.58	2.29E-06	Tchh	1.32	6.72E-12	Smpd3	-1.14	1.13E-11
Fam65b	-2.49	2.95E-22	Slc5a3	1.54	1.20E-30	Fbxl22	1.28	1.73E-13	2810403D21Rik	-1.14	5.64E-13
Lmod2	2.35	7.65E-14	Hist2h3c2	1.54	9.78E-20	Syt11	1.27	4.28E-26	P4ha2	1.14	6.88E-26
Hist1h1e	2.34	3.56E-22	Cirbp	-1.53	8.26E-32	Esrrg	-1.26	1.86E-15	Cand2	-1.12	2.97E-10
Zbtb16	2.22	1.24E-21	1810046K07Rik	-1.53	2.77E-25	Cyp2b10	1.25	5.87E-07	Syn3	-1.12	9.65E-20
Herc3	-2.01	4.90E-29	Lrtm1	-1.52	3.48E-21	Mt2	1.22	2.94E-08	Unc13a	-1.12	2.40E-10
Nfe2	-1.96	4.64E-14	Rps6kl1	-1.49	2.82E-10	Pde4d	-1.21	1.55E-17	AA986860	1.11	4.68E-13
Slc51b	-1.95	7.85E-13	Tnfaip8l3	-1.49	3.28E-15	Plin4	1.21	2.44E-11	Nlrp12	1.11	5.10E-20
1700092C10Rik	-1.92	6.44E-20	Themis	-1.46	5.95E-13	Fam171b	-1.21	7.84E-08	1700030J22Rik	1.11	2.26E-11
Mroh6	1.81	7.37E-14	Hist2h3c1	1.46	1.71E-27	Chrna4	-1.20	4.67E-07	Cd79b	-1.11	4.73E-14
Rad51c	1.81	7.22E-22	Pfkfb3	1.45	5.04E-25	Tmc7	1.20	1.19E-14	Dnah5	-1.10	5.53E-08
Dmbt1	-1.77	4.85E-02	Gm16063	-1.44	5.51E-16	4931408D14Rik	-1.20	4.93E-22	Nr4a2	1.10	6.65E-12
5930430L01Rik	-1.73	1.06E-21	Map3k6	1.44	3.54E-16	1700040L02Rik	-1.19	2.11E-20	Fos	1.10	6.83E-12
D130043K22Rik	-1.72	1.88E-21	Lrrc16a	-1.43	1.02E-28	Snhg11	-1.18	1.51E-14	C330021F23Rik	-1.09	1.46E-18
Hspa1a	1.70	7.72E-12	Ell3	-1.40	1.46E-19	Inhbb	1.18	4.01E-16	Gnat1	-1.09	7.00E-22
Fam83f	-1.70	2.05E-19	Ctgf	1.39	1.31E-16	Dio3os	1.18	7.46E-13	Pde6c	-1.08	8.21E-09
Hspa1b	1.68	9.77E-20	4732491K20Rik	-1.37	4.91E-14	Prss53	1.17	6.03E-23	Cdkn1a	1.08	2.91E-07
Chd3os	1.66	1.21E-22	1810053B23Rik	1.36	3.82E-08	Esco2	1.16	4.27E-15	B930025P03Rik	1.08	3.27E-11
E2f8	1.65	4.35E-19	Ms4a1	-1.36	4.55E-08	Mfsd2a	1.16	1.45E-17	Dlgap1	-1.08	2.13E-16
Pnpla3	-1.65	4.32E-08	Scara5	1.34	2.38E-20	Fam184b	-1.16	1.09E-10	Eif4ebp3	1.08	9.96E-25
Derl3	1.64	5.23E-16	Zfp618	-1.34	6.12E-14	Dnah17	-1.15	8.47E-10	Adamts7	-1.07	9.37E-21
Cd79a	-1.60	3.84E-11	Ppp1r3g	1.34	2.48E-08	Slc1a4	-1.15	2.18E-18	Gramd1c	-1.07	8.88E-26
Tes	1.59	8.16E-21	Fkbp5	1.34	6.55E-21	Gng7	-1.15	5.33E-14	Tbc1d8	1.06	1.20E-23
Nptx1	-1.59	1.29E-12	Hspb1	1.34	2.06E-23	1700056E22Rik	-1.15	2.29E-14	Fam46c	1.06	2.02E-18

Mongrain V, Hernandez SA, Pradervand S, Dorsaz S, Curie T, Hagiwara G, Gip P, Heller HC, Franken P (2010) Separating the contribution of glucocorticoids and wakefulness to the molecular and electrophysiological correlates of sleep homeostasis. *Sleep* **33**: 1147-1157

Chapter 3. A bioinformatics strategy to transform systems genetics data into a digital research object.

3.1 Results summary

In this chapter, we present a more detailed the bioinformatics strategies that have been applied to the systems genetics of sleep project and more particularly the techniques that have been used to enable the repeatability and reusability of the workflow, analyses and data. Much of our analytical pipeline were made of novel tools and methodologies, such as the EEG/EMG annotation approach, the sleep phenotypes nomenclature or the gene prioritization. To improve the reproducibility of these methods by the scientific community as well as the laboratory members, an important part of the analyses were moved to R and implemented within Rmarkdown reports. Segments of the data processing steps and analyses were done using more standard pipelines, but that needed to be handled in the context of a complete workflow to be truly reproductible. Thus, a workflow metadata was generated to further describe the comprehensive data processing and analyses procedures. Finally, the Swiss-BXD web application is presented in detail with tutorials for exploring the data and of our results.

3.2 Contribution

In this research paper, I developed part of the back-end scripts to mine the systems genetics and optimized code to allow more users at the time on the web application. I generated all markdown reports and workflow related metadata. I write the paper and generated all figures.

3.3 Publication

Maxime Jan, Nicolas Guex, Mark Ibberson, Frederic Burdet, Paul Franken, Alan Bridge, Lou Gotz, Martial Sankar, Robin Liechti, Ioannis Xenarios. *A bioinformatics application of the reusability and reproducibility principles for systems genetics* **In Preparation.**

A bioinformatics application of the reusability and reproducibility principles for systems genetics

Maxime Jan, Nicolas Guex, Mark Ibberson, Frederic Burdet, Paul Franken, Alan Bridge, Lou Gotz, Martial Sankar, Robin Liechti, Ioannis Xenarios

Abstract

The principles of "data reusability" and "reproducible research" are currently two core doctrines for better science. Reusability and reproducibility ensure that a dataset as well as the associated methods are exploitable by the scientific community. Many directives were proposed to improve these two aspects such as the *FAIR* guidelines, however their practical applications on large state-of-the-art, multi -omics & -phenome data-sets were not particularly well defined. Here, we describe from a bioinformatic point of view the approaches that were adopted to advocate these principles in the context of a systems genetics design of the sleep homeostasis. We aimed to transform the mouse sleep biological resource into an advanced digital research object favoring open-source and markup language along with a rich data and workflow description using meta-data. Finally, the exposure of this digital-object as well as the tools for its data-mining were provided by latest cloud-technologies with data interactive visualization using data driven document (D3).

Keyword: digital research object, reproducibility, reusability, systems genetics, knitr, data driven document, sleep homeostasis

Introduction

Over the last years, *omics technologies evolved rapidly, with broader detection's range, higher sensitivity and cost reduction which facilitate the generation of BigData. With the advent of web, data-bases and repositories were created to facilitate data-sharing and thus improving the scientific discoveries through data-integration. It becomes now clearer that data alone are hard to work with and have yet to be accompanied with other valuable information such as interpretations, results, workflows, publications and meta-data to produce a complete and operable digital research object (DRO) (Bechhofer, Buchan et al. 2013). In the same way, the reproducibility of the data-processing pipeline within the shared DRO is essential for scientific reliability of the results (Munafò, Nosek et al. 2017), but also to save time for the community, the lab and yourself when a workflow can be reiterated.

However, both objectives for the DRO reusability and reproducibility are nearly impossible to attain with journal publications and thesis documents only, which requires additional time and effort to achieve compliance, e.g. changing working habits to move to a fully reproducible workflow can take up to many years (Lowndes, Best et al. 2017). Currently, many challenges confront the sharing of DRO (Figueiredo 2017) and some guidelines were provided to improve some aspects. Among popular initiatives, the FAIR principles by FORCE 11 (Wilkinson, Dumontier et al. 2016) proposed 4 basic principles to improve human or computer knowledge sharing. The DRO should be Findable, Accessible, Interoperable in order to be Reusable. The Big Data to Knowledge framework (BD2K) is another initiative that provide guidelines for biomedical workflow, meta-data structures and computer infrastructures to facilitate reusability and interoperability of digital resources (Jagodnik, Koplev et al. 2017). For a more complete list of available initiatives, see http://www.researchobject.org. This path is now followed by some funding organizations, like the Swiss National Science Foundation (SNSF) with the "Open Research Data" directives which now requires the integration of complete Data Management Plan (DMP) within funds requests for better data analysis, preservation and reusability. Whereas only a few biomedical journals currently have an explicit data-sharing requirement (Vasilevsky, Minnier et al. 2017) or reproducibility (Iqbal, Wallach et al. 2016) policy.

These guidelines were often described and applied in the context of single assay repository structure (Wilkinson, Verborgh et al. 2017) but their practical application to propose a coherent

and united DRO over trans-disciplinary research project containing the integrations of heterogeneous data-type and novel methodologies was revealed to be challenging (Sansone, Rocca-Serra et al. 2012).

The case of systems genetics

The systems genetics method is described as a population based approach for deciphering the molecular mechanisms underlying the relationship between genetics and complex trait, with the integration of multiple *omics resources (Civelek and Lusis 2014) generally mentioned as intermediate phenotype (e.g. chromatin accessibility, transcriptomics, proteomics, metabolomics, microbiome). Data integration is the keystone of systems genetics and can greatly benefits from the creation of DRO. Many improvements are still needed (Baliga, Björkegren et al. 2017), from a methodological point of view with the development of novel mathematical, statistical, computational (Gligorijević and Pržulj 2015) and visualization (Nekrutenko and Taylor 2012) strategies, but also to facilitate the data discovery and reusability amongst peers.

In our case, the experiment was performed on recombinant inbred mice: the BXDs. At the biological level, this mouse reference panel is an example of suitable organism for systems genetics design as it is genetically diverse and can be used to investigate the interaction between genetics and the environment. Furthermore, the lines are maintained to be genetically long-time stable and reflect the efforts to produce biological resources reusable among labs. Many complex traits were observed in these mice like mitochondria proteomics (Williams, Wu et al. 2016), glucose regulation (Picard, Soyer et al. 2016) or cognitive aging (Neuner, Garfinkel et al. 2016). In our case, to measure the genetics contribution to the sleep homeostatic process, the data were sampled from 33 different BXD lines, (plus the 2 parental strains C57Bl6/2J & DBA/2J and the 2 F1). Furthermore, the mice underwent 6h of Sleep Deprivation (SD) to challenge the sleep homeostat and evaluate the effect of prolonged wakefulness over mice transcriptome, metabolome and phenome in addition to the interaction between SD and genetics. The complete raw dataset was distributed among: (i) 156 *illumina* single reads RNA-seq from cortex and liver tissues, (ii) 250 LC-MS blood metabolomic profiles, (iii) 256 EEG/EMG spectral signals of 4 days recording, for sleep and brain activity measurement, and finally (iv) 256 PIR (Passive InfraRed) motion sensor for physical activity quantification (for

details, refer to our paper). The genes expression from RNA-seq and metabolites level were referred as intermediate phenotypes, and the physical activity, sleep aspects and EEG/EMG spectra were referred as end-phenotypes.

Here we present the general bioinformatic strategy we applied to convert this systems genetics dataset into a better DRO. Our approach aimed at improving: (I) the exposure of the dataset, methods and results, (ii) the reproducibility of our pipeline, (iii) the reusability of the dataset to facilitate knowledge-sharing between scientists with different skills, from bioinformatician to geneticist and sleep scientist with little-to-no bioinformatics expertise.

Summary of the bioinformatic process

To facilitate the interpretation of the complete bioinformatic workflow that was performed for data processing in addition to their different sharing-problematics, the main steps involved as quickly summarized here. The analyses can be separated into 3 principal layers that each increased the level of data abstraction and interpretation. The first *low-level* layer contained the procedures we did to reduce and transform raw-data (i.e. RNA-seq reads alignment, machine learning approach for EEG/EMG annotation, sleep phenotyping or mouse genotyping) which allowed for the data exploitation by further analytical steps. This layer is characterized by long and quite computational intensive procedures that need multiple processor cores, large I.O. bandwidth and where most large intermediate files that were generated were deleted. When available, dedicated standard software were used but the development of novel non-standard methods was also needed.

An *intermediate* layer contained some classical analyses that could be performed on the data such as gene expression normalization followed by differential expression or Quantitative Trait Locus (QTL) mapping. We explored the effect of sleep deprivation, genetic variations as well as their interaction over sleep phenotypes and intermediate phenotypes.

The *high-level* layer contained the novel integrative methods that we developed, to prioritized gene driving the sleep regulation and the meta-dimensional visualization method to represent the multi-omics network underlying sleep phenotypes.

Standard workflow and semantics

The processing of the raw molecular data within the *low-level* layer, could be performed using standard and well-established pipeline (Conesa, Madrigal et al. 2016). For example, the RNA-seq illumina reads alignment was done combining the splice aligner *STAR* (Dobin and C. 2013) and the python read counter *htseq-count* (Anders, Pyl et al. 2015). To maximize the reusability of the alignment with other BXD projects in the department, we used common annotation files, i.e. the reference genome assembly MGSCv37 (mm9) and the RefSeq annotation for gene symbol & boundaries to quantify gene expression. The common genome assembly permitted to increase the genetic variant detection confidence and resolution during the mouse genotyping using the variant caller *GATK* (McKenna, Hanna et al. 2010). In total, 6 RNA-seq samples per BXD lines were integrated, from our cortex (2x) and liver (2x) tissues, in addition of hypothalamus and brain-stem tissues from 2 different projects.

The use of curated symbols for genes nomenclature by RefSeq allows a better semantic interoperability with other resources like the Uniprot protein ID, Mouse Genome Informatics (MGI), or GO terms using solution like R/biomaRt (Durinck, Spellman et al. 2009). Furthermore, it minimizes discordance between expression level among our different projects, thus increasing the reusability and comparison of the data and limiting mice sacrifices in the incoming BXD studies. This single RNA-seq workflow involved many other different scripts and software to format the files and to run the alignment and count algorithms properly and efficiently on a computer cluster. All the computation was performed at the Vital-IT Center for high-performance compute (HPC) of the Swiss Institute of Bioinformatics (https://www.vital-it.ch/). This HPC structure provides a version control of the software used, thus enhancing the pipeline reproducibility. Moreover, an archiving system is available to ensure the sustainability of the data and results generated.

Non-standard workflow and semantics

The use of standardized workflow, format and terminologies are recommended for the generation of DRO (Bechhofer, Buchan et al. 2013) and could be achieved in most case for molecular data processing where lot of pipelines exist. However, our sleep phenotyping procedures could not be carried out using a standard computational workflow or rely on common semantics.

The sleep quantification is usually performed in our lab using an EEG/EMG spectra visual annotation by an expert using 4-seconds window (epoch). For this dataset a supervised classification algorithm had to be developed to annotate the 22.5 million epochs divided among the 262 recorded mice. The specificity of the hardware amongst sleep labs and by consequence the raw output format prevented the release of a packaged workflow for a public usage. However, our algorithms and scripts were versioned with git and transformed into a single user-friendly command on our informatic infrastructure that permit our sleep colleagues to easily reuse this pipeline for other experiment. In total, 341 physiological phenotypes were derived from the EEG/EMG signal and the annotations, representing e.g. sleep duration or architecture. Currently, none existing nomenclature was defined for normal mouse sleep phenotypes. The nomenclature that was chosen for unique phenotype-IDs were a combination of the phenotype observed (e.g. EEG power during NREM sleep) and the feature observed in this phenotype (e.g. in delta bands, 1-4 Hz). Some features could have identical name but the phenotype names were always unique, thus phenotype-IDs were unique. A complete meta-data document was created, categorizing and defining each of the 341 physiological phenotypes that were recorded.

Generate reports with markdown language

Once the data processed within *low-level* layer, the effect of sleep deprivation, genetics and their interaction were measured using different statistical models and computational methods. The programming language *R* was the best suited tools for the statistical analyses and figure generation that were required on these following steps. Beside the advantages of a license-free and portable language, *R* was already recommended as main tool for systems genetics analysis (Durrant, Swertz et al. 2012). Many available packages are 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*), parallel computing (*parallel*) and many others. However, a few analyses were performed using others software, principally for efficiency reasons in eQTL analysis where the number of model to test is quite large (Schupbach, Xenarios et al. 2010, H., Buil et al. 2015).
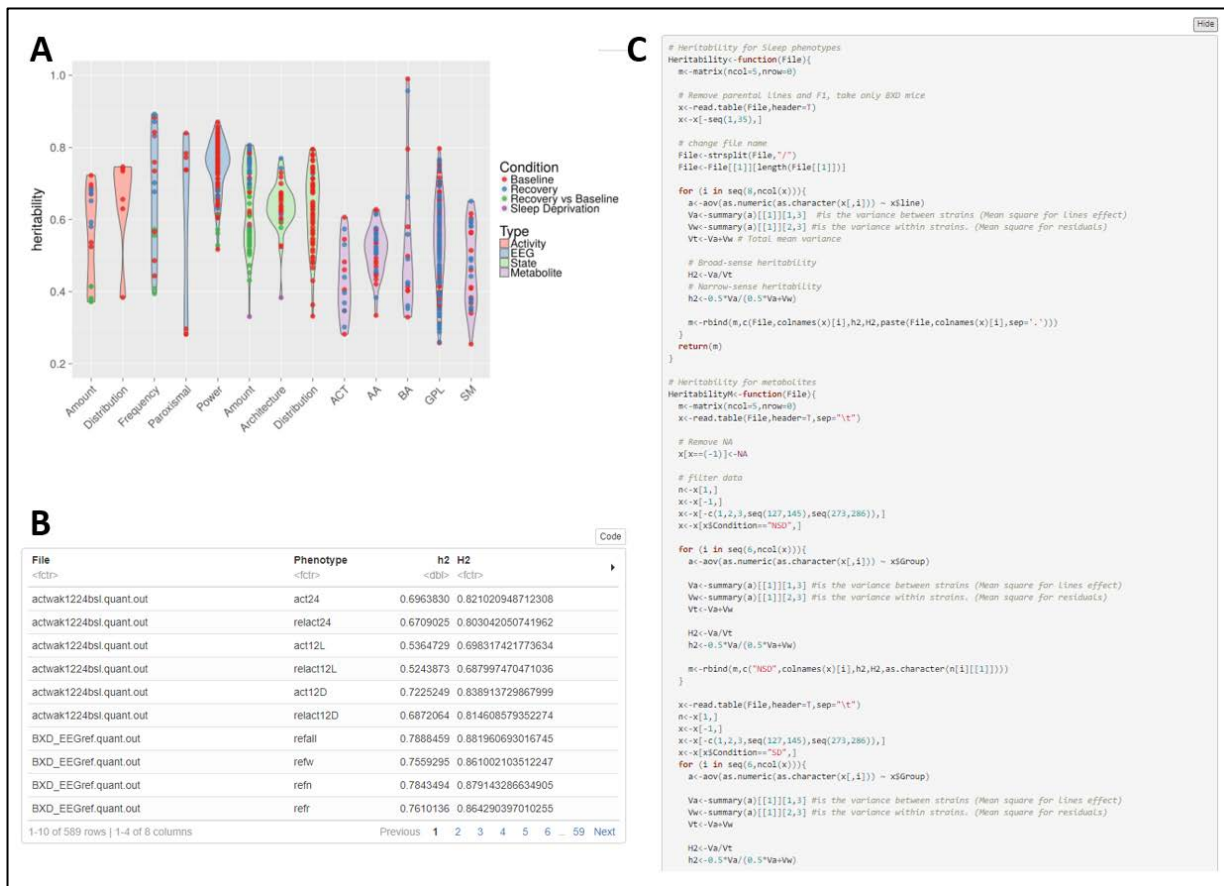
Figure 1: Rmarkdown for traits heritability

(**A**) The figure illustrated in a published paper is displayed within the Rmakrdown. (**B**) Data behind the plot are displayed in a table format, containing the phenotypes file and name, the narrow-sense heritability and the broad-sense heritability. (**C**) the code used to compute the heritability values is display in a Rmarkdown chunk that can be hide or showed.

With a reviewable source code, *R* is one of the flagship of open science and reproducible research. The development of 2 packages: *knitr* (Xie 2014) and *rmarkdown* (Baumer, Cetinkaya-Rundel et al. 2014) enabled the generation of human-friendly report known as '*Rmarkdown*' (Figure 1). This report format contains nice solutions for reproducible results (Sandve, Nekrutenko et al. 2013) like the combination of code (Figure 1A), figures (Figure 1B) and comments contained within a single *markdown* document that can be easily converted into *pdf* or *html* format. The data used for figure generation could also be easily displayed into tables (Figure 1C), where, in this example, the narrow-sense heritability (Hegmann and Possidente 1981) was calculated for physiological and metabolic phenotypes.

The production of *Rmarkdown* reports on a remote cluster infrastructure requires the use of a second script that will generate the document with the *rmarkdown::render()* function and pass the expected function arguments. Common functions shared between *Rmarkdowns* did not need to be copy/paste for each reports, they could be sourced within the code chunks and reported with the *readLines()* function. Finally, the use of *sessionInfo()* function at the end of the document allows to keep track of each packages version and environment variable used to perform an analysis. The *Rmarkdowns* were transformed into *html* document to gain interactivity with the information displayed. 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 report can be visualized using any web-browser. However, on github (https://github.com/) or gitlab, these documents must be downloaded locally or visualized using rawgit (https://rawgit.com/) as the current github's version does not allows the direct visualization of html document.

Digital Research Object & workflow metadata

Systems genetics was an integrative project that implicated multiple collaborators, that each contributed to the final results, with their own working habit related to their field expertise. For a better reproducibility of the generated DRO, a critical goal was to keep track of the different files created, associated documents or analytical steps that were produced. For example, physiological phenotypes could be found within many files and reports, from *low-level* to *high-level* layers, but their nomenclatures were still hard to interpret, in particular for non-sleep scientists (see *File* column in Figure 1B) or persons unrelated to this project. A new comer 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 exists, where to find it or who to ask for it). The notion to establish what was exactly performed, the inputs/outputs and where to locate the information distributed among different person or different directories on a digital infrastructure could be extended to each data, figures, results or scripts. A complete workflow meta-data would improve the reproducibility of the DRO (Cohen-Boulakia, Belhajjame et al. 2017).

A core meta-data file was generated containing the essential information and relationships between all the files, scripts, Rmarkdown, small workflow or database annotation (referred here simply as file-objects) used and produced in this project (Figure 2).
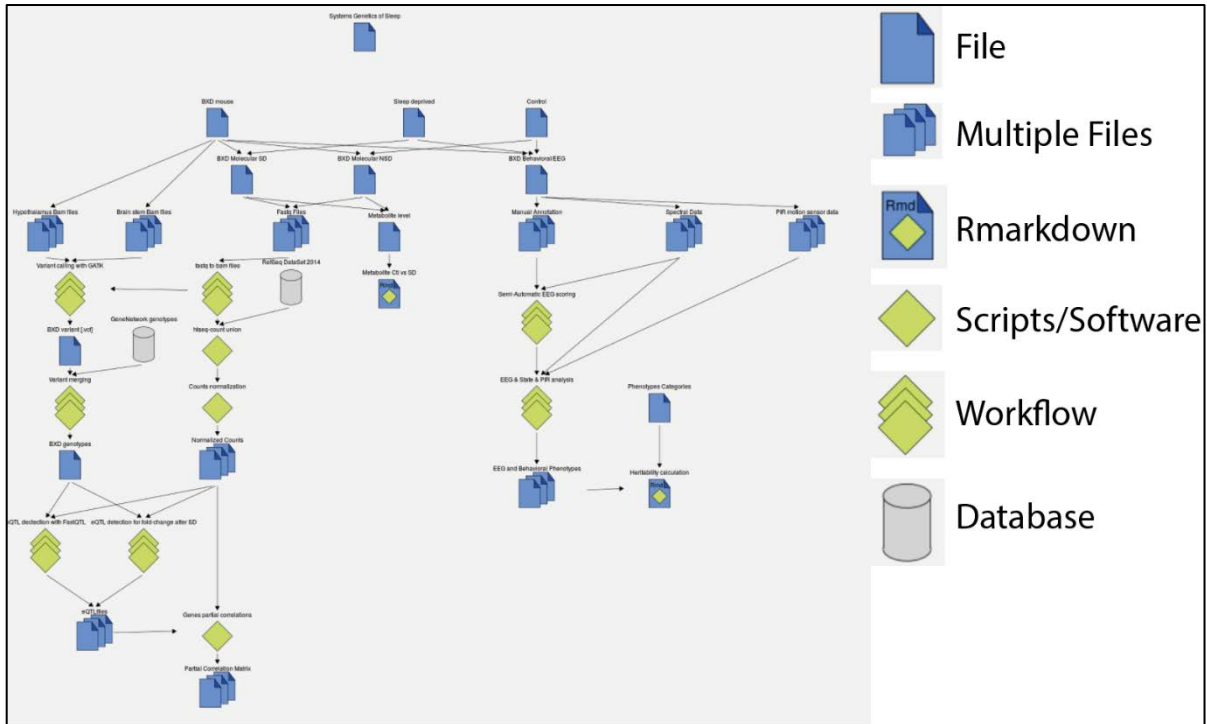

Figure 2. Bioinformatics workflow metadata

File: Data that can be described as a single text file. **Multiple Files**: Data data are splited into different text files (i.e. fastq reads for each BXD lines). **Rmarkdown**: Analysis that are performed in R and reported using Rmarkdown. **Scripts/Software**: A process that is run using a single command lines. **Workflow**: A process that is run using multiple command lines and softwares. **Database**: Public references (i.e. genome reference or BXD genotypes). Each file-object is annotated with its complete name and directory path, input/output, arguments passed, a short description, version and authors.

The markdown format was kept as it was easy to write/read by a human or to generate via simple python script (so collaborators could easily understand it and integrate their own pipeline, comments, description and cross-reference). The metadata file was formatted into a simplified RDF-like triples structure, were each files-objects (subject) were linked to information (objects) 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 (figure 2, only input/output) to display the important files and pipeline used. This meta-data is useful for an external user to understand how exactly was generated the final DRO. But also for a collaborator to recover the scripts and input files that were used even after prolonged periods and to use them again, which permit 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 metadata file is accessible here.

Accessing data through web application & data-mining

The DRO that was built for the systems genetics of sleep is constituted of the following collection: raw-data, processed data, Rmarkdown reports, results & interpretation (via publication), workflow, scripts and metadata. A web application was developed to further apply the FAIR principles, which promoted the exposure of our results, interpretations and associated method through findable and accessible web pages. Furthermore, a set of tools were developed for the exploratory data-mining of this dataset and reusability improvement of our results.

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 (Figure 3D). This information is formatted with hiveplots (Figure 3G). The following tools allow to change the filtering parameters, explore and compare the hiveplots, explore the EEG/behavioral phenotypes and finally to search for specific genes:

(Figure 3A) Search parameters - Tissue & cutoff: Sets the parameters for selection of genes in the specified tissue (cortex or liver) according to the specified cutoff for the pearson correlation coefficient and/ or a *cis-e*QTL q-value.

(Figure 3B) Search parameters - Phenotypes or genes: One can search by phenotype(s) or by gene(s). A search by phenotype will output the genes that correlate to the submitted phenotype(s) with a correlation coefficient larger than the cutoff set in A). Search by gene(s) will output the most correlated phenotype(s) to the submitted gene(s).

(Figure 3C) Results - Heatmap: The results are displayed as heatmaps according to the selections made under Search parameters in **A** and **B**.

(Figure 3D) Results - Hiveplots: For each phenotype present in the heatmap in **C**, the corresponding hiveplot is displayed.

(Figure 3E) Hiveplot-Filtering: Filtering options specific for the hive plots.

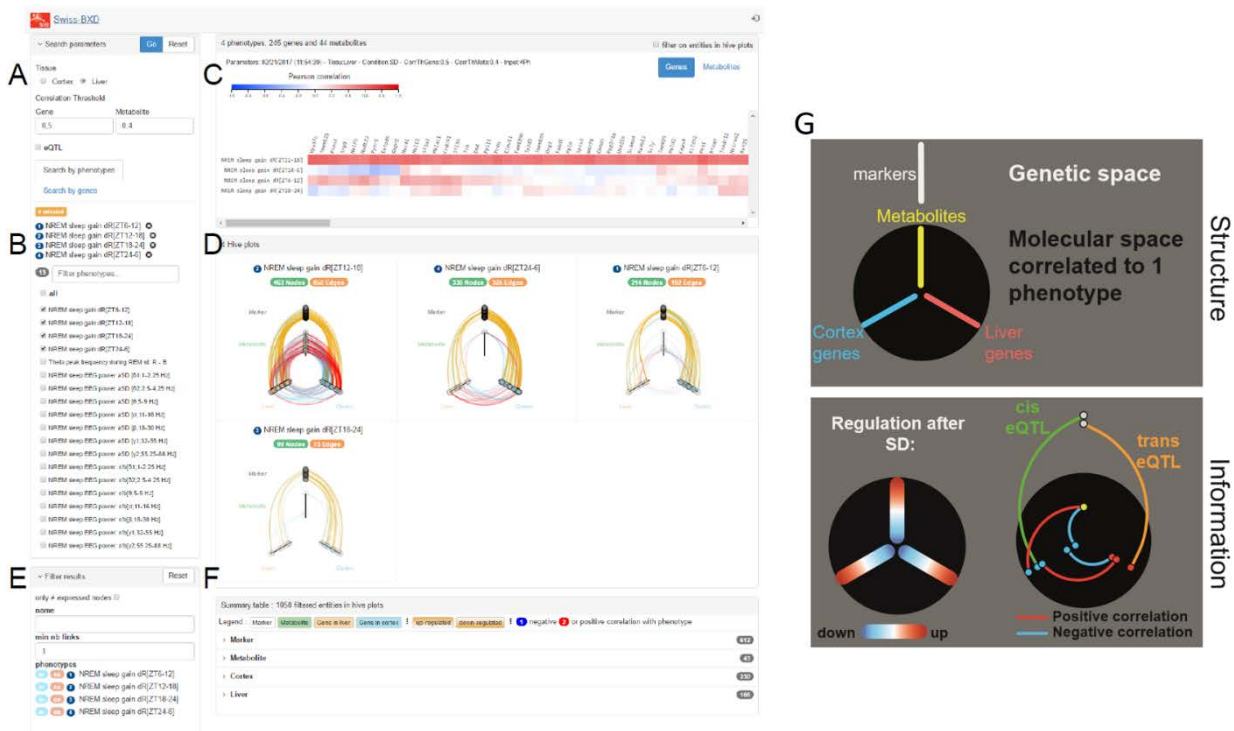**(Figure 3F) Results – Table:** Tables list all genes, markers, and metabolites and their relations for each hive-plot in **D**.

Figure 3: Web interface home page

(**A**) Search parameters - Tissue & *cutoff,* (**B**) Search parameters - Phenotypes or genes, (**C**) Results – Heatmap, (**D**) Results – Hiveplots, (**E**) Hiveplot-Filtering, (**F**) Results – Table, (**G**) hiveplot structure: Each plot represents one EEG/behavioral phenotype and its underlying associated molecular network; i.e., only the genes and metabolites strongly associated with a given phenotype are displayed. Each hiveplot is composed of 3 radial axes containing the molecular data with nodes assigned to the 2 bottom axis for genes expressed in the cortex (in blue) and liver (in red) and nodes on the vertical axis (in yellow) represent metabolites. On top, we added a separate 'genetic' axis (white) containing the genotypes. The node position on the 3 (molecular) radial axes was determined by the response to sleep deprivation; i.e., molecules closer to the center were down-regulated more strongly, while more upregulated genes/metabolites were closer to the axis' edge. Edges connecting nodes represent positive/negative correlations (red/blue, respectively) between molecules using expression values. Genetic markers linked to genes by *e*QTLs connect the genetic and molecular space.

Three data-mining tutorials are described here and in more details on the web interface to: (i) mine a single phenotype, (ii) search for gene and (iii) compare hiveplots.

Single phenotype mining:

In this section, we assume that we are interested in a single sleep phenotype. We will approach the data by multiple filtering steps. Compared to the 3 other sleep intervals, NREM gain was strongly associated with metabolites during the 2nd 6h interval; i.e., the 1st 6h of the recovery dark period (ZT12-18, Figure 3D). We will focus on this phenotype only, other phenotypes can be removed using filtering options in Fig.3B. The heatmap (Figure 3C) displays all genes for which the expression is correlated to this sleep phenotype (note that liver is the default tissue). You can explore genes in cortex by changing the filtering parameters (Figure 3A). You can sort genes by correlation strength by clicking on the phenotype name in the heatmap. The default thresholds are: absolute pearson r equal or above 0.5 for genes and absolute pearson r equal or above 0.4 for metabolites. Given the high number of genes displayed in the heatmap, the pearson r coefficient cutoff can be increased under Search parameters (Figure 3A). In this case, we select an absolute pearson r of 0.6 for genes and 0.5 for metabolites. The user might be interested in genes that are driven by genetic variation (cis-eQTL). Genes are then also filtered according to cis-eQTL association with an FDR adjusted p-value equal or below 1e-5. Five genes remain in cortex and liver: *Acot11*, *Tceanc2*, *Aldh9a1*, *Mroh7*, and *Btf3l4* (highlighted in Figure 4A). Further filtering can be applied to the hiveplot. We can select for genes that have a minimum number of links (edges), e.g. 6 edges on our unfiltered hiveplot (Figure 3E). Acot11 is the only highlighted gene that is kept after applying these filtering options (Figure 4B). Other options are available, such as keeping only top differentially expressed genes in the selection: The hiveplots can be investigated using 'mouseover' the different edges and nodes. This will display the node name (gene name, metabolite name, or marker name), or the nodes connected to an edge. All nodes and edges within the displayed hiveplot are listed in the table (Figure 3F). The table gives the following information: i) the type of node: genes in cortex, genes in liver, metabolites, and markers, ii) effects of sleep deprivation on genes and metabolites are indicated by fading color. A new tab is created for each highlighted gene selection and shows additional information concerning gene function, i.e. the 5 genes with the highest partial-correlation (Figure 4D), within the same tissue, across-tissue, and metabolite. The tab (Figure 4C) contains the gene name (1), the Uniprot description of the gene (2),

Synonyms (3), a Uniprot and MGI access ID links (4) and related GO terms (5). Here the network of the highlighted gene (*Acot11*) with the 5 top genes partially-correlated to within the same tissue (blue), 2 top genes partially-correlated to the highlighted gene within the other tissue (orange) and the top correlated metabolite (green). Red edges indicate positive correlation and blue edges indicates negative correlation. The size of the nodes indicates the correlation coefficient with the phenotype.
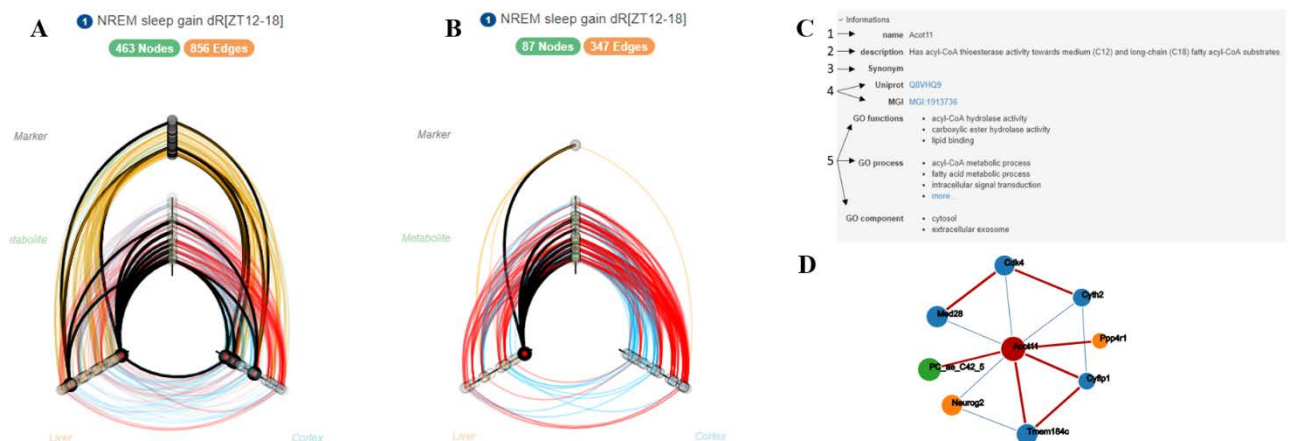

Figure 4: Single phenotype filtering results

(A) Highlight selected genes (in black). **(B)** Hiveplot filtering, **(C)** Selected gene additional information, **(D)** close neighbor of the selected gene.

Gene search:

The interface allows also to search for specific genes associated with a sleep phenotype at a specific correlation cutoff (Figure 3E). Here are a 2 examples of sleep related genes:

1. *Per2* is a well known circadian gene. It is associated in the BXD set with NREM sleep gain after SD [ZT12-18], and is associated with 2 metabolites: PC aa C36:3 & PC ae C40:0.

2. *Sik3* KO mice were associated with the increase total of NREM sleep (minutes/24h) and a decrease in total wake in mouse by (Funato, Miyoshi et al. 2016) using random genetic screening. With our extended database, we found that: Sik3 cortical expression in baseline was correlated (r=0.55) with the wake amount at the beginning of light (BXDw24bsl.h1).

Compare hiveplots with edges filtering:

Hiveplot filtering parameters allow to recover common edges among hiveplots or to exclude common edges. In the first example, we select phenotypes of EEG power after SD for all the bands using search option in Figure 3B. Here we are interested in slow delta power after sleep deprivation (δ1: 1.0-2.25Hz) during NREM sleep, and we want to remove all genes, metabolites that are present also within other bands. Therefore, we exclude other phenotypes using the "ex" button in the hiveplot filtering options (Figure 5A).
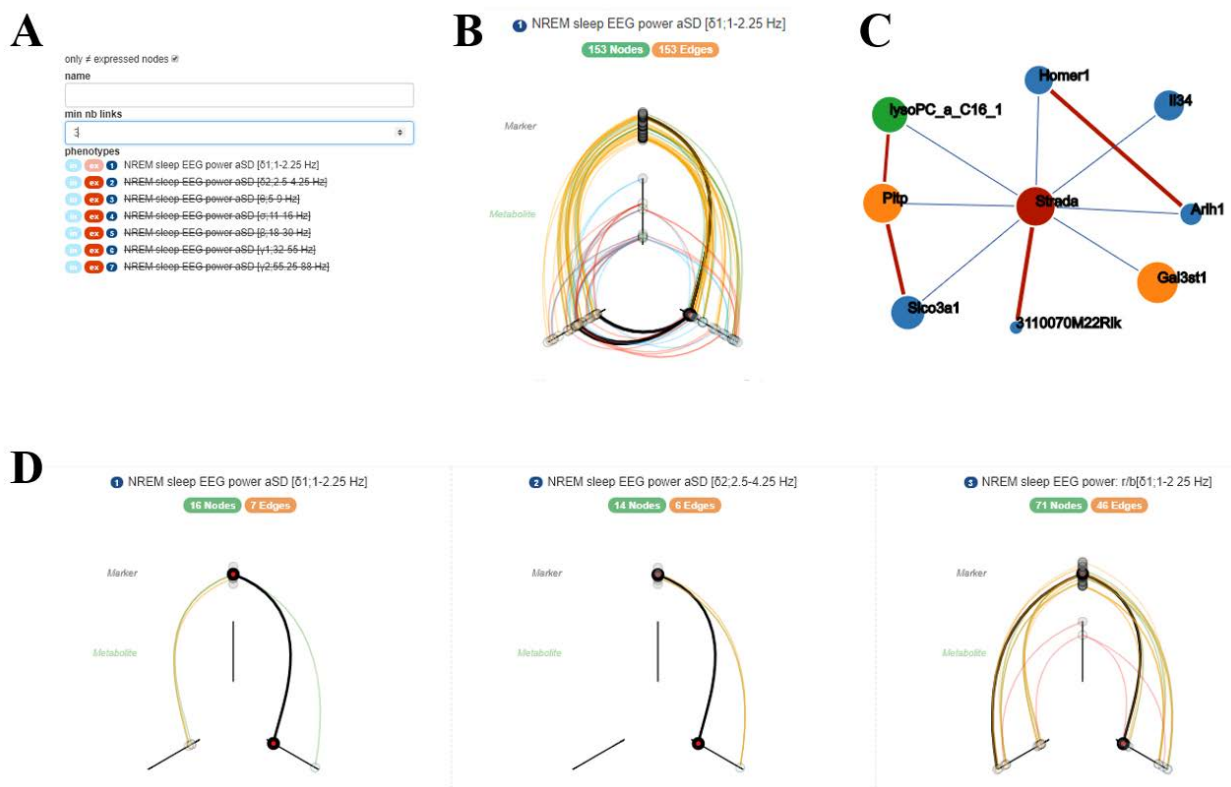
Figure 5: Hiveplot comparison

(A) Options to select common 'in' edges and nodes, or to exclude 'ex' common edges and nodes **(B)** selected gene after excluding common edge for EEG slow delta power after SD **(C)** Neighbor of *Strada*, **(D)** Filter only common edges for delta power bands gain and after sleep deprivation. *Wrn* is the highlighted common gene to 3 phenotypes

Finally, we keep only genes differentially expressed and genes connected to 3 other elements, the remaining genes are *Strada*, *Mapk4* and *Fhit* in cortex and *Susd4* in liver. These genes thus correlate only to slow delta and not to other EEG frequency bands of the EEG spectra during NREM sleep after sleep deprivation. Inspection of Strada (Figure 5B) in cortex informed

us that *Homer1* is among the 5 closest genes of Strada (Figure 5C). *Homer1* has been implicated in the homeostatic regulation of EEG delta power. The second example shows how to inspect common edges: In this case we are interested in genes common to slow delta gain (r/b) during NREM sleep correlated within other delta bands. Using the "in" button in the hiveplot filtering option for slow delta gain after SD, the common edges and nodes within other bands will be kept. We also filter for genes differentially expressed after SD. Inspecting the table (F), we see that only 1 gene is connected to 3 other bands. *Wrn* is present in the hiveplot 1, 2, and 3 [NREM sleep EEG power aSD δ1, NREM sleep power aSD δ2 and NREM sleep EEG power gain (r/b) δ1] (Figure 5D).

Discussion:

We presented here a data-management strategy for the application of the reproducibility and reusability principles in a context of a systems genetics dataset. The reproducibility of our workflow to create a complete digital research object from a mouse biological resource was supported with the use of dynamic reports using Rmarkdown when it was feasible, but also version control with git and rich metadata description for the total data-processing pipeline. With the development of web-interface, the systems genetics digital research object is findable and accessible with web-browser. Moreover, the dataset is reusable with data-mining tools available on the web-interface. This bioinformatic strategy stands as a real case application example for trans-disciplinary projects, but which is likely to change over the next years with novel annotation version, standard workflow and novel integrative methods.

References

Anders, S., et al. (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." <u>Bioinformatics</u> **31**(2): 166-169.

Baliga, N. S., et al. (2017). "The State of Systems Genetics in 2017." <u>Cell Systems</u>.

Baumer, B., et al. (2014). "R Markdown: Integrating a reproducible analysis tool into introductory statistics." <u>arXiv preprint arXiv: …</u>.

Bechhofer, S., et al. (2013). "Why linked data is not enough for scientists." <u>Future Generation …</u>.

Civelek, M. and A. J. Lusis (2014). "Systems genetics approaches to understand complex traits." <u>Nature reviews. Genetics</u> **15**(1): 34-48.

Cohen-Boulakia, S., et al. (2017). "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities." <u>Future Generation Computer Systems</u> **75**: 284-298.

Conesa, A., et al. (2016). "A survey of best practices for RNA-seq data analysis." <u>Genome biology</u> **17**: 13.

Dobin, A. and D. C. (2013). "STAR: ultrafast universal RNA-seq aligner."

Durinck, S., et al. (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." <u>Nature protocols</u> **4**(8): 1184-1191.

Durrant, C., et al. (2012). "Bioinformatics tools and database resources for systems genetics analysis in mice--a short review and an evaluation of future needs." <u>Briefings in bioinformatics</u> **13**(2): 135-142.

Figueiredo, A. S. (2017). "Data Sharing: Convert Challenges into Opportunities." <u>Frontiers in public health</u> **5**: 327.

Funato, H., et al. (2016). "Forward-genetics analysis of sleep in randomly mutagenized mice." <u>Nature</u> **539**(7629): 378-383.

Gligorijević, V. and N. Pržulj (2015). "Methods for biological data integration: perspectives and challenges." <u>Journal of the Royal Society, Interface</u> **12**(112).

H., O., et al. (2015). "Fast and efficient QTL mapper for thousands of molecular phenotypes."

Hegmann, J. P. and B. Possidente (1981). "Estimating genetic correlations from inbred strains." <u>Behavior genetics</u> **11**(2): 103-114.

Iqbal, S. A., et al. (2016). "Reproducible Research Practices and Transparency across the Biomedical Literature." <u>PLoS biology</u> **14**(1).

Jagodnik, K. M., et al. (2017). "Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop." <u>Journal of biomedical informatics</u> **71**: 49-57.

Lowndes, J. S. S. S. S., et al. (2017). "Our path to better science in less time using open data science tools." <u>Nature ecology & evolution</u> **1**(6): 160.

McKenna, A., et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." <u>Genome Res</u> **20**(9): 1297-1303.

Munafò, M. R., et al. (2017). "A manifesto for reproducible science." <u>Nature Human Behaviour</u> **1**(1).

Nekrutenko, A. and J. Taylor (2012). "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility." <u>Nature Reviews Genetics</u>.

Neuner, S. M., et al. (2016). "Systems genetics identifies Hp1bp3 as a novel modulator of cognitive aging." <u>Neurobiology of aging</u> **46**: 58-67.

Picard, A., et al. (2016). "A Genetic Screen Identifies Hypothalamic Fgf15 as a Regulator of Glucagon Secretion." <u>Cell reports</u> **17**(7): 1795-1806.

Sandve, G. K., et al. (2013). "Ten simple rules for reproducible computational research." <u>PLoS Comput Biol</u>.

Sansone, S.-A. A., et al. (2012). "Toward interoperable bioscience data." <u>Nature genetics</u> **44**(2): 121-126.

Schupbach, T., et al. (2010). "FastEpistasis: a high performance computing solution for quantitative trait epistasis." <u>Bioinformatics</u> **26**(11): 1468-1469.

Vasilevsky, N. A., et al. (2017). "Reproducible and reusable research: are journal data sharing policies meeting the mark?" PeerJ **5**.

Wilkinson, M. D., et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data **3**: 160018.

Wilkinson, M. D., et al. (2017). "Interoperability and FAIRness through a novel combination of Web technologies." PeerJ Computer Science **3**.

Williams, E. G., et al. (2016). "Systems proteomics of liver mitochondria function." Science (New York, N.Y.) **352**(6291).

Xie, Y. (2014). "knitr: a comprehensive tool for reproducible research in R." Implement Reprod Res **1**: 20.

Discussion & Perspectives

The systems genetics approach is still a young scientific strategy with great potential for dissecting the complexity of the molecular machinery underlying phenotypic variation. Currently, no accepted analytical directives have been adopted to guide the researcher through this *Terra Incognita*, and no gold-standard dataset exists to accurately compare methods beyond the single gene regulatory network inference (Pinna, Soranzo et al. 2011). Accordingly, many experimental protocols and methodological tools have yet to be developed, considered, tested and evaluated (Baliga, Björkegren et al. 2017). In the following section I will first discuss our strategy to systems genetics. In a second part, I will discuss our newly developed analytical, integrative, and communicative schemes with their strengths and limitations compared to existing methods, as well as the biological resource we used, i.e. the BXD panel. Finally, I will highlight some of our findings., linking metabolism to sleep regulation and the perspectives for novel regulatory network research.

Systems genetics, a multifaceted approach

The systems genetics approach stands at the intersection of many disciplines, from the experimental/biological to data/computational science. In our study, we aimed at gaining insight into the molecular factors driving sleep. The extraction and transformation of the relevant biological information into the findable, understandable and exploitable knowledge resource that we aimed for involved multidisciplinary expertise for each of the steps of the project, with its own challenges (Figure 6).

First, the experimental design aimed at capturing factors driving sleep within a multi-dimensional molecular network, nested within different tissues and modulated by environmental, genetic, and time factors. Therefore, a data-driven type of study was selected necessitating a large sample number, because of the dimensional aspects mentioned earlier and to obtain the statistical power needed to attain reproducibility of the results (Halsey, Curran-Everett et al. 2015). Among the challenges faced were the ability to scale the experiment over 250 mice divide among 33 different genotypic backgrounds. To limit potential technical and experimental bias, and guarantee an optimal data quality over 2 years of data acquisition, a
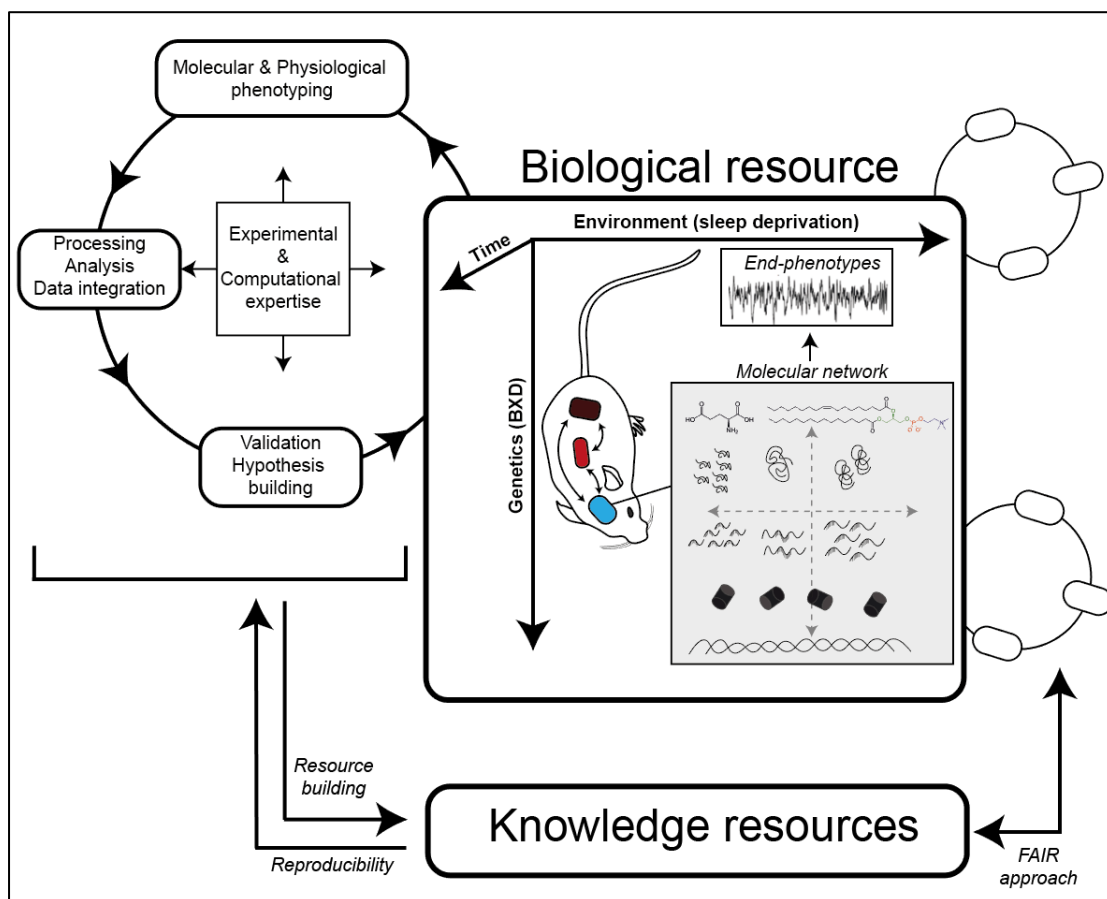
Figure 6 : Schematic design to transform biological resource into knowledge resource with systems genetics.

Schematic representation of our systems genetics approach to generate hypothesis, using experimental and computational expertise (big circle). A knowledge resource was built to access and data mine our results, which allows other research studies to reuse these data (small circles). And the implementation of a reproducible strategy allows to replicate the workflow that was developed.

randomization strategy was applied between the experimental cohorts for the sleep deprivation, tissues sampling and EEG implementation/recording. Furthermore, experiment on parental strains were replicated at the beginning and end of the experiment. On the computational side, the capacity to scale the EEG/EMG annotation for behavioral phenotyping, while keeping a high accuracy compared to manual annotation was covered by the development of a supervised machine learning algorithm. The use of multiple SVMs with a training process using partially annotated signal by our expert and the high resolution of the annotation (1 second epoch) ensured that this algorithm was generalizable for the high EEG/EMG signal diversity that was found the BXD panel.

Among the existing systems genetics publications, the focus was mostly on specific mechanical aspect of cellular regulation, thus, often limiting the dimensionality of the resulting datasets. While the genetic diversity component is a constant in each systems genetics design, some adopted single tissue exploration (Andreux, Williams et al. 2012, Taneera, Lang et al. 2012), single condition (Krishna, Biryukov et al. 2014) or single physiological phenotype (Chintalapudi, Maria et al. 2017). At this moment, this dataset is unique, at least in the mouse sleep field given its multi-dimensionality over genetics, tissues, conditions, time, and large physiological phenotypes within a single project. The challenge of data integration on so many levels necessitated the development of novel integrative strategies besides the more traditional data analysis. Our strategy for data-integration and suggestion for improvement are discussed below. In addition, an effort was made to guarantee the reproducibility of the methods used with the implementation of open and dynamic reports containing the programming code used. This ensured that the workflow at the origin of our knowledge resource was reviewable.

Among the intriguing systems genetics relationships detected for the phenotypic response to sleep deprivation, we identified effects on brain slow-waves dynamics by cortical glutamate receptor recycling and DNA-helicase enzyme, specifically after sleep loss. But also, an association between NREM sleep recovery dynamics after sleep loss in the dark phase, with lipids metabolism through the action of acyl-thioesterase enzyme *Acot11* in the liver. A validation procedure using complementary or orthogonal experimental designs is recommended to strengthen the overall quality of the dataset given the many possible and different steps taken in processing the data. But also, to evaluate the robustness of the conclusion using novel methods, as mentioned earlier. Accordingly, we tested the hypothesis of *Acot11* to be part of the sleep homeostatic regulatory network using knock-out mice, and could confirm the differential NREM sleep recovery dynamics of the KO mice in the dark phase following sleep deprivation. This validation step confirmed that the relation we discovered was likely to be causative and opens the door for other hypotheses testing like the cross-relationship between the homeostatic regulation of sleep alongside with fatty-acid turnover and possible metabolic disorder like obesity or type-2 diabetes.

The systems genetics phenotypes that we highlighted in Chapter 2 are only the tip of an iceberg given the scale of the dataset and the numerous other examples that still await further exploration. Tools were developed to allows the exploration of all end-phenotypes and the data-

mining of the underlying molecular networks, therefore transforming raw and processed data into a knowledge resource that can be interrogated. Using cloud-based technologies, a set of web-based tools were developed to enable a user to search for a specific gene or sleep phenotypes and perform interactive filtering on the data-base and on our novel visualization methods. The resulting database and associated web interface ensures the reusability of the data and associated methodologies and facilitates the distribution of this unique integrated system-level genetics knowledge base to the public. This strategy of resource building which respects the principles of the FAIR approach enables researchers working on the same topic or using the same mouse lines to reuse this dataset for their own research.

Phenomics integration, methods and limitations

A crucial step in the systems genetics general strategy is the integration of large, heterogenous information with mathematical and computational frameworks to strengthen *data-exploitation*. It is one of the greatest challenges for multi-omic datasets where increasing numbers of data-types will also highly increase analyses effort (Palsson and Zengler 2010). The addition of supplementary and broader predictive variables within a model is expected to yield a more comprehensive insight into the biological processes at work. Two general categories of integrative approaches are regularly followed in multi-omics design: multi-staged and meta-dimensional methods (Ritchie, Holzinger et al. 2015). In <u>multi-staged</u> methods the analytical workflow is divided into multiple and independent analysis that are later hierarchically integrated. Our approach of data integration in this project can be viewed as following this multi-staged approach with: QTL mapping, statistical analysis of sleep deprivation consequences or hierarchical integration of analysis outcome within a prioritization strategy. <u>Meta-dimensional</u> methods are larger, more complex, multivariate models that can stand on machine learning algorithms (based on supervised or unsupervised methods, i.e. that can use or not prior knowledge as learning step) or network-based approaches (Li, Wu et al. 2016, Huang, Chaudhary et al. 2017). In our project, the visualization method we implemented can be viewed as a meta-dimensional method, integrating all genetic, transcriptomic, metabolomic and end-phenomic layers within a single figure. Both methods are further described in the two following sections.

Multi-staged approaches, improvement of our methods:

One of the main advantages of a multi-staged approach is an easier integration of the state-of-the-art methods for data-processing or data-analysis committed to simpler (i.e. with less dimensions) problems. The use of outdated software, annotation or methods strongly affect reproducibility of the results (Offord 2018) and is therefore to be avoided as much as possible as it generates unwanted technical noise and impedes the assembly of a serviceable resource. Each processing step we performed continues to be developed further with periodical updates, reviews and evaluations like RNA-seq (Williams, Baccarella et al. 2017), and some parts of our analyses pipeline can be improved, like reads alignments, association mapping, trans-eQTL or epistatic detection.

Reads alignment and transcript expression

For the BXD panel, a common pipeline for RNA-seq analysis is to align reads on the mouse reference assembly MGSCv37 (mm9) or more recently GRCm38.p6 (mm10) (Fan, Waizenegger et al. 2017). The mouse strain used for the assembly is C57BL/6J, which may introduce a reference bias during the alignment and change the read mapping success rate for some lines, considering the C57BL/6J and DBA/2J genome admixture of the BXD. This bias is often observed for region with indels, were splice aligners have most alignment difficulty, even when using the STAR methods (Sun, Bhagwate et al. 2017). The use of an individualized genome reference for each BXD line would improve gene expression analysis (Munger, Raghupathy et al. 2014). The haplotype sequence reconstruction and genome assembly of the BXD panel could be performed using PacBio reads with the use of phasing algorithm like FALCON (Chin, Peluso et al. 2016) or a reference guided assembly like CrossStitch (https://github.com/schatzlab/crossstitch).

In this project we quantified gene expression, which reduced the quantity of parameters to evaluate within integrative methods and facilitated the validation procedure by selecting mice with gene KO compared to condition allele KO. However, recent transcript quantification methods like Kallisto (Bray, Pimentel et al. 2016) show good accuracy. It could be used in combination with gene & transcript aggregating approach (Van den Berge, Soneson et al. 2017,

Yi, Pimentel et al. 2017) which would increase the accuracy of eQTL analysis or sleep deprivation effect on transcriptome.

Association mapping:

For QTL mapping, we used a linear model by associating the gene expression, metabolite levels or EEG/behavioral phenotypes with genetic markers. Other statistical models were developed and recently used for the BXD panel to take into account the stratification and genetic covariance that can be found in this panel (Kang, Zaitlen et al. 2008, Segura, Vilhjálmsson et al. 2012, Li, Wang et al. 2017). However, in such mixed models the estimation of the degree of freedom and therefore the significance of fixed effect using p-values can be difficult, and the methods used tend to be non-conservative (high type I error) when sample size is small (Luke 2017). In our project, this type of approach could be used for end-phenotypes and metabolites where replicates are used but would be inappropriate for genes expression that were evaluated using pooled samples.

Trans-eQTL:

Cis-eQTL associations and interactions with sleep deprivation were described in details in our resource and used for gene prioritization within associated ph-QTLs. However trans-eQTL remained mostly descriptive and primarily used for visualization purposes. Still, trans-eQTL are expected to importantly contribute to the gene expression heritability (Grundberg, Small et al. 2012) and could be integrated within our prioritization method to evaluate the effects of transcriptome-wide variation. Clustering methods between cis- and trans-eQTLs were proposed to assess their combined effects using bipartite community networks (Platig, Castaldi et al. 2016).

Epistasis:

Epistasis (GxG) is considered as an important contributor of the regulatory network complexity (Mackay and Moore 2014). But epistatic analysis tends to be under-represented compared to single association in science (Carlborg 2004) because of the related computational

and statistical complexity (Upton, Trelles et al. 2015) resulting from the billion possible associations that have to be considered if no filtering is performed before (Sun, Lu et al. 2014).

In this project, epistasis is partially considered for gene co-expression within hiveplots. A partial correlation method was used to disentangle the cis-eQTL effects between pairs of genes. This method allows to 'decorrelate' genes that are functionally unrelated but located within the same haplotype region and under cis-eQTL effect, or to associate co-expressed genes under different cis-eQTL effect. Similar strategies were used in combination with gene co-expression network (Vanderlinden, Saba et al. 2013) or using small bayesian networks approach (Wang and Michoel 2017). However, both methods model single marker effects, whereas it is highly likely that within our list of 5000+ genes with a significant cis-eQTL effect, many of them will interact.

More advanced methods are now being developed using machine learning approachrd which optimize the number of Fisher exact tests, chi-square tests or anova F-test performed. It restricts the computation and statistical tests to SNP-pairs that can possibly reach the required adjusted p-value (Zhang, Zou et al. 2008, Zhang, Zou et al. 2009). These methods work with binary SNPs which fit perfectly inbred lines like BXD, and thus could be applied to our design (Zhang, Huang et al. 2011) to build epistatic network (Tyler, Ji et al. 2017).

Meta-dimensional analysis:

Deep supervised learning:

The stratification of the analyses for each intermediate phenotypic layer as it is performed with the multiple staged approach is in general easier to implement and interpret. It is often used for its biological logic where the information flows directly from genotype toward end-phenotypes (Figure 7A). Meta-dimensional analysis allows for detection of non-linear interaction between all omics layers (Figure 7B) and more likely reflects reality of complex trait.

Many machine learning methods can be used for their predictive capacity and successful integration of non-linear relationship (Li, Wu et al. 2016). Frameworks like MOMA (Kim, Rai et al. 2016) or ATHENA (Kim, Li et al. 2013) could integrate multi-omics layers for *E.coli* end-phenotype prediction using recurrent neural network or cancer clinical outcome using grammatical evolution neural networks. But they often come with costly limitations for the exploratory approach like our systems genetics design.

Figure 7 : Complex trait models:

Hypothesis from complex trait investigation. (Adapted from (Ritchie, Holzinger et al. 2015)

These methods often require large number of samples in combination with a good filtering strategie for dimensionality reduction to avoid overfitting, a recurrent problem for predictive methods in data-driven science considering thousands of parameters extracted. Another difficulty of deep machine learning algorithms is the interpretation of the results. Knowing which parameters were selected and how they interact with other parameters is currently difficult as these algorithms are mostly black-boxes. Until white-box algorithms will be developed, these methods seems limiting for mechanistic inference and thus limiting for explorative and hypothesis-building approaches.

Network-based models:

Network-based models are other powerful investigative methods for meta-dimensional designs. In the BXD, the Similarity Network Fusion (SNF) R package (Wang, Mezlini et al. 2014) combined with unsupervised clustering shows good predictive capacity for mitochondrial metabolism (Tini, Marchetti et al. 2017). SNF is specifically designed to integrate and merge multiple data-types that are in a network format, into a meta-network without supervision. These network based approaches that are based on gene co-expression like WGCNA (Langfelder and Horvath 2008) could be used in combination with our partial correlation method. Finally, many improvements can be implemented for the analysis of metabolic data with e.g. genome scale metabolic network models that could greatly clarify some of the many metabolic change that were observed after sleep deprivation.

Data & knowledge sharing for systems genetics:

Data accessibility is a second big challenge in systems genetics. Currently, there are no centralized repositories that contains all type of phenotypic data that could be extracted within a project. Most maintained repositories are dedicated to a single data-type, like the NCBI Gene Expression Omnibus (GEO) for sequencing data. Such a repository would need to be highly flexible to accommodate highly diverse phenotypes, and deeply metadata centered to permit and facilitate data-integration while controlling for similar experimental condition (e.g. sampling time). Finally, setting up a software infrastructure for data-process standardization would generate better analysis and better reproducibility of results.

Present data-bases for systems genetics are mostly designed to map and mine genotype-phenotype association. GeneNetwork is one good example of a web-based repositoriy that provides tools for QTL mapping and offer a structure for processed mouse phenotypes. An extension of GeneNetwork is http://www.systems-genetics.org (Li, Wang et al. 2017), that provides more developed tools to mine the QTL collection of GeneNetwork and also detect PheWas association (i.e. the association between one selected marker across many phenotypes) (Flintoft 2013). The UCLA Systems Genetics Resource (https://systems.genetics.ucla.edu/) is another web resource for mouse and human studies that contains tools for transcripts and eQTL mining, but also contains meta-data for samples and protocols used to generate the data.

Swiss-BXD interface:

In our web resource, we provide an advanced interactive interface for end-phenotypes and their associated molecular network data-mining. Compared to other web-resources in systems genetics, the focus exceeded the QTL mining and allows to move within and between our novel visualization models using hiveplots. The advantage of end-results data-mining is mainly for the user to easily access processed data without having to regenerate the complete dataset that can be computationally demanding. Furthermore, the direct interaction with novel methods is likely to improve the comprehension of the results, and thus improve knowledge distribution.

Some of our analyses are not directly reproducible on this web site, like the prioritization strategy we developed. However, the code is available via the Rmarkdown format and could be easily inspected.

Fatty acid metabolism & sleep:

Acot11 is an interesting target for the bidirectional communicative pathways between liver and cortex in the context of sleep regulation via blood metabolite transport. This acyl-CoA enzyme is differentially regulated in cortex and liver: the sleep deprivation will upregulate *Acot11* cortical expression but downregulate liver expression. At the genetic level, only *Acot11* expression in the liver is driven by a cis-eQTL and predict the NREM sleep gain. With an Acot11 KO, we showed a strong GxE effect on NREM sleep gain after sleep deprivation (Figure 8).

The exact function of *Acot11* is not yet totally clear, but it was associated with cell proliferation (Hung, Chiang et al. 2017), body temperature (Lau, Tuong et al. 2015) and mainly with lipid metabolic functions. *Acot11* KO mice displayed increased energy expenditure and obesity resistance (Zhang, Li et al. 2012). In brown adipose tissue of C57BL/6J strains, *Acot11* is upregulated during short and prolonged period of high fat diet (HFD) and low fat or standard diet (LFD). However the response was reversed in ddY, ICR, and KK-A[y] strains after HFD, with a downregulation of *Acot11* (Ohtomo, Ino et al. 2017). The strains ddY stain is a model for postprandial hypertriglyceridemia and KK-A[y] mice for type-2 diabetes. These results show

a rather complex GxE effect in BAT but also in our own data with a tissue specific regulation of *Acot11* expression.
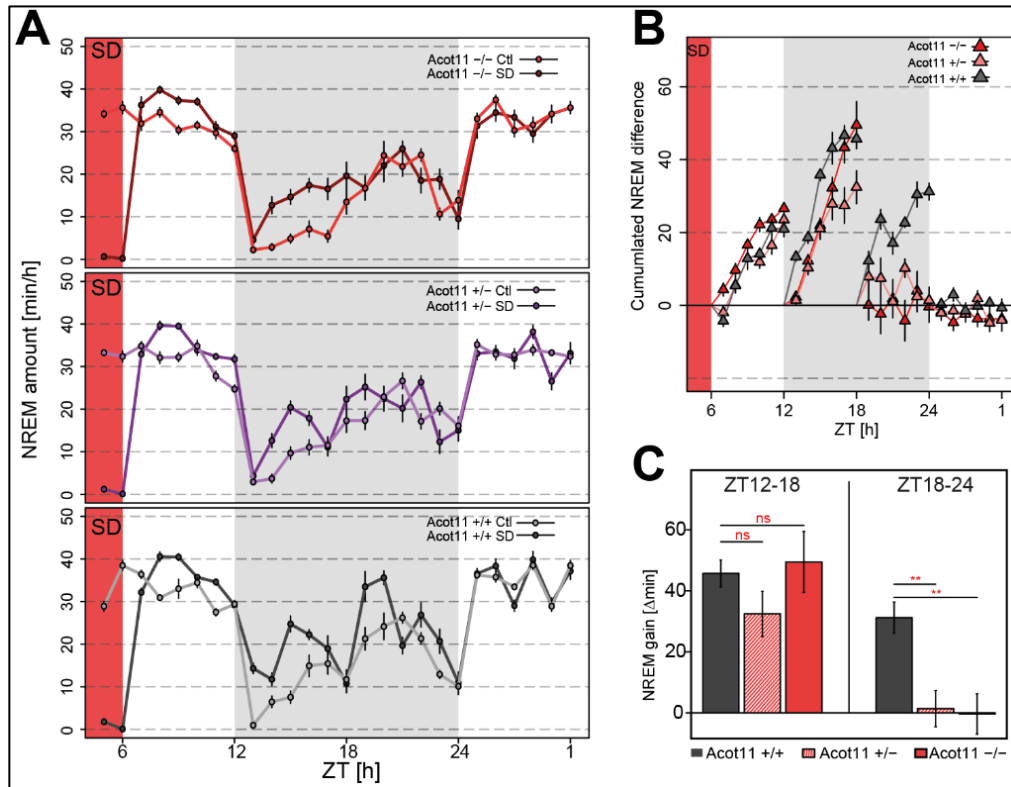
Figure 8 : Reduced NREM gain in Acot11 KO.

(A) NREM sleep amount for Acot11 KO, HT, WT in baseline and after sleep deprivation. **(B),** cumulated NREM sleep gain per 6h bin, for Acot11 KO, HT and WT. **(C)** Significant NREM sleep gain difference between WT mice vs KO or HT in second phase of the dark.

Acot11 is not the only gene linking sleep and fatty acid metabolism identified in our analysis. *Cpy4a32* is involved in the arachidonic acid metabolism and is suggested to modulate the REM sleep theta peak frequency shift after sleep deprivation in our data. The Arachidonic acid is one of the main free fatty acids that can be found in the brain and some of the derived forms like prostaglandin D2 play a role in sleep-wake regulation (Urade and Hayaishi 2011).

Phenotypic variability in the BXD panel

The BXD mice are very convenient for studying complex traits, with a large choice of lines, available genotypes and dedicated tools for phenotype mapping, however it also comes with some caveats. Even more advanced lines recombinant inbred lines were developed like the Collaborative Cross (CC) panel (Threadgill, Miller et al. 2011), were parental lines are composed of the classic C57BL/6J but also 7 other strains (A/J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ). This greater genetic diversity is also reflected on behavioral and physiological phenotypes, with greater variability in the CC panel than the BXD panel that is composed of B6 and D2 allele only (Philip, Sokoloff et al. 2011).

The BXD panel is also heterogeneous for its Y and mitochondrial chromosomes which should be considered in further analysis including population stratification. The parental origin of the BXD mitochondrial chromosome was typed using SNP marker and a majority of BXD lines inherited the maternal C57BL/6J mitochondrial chromosome. However, 10 lines inherited a DBA2/J maternal chromosome (BXD32, 61, 74, 76, 82, 89, 90, 91, 95, and BXD99) and should be considered as DXB. This difference can be explained with different methods for the generation of AIL in Princeton and UTHSC, that used fixed paternal/maternal strains or used reciprocal crosses (Peirce, Lu et al. 2004).

The true genetic nature of the phenotypic variability that is found within the BXD panel may have multiple origins that is not always evident to untangle. The molecular and physiological regulatory difference might be the result of single allelic effect that we measured with QTL analysis. It might result from higher order interaction (GxG) due to the different B6-D2 combination in the BXD allelic background. The different parental D2 sub-strains used in 'old' and 'new' BXD may also affect the reproducibility of the results. Finally, the chromosomal Y and mt heterogeneity can also introduce phenotypic variability in the panel.

The purpose of systems genetics

The exact definition and purpose of the systems genetics approach seems to depend on the biological question asked. Among the various definitions the one by Michael Boutros particularly suits our approach:

"Systems genetics in my view is an exciting and emerging field that combines many state-of-the art approaches in genetics and genomics, systems biology and what can be termed "phenomics"—methods to capture phenotypes, either at the molecular or organismal level. The field builds upon recent advances in genomics, imaging, and related fields. System genetics studies can rely on natural variation as a source of genetic differences, or systematic perturbation studies that are done in model organisms …. Together, these advances will allow a better understanding of the link between genotypes and phenotypes—in my view the core question in systems genetics." (Baliga, Björkegren et al. 2017)

In our design, systems genetics is used to uncover novel regulatory pathways, with a mechanistic approach focused on gene expression and metabolic levels. The main purpose of systems genetics is not only a better characterization of the existing genotype-phenotype relationship in BXD, but mostly a better characterization of the molecular systems underlying end-phenotypes (the links). In this case, genetics should be considered as a tool or a method to perturb biological systems, to observe where variation effect(s) propagate within the system and identify links that are persistent under different conditions and genetic backgrounds. The genetic marker composition is prone to change among individuals, which will change the way a molecular system is perturbed or will adapt, but the nature of the systems should remain stable.

In the context of our study, systems genetics is used to aggregate liver, plasma blood and cortical systems, we could summarize this into a *biological system scaffolding*. Where the organism is examined more broadly with different regulatory structures that at some point will contribute to sleep regulation (given in this case a highly interconnected phenotype: sleep).

Deeper mechanistic exploration using chromatin accessibility

A better description of the relation between marker and gene expression regulation can be conducted with the integration of epigenetic signal. In the next months we aimed at measuring the chromatin accessibility using ATAC-sequencing (Buenrostro, Giresi et al. 2013) in the BXD. The set up of the analysis tools is currently under investigation using ATAC-seq time-course and sleep deprivation effect on chromatin accessibility on C57BL/6J strains. This project is under preparation and further described in Annex.

Conclusion

In conclusion, systems genetics is a trans-disciplinary approach that need novel methodologies at many levels for proper data interpretation. This approach should greatly benefit from the joint effort to develop workflows and datasets that can be shared. Open questions have still to be answered from the results we found such as the contribution of epigenetics to gene regulation, the dynamic contribution of free fatty acid to sleep regulation or the different pathways underlying fast and slow EEG delta power bands. Often defined as a holistic approach (Civelek and Lusis 2014) opposed to reductionist approaches, systems genetics and associated methods currently still stand on discrete and selected sampling strategy and should be considered to certain extend as an advanced reductionist approach, on a continuous scale toward holistic science. The reductionist and systems approaches should be considered as complementary (Fang and Casadevall 2011) and not exclusive, as exposed with our *Acot11* example, merging systems genetics and *reverse* genetics.

Lists of abbreviations

QTL : Quantitative Trait Locus

eQTL : Expression QTL

mQTL : Metabolite QTL

phQTL : Phenotypic QTL (EEG/Behavioral)

SD : Sleep Deprivation

rec : Recovery time

bsl : Baseline time

Ctl : Control

B1 : Baseline Day 1

B2 : Baseline Day 2

R1 : Recovery Day 1

R2 : Recovery Day 2

EEG : ElectroEncephaloGram

TPF : Theta Peak Frequency

D(G)E : Differential (Gene) Expression

LOD : Logarithm Of Odds

FDR : False Discovery Rate

FC : Fold-Change

ZT : Zeitgeber Time

REM sleep : Rapid Eye Movement sleep

NREM sleep : Non-Rapid Eye Movement sleep

TDW : Theta Dominated Wakefulness

h^2 : Narrow-sense heritability

H^2 : Broad-sense heritability

SVM : Support Vector Machine

LDLA : Diagonal Linear Discriminant Analysis

D3 : Data Driven Document

A.A. : Amino-Acids

ACT : Acylcarnitines

BA : Biogenic Amines

GPL : Glycerophospholipids

SM : Sphingomyelins

Mb : Mega-base

B6 : C57BL6/J

D2 : DBA2/J

n.s. : not significant

CPM : Count Per Million

FPKM : Fragment Per Kilobase Million

PC : PhosphatidylCholine

SNP : Single Nucleotide Polymorphism

Indel : Insertion/Deletion

LMA : LocoMotor Activity

GATK : Genome Analysis ToolKit

MGI : Mouse Genome Informatics

LRS : Likelihood Ratio Statistics

GxG : Gene – Gene Interaction

GxE : Gene – Environment Interaction

BXD : B6 X D2 Recombinant inbred line

GRP : Genetic Reference Population

GWAS : Genome-wide association study

HFD: High Fat Diet

LFD: Low Fat Diet

Online resources:

Systems genetics of sleep web interface: https://bxd.vital-it.ch

Open source code, Rmarkdown report and project meta-data : https://gitlab.isb-sib.ch/majan/Systems_Genetics_of_Sleep

References:

Adamantidis, A. and L. de Lecea (2008). "Sleep and metabolism: shared circuits, new connections." Trends in endocrinology and metabolism: TEM **19**(10): 362-370.

Ala-Korpela, M., A. J. Kangas and M. Inouye (2011). "Genome-wide association studies and systems biology: together at last." Trends in genetics : TIG **27**(12): 493-498.

Altevogt, B. M. and H. R. Colten (2006). Sleep disorders and sleep deprivation: an unmet public health problem, National Academies Press.

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

Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Andretic, R., P. Franken and M. Tafti (2008). "Genetics of sleep." Annual review of genetics **42**: 361-388.

Andreux, P. A. A., E. G. Williams, H. Koutnikova, R. H. Houtkooper, M.-F. F. Champy, H. Henry, K. Schoonjans, R. W. Williams and J. Auwerx (2012). "Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits." Cell **150**(6): 1287-1299.

Archer, S. N., E. E. Laing, C. S. Möller-Levet, D. R. van der Veen, G. Bucca, A. S. Lazar, N. Santhi, A. Slak, R. Kabiljo, M. von Schantz, C. P. Smith and D.-J. J. Dijk (2014). "Mistimed sleep disrupts circadian regulation of the human transcriptome." Proceedings of the National Academy of Sciences of the United States of America **111**(6): 91.

Arnar, D. O., K. Andersen and G. Thorgeirsson (2016). "Genetics of cardiovascular diseases: lessons learned from a decade of genomics research in Iceland." Scandinavian cardiovascular journal : SCJ **50**(5-6): 260-265.

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. Anholt and T. F. Mackay (2009). "Systems genetics of complex traits in Drosophila melanogaster." Nature genetics **41**(3): 299-307.

Baliga, N. S., J. L. M. Björkegren, J. D. Boeke and M. Boutros (2017). "The State of Systems Genetics in 2017." Cell Systems.

Basnet, R. K., D. P. Del Carpio, D. Xiao, J. Bucher, M. Jin, K. Boyle, P. Fobert, R. G. Visser, C. Maliepaard and G. Bonnema (2016). "A Systems Genetics Approach Identifies Gene Regulatory Networks Associated with Fatty Acid Composition in Brassica rapa Seed." Plant Physiol **170**(1): 568-585.

Bass, J. and J. S. Takahashi (2010). "Circadian integration of metabolism and energetics." Science (New York, N.Y.) **330**(6009): 1349-1354.

Baumer, B., M. Cetinkaya-Rundel, A. Bray and L. Loi (2014). "R Markdown: Integrating a reproducible analysis tool into introductory statistics." arXiv preprint arXiv: ….

Bayon, V., D. Leger, D. Gomez-Merino, M.-F. F. Vecchierini and M. Chennaoui (2014). "Sleep debt and obesity." Annals of medicine **46**(5): 264-272.

Bechhofer, S., I. Buchan, D. D. Roure and M.-P. Generation … (2013). "Why linked data is not enough for scientists." Future Generation ….

Bendesky, A., Y.-M. M. Kwon, J.-M. M. Lassance, C. L. Lewarch, S. Yao, B. K. Peterson, M. X. He, C. Dulac and H. E. Hoekstra (2017). "The genetic basis of parental care evolution in monogamous mice." Nature **544**(7651): 434-439.

Berger, N. A. and S. Redline (2014). "Impact of sleep and sleep disturbances on obesity and cancer." Impact of sleep and sleep disturbances on obesity and cancer.

Blanco-Gómez, A., S. Castillo-Lluva, M. Del Mar Sáez-Freire, L. Hontecillas-Prieto, J. H. Mao, A. Castellanos-Martín and J. Pérez-Losada (2016). "Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes." BioEssays : news and reviews in molecular, cellular and developmental biology **38**(7): 664-673.

Borbély, A. A. (1982). "A two process model of sleep regulation." Human neurobiology **1**(3): 195-204.

Borbély, A. A., S. Daan, A. Wirz-Justice and T. Deboer (2016). "The two-process model of sleep regulation: a reappraisal." Journal of sleep research **25**(2): 131-143.

Bourdi, M., J. S. Davies and L. R. Pohl (2011). "Mispairing C57BL/6 substrains of genetically engineered mice and wild-type controls can lead to confounding results as it did in studies of JNK2 in acetaminophen and concanavalin A liver injury." Chemical research in toxicology **24**(6): 794-796.

Boyce, R., S. D. Glasgow, S. Williams and A. Adamantidis (2016). "Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation." Science (New York, N.Y.) **352**(6287): 812-816.

Bray, N. L., H. Pimentel, P. Melsted and L. Pachter (2016). "Near-optimal probabilistic RNA-seq quantification." Nature biotechnology **34**(5): 525-527.

Brown, R. E., R. Basheer, J. T. McKenna, R. E. Strecker and R. W. McCarley (2012). "Control of sleep and wakefulness." Physiological reviews **92**(3): 1087-1187.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang and W. J. Greenleaf (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nature methods **10**(12): 1213-1218.

Buil, A., A. A. Brown, T. Lappalainen, A. Vinuela, M. N. Davies, H. F. Zheng, J. B. Richards, D. Glass, K. S. Small, R. Durbin, T. D. Spector and E. T. Dermitzakis (2015). "Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins." Nat Genet **47**(1): 88-91.

Büttner, J. (1998). "Biological variation and quantification of health: the emergence of the concept of normality." Clinical chemistry and laboratory medicine **36**(1): 69-73.

Carlborg, O. (2004). "Epistasis: too often neglected in complex trait studies?".

Carlsten, C., M. Brauer, F. Brinkman, J. Brook, D. Daley, K. McNagny, M. Pui, D. Royce, T. Takaro and J. Denburg (2014). "Genes, the environment and personalized medicine: We need to harness both environmental and genetic data to maximize personal and population health." EMBO reports **15**(7): 736-739.

Chin, C.-S. S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. R. Cramer, M. Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank and M. C. Schatz (2016). "Phased diploid genome assembly with single-molecule real-time sequencing." Nature methods **13**(12): 1050-1054.

Chintalapudi, S. R., D. Maria, X. Di Wang, J. N. C. N. C. Bailey, N. consortium, I. consortium, P. G. Hysi, J. L. Wiggs, R. W. Williams and M. M. Jablonski (2017). "Systems genetics identifies a role for Cacna2d1 regulation in elevated intraocular pressure and glaucoma susceptibility." Nature communications **8**(1): 1755.

Civelek, M. and A. J. Lusis (2014). "Systems genetics approaches to understand complex traits." Nature reviews. Genetics **15**(1): 34-48.

Cohen-Boulakia, S., K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Bras and F. Lemoine (2017). "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities." Future Generation Computer Systems **75**: 284-298.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang and A. Mortazavi (2016). "A survey of best practices for RNA-seq data analysis." Genome biology **17**: 13.

Coveney, P. V., E. R. Dougherty and R. R. Highfield (2016). "Big data need big theory too." Philosophical transactions. Series A, Mathematical, physical, and engineering sciences **374**(2080).

Crawford, N. P. (2016). "Deciphering the Dark Matter of Complex Genetic Inheritance." Cell systems **2**(3): 144-146.

Crocker, A. and A. Sehgal (2010). "Genetic analysis of sleep." Genes & development **24**(12): 1220-1235.

Daan, S., D. G. Beersma and A. A. Borbély (1984). "Timing of human sleep: recovery process gated by a circadian pacemaker." The American journal of physiology **246**(2 Pt 2): 83.

Davies, S. K., J. E. Ang, V. L. Revell, B. Holmes, A. Mann, F. P. Robertson, N. Cui, B. Middleton, K. Ackermann, M. Kayser, A. E. Thumser, F. I. Raynaud and D. J. Skene (2014). "Effect of sleep deprivation on the human metabolome." Proceedings of the National Academy of Sciences of the United States of America **111**(29): 10761-10766.

De Gennaro, L., C. Marzano, F. Fratello, F. Moroni, M. C. Pellicciari, F. Ferlazzo, S. Costa, A. Couyoumdjian, G. Curcio, E. Sforza, A. Malafosse, L. A. Finelli, P. Pasqualetti, M. Ferrara, M. Bertini and P. M. Rossini (2008). "The electroencephalographic fingerprint of sleep is genetically determined: a twin study." Annals of neurology **64**(4): 455-460.

Dibner, C., U. Schibler and U. Albrecht (2010). "The mammalian circadian timing system: organization and coordination of central and peripheral clocks." Annual review of physiology **72**: 517-549.

Diering, G. H., R. S. Nirujogi, R. H. Roth, P. F. Worley, A. Pandey and R. L. Huganir (2017). "Homer1a drives homeostatic scaling-down of excitatory synapses during sleep." Science (New York, N.Y.) **355**(6324): 511-515.

Dijk, D.-J. and P. Franken (2005). "Principles and Practice of Sleep Medicine (Fourth Edition)." Part I: Principles of Sleep Medicine: Section 4Chronobiology: Section 4: Chronobiology: 418-434.

Dijk, D. J., J. F. Duffy and C. A. Czeisler (2000). "Contribution of circadian physiology and sleep homeostasis to age-related changes in human sleep." Chronobiology international **17**(3): 285-311.

Ding, D., K. Rogers, H. van der Ploeg, E. Stamatakis and A. E. Bauman (2015). "Traditional and Emerging Lifestyle Risk Behaviors and All-Cause Mortality in Middle-Aged and Older Adults: Evidence from a Large Population-Based Australian Cohort." PLoS medicine **12**(12).

Dobin, A. and D. C. (2013). "STAR: ultrafast universal RNA-seq aligner."

Durinck, S., P. T. Spellman, E. Birney and W. Huber (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." Nature protocols **4**(8): 1184-1191.

Durrant, C., M. A. Swertz, R. Alberts, D. Arends, S. Möller, R. Mott, P. Prins, K. J. van der Velde, R. C. Jansen and K. Schughart (2012). "Bioinformatics tools and database resources for systems genetics analysis in mice--a short review and an evaluation of future needs." Briefings in bioinformatics **13**(2): 135-142.

Ehlen, J. C., A. J. Brager, J. Baggs, L. Pinckney, C. L. Gray, J. P. DeBruyne, K. A. Esser, J. S. Takahashi and K. N. Paul (2017). "Bmal1 function in skeletal muscle regulates sleep." eLife **6**.

eLife (Jul 14, 2017). "Innovation: Understanding the demand for reproducible research articles." INSIDE ELIFE **https://elifesciences.org/inside-elife/e832444e**.

eLife (Sep 7, 2017). "eLife supports development of open technology stack for publishing reproducible manuscripts online." elife **PRESS PACK**.

Fadista, J., A. K. Manning, J. C. Florez and L. Groop (2016). "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants." European journal of human genetics : EJHG **24**(8): 1202-1205.

Falbe, J., K. K. Davison, R. L. Franckle, C. Ganter, S. L. Gortmaker, L. Smith, T. Land and E. M. Taveras (2015). "Sleep duration, restfulness, and screens in the sleep environment." Pediatrics **135**(2).

Fan, W., W. Waizenegger, C. S. Lin, V. Sorrentino, M.-X. X. He, C. E. Wall, H. Li, C. Liddle, R. T. Yu, A. R. Atkins, J. Auwerx, M. Downes and R. M. Evans (2017). "PPARδ Promotes Running Endurance by Preserving Glucose." Cell metabolism **25**(5): 1186-11930000.

Fang, F. C. and A. Casadevall (2011). Reductionistic and holistic science. Reductionistic and holistic science, Am Soc Microbiol.

Fifel, K., J. H. Meijer and T. Deboer (2018). "Long-term effects of sleep deprivation on neuronal activity in four hypothalamic areas." Neurobiology of Disease **109**(Pt A): 54-63.

Figueiredo, A. S. (2017). "Data Sharing: Convert Challenges into Opportunities." Frontiers in public health **5**: 327.

Flint, J. and E. Eskin (2012). "Genome-wide association studies in mice." Nature reviews. Genetics **13**(11): 807-817.

Flintoft, L. (2013). "Disease genetics: Phenome-wide association studies go large." Nature Reviews Genetics **15**(1).

Fox, C. S., Y. Liu, C. C. White, M. Feitosa, A. V. Smith, N. Heard-Costa, K. Lohman, G. Consortium, M. Consortium, G. Consortium, A. D. Johnson, M. C. Foster, D. M. Greenawalt, P. Griffin, J. Ding, A. B. Newman, F. Tylavsky, I. Miljkovic, S. B. Kritchevsky, L. Launer, M. Garcia, G. Eiriksdottir, J. J. Carr, V. Gudnason, T. B. Harris, L. A. Cupples and I. B. Borecki (2012). "Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women." PLoS genetics **8**(5).

Frank, M. (2013). "Why I am not shy: a reply to Tononi and Cirelli." Neural plasticity **2013**.

Frank, M. G. (2012). "Erasing synapses in sleep: is it time to be SHY?" Neural plasticity **2012**: 264378.

Franken, P. (2013). "A role for clock genes in sleep homeostasis." Current opinion in neurobiology **23**(5): 864-872.

Franken, P., D. Chollet and M. Tafti (2001). "The homeostatic regulation of sleep need is under genetic control." The Journal of neuroscience : the official journal of the Society for Neuroscience **21**(8): 2610-2621.

Franken, P., D. J. Dijk, I. Tobler and A. A. Borbély (1991). "Sleep deprivation in rats: effects on EEG power spectra, vigilance states, and cortical temperature." The American journal of physiology **261**(1 Pt 2): 208.

Franken, P. and D. J. J. Dijk (2009). "Circadian clock genes and sleep homeostasis." The European journal of neuroscience **29**(9): 1820-1829.

Franken, P., A. Malafosse and M. Tafti (1998). "Genetic variation in EEG activity during sleep in inbred mice." The American journal of physiology **275**(4 Pt 2): 37.

Franken, P., A. Malafosse and M. Tafti (1999). "Genetic determinants of sleep regulation in inbred mice." Sleep **22**(2): 155-169.

Franken, P. and M. Tafti (2003). "Genetics of sleep and sleep disorders." Frontiers in bioscience : a journal and virtual library **8**: 97.

Funato, H., C. Miyoshi, T. Fujiyama, T. Kanda, M. Sato, Z. Wang, J. Ma, S. Nakane, J. Tomita, A. Ikkyu, M. Kakizaki, N. Hotta-Hirashima, S. Kanno, H. Komiya, F. Asano, T. Honda, S. J. Kim, K. Harano, H. Muramoto, T. Yonezawa, S. Mizuno, S. Miyazaki, L. Connor, V. Kumar, I. Miura, T. Suzuki, A. Watanabe, M. Abe, F. Sugiyama, S. Takahashi, K. Sakimura, Y. Hayashi, Q. Liu, K. Kume, S. Wakana, J. S. Takahashi and M. Yanagisawa (2016). "Forward-genetics analysis of sleep in randomly mutagenized mice." Nature **539**(7629): 378-383.

Gao, V., F. Turek and M. Vitaterna (2016). "Multiple classifier systems for automatic sleep scoring in mice." Journal of neuroscience methods **264**: 33-39.

Gligorijević, V. and N. Pržulj (2015). "Methods for biological data integration: perspectives and challenges." Journal of the Royal Society, Interface **12**(112).

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa and J. Tegnér (2014). "Data integration in the era of omics: current and future challenges." BMC systems biology **8 Suppl 2**.

Gronfier, C., R. Luthringer, M. Follenius, N. Schaltenbrand, J. P. Macher, A. Muzet and G. Brandenberger (1997). "Temporal relationships between pulsatile cortisol secretion and electroencephalographic activity during sleep in man." Electroencephalography and clinical neurophysiology **103**(3): 405-408.

Großkinsky, D. K., J. Svensgaard, S. Christensen and T. Roitsch (2015). "Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap." Journal of experimental botany **66**(18): 5429-5440.

Grundberg, E., K. S. Small, Å. K. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S.-Y. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector and M. Consortium (2012). "Mapping cis- and trans-regulatory effects across multiple tissues in twins." Nature genetics **44**(10): 1084-1089.

Gurumurthy, C. B., M. h. Grati, M. Ohtsuka, S. L. Schilit, R. M. Quadros and X. Z. Liu (2016). "CRISPR: a versatile tool for both forward and reverse genetics research." Human genetics **135**(9): 971-976.

H., O., A. Buil, A. A. Brown and E. T. Dermitzakis (2015). "Fast and efficient QTL mapper for thousands of molecular phenotypes."

Halsey, L. G., D. Curran-Everett, S. L. Vowler and G. B. Drummond (2015). "The fickle P value generates irreproducible results." Nature methods **12**(3): 179-185.

He, H., D. Sun, Y. Zeng, R. Wang, W. Zhu, S. Cao, G. A. Bray, W. Chen, H. Shen, F. M. Sacks, L. Qi and H.-W. W. Deng (2017). "A Systems Genetics Approach Identified GPD1L and its Molecular Mechanism for Obesity in Human Adipose Tissue." Scientific reports **7**(1): 1799.

He, K. Y., D. Ge and M. M. He (2017). "Big Data Analytics for Genomic Medicine." International journal of molecular sciences **18**(2).

Hegmann, J. P. and B. Possidente (1981). "Estimating genetic correlations from inbred strains." Behavior genetics **11**(2): 103-114.

Hirotsu, C., S. Tufik and M. L. Andersen (2015). "Interactions between sleep, stress, and metabolism: From physiological to pathological conditions." Sleep science (Sao Paulo, Brazil) **8**(3): 143-152.

Hu, J. X., C. E. Thomas and S. Brunak (2016). "Network biology concepts in complex disease comorbidities." Nature reviews. Genetics **17**(10): 615-629.

Huang, S., K. Chaudhary and L. X. Garmire (2017). "More Is Better: Recent Progress in Multi-Omics Data Integration Methods." Frontiers in Genetics **8**: 84.

Huang, W., K. M. Ramsey, B. Marcheva and J. Bass (2011). "Circadian rhythms, sleep, and metabolism." The Journal of clinical investigation **121**(6): 2133-2141.

Hung, J.-Y. Y., S.-R. R. Chiang, K.-T. T. Liu, M.-J. J. Tsai, M.-S. S. Huang, J.-M. M. Shieh, M.-C. C. Yen and Y.-L. L. Hsu (2017). "Overexpression and proliferation dependence of acyl-CoA thioesterase 11 and 13 in lung adenocarcinoma." Oncology letters **14**(3): 3647-3656.

Hunter, R. G. (2012). "Epigenetic effects of stress and corticosteroids in the brain." Frontiers in cellular neuroscience **6**: 18.

Iqbal, S. A., J. D. Wallach, M. J. Khoury, S. D. Schully and J. P. Ioannidis (2016). "Reproducible Research Practices and Transparency across the Biomedical Literature." PLoS biology **14**(1).

Issurin, V. B. (2017). "Evidence-Based Prerequisites and Precursors of Athletic Talent: A Review." Sports medicine (Auckland, N.Z.) **47**(10): 1993-2010.

Jagodnik, K. M., S. Koplev, S. L. Jenkins, L. Ohno-Machado, B. Paten, S. C. Schurer, M. Dumontier, R. Verborgh, A. Bui, P. Ping, N. J. McKenna, R. Madduri, A. Pillai and A. Ma'ayan (2017). "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.

Jensen, A. B., P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen and S. Brunak (2014). "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients." Nature communications **5**: 4022.

Jeong, C., G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. B. Witonsky, J. K. Pritchard, C. M. Beall and A. Di Rienzo (2014). "Admixture facilitates genetic adaptations to high altitude in Tibet." Nature communications **5**: 3281.

Jiang, P., J. R. Scarpa, K. Fitzpatrick, B. Losic, V. D. Gao, K. Hao, K. C. Summa, H. S. Yang, B. Zhang, R. Allada, M. H. Vitaterna, F. W. Turek and A. Kasarskis (2015). "A systems approach identifies networks and genes linking sleep and stress: implications for neuropsychiatric disorders." Cell reports **11**(5): 835-848.

Jiménez, R. C., M. Kuzak, M. Alhamdoosh, M. Barker, B. Batut, M. Borg, S. Capella-Gutierrez, N. Chue Hong, M. Cook, M. Corpas, M. Flannery, L. Garcia, J. L. Gelpí, S. Gladman, C. Goble, M. González Ferreiro, A. Gonzalez-Beltran, P. C. Griffin, B. Grüning, J. Hagberg, P. Holub, R. Hooft, J. Ison, D. S. Katz, B. Leskošek, F. López Gómez, L. J. Oliveira, D. Mellor, R. Mosbergen, N. Mulder, Y. Perez-Riverol, R. Pergl, H. Pichler, B. Pope, F. Sanz, M. V. Schneider, V. Stodden, R. Suchecki, R. Svobodová Vařeková, H.-A. A. Talvik, I. Todorov, A. Treloar, S. Tyagi, M. van Gompel, D. Vaughan, A. Via, X. Wang, N. S. Watson-Haigh and S. Crouch (2017). "Four simple recommendations to encourage best practices in research software." F1000Research **6**.

Johnson, M. R., J. Behmoaras, L. Bottolo, M. L. Krishnan, K. Pernhorst, P. L. Santoscoy, T. Rossetti, D. Speed, P. K. Srivastava, M. Chadeau-Hyam, N. Hajji, A. Dabrowska, M. Rotival, B. Razzaghi, S. Kovac, K. Wanisch, F. W. Grillo, A. Slaviero, S. R. Langley, K. Shkura, P. Roncon, T. De, M. Mattheisen, P. Niehusmann, T. J. O'Brien, S. Petrovski, M. von Lehe, P. Hoffmann, J. Eriksson, A. J. Coffey, S. Cichon, M. Walker, M. Simonato, B. Danis, M. Mazzuferi, P. Foerch, S. Schoch, V. De Paola, R. M. Kaminski, V. T. Cunliffe, A. J. Becker and E. Petretto (2015). "Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus." Nat Commun **6**: 6031.

Johnson, M. R., K. Shkura, S. R. Langley, A. Delahaye-Duriez, P. Srivastava, W. D. Hill, O. J. Rackham, G. Davies, S. E. Harris, A. Moreno-Moral, M. Rotival, D. Speed, S. Petrovski, A. Katz, C. Hayward, D. J. Porteous, B. H. Smith, S. Padmanabhan, L. J. Hocking, J. M. Starr, D. C. Liewald, A. Visconti, M. Falchi, L. Bottolo, T. Rossetti, B. Danis, M. Mazzuferi, P. Foerch, A. Grote, C. Helmstaedter, A. J. Becker, R. M. Kaminski, I. J. Deary and E. Petretto (2015). "Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease." Nat Neurosci.

Kammermeier, P. J. and P. F. Worley (2007). "Homer 1a uncouples metabotropic glutamate receptor 5 from postsynaptic effectors." Proceedings of the National Academy of Sciences of the United States of America **104**(14): 6055-6060.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly and E. Eskin (2008). "Efficient control of population structure in model organism association mapping." Genetics **178**(3): 1709-1723.

Katsageorgiou, V.-M., G. Lassi, V. Tucci, V. Murino and D. Sona (2015). "Sleep-Stage Scoring in Mice: The Influence of Data Pre-Processing on a System's Performance." 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) **2015**: 598-601.

Kim, D., R. Li, S. M. Dudek and M. D. Ritchie (2013). "ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network." BioData mining **6**(1): 23.

Kim, M., N. Rai, V. Zorraquino and I. Tagkopoulos (2016). "Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli." Nature Communications **7**: 13090.

Kimura, M. and J. Winkelmann (2007). "Genetics of sleep and sleep disorders." Cellular and molecular life sciences : CMLS **64**(10): 1216-1226.

Kollmus, H., E. Wilk and K. Schughart (2014). "Systems biology and systems genetics - novel innovative approaches to study host-pathogen interactions during influenza infection." Curr Opin Virol **6**: 47-54.

Kopp, M. S., A. Stauder, G. Purebl, I. Janszky and A. Skrabski (2008). "Work stress and mental health in a changing society." European journal of public health **18**(3): 238-244.

Krishna, A., M. Biryukov, C. Trefois, P. M. Antony, R. Hussong, J. Lin, M. Heinäniemi, G. Glusman, S. Köglsberger, O. Boyd, B. H. van den Berg, D. Linke, D. Huang, K. Wang, L. Hood, A. Tholey, R. Schneider, D. J. Galas, R. Balling and P. May (2014). "Systems genomics evaluation of the SH-SY5Y neuroblastoma cell line as a model for Parkinson's disease." BMC genomics **15**: 1154.

Krueger, J. M., M. G. Frank, J. P. Wisor and S. Roy (2016). "Sleep function: Toward elucidating an enigma." Sleep medicine reviews **28**: 46-54.

Kume, Y., S. Makabe, N. Singha-Dong, P. Vajamun, H. Apikomonkon and J. Griffiths (2017). "Seasonal effects on the sleep–wake cycle, the rest–activity rhythm and quality of life for Japanese and Thai older people." Chronobiology International: 1-11.

Lampert, T., A. Plano, J. Austin and B. Platt (2015). "On the identification of sleep stages in mouse electroencephalography time-series." Journal of Neuroscience Methods **246**: 52-64.

Lane, J. M., J. Liang, I. Vlasac, S. G. Anderson, D. A. Bechtold, J. Bowden, R. Emsley, S. Gill, M. A. Little, A. I. Luik, A. Loudon, F. A. Scheer, S. M. Purcell, S. D. Kyle, D. A. Lawlor, X. Zhu, S. Redline, D. W. Ray, M. K. Rutter and R. Saxena (2017). "Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits." Nature genetics **49**(2): 274-281.

Langfelder, P., L. W. Castellani, Z. Zhou, E. Paul, R. Davis, E. E. Schadt, A. J. Lusis, S. Horvath and M. Mehrabian (2012). "A systems genetic analysis of high density lipoprotein metabolism and network preservation across mouse models." Biochimica et biophysica acta **1821**(3): 435-447.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**: 559.

Lanktree, M. B. and R. A. Hegele (2009). "Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease." Genome medicine **1**(2): 28.

Lapatas, V., M. Stefanidakis, R. C. Jimenez, A. Via and M. V. Schneider (2015). "Data integration in biological research: an overview." Journal of biological research (Thessalonike, Greece) **22**(1): 9.

Lau, P., Z. K. Tuong, S.-C. C. Wang, R. L. Fitzsimmons, J. M. Goode, G. P. Thomas, G. J. Cowin, M. A. Pearen, K. Mardon, J. L. Stow and G. E. Muscat (2015). "Rorα deficiency and decreased adiposity are associated with induction of thermogenic gene expression in subcutaneous white adipose and brown adipose tissue." American journal of physiology. Endocrinology and metabolism **308**(2): 71.

Li, H., X. Wang, D. Rukina, Q. Huang, T. Lin, V. Sorrentino, H. Zhang, M. Sleiman, D. Arends, A. McDaid, P. Luan, N. Ziari, L. A. Velázquez-Villegas, K. Gariani, Z. Kutalik, K. Schoonjans, R. A. Radcliffe, P. Prins, S. Morgenthaler, R. W. Williams and J. Auwerx (2017). "An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function." Cell Systems.

Li, Y., F.-X. Wu and A. Ngom (2016). "A review on machine learning principles for multi-view biological data integration." Briefings in Bioinformatics.

Linkowski, P. (1999). "EEG sleep patterns in twins." Journal of sleep research **8 Suppl 1**: 11-13.

Llinares-López, F., L. Papaxanthos, D. Bodenham, D. Roqueiro, C. Investigators and K. Borgwardt (2017). "Genome-wide genetic heterogeneity discovery with categorical covariates." Bioinformatics (Oxford, England) **33**(12): 1820-1828.

Lowndes, J. S. S. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang and B. S. Halpern (2017). "Our path to better science in less time using open data science tools." Nature ecology & evolution **1**(6): 160.

Luke, S. G. (2017). "Evaluating significance in linear mixed-effects models in R." Behavior research methods **49**(4): 1494-1502.

Mackay, T. F. and J. H. Moore (2014). "Why epistasis is important for tackling complex human disease genetics." Genome medicine **6**(6): 124.

Mackiewicz, M., B. Paigen, N. Naidoo and A. I. Pack (2008). "Analysis of the QTL for sleep homeostasis in mice: Homer1a is a likely candidate." Physiological genomics **33**(1): 91-99.

137

Macklem, P. T. (2008). "Emergent phenomena and the secrets of life." Journal of applied physiology (Bethesda, Md. : 1985) **104**(6): 1844-1846.

Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature **456**(7218): 18-21.

Mang, G. M. and P. Franken (2013). Genetic dissection of sleep homeostasis, Springer.

Mang, G. M. M. and P. Franken (2015). "Genetic dissection of sleep homeostasis." Current topics in behavioral neurosciences **25**: 25-63.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Maret, S., S. Dorsaz, L. Gurcel, S. Pradervand, B. Petit, C. Pfister, O. Hagenbuchle, B. F. O'Hara, P. Franken and M. Tafti (2007). "Homer1a is a core brain molecular correlate of sleep loss." Proceedings of the National Academy of Sciences of the United States of America **104**(50): 20090-20095.

Margolis, R., L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, M. Guyer and E. D. Green (2014). "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data." Journal of the American Medical Informatics Association : JAMIA **21**(6): 957-958.

McCue, M. E. and A. M. McCoy (2017). "The Scope of Big Data in One Medicine: Unprecedented Opportunities and Challenges." Frontiers in veterinary science **4**: 194.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

Merkwirth, C., V. Jovaisaite, J. Durieux, O. Matilainen, S. D. Jordan, P. M. Quiros, K. K. Steffen, E. G. Williams, L. Mouchiroud, S. U. Tronnes, V. Murillo, S. C. Wolff, R. J. Shaw, J. Auwerx and A. Dillin (2016). "Two Conserved Histone Demethylases Regulate Mitochondrial Stress-Induced Longevity." Cell **165**(5): 1209-1223.

Mieth, B., M. Kloft, J. A. Rodríguez, S. Sonnenburg, R. Vobruba, C. Morcillo-Suárez, X. Farré, U. M. Marigorta, E. Fehr, T. Dickhaus, G. Blanchard, D. Schunk, A. Navarro and K.-R. R. Müller (2016). "Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies." Scientific reports **6**: 36671.

Mignot, E. (2008). "Why we sleep: the temporal organization of recovery." PLoS biology **6**(4).

Mongrain, V., S. A. Hernandez, S. Pradervand, S. Dorsaz, T. Curie, G. Hagiwara, P. Gip, H. C. Heller and P. Franken (2010). "Separating the contribution of glucocorticoids and wakefulness to the molecular and electrophysiological correlates of sleep homeostasis." Sleep **33**(9): 1147-1157.

Muir, P., S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky and M. Gerstein (2016). "The real cost of sequencing: scaling computation to keep pace with data generation." Genome biology **17**: 53.

Munafò, M. R., B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware and J. P. A. Ioannidis (2017). "A manifesto for reproducible science." Nature Human Behaviour **1**(1).

Munger, S. C., N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti, D. A. Hinerfeld, K. L. Svenson, M. P. Keller, A. D. Attie, M. A. Hibbs, J. H. Graber, E. J. Chesler and G. A. Churchill (2014). "RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations." Genetics **198**(1): 59-73.

Narimatsu, H. (2017). "Gene-Environment Interactions in Preventive Medicine: Current Status and Expectations for the Future." International journal of molecular sciences **18**(2).

Nekrutenko, A. and J. Taylor (2012). "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility." Nature Reviews Genetics.

Neuner, S. M., B. P. Garfinkel, L. A. Wilmott, B. M. Ignatowska-Jankowska, A. Citri, J. Orly, L. Lu, R. W. Overall, M. K. Mulligan, G. Kempermann, R. W. Williams, K. M. O'Connell and C. C. Kaczorowski (2016). "Systems genetics identifies Hp1bp3 as a novel modulator of cognitive aging." Neurobiology of aging **46**: 58-67.

Nunn, C. L., D. R. Samson and A. D. Krystal (2016). "Shining evolutionary light on human sleep and sleep disorders." Evolution, medicine, and public health **2016**(1): 227-243.

Offord, C. (2018). "Scientists Continue to Use Outdated Methods." The Scientist **https://www.the-scientist.com/?articles.view/articleNo/51260/**.

Ollila, H. M., J. Kettunen, O. Pietiläinen, V. Aho, K. Silander, E. Kronholm, M. Perola, J. Lahti, K. Räikkönen, E. Widen, A. Palotie, J. G. Eriksson, T. Partonen, J. Kaprio, V. Salomaa, O. Raitakari, T. Lehtimäki, M. Sallinen, M. Härmä, T. Porkka-Heiskanen and T. Paunio (2014). "Genome-wide association study of sleep duration in the Finnish population." <u>Journal of Sleep Research</u> **23**(6): 609-618.

Palsson, B. and K. Zengler (2010). "The challenges of integrating multi-omic data sets." <u>Nature Chemical Biology</u> **6**(11).

Park, C. C., G. D. Gale, S. de Jong, A. Ghazalpour, B. J. Bennett, C. R. Farber, P. Langfelder, A. Lin, A. H. Khan, E. Eskin, S. Horvath, A. J. Lusis, R. A. Ophoff and D. J. Smith (2011). "Gene networks associated with conditional fear in mice identified using a systems genetics approach." <u>BMC Syst Biol</u> **5**: 43.

Peirce, J. L., L. Lu, J. Gu, L. M. Silver and R. W. Williams (2004). "A new set of BXD recombinant inbred lines from advanced intercross populations in mice." <u>BMC genetics</u> **5**: 7.

Philip, V. M., S. Duvvuru, B. Gomero, T. A. Ansah, C. D. Blaha, M. N. Cook, K. M. Hamre, W. R. Lariviere, D. B. Matthews, G. Mittleman, D. Goldowitz and E. J. Chesler (2010). "High-throughput behavioral phenotyping in the expanded panel of BXD recombinant inbred strains." <u>Genes, brain, and behavior</u> **9**(2): 129-159.

Philip, V. M., G. Sokoloff, C. L. Ackert-Bicknell, M. Striz, L. Branstetter, M. A. Beckmann, J. S. Spence, B. L. Jackson, L. D. Galloway, P. Barker, A. M. Wymore, P. R. Hunsicker, D. C. Durtschi, G. S. Shaw, S. Shinpock, K. F. Manly, D. R. Miller, K. D. Donohue, C. T. Culiat, G. A. Churchill, W. R. Lariviere, A. A. Palmer, B. F. O'Hara, B. H. Voy and E. J. Chesler (2011). "Genetic analysis in the Collaborative Cross breeding population." <u>Genome research</u> **21**(8): 1223-1238.

Picard, A., J. Soyer, X. Berney, D. Tarussio, S. Quenneville, M. Jan, E. Grouzmann, F. Burdet, M. Ibberson and B. Thorens (2016). "A Genetic Screen Identifies Hypothalamic Fgf15 as a Regulator of Glucagon Secretion." <u>Cell reports</u> **17**(7): 1795-1806.

Pinna, A., N. Soranzo, I. Hoeschele and A. de la Fuente (2011). "Simulating systems genetics data with SysGenSIM." <u>Bioinformatics (Oxford, England)</u> **27**(17): 2459-2462.

Plaisier, C. L., S. Horvath, A. Huertas-Vazquez, I. Cruz-Bautista, M. F. Herrera, T. Tusie-Luna, C. Aguilar-Salinas and P. Pajukanta (2009). "A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia." <u>PLoS genetics</u> **5**(9).

Platig, J., P. J. Castaldi, D. DeMeo and J. Quackenbush (2016). "Bipartite Community Structure of eQTLs." <u>PLoS computational biology</u> **12**(9).

Riedel, S. (2005). "Edward Jenner and the history of smallpox and vaccination." <u>Proceedings (Baylor University. Medical Center)</u> **18**(1): 21-25.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim (2015). "Methods of integrating data to uncover genotype-phenotype interactions." <u>Nature Reviews Genetics</u> **16**(2): 85-97.

Ryu, D., L. Mouchiroud, P. A. A. Andreux, E. Katsyuba, N. Moullan, A. A. Nicolet-Dit-Félix, E. G. Williams, P. Jha, G. Lo Sasso, D. Huzard, P. Aebischer, C. Sandi, C. Rinsch and J. Auwerx (2016). "Urolithin A induces mitophagy and prolongs lifespan in C. elegans and increases muscle function in rodents." <u>Nature medicine</u> **22**(8): 879-888.

Sabag, A., K. L. Way, S. E. Keating, R. N. Sultana, H. T. O'Connor, M. K. Baker, V. H. Chuter, J. George and N. A. Johnson (2017). "Exercise and ectopic fat in type 2 diabetes: A systematic review and meta-analysis." <u>Diabetes & metabolism</u> **43**(3): 195-210.

Sandoval-Motta, S., M. Aldana, E. Martínez-Romero and A. Frank (2017). "The Human Microbiome and the Missing Heritability Problem." <u>Frontiers in genetics</u> **8**: 80.

Sandve, G. K., A. Nekrutenko, J. Taylor and E. Hovig (2013). "Ten simple rules for reproducible computational research." <u>PLoS Comput Biol</u>.

Sansone, S.-A. A., P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. Ho Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios and W. Hide (2012). "Toward interoperable bioscience data." <u>Nature genetics</u> **44**(2): 121-126.

Schrodi, S. J., S. Mukherjee, Y. Shan, G. Tromp, J. J. Sninsky, A. P. Callear, T. C. Carter, Z. Ye, J. L. Haines, M. H. Brilliant, P. K. Crane, D. T. Smelser, R. C. Elston and D. E. Weeks (2014). "Genetic-

based prediction of disease traits: prediction is very difficult, especially about the future." Frontiers in genetics **5**: 162.

Schughart, K., C. Libert, S. consortium and M. J. Kas (2013). "Controlling complexity: the clinical relevance of mouse complex genetics." European journal of human genetics : EJHG **21**(11): 1191-1196.

Schupbach, T., I. Xenarios, S. Bergmann and K. Kapur (2010). "FastEpistasis: a high performance computing solution for quantitative trait epistasis." Bioinformatics **26**(11): 1468-1469.

Segura, V., B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long and M. Nordborg (2012). "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations." Nature genetics **44**(7): 825-830.

Seok, J., H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. López, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H. Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S. F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, R. G. Tompkins and L. and to Injury (2013). "Genomic responses in mouse models poorly mimic human inflammatory diseases." Proceedings of the National Academy of Sciences of the United States of America **110**(9): 3507-3512.

Shah, A. N., C. F. Davey, A. C. Whitebirch, A. C. Miller and C. B. Moens (2015). "Rapid reverse genetic screening using CRISPR in zebrafish." Nature methods **12**(6): 535-540.

Shan, Z., H. Ma, M. Xie, P. Yan, Y. Guo, W. Bao, Y. Rong, C. L. Jackson, F. B. Hu and L. Liu (2015). "Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies." Diabetes care **38**(3): 529-537.

Shepherd, J. D. (2012). "Memory, plasticity and sleep - A role for calcium permeable AMPA receptors?" Frontiers in molecular neuroscience **5**: 49.

Shin, D.-L. L., A. K. Pandey, J. D. Ziebarth, M. K. Mulligan, R. W. Williams, R. Geffers, B. Hatesuer, K. Schughart and E. Wilk (2014). "Segregation of a spontaneous Klrd1 (CD94) mutation in DBA/2 mouse substrains." G3 (Bethesda, Md.) **5**(2): 235-239.

Simon, M. M., S. Greenaway, J. K. White, H. Fuchs, V. Gailus-Durner, S. Wells, T. Sorg, K. Wong, E. Bedu, E. J. Cartwright, R. Dacquin, S. Djebali, J. Estabel, J. Graw, N. J. Ingham, I. J. Jackson, A. Lengeling, S. Mandillo, J. Marvel, H. Meziane, F. Preitner, O. Puk, M. Roux, D. J. Adams, S. Atkins, A. Ayadi, L. Becker, A. Blake, D. Brooker, H. Cater, M.-F. F. Champy, R. Combe, P. Danecek, A. di Fenza, H. Gates, A.-K. K. Gerdin, E. Golini, J. M. Hancock, W. Hans, S. M. Hölter, T. Hough, P. Jurdic, T. M. Keane, H. Morgan, W. Müller, F. Neff, G. Nicholson, B. Pasche, L.-A. A. Roberson, J. Rozman, M. Sanderson, L. Santos, M. Selloum, C. Shannon, A. Southwell, G. P. Tocchini-Valentini, V. E. Vancollie, H. Westerberg, W. Wurst, M. Zi, B. Yalcin, R. Ramirez-Solis, K. P. Steel, A.-M. M. Mallon, M. H. de Angelis, Y. Herault and S. D. Brown (2013). "A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains." Genome biology **14**(7).

Singh, A., B. Ganapathysubramanian, A. K. Singh and S. Sarkar (2016). "Machine Learning for High-Throughput Stress Phenotyping in Plants." Trends in plant science **21**(2): 110-124.

Spiegel, K., K. Knutson, R. Leproult, E. Tasali and E. Van Cauter (2005). "Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes." Journal of applied physiology (Bethesda, Md. : 1985) **99**(5): 2008-2019.

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

Spiro, J. (2005). "Introduction: Sleep." Nature **437**(7063).

Stančáková, A. and M. Laakso (2016). "Genetics of Type 2 Diabetes." Endocrine development **31**: 203-220.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha and G. E. Robinson (2015). "Big Data: Astronomical or Genomical?" PLoS biology **13**(7).

Sun, X., Q. Lu, S. Mukherjee, P. K. Crane, R. Elston and M. D. Ritchie (2014). "Analysis pipeline for the epistasis search - statistical versus biological filtering." Front Genet **5**: 106.

Sun, Z., A. Bhagwate, N. Prodduturi, P. Yang and J.-P. A. Kocher (2017). "Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations." Briefings in bioinformatics **18**(6): 973-983.

Sunagawa, G. A., H. Séi, S. Shimba, Y. Urade and H. R. Ueda (2013). "FASTER: an unsupervised fully automated sleep staging method for mice." Genes to Cells **18**(6): 502-518.

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen and C. von Mering (2015). "STRING v10: protein-protein interaction networks, integrated over the tree of life." <u>Nucleic acids research</u> **43**(Database issue): 52.

Tafti, M., B. Petit, D. Chollet, E. Neidhart, F. de Bilbao, J. Z. Kiss, P. A. Wood and P. Franken (2003). "Deficiency in short-chain fatty acid beta-oxidation affects theta oscillations during sleep." <u>Nature genetics</u> **34**(3): 320-325.

Takao, K. and T. Miyakawa (2015). "Genomic responses in mouse models greatly mimic human inflammatory diseases." <u>Proceedings of the National Academy of Sciences of the United States of America</u> **112**(4): 1167-1172.

Talukdar, H. A., H. Foroughi Asl, R. K. Jain, R. Ermel, A. Ruusalepp, O. Franzén, B. A. Kidd, B. Readhead, C. Giannarelli, J. C. Kovacic, T. Ivert, J. T. Dudley, M. Civelek, A. J. Lusis, E. E. Schadt, J. Skogsberg, T. Michoel and J. L. Björkegren (2016). "Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease." <u>Cell systems</u> **2**(3): 196-208.

Taneera, J., S. Lang, A. Sharma, J. Fadista, Y. Zhou, E. Ahlqvist, A. Jonsson, V. Lyssenko, P. Vikman, O. Hansson, H. Parikh, O. Korsgren, A. Soni, U. Krus, E. Zhang, X.-J. J. Jing, J. L. Esguerra, C. B. Wollheim, A. Salehi, A. Rosengren, E. Renström and L. Groop (2012). "A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets." <u>Cell metabolism</u> **16**(1): 122-134.

Taylor, B. A., C. Wnek, B. S. Kotlus, N. Roemer, T. MacTaggart and S. J. Phillips (1999). "Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps." <u>Mammalian genome : official journal of the International Mammalian Genome Society</u> **10**(4): 335-348.

Theodoratou, E., M. Timofeeva, X. Li, X. Meng and J. P. A. P. A. Ioannidis (2017). "Nature, Nurture, and Cancer Risks: Genetic and Nutritional Contributions to Cancer." <u>Annual review of nutrition</u> **37**: 293-320.

Thimgan, M. S., Y. Suzuki, L. Seugnet, L. Gottschalk and P. J. Shaw (2010). "The perilipin homologue, lipid storage droplet 2, regulates sleep homeostasis and prevents learning impairments following sleep loss." <u>PLoS biology</u> **8**(8).

Threadgill, D. W., D. R. Miller, G. A. Churchill and F. P. de Villena (2011). "The collaborative cross: a recombinant inbred mouse population for the systems genetic era." <u>ILAR journal</u> **52**(1): 24-31.

Tini, G., L. Marchetti and P.-C. in … (2017). "Multi-omics integration—a comparison of unsupervised clustering methodologies." <u>Briefings in …</u>.

Tobler, I. and B.-A. A. and clinical (1986). "Sleep EEG in the rat as a function of prior waking." <u>Electroencephalography and clinical …</u>.

Tononi, G. and C. Cirelli (2003). "Sleep and synaptic homeostasis: a hypothesis." <u>Brain research bulletin</u> **62**(2): 143-150.

Tononi, G. and C. Cirelli (2006). "Sleep function and synaptic homeostasis." <u>Sleep medicine reviews</u> **10**(1): 49-62.

Tononi, G. and C. Cirelli (2014). "Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration." <u>Neuron</u> **81**(1): 12-34.

Toth, L. A. and P. Bhargava (2013). "Animal models of sleep disorders." <u>Comparative medicine</u> **63**(2): 91-104.

Trerotola, M., V. Relli, P. Simeone and S. Alberti (2015). "Epigenetic inheritance and the missing heritability." <u>Human genomics</u> **9**: 17.

Tyler, A. L., B. Ji, D. M. Gatti, S. C. Munger, G. A. Churchill, K. L. Svenson and G. W. Carter (2017). "Epistatic Networks Jointly Influence Phenotypes Related to Metabolic Disease and Gene Expression in Diversity Outbred Mice." <u>Genetics</u> **206**(2): 621-639.

Upton, A., O. Trelles, J. A. Cornejo-Garcia and J. R. Perkins (2015). "Review: High-performance computing to detect epistasis in genome scale data sets." <u>Brief Bioinform</u>.

Urade, Y. and O. Hayaishi (2011). "Prostaglandin D2 and sleep/wake regulation." <u>Sleep medicine reviews</u> **15**(6): 411-418.

Van den Berge, K., C. Soneson, M. D. Robinson and L. Clement (2017). "stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage." <u>Genome biology</u> **18**(1): 151.

Vanderlinden, L. A., L. M. Saba, K. Kechris, M. F. Miles, P. L. Hoffman and B. Tabakoff (2013). "Whole brain and brain regional coexpression network interactions associated with predisposition to alcohol consumption." <u>PloS one</u> **8**(7).

Vasilevsky, N. A., J. Minnier, M. A. Haendel and R. E. Champieux (2017). "Reproducible and reusable research: are journal data sharing policies meeting the mark?" PeerJ **5**.

Viola, A. U., S. N. Archer, L. M. James, J. A. Groeger, J. C. Lo, D. J. Skene, M. von Schantz and D.-J. J. Dijk (2007). "PER3 polymorphism predicts sleep structure and waking performance." Current biology : CB **17**(7): 613-618.

Wang, B., A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg (2014). "Similarity network fusion for aggregating data types on a genomic scale." Nature Methods **11**(3).

Wang, L. and T. Michoel (2017). "Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data." PLoS computational biology **13**(8).

Wang, X., A. K. Pandey, M. K. Mulligan, E. G. Williams, K. Mozhui, Z. Li, V. Jovaisaite, L. D. Quarles, Z. Xiao, J. Huang, J. A. Capra, Z. Chen, W. L. Taylor, L. Bastarache, X. Niu, K. S. Pollard, D. C. Ciobanu, A. O. Reznik, A. V. Tishkov, I. B. Zhulin, J. Peng, S. F. Nelson, J. C. Denny, J. Auwerx, L. Lu and R. W. Williams (2016). "Joint mouse-human phenome-wide association to test gene function and disease risk." Nature communications **7**: 10464.

Ward, M. R., R. Schmieder, G. Highnam and D. Mittelman (2013). "Big data challenges and opportunities in high-throughput sequencing." Systems Biomedicine **1**(1): 29-34.

Wilkinson, K. and C. Shapiro (2012). "Nonrestorative sleep: Symptom or unique diagnostic entity?" Sleep Medicine **13**(6): 561-569.

Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data **3**: 160018.

Wilkinson, M. D., R. Verborgh, L. Santos, T. Clark, M. A. Swertz, F. D. L. Kelpin, A. J. G. Gray, E. A. Schultes, E. M. Mulligen and P. Ciccarese (2017). "Interoperability and FAIRness through a novel combination of Web technologies." PeerJ Computer Science **3**.

Williams, C. R., A. Baccarella, J. Z. Parrish and C. C. Kim (2017). "Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq." BMC bioinformatics **18**(1): 38.

Williams, E. G., Y. Wu, P. Jha, S. Dubuis, P. Blattmann, C. A. Argmann, S. M. Houten, T. Amariuta, W. Wolski, N. Zamboni, R. Aebersold and J. Auwerx (2016). "Systems proteomics of liver mitochondria function." Science (New York, N.Y.) **352**(6291).

Wu, Y., E. G. Williams, S. Dubuis, A. Mottis, V. Jovaisaite, S. M. Houten, C. A. Argmann, P. Faridi, W. Wolski, Z. Kutalik, N. Zamboni, J. Auwerx and R. Aebersold (2014). "Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population." Cell **158**(6): 1415-1430.

Xie, X., T. Dumas, L. Tang, T. Brennan and R.-T. research (2005). "Lack of the alanine–serine–cysteine transporter 1 causes tremors, seizures, and early postnatal death in mice." Brain research.

Xie, Y. (2014). "knitr: a comprehensive tool for reproducible research in R." Implement Reprod Res **1**: 20.

Yeadon, J. (2014). "6 FACTS YOU SHOULD KNOW ABOUT NNT, C57BL/6J AND DIABETES." Jax Blog Post.

Yi, L., H. Pimentel, N. L. Bray and P.-L. bioRxiv (2017). "Gene-level differential analysis at transcript-level resolution." bioRxiv.

Zeggini, E. (2011). "Next-generation association studies for complex traits." Nature genetics **43**(4): 287-288.

Zhang, X., S. Huang, F. Zou and W. Wang (2011). "Tools for efficient epistasis detection in genome-wide association study." Source code for biology and medicine **6**(1): 1.

Zhang, X., F. Zou and W. Wang (2008). "FastANOVA: an Efficient Algorithm for Genome-Wide Association Study." KDD : proceedings. International Conference on Knowledge Discovery & Data Mining: 821-829.

Zhang, X., F. Zou and W. Wang (2009). "FastChi: an efficient algorithm for analyzing gene-gene interactions." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 528-539.

Zhang, Y., Y. Li, M. W. Niepel, Y. Kawano, S. Han, S. Liu, A. Marsili, P. R. Larsen, C.-H. H. Lee and D. E. Cohen (2012). "Targeted deletion of thioesterase superfamily member 1 promotes energy expenditure and protects against obesity and insulin resistance." Proceedings of the National Academy of Sciences of the United States of America **109**(14): 5417-5422.

Zierer, J., C. Menni, G. Kastenmüller and T. D. Spector (2015). "Integration of 'omics' data in aging research: from biomarkers to systems biology." Aging cell **14**(6): 933-944.

Zuk, O., E. Hechter, S. R. Sunyaev and E. S. Lander (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." Proceedings of the National Academy of Sciences of the United States of America **109**(4): 1193-1198.

Appendix 1: Web interface usage

BXD web interface informations:

Home page displays the information for the NREM sleep gain during the 24 hours (in four 6- hour intervals) after sleep deprivation.

(A) Search parameters - Tissue & cutoff: Sets the parameters for selection of genes in the specified tissue (cortex or liver) according to the specified cutoff for the pearson correlation coefficient and/ or a *cis-e*QTL q-value.

(B) Search parameters - Phenotypes or genes: One can search by phenotype(s) or by gene(s). A search by phenotype will output the genes that correlate to the submitted

phenotype(s) with a correlation coefficient larger than the cutoff set in A). Search by gene(s) will output the most correlated phenotype(s) to the submitted gene(s).

(C) Results - Heatmap: The results are displayed as heatmaps according to the selections made under Search parameters in **A** and **B**.

(D) Results - Hiveplots: For each phenotype present in the heatmap in **C**, the corresponding hiveplot is displayed.

(E) Hiveplot-Filtering: Filtering options specific for the hive plots.

(F) Results – Table: Tables list all genes, markers, and metabolites and their relations for each hive-plot in **D**.

Hiveplots details:

Each plot represents one EEG/behavioral phenotype and its underlying associated molecular network; i.e., only the genes and metabolites strongly associated with a given phenotype are displayed. Each hiveplot is composed of 3 radial axes containing the molecular data with nodes assigned to the 2 bottom axis for genes expressed in the cortex (in blue) and liver (in red) and nodes on the vertical axis (in yellow) represent metabolites. On top, we added a separate 'genetic' axis (white) containing the genotypes. The node position on the 3 (molecular) radial

axes was determined by the response to sleep deprivation; i.e., molecules closer to the center were down-regulated more strongly, while more upregulated genes/metabolites were closer to the axis' edge. Edges connecting nodes represent positive/negative correlations (red/blue, respectively) between molecules using expression values. Genetic markers linked to genes by *e*QTLs connect the genetic and molecular space.

Notes:

- URLs can be easily send to share search options and results among users
- History of your search is kept during your working session
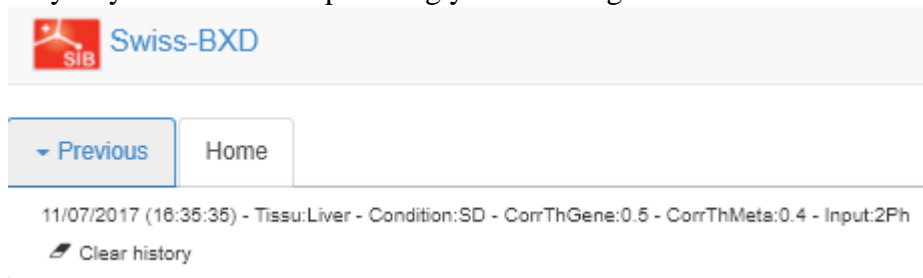
Tutorial 1. Single phenotype mining

In this section, we assume we are interested in a single sleep phenotype only. We will approach the data by multiple filtering steps.

1.   **Visual comparison of hiveplot:**
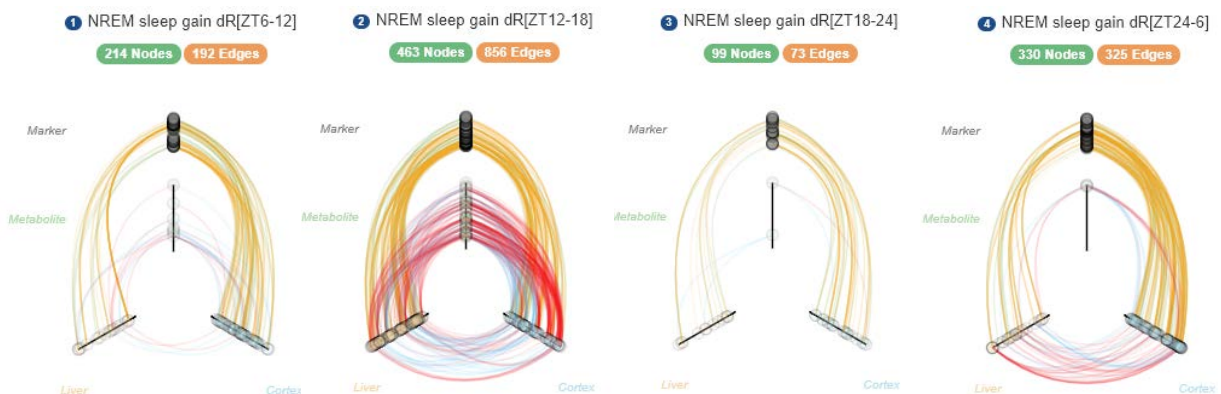
Hive-plots for NREM sleep gain in 4 consecutive 6h recovery intervals after the sleep deprivation ending at ZT6. Compared to the other 3 intervals, NREM gain was strongly associated with metabolites during the 2nd 6h interval; i.e., the 1st 6h of the recovery dark

period (ZT12-18). We will focus on this phenotype. These hiveplots (and phenotypes) are set as the default on the homepage.

2.   Remove other phenotypes

Only the NREM sleep gain between ZT12-18 is kept. Other phenotypes can be ⊗ removed clicking

To apply filtering selection, Go press

3.   Heatmap:

Now we only have a single hiveplot that we can explore further. The heatmap **(C)** displays all genes for which the expression is correlated to this sleep phenotype (note that liver is the default tissue). You can explore genes in cortex by changing the filtering parameters (see below). You can sort genes by correlation strength by clicking on the phenotype name in the heatmap. A similar heatmap can be displayed for the metabolites.
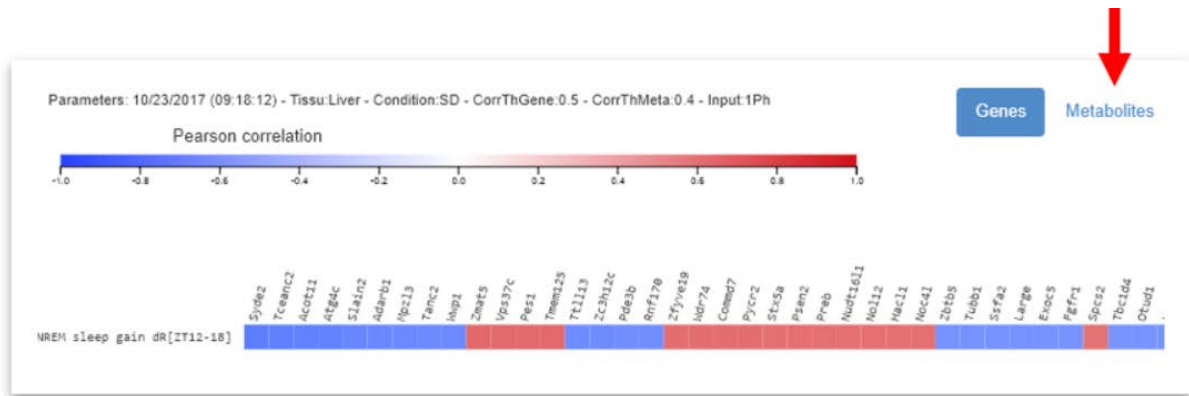
4. Increase filtering stringency:

Default thresholds are: absolute pearson r equal or above 0.5 for genes and absolute pearson r equal or above 0.4 for metabolites.

Given the high number of genes displayed in the heatmap, the pearson r coefficient cutoff can be increased under Search parameters **(A)**. Here we select an absolute pearson r of 0.6 for genes and 0.5 for metabolites. The user might be interested in genes that are driven by genetic variation (*cis-e*QTL). In the example, genes are also [Go] filtered according to *cis-e*QTL association with an FDR adjusted p-value equal or below 1e-5. Apply the filtering option by clicking on

Heatmap results for liver genes: Heatmap results for cortex genes:

Notes: More genes are displayed in the hiveplot than in the heatmap. The heatmap contains all the genes that pass the search parameters selected. The hiveplot is filtered according to the following rules: i) the genes and metabolites must pass the pearsons r cutoff, ii) the nodes (genes/metabolites/ markers) that are unconnected (no edges) are removed, and iii) the eQTL threshold is not applied in order to keep cross-tissue associations.

5.   Save and highlight potential targets:

Once you applied your filtering parameters, you can save a target of interest to highlight it within the hiveplot. Double-clicking on genes within the heatmap will select the genes to highlight. A new tab will be created for each target saved. You can remove these saved targets

by again double clicking the gene within the heatmap or pressing

Here we highlight all genes that passed our filtering parameters:

These genes can be highlighted in the filtered hiveplot (**1**) or the user can reset the filtering parameters (**2**).
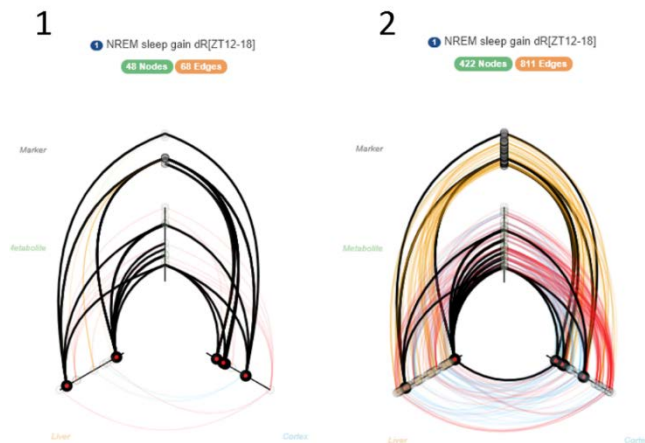
6.   Hiveplot filtering:

Further filtering can be applied to the hiveplot. We can select for genes that have a minimum number of links (edges), e.g. 6 edges on our unfiltered hiveplot.

Acot11 is the only highlighted gene that is kept after applying these filtering options. Other options are available, such as keeping only top differentially expressed genes in the selection:

only ≠ expressed nodes ✔

7.   Hiveplot investigation:

The hiveplots can be investigated using 'mouseover' the different edges and nodes. This will display the node name (gene name, metabolite name, or marker name), or the nodes connected to an edge. All nodes and edges within the displayed hiveplot are listed in the table **(F)**.

The table gives the following information: i) the type of node: genes in cortex (blue), genes in liver (orange), metabolites (green), and markers (white), ii) effects of sleep deprivation on genes and metabolites are indicated by fading color: upregulated (solid color on top, fading down to white), downregulated (solid color on bottom)

up-regulated   down-regulated

Acot11   Ovgp1 **1** rs28135130 **1** rs13477883 **1** rs28150397 **1** PC_aa_C36_2 **1** PC_aa_C36_5 **1** PC_aa_C38_0 **1** PC_ae_C34_2 **1**
PC_ae_C36_2 **1** PC_ae_C36_3 **1** PC_ae_C38_2 **1** PC_ae_C38_3 **1** PC_ae_C40_2 **1** PC_ae_C40_3 **1** PC_ae_C40_4 **1**
PC_ae_C40_6 **1** PC_ae_C42_5 **1**

Expression of *Acot11* in liver is downregulated after sleep deprivation and is highly correlated with 13 metabolites, one cortical transcript, and 3 SNPs. The correlation was positive for all (red symbol with phenotype number). Negative correlations would have been indicated in blue.

8.   Target investigation:
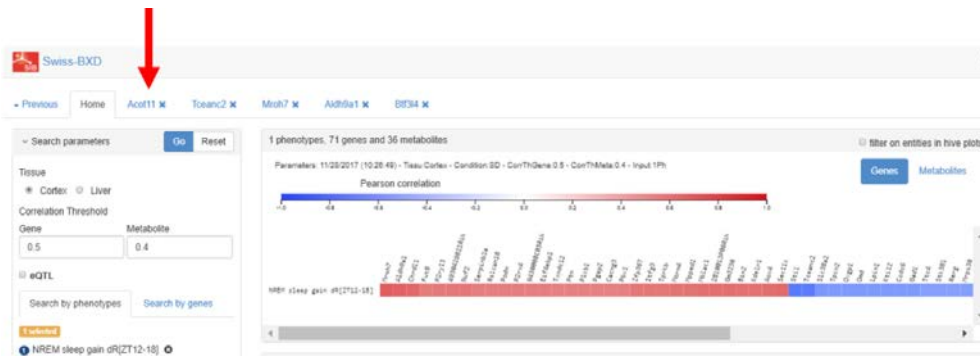
The tab containing the highlighted genes shows additional information concerning gene function, the 5 genes with the highest partial-correlation, within the same tissue, across-tissue, and metabolite. The tab contains the gene name (1), the Uniprot description of the gene (2), Synonyms (3), a Uniprot and MGI access ID links (4) and related GO terms (5).

Here the network of the highlighted gene (*Acot11*) with the 5 top genes partially-correlated to within the same tissue (blue), 2 top genes partially-correlated to the highlighted gene within the other tissue (orange) and the top correlated metabolite (green). Red edges indicate positive correlation and blue edges indicates negative correlation. The size of the nodes indicates the correlation coefficient with the phenotype.

Cortex - NREM sleep gain during recovery [ZT12-18]

Liver - NREM sleep gain during recovery [ZT12-18]
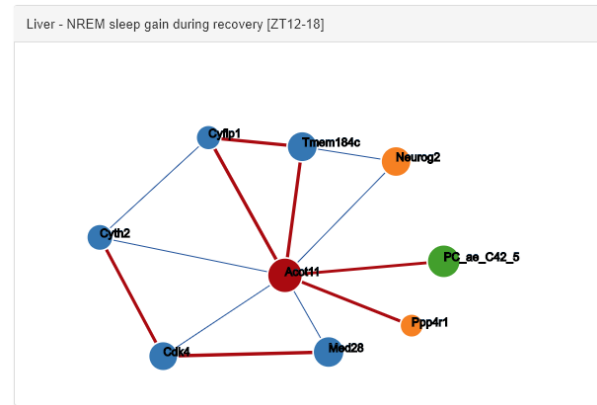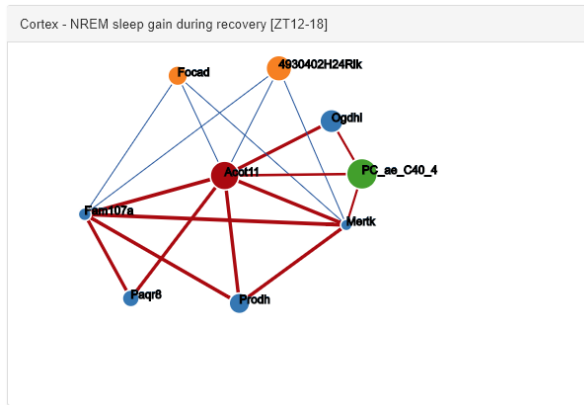
Tutorial 2. Gene search:

The interface allows you also to search for specific genes associated with a sleep phenotype at a specific correlation cutoff.

If you previously made some filtering and research, you can reset your settings by pressing the RESET button.
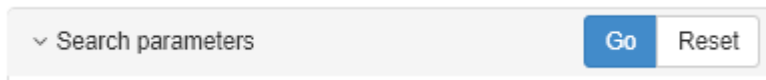
Then select the relevant "search by genes" area
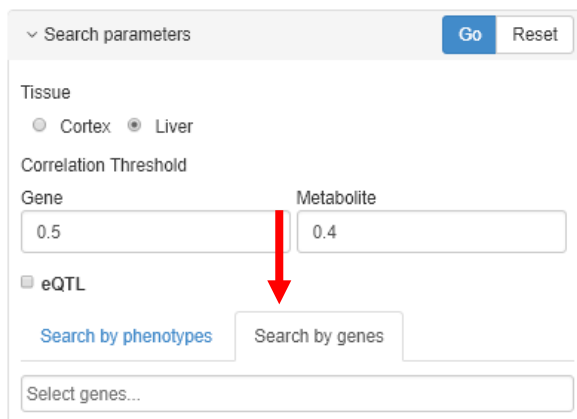
Below are a 2 examples of sleep related genes:

1. [GWAS search for Per2](#): *Per2* is a well known circadian gene. It is associated in the BXD set with NREM sleep gain after SD [ZT12-18] ([Per2 BXD search)](#), and is associated with 2 metabolites: PC aa C36:3 & PC ae C40:0
2. *Sik3* KO mice were associated with the increase total of NREM sleep (minutes/24h) and a decrease in total wake in mouse by (Funato, Miyoshi et al. 2016) using random genetic screening. With our extended database, we found that: *Sik3* cortical expression in baseline was correlated (r=0.55) with the wake amount at the beginning of light (BXDw24bsl.h1).

Cortical Sik3 in baseline
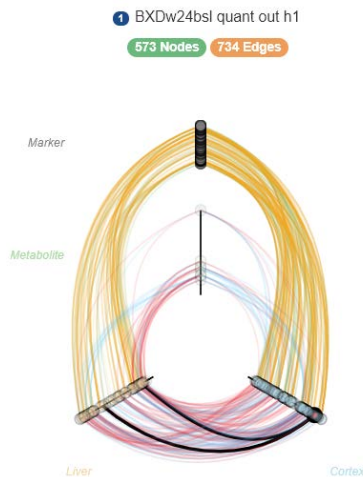
Tutorial 3. Edges filtering:

Hiveplot filtering parameters allow to recover common edges among hiveplots or to exclude common edges.

The first example shows how to exclude common edges:

1.   First we [select](#) phenotypes of EEG power after SD for all the bands.

2.   Here we are interested in slow delta power after sleep deprivation (δ1: 1.0-2.25Hz) during NREM sleep, and we want to remove all genes, metabolites that are present also

within other bands. Therefore, we exclude other phenotypes using the "x" button in the hiveplot filtering options (**E**):

3.   Finally, we keep only genes differentially expressed and genes connected to 3 other elements.

4.   Remaining known genes are *Strada*, *Mapk4* and *Fhit* in cortex and *Susd4* in liver.

These genes thus correlate only to slow delta and not to other EEG frequency bands of the EEG spectra during NREM sleep after sleep deprivation.

Inspection of *Strada* in cortex informed us that *Homer1* is among the 5 closest genes of *Strada*. *Homer1* has been implicated in the homeostatic regulation of EEG delta power.

Cortex_BXD_NREMSspecSD.quant.out.del1_r
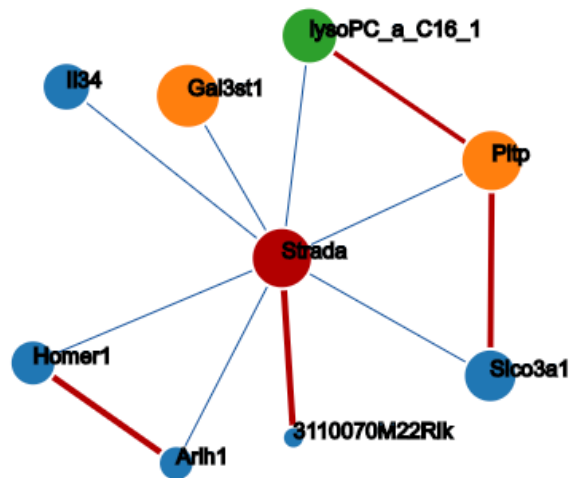
The second example shows how to inspect common edges:

1.   In this case we are interested in genes common to slow delta gain (r/b) during NREM sleep correlated within other delta bands.

1 NREM sleep EEG power aSD [δ1;1-2.25 Hz] ⊗
2 NREM sleep EEG power aSD [δ2;2.5-4.25 Hz] ⊗
3 NREM sleep EEG power: r/b[δ1;1-2 25 Hz] ⊗
4 NREM sleep EEG power: r/b[δ2;2 5-4 25 Hz] ⊗

2.   Using the "in" button in the hiveplot filtering option **(E)** for slow delta gain after SD, the common edges and nodes within other bands will be kept. We also filter for genes differentially expressed after SD.
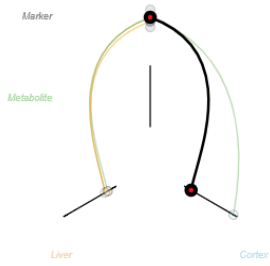
phenotypes

in ex 1 NREM sleep EEG power aSD [δ1;1-2.25 Hz]
in ex 2 NREM sleep EEG power aSD [δ2;2.5-4.25 Hz]
in ex 3 **NREM sleep EEG power: r/b[δ1;1-2 25 Hz]**
in ex 4 NREM sleep EEG power: r/b[δ2;2 5-4 25 Hz]

3.   Inspecting the table (F), we see that only 1 gene is connected to 3 other bands. *Wrn* is present in the hiveplot 1, 2, and 3 [NREM sleep EEG power aSD δ1, NREM sleep power aSD δ2 and NREM sleep EEG power gain (r/b) δ1].

Cortex

Prss8	rs32616308 **1 3**
Wrn	rs51740715 **1 2 3**
Tspan11	rs30384178 **2 3** rs30955623 **2 3** rs32957022 **2 3** rs36187574 **2 3** rs50472734 **2 3**
Lrp8	novel_6_130057653 **3** rs51691312 **3**
Eif4ebp1	rs13464206 **3**
Chmb3	rs13479873 **3** rs33257768 **3**
Fut10	rs36861932 **3** rs32818574 **3**
Krt1	PC_ss_C30_2 **3**
Fam211a	rs13481062 **3** Affy_10352113 **3** rs13474599 **3** rs13476287 **3** rs29843649 **3** rs8280764 **3**
Tbcd	novel_1_173250231 **3 4** rs8255229 **3 4**
Cenpp	PC_ss_C30_2 **3**

1 NREM sleep EEG power aSD [δ1;1-2.25 Hz]   **2** NREM sleep EEG power aSD [δ2;2.5-4.25 Hz]   **3** NREM sleep EEG power: r/b[δ1;1-2 25 Hz]

16 Nodes   7 Edges   14 Nodes   6 Edges   71 Nodes   46 Edges

157
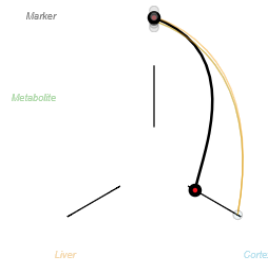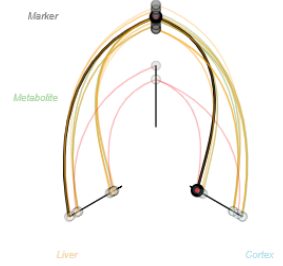
Annex 1: The epigenetic consequences of sleep loss.

<u>Preliminary results:</u>

In this project, we asked which regulatory elements were involved in the response to sleep deprivation. The cortical chromatin accessibility and transcriptomic landscape was monitored in C57BL/6J mice using ATAC-seq and RNA-seq. We identified a total of 215'045 ATAC-positive regions (peaks) over all time points using Macs2. We examined the differential accessibility at each time point (Figure 9), 1542 peaks were differentially accessible in SD at ZT3 (T27, after 3h SD), 1906 in SD at ZT6 (T30, end of 6h SD), and 678 at ZT12 (T36, after 6h recovery.
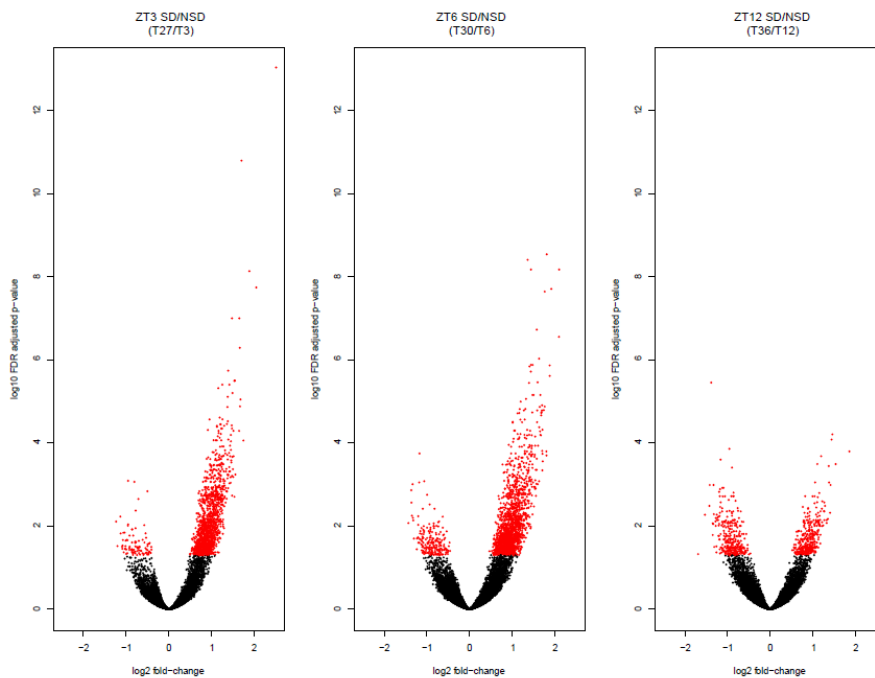

Figure 9 : Differential chromatin accessibility after sleep deprivation

Volcano plot of ATAC differential activity at T27, T30 and T36 (sleep deprivation is performed from T24-T30). Red points: Significant differentially accessible peak

To associate ATAC-seq peaks with target genes, we looked for gene-peak correlations within topologically associating domains (TADs) defined in cortex. Genes associated with slow delta power gain after SD (*Wrn*) and fast delta power after SD (*Kif16b*) have a temporal expression that is significantly associated with a temporal chromatin accessibility (Figure 10).
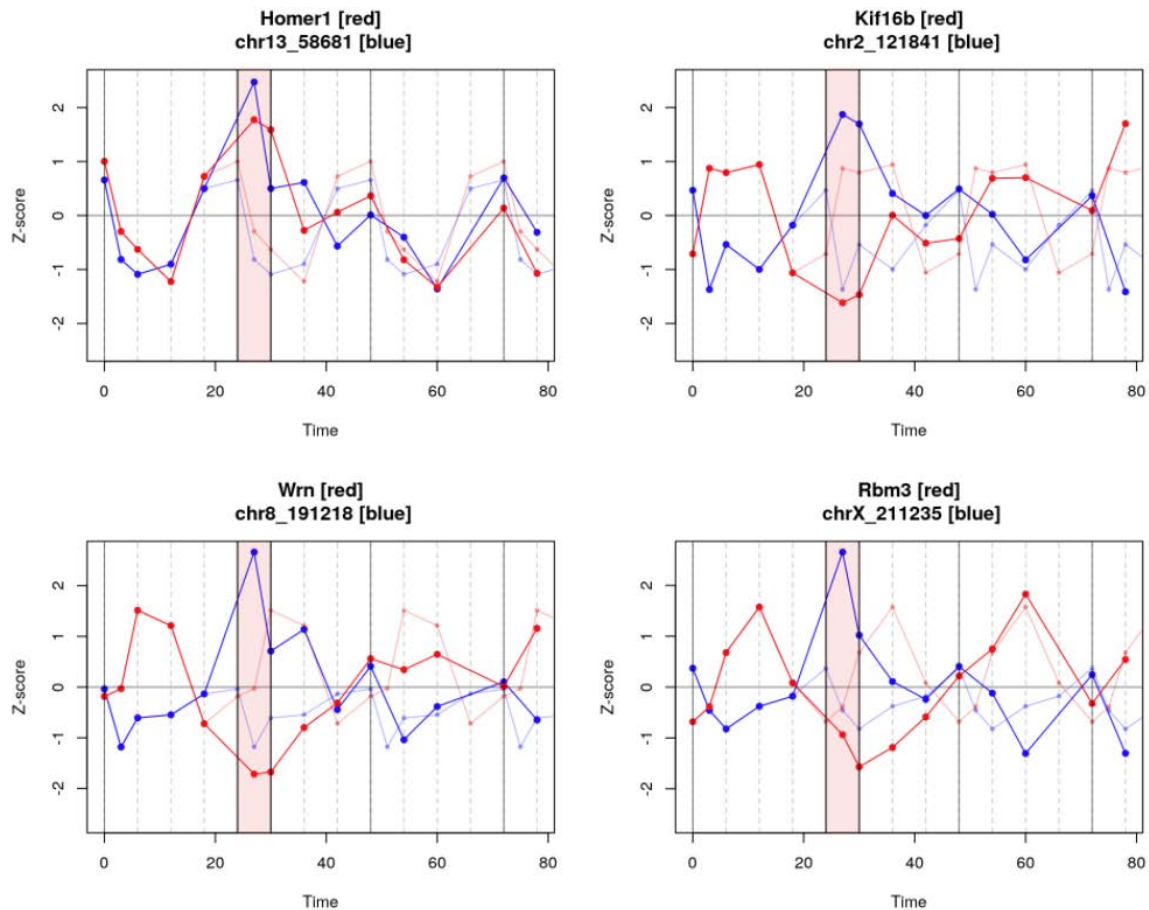

Figure 10 : Example of temporal patterns association between ATAC-seq peaks and mRNA

Temporal patterns of correlated gene-peak pairs. Red curves: gene; blue curves: ATAC peak. Red shaded box: SD, shade lines are replicated baseline values (Time 0-24).

Contribution:

I processed and analysis ATAC-seq data, integrated ATAC-seq peaks and mRNA expression, participated to the data interpretation and figure generation.

Annex 2: Fgf15 drives glucagon secretion in BXD

Publication:

Picard, A., J. Soyer, X. Berney, D. Tarussio, S. Quenneville, M. Jan, E. Grouzmann, F. Burdet, M. Ibberson and B. Thorens (2016). *A Genetic Screen Identifies Hypothalamic Fgf15 as a Regulator of Glucagon Secretion.* Cell reports 17(7): 1795-1806.

Contribution:

I mapped glucagon secretion phenotype and generated QTL figures.