PhD-FSTC-2015-43

The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on the 28/08/2015 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN BIOLOGIE

by

Mafalda Sofia RAMOS GALHARDO

Born on the 1st of November 1988 in Santarém (Portugal)

INTEGRATED ANALYSIS OF TRANSCRIPT-LEVEL REGULATION OF HUMAN ADIPOGENESIS AND CELL TYPE-SELECTIVE DISEASE ASSOCIATION OF HIGH REGULATORY LOAD GENES

Dissertation Defense Committee

Dr. Thomas Sauter, dissertation supervisor

Professor, Université du Luxembourg

Dr. Jean-Luc Bueb, chairman

Professor, Université du Luxembourg

Dr. Francisco Pinto

Assist. Professor, University of Lisbon

Dr. Yvan Devaux

Medical Doctor, Luxembourg Institute of Health

Dr. Alexander Skupin

Research Scientist, Luxembourg Centre for Systems Biomedicine

Affidavit

I hereby confirm that the PhD thesis entitled "Integrated analysis of transcriptional regulation of human adipogenesis and cell-type selective disease association of high regulatory load genes" has been written independently and without any other sources than cited.

Luxembourg, 27.07.2015

Acknowledgments

My profound gratitude goes to Prof. Thomas Sauter, my supervisor, for his positive and encouraging character, together with his interest for all problems one poses him. He could always explain problems in a simple and understandable way.

I am deeply thankful to Dr. Lasse Sinkkonen, for all daily advise, rich ideas and enthusiasm.

I am thankful to both for being there through the uncertain times when doubt and disbelief fell my mind. For everything, thank you Thomas and thank you Lasse.

To Dr. Karsten Hiller, I am very thankful for the support and problem solving character. Equally, I am also thankful to Dr. Thekla Cordes, Dr. Christian Jäger, Dr. Daniel Weindl, Jenny Ghelfi and all other members from the metabolomics group at the LCSB, for their constant help and welcoming environment.

To Dr. Alexander Skupin I express my gratitude. He is a great example encouraging young scientists to continue pursuing research questions.

My thankfulness goes to Jake Lin as well. His dedication and simplicity amaze me constantly and I could only wish more people would have similar ways.

I am uttermost grateful to Mr. Hippert, for his kind gift through the Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg.

I thank the friends who accompanied me through this journey and were able to cheer me up, among them, Joana Corte-Real, Adina Lupu, Swagatika Sahoo, Maike Aurich, Komal Baig, Maiti Lommel, Pit Ullmann, Isabelle Naegelen, Giulia Cesi, Andreas Zimmer and Susann Mudrich (to whom I also thank for the LaTeX prime!). A big thank you to the HW group, for their good vibe! And to all of those who directly or indirectly helped me in some way.

Most of all, I am deeply thankful to all my family, for their effort to educate me, believing and encouraging me at all times. To my dear mother *Gloríssima* for her constant presence.

Para a Mãe Glória

Os meus olhos são uns olhos.
E é com esses olhos uns
que eu vejo no mundo escolhos
onde outros, com outros olhos,
não vêem escolhos nenhuns.

Quem diz escolhos diz flores.
De tudo o mesmo se diz.
Onde uns vêem luto e dores,
uns outros descobrem cores
do mais formoso matiz.

Pelas ruas e estradas
onde passa tanta gente,
uns vêem pedras pisadas,
mas outros gnomos e fadas
num halo resplandescente.

Inútil seguir vizinhos,
querer ser depois ou ser antes.
Cada um é seus caminhos.
Onde Sancho vê moinhos
D. Quixote vê gigantes.

Vê moinhos? São moinhos.
Vê gigantes? São gigantes.

Impressão digital, António Gedeão, in "Movimento Perpétuo", 1956

Contents

Abbreviations

3’UTR	3’-untranslated-region
AD	Alzheimer’s disease
Ago	Argonaute protein family
BAT	brown adipose tissue
BCAAs	branched chain amino acids
BMI	body mass index
cAMP	cyclic adenosine monophosphate
CBM	constraint-based modelling
C/EBP	CCAAT/enhancer-binding protein
ChIP	chromatin immunoprecipitation
coA	coenzyme A
CTD	Comparative Toxicogenomics Database
DBD	DNA binding domain
DEGs	differentially expressed genes
ER	endoplasmic reticulum
FAs	fatty acids
FDA	US Food and Drug Administration
FFLs	feedback and feedforward loops
FVA	flux variability analysis
GO	gene ontology
GWAS	genome-wide association studies
HRL	high regulatory load
HPRD	Human Protein Reference Database
HUVEC	human umbilical vein endothelial cells
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LBD	ligand binding domain
LXR	liver X receptor
MILP	mixed integer linear programming
miRNA	microRNA, also miR
MSCs	mesenchymal stem cells
NEFAs	non-esterified fatty acids
NRs	nuclear receptors
PPAR	peroxisome proliferator-activated receptor
PPREs	peroxisome proliferator-activated receptor response elements

pri-miRNA	primary-miRNA
PTMs	post-translational modifications
RBD	RNA-binding-domain
RBPs	RNA-binding-proteins
REs	response elements
RISC	RNA-induced silencing complex
RRM	RNA-recognition-motif
RXR	retinoid X receptor
SBML	Systems Biology Markup Language
SE	super-enhancer
SGBS	Simpson-Golabi-Behmel syndrome
SOMs	self-organizing maps
SSD	signal-sensing-domain
T2DM	type 2 diabetes mellitus
TAD	trans-activating-domain
TCA	tricarboxilic acid cycle
TF	transcription factor
TSS	transcription start site
TZD	thiazolidinedione
WAT	white adipose tissue
WHO	World Health Organization

List of Figures

Summary

Most diseases are characterized by altered epigenetic and metabolic states, pointing to the need of a global and combined understanding of mechanisms underlying epigenetic and metabolic changes as an important piece to enable disease eradication. Adipocytes impact systemic homeostasis and their differentiation encompasses a phenotypic change whose function becomes impaired with diseases such as obesity and metabolic syndrome.

Following an integrative systems biology approach, we combined different *omics* data from the differentiation of Simpson-Golabi-Behmel syndrome (SGBS) adipocytes with a human metabolic model to observe key metabolic changes upon differentiation, their regulation and relevance for disease.

Pursuing the link to disease, we used public data from the genome-wide binding of TFs and location of active enhancers to test for disease association in function of the regulatory load, revealing a cell type-selective enrichment for disease of the high regulatory load genes.

Diverse experimental data were collected, covering a gene expression time-course during adipogenesis, with identification of miR-27a, miR-29a and miR-222 target genes, the genome-wide binding profiles of PPARγ, C/EBPα and LXRα, and the H3K4me3 histone modification mark for actively transcribed transcription start sites (TSSs).

Metabolic genes showed a highly dynamic expression pattern during adipogenesis, most being targeted by PPARγ and C/EBPα. Lipid metabolism pathways including triacylglyceride synthesis showed extensive and combinatorial regulation by TFs and miRNAs, converging on known dyslipidemia genes.

For data visualization, we developed a web portal that interactively renders metabolic pathways with *omics* data overlaid (IDARE, `http://systemsbiology.uni.lu/idare.html`).

Public ChIP-seq data revealed a general principle of higher disease association of genes under higher regulatory control in a cell type-selective manner.

First, data from the genome-wide binding of 10 TFs in HUVEC cells showed an enrichment for vascular diseases on metabolic genes targeted by $\geqslant$ 6 TFs.

Second, data from the binding of a total of 93 TFs confirmed the enrichment for disease association of genes with the top TF load in 8 additional cell lines.

Finally, active enhancer data from 139 samples spanning 96 cell types and tissues demonstrated the cell type-selective disease enrichment of the genes with the highest active enhancer load.

High regulatory load genes enriched for disease association beyond genetic variation, including association types like “altered expression” and “biomarker”, among others.

Additionally, high regulatory load genes appeared on average in more KEGG pathways and had higher betweenness centrality in a liver disease network than other genes, showing longer 3’UTRs harboring more binding sites for diverse microRNA families, suggesting also a higher post-transcriptional regulatory load and a role as signal integrators within biological networks.

Our results point to the pertinence of including high regulatory load genes for unbiased prioritization of novel candidate genes for disease association.

Preface

Here I describe the different chapters of the current thesis and how they are organized, as a reference point. The thesis is cumulative, as allowed by the Doctoral School in Systems and Molecular Biomedicine, and includes three entire manuscripts within the main text body, **Manuscript I**, **Manuscript II** and **Manuscript III**, presented in **Chapter 4**, **Results** (page 57).

Additionally, two other manuscripts are presented in appendix, **Manuscript IV** and **Manuscript V**, starting on page 223. These two are technical summaries of **Manuscript I**.

The general structure of the thesis comprises an overall introduction (Chapter 1, page 1), followed by two short chapters, the first describing the scope and aims of the thesis (Chapter 2, page 45) and the second a brief desciption of the material and methods employed within the thesis (Chapter 3, page 49). Following, a chapter with results is presented, containing the three entire manuscripts that compose the main thesis body (Chapter 4, page 57). Finally, the presented results are discussed and future perspectives presented in chapter 5 (page 151).

Chapter 1, **Introduction**, draws a line from **Cellular Biology to Systems Biology**, enabled by all discoveries from classical sciences and the technological advances observed by the end of the 20th century. Moving from single components to networks and systems, we will see how complex the human organism is and that we are still missing many of the pieces to complete the puzzle. Individual components interact within cellular networks, which sustain life processes and are responsible for the large diversity of cellular functions and responses, and often disrupted in diseases.

A considerable portion of the **Introduction** is dedicated to describe **genome functioning**. The genome is a supramolecular entity composed of chromatin, which encloses DNA and many associated proteins. Its usage is highly regulated within the nucleus of eukaryotic cells in order to achieve a precise control of gene expression. The precise control of gene expression is essential to maintain cellular viability and respond to diverse stimuli. Within cells, highly dynamic and context-dependent biological networks arise from a lot of combinatorial molecular interactions between many interacting partners, resulting in a precise control of chromatin accessibility, gene expression and cellular function.

In my thesis, extensive focus was given to the genome and gene regulatory networks both in context of adipocyte differentiation and the link between the regulatory load and disease association.

Moving then from genome to metabolism, which much promptly reflects the cellular state and accounts for the biochemical reactions that characterize life, we see that the **genome and metabolism** largely interact, and regulate each other for instance via transriptional control of enzymatic activity or by contributing to chromatin remodelling through histone modifications dependent on the metabolic availability of chemical groups.

Perturbations in biological networks and their interplay underly diseases and we will dive into **adipocyte differentiation** as an example of a phenotypic transition useful to study the interplay between the gene regulatory and metabolic networks and relevant for diseases such as obesity, type 2 diabetes and metabolic syndrome.

At the end of the introduction, we return to **Systems Biology** as a framework to study biological processes at a systems level and attempting to consider their complexity as a unit rather than isolated parts. Systems Biology relies on the ***omics*** **techniques** to extensively acquire evidence on biological components and their interactions at different time points and conditions, currently allowing us to move away from the analysis of single or few components to networks and their relationships. Data integration and modelling of the biological processes in which they are involved serves to better describe or predict how they function.

Transitioning from the Introduction to the Results, a **Scope and Aims** chapter (page 45) contextualizes the thesis within Biology, stating current challenges, introducing the undertaken work, describing its aims and introducing the resulting scientific publications and my contribution. Following, the materials and methods employed in the thesis are presented (Chapter 3, **Materials and Methods**, page 49), not very detailed in context of a cumulative thesis. The respective detailed descriptions can be found within each manuscript.

The **Results** chapter contains one general overview introducing each of the manuscripts presented within the thesis and then is divided in three sections, one per manuscript.

Manuscript I (section 4.2, page 59), "Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network", deals with transcript-level regulation of human adipocyte differentiation and associated metabolic changes as well as the relation between transcription factor load and disease association in HUVEC cells. Manuscript I has been published in *Nucleic Acids Research* in 2014 (PMID: 24198249).

Manuscript II (section 4.3, page 117), extends from the first one by looking at the relation between regulatory load and disease association in several different cell and tissue types, showing also properties of the high regulatory load genes that distinguish them from other genes, in a computational analysis of public data. It is entitled "Cell type-selective disease-association of genes under high regulatory load" and was accepted for publishing in *Nucleic Acids Research* on 14.08.2015 (published online on 03.09.2015, PMID:26338775). The published version of Manuscript II has been included in the thesis after defence.

Manuscript III (section 4.4, page 147), describes a tool for automated generation of image metanodes and a Cytoscape app for multi-*omics* integration and visualization within Cystocape networks, "IDARE2 - Simultaneous visualization of multi-*omics* data in Cytoscape". The main author of Manuscript III is Thomas Pfau (thomas.pfau@uni.lu) and it is in preparation for a soon submission.

Finally, in the **Discussion and perspectives** chapter (page 151), I comment on the presented results with respect to the literature and discuss their relevance. Perspectives and suggestions for changes or improvements are also given.

An appendix chapter contains **Manuscript IV** (page 224) and **Manuscript V** (page 228), two technical summaries produced in context of Manuscript I and published in *Genomics Data*. The IDARE2 user guide (page 236) is also presented in appendix.

1 Introduction

1.1 From Cell Biology to Systems Biology

Ever since Robert Hooke looked at a cork slice under a microscope in 1665 and observed walls and compartments resembling "cells" of a monastery that our advent through Cellular Biology started [1]. Cells are the units of life, being the smallest independent and self-perpetuating entities. In multicellular organisms, they operate concertedly in order to achieve multiple functions that allow survival. With increasing knowledge on cellular biology, the existence of a microscopic sub-cellular world governing the processes of life and becoming disrupted in case of disease became clear [2, 3]. As more disease causes and treatments were discovered, the realization that a single disease cause (genetic, environmental or other) is rarely the case in place led researchers to hypothesize on interacting partners and effectors that would overall cause the disease. The complexity is huge, with around 20000 genes [4] and an even larger number of proteins, metabolites and macromolecules [5–7]. An extra layer of complexity is raised by the interactions and interplay between all different biological entities, making it impossible for human cognition to handle and understand without computational support. This complexity reflects the need for models and modelling frameworks in biology that can help us to *in silico* reproduce and test biological phenomena [8].

More than 50 years ago, F. Crick first stated the “central dogma of Biology”, with an information flow from DNA to RNA to proteins. More recently, evidence for a complex multidirectional relationship among biological components and processes not following such simple hierarchical structure has been recognized. The order and nature of processes rather follow a cyclic pattern, where indeed information stored in DNA is transcribed into RNA and from therein translated to proteins. However, there is neither DNA transcription to RNA nor RNA translation to protein without proteins, metabolites and other molecules as well as their interactions, in a cyclic self-regulatory system [9]. Proteins are in general the cellular effectors, while DNA contains a vast and versatile library of "code" that can be used upon stimuli or need to generate proteins. RNA is transcribed from DNA and presents a vast set of functions that operate between the "two extremes", including mRNAs which can be translated to proteins but also many other RNA types, including microRNAs that are involved in post-transcriptional gene silencing. Figure 1.1 (page 2) represents the “central dogma of Biology” in light of this cyclic and interconnected dependency between cellular components and processes not following a unidirectional flow.

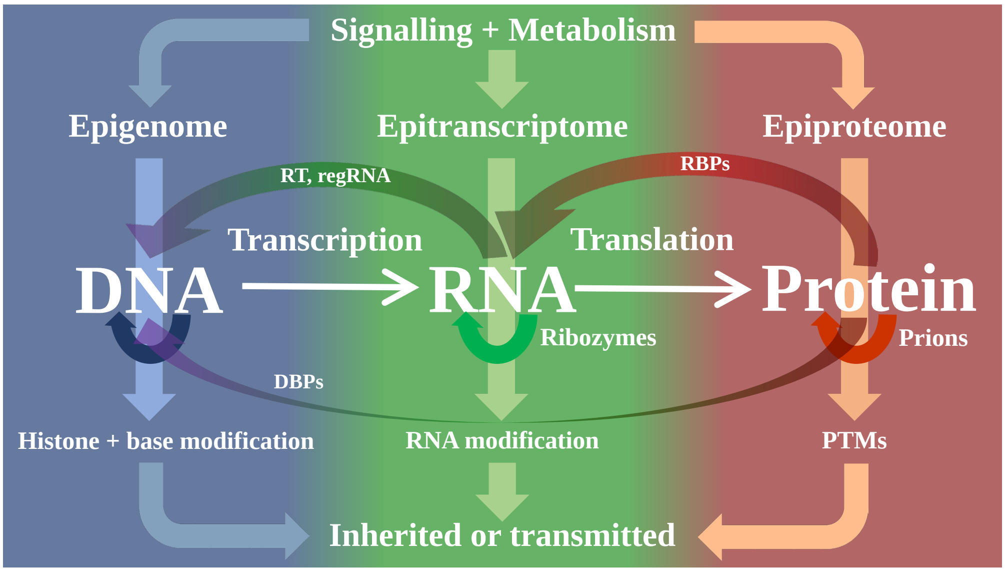


Figure 1.1: The “central dogma of Biology” in light of today’s knowledge. A complex interrelationship among many processes and components underlies Biology. The information flow is multidirectional, sustained by processes such as replication, reverse transcription (RT) and transposition. Ribozymes and prions confer an increased diversity of cellular processes and complexity. Chemical modifications impacting DNA, RNA and proteins that define the epigenome, epitranscriptome and epiproteome fine tune the course of cellular responses. DNA- and RNA-binding proteins represent another layer of interdependency between the proteome, genome and transcriptome, and sustain all the basic processes that end up in translation, including the transcriptional machinery, the binding of specific TFs and RNA processing. The many regulatory RNAs (regRNA) define an additional system for controlling gene expression. Signalling and metabolism closely relate with all cellular dimensions, through many diverse molecular interactions and transferred chemical groups, exemplifying the highly interconnected nature of cellular functioning. Systems Biology uses *panomics* approaches to attempt a higher understanding of the complex interactions of all biological processes within cells and how they are integrated within organisms in health and disease. Adapted from Saletore *et al.* [10].

Proteins, resulting from the translation of mRNAs, are cellular effectors by excellence, performing tasks as diverse as cellular structure and support (e.g. involved in cellular shape), cellular events requiring enzymatic catalysis (e.g. anabolic & catabolic metabolism, signaling cascades, active transport), non-enzymatic mediated cellular transport (e.g. facilitated diffusion or electrochemical potential-driven transport, in which proteins form channels and pores) and regulatory functions in which they are generally called “factors” (e.g. transcription factor (TF) binding specific DNA sequences and thereby controlling the flow of genetic information to mRNA - transcription) [2, 7, 11].

With only about 2% of the genome encoding for proteins, the remaining consists of non-coding RNA genes, regulatory sequences (to which regulatory proteic factors bind and regulate genetic events), introns, and non-coding DNA. The "Genome" encloses the code with the basis for life perpetuation, which is transmitted to the progeny and contains the necessary information for a fully functional organism, being physically composed by DNA and associated proteins, which compose a 3D structure named "chromatin".

"Genes" within the genome are "organized pieces of code" serving as template to build effector gene products, such as proteins, which we have studied and understand relatively well. Beyond the nucleotide sequence of the genome is the epigenome, represented by the chemical modifications to the DNA and associated histones and involved in the control of the chromatin structure and genome accessibility, which tune gene expression programs together with specific TFs.

The ENCODE project [12] attributed a biochemical function to 80.4% of the human genome, much of which owing to non-coding DNA [12], with a recent study estimating 7.1-9.2% of the human genome to be presently under negative selection with respect to indels [13], meaning incurring removal of deleterious mutations (disadvantageous or harmful insertions and deletions).

Biochemistry and Molecular Biology generated a legacy of data and knowledge about cellular components, processes and their regulation that allowed to depict metabolic pathways of the cell (series of reactions in which the products of one reaction are the substrates for the following reactions). Although not comprehensive, this legacy allows to build biochemically, genetically and genomically (BiGG) structured reconstructions that can be mathematically represented (models) and integrated with *omics* data to e.g. predict metabolic flux distributions of cells / tissues, a task experimentally complex at a large-scale level [14, 15] (more details in section 1.7.1).

In the “genome era”, increasing numbers of genomes have been sequenced, including a first version of the human genome in 2001 [16, 17]. It soon became clear that knowing the genome’s nucleotide sequence by itself would not provide much additional insights into the interconnection of genetic information and cellular processes. The need for an integrated approach considering the multiple levels of cellular organization and regulation became clear. This further stimulated the expansion of additional high-throughput technologies towards global molecular profiling (*omics*), which provide the biological data needed to connect observable phenotypes with realistic genome-scale network reconstructions [18].

1.2 Genome and gene regulatory networks

The study of the human genome and its regulation is of high importance because it contains the information that allows cellular multiplication and life sustainment. Understanding it is therefore crucial for both medical and purely research purposes [19]. In the context of this thesis, genome regulation and changes in gene expression have a central role. For this reason, in the current chapter, I give an overview of several aspects relating with the genome and how its usage is regulated in order to achieve a precise control of gene expression.

DNA, the genome component containing the "code of life", is a complex macromolecule formed by two nucleotide strands with the particular feature of intertwining in each other from opposite directions and creating an anti-parallel double helix around the same axis. This complex structure provides additional resistance to cleavage with repeated phosphate and sugar residues filling the outter surface (backbone) while the four nucleotide bases Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) fill the double helix core further increasing its stereochemical stability through complementary base-pairing (hydrogen bonds) and adjacent base-stacking (non-covalent attractive interactions between aromatic rings) [20, 21].

Most cells in the human body contain one equal copy of the genome (diploid somatic cells, 46 chromosomes), except gametes (haploid cells, 23 chromosomes) and enucleated cells (erythrocytes and platelets). The human adult body contains an estimated 3.72×10^{13} cells [22], spanning $>$ 400 cell types [23] that originate from one zygote, through cellular divisions replicating an identical genome through each generation. Such diversity of distinct cell types arises from a complex regulation of gene expression programs throughout development and adult life. Despite one largely identical genome across cell types, they are characterized by a distinct integration of signals via the epigenome. Many variable signals and chemical modifications confer a highly dynamic chromatin structure and modulate the accessibility to genes, giving rise to diverse epigenetic signatures and gene expression programs.

Gene expression is the process by which inert information encoded in genes is used to synthesize a gene product, commonly a protein, or a functional RNA (e.g. tRNA or snRNA). It involves transcription, RNA processing and, in case the gene encodes a protein, translation and post-translational modification.

Regulation of gene expression is a complex process involving spatial, physical and sequence features that occur at all steps leading to the synthesis of a gene product, including **genome accessibility** (section 1.2.1), **when and how much genes are transcribed** (section 1.2.2), the resulting **RNA properties and processing** (section 1.2.3) and the resulting **protein translation and properties** (section 1.2.4). Overall, it reflects **combinatorial molecular interactions** within (section 1.2.5) and between (section 1.4) **sub-cellular networks** (e.g. gene regulatory, signalling and metabolic networks, section 1.2.6 and 1.3) with varying amount, strength and nature which are dynamically integrated with external stimuli at the organism-level and whose **perturbation is associated with diseases** (section 1.5).

Conceptually, there are two modes by which differences in gene expression arise: i) DNA sequence variation, such as single nucleotide polymorphisms (SNPs) in a population (traditionally studied by genetics), and ii) epigenetic factors impacting genome usability without differences in the nucleotide sequence, e.g. DNA methylation or histone modifications.

While many sequence variants affecting gene expression and associated to disease have been discovered [24–26], here more focus is given to epigenetic processes involved in gene expression control [27, 28]. Of note, nuclear organization and sub-nuclear location have also been reported to play an important role in determining gene expression levels [29, 30]. Fertilization marks the beginning of the long developmental process that results in the formation of a new individual and involves time- and spatially coordinated gene regulatory mechanisms precisely controlling gene expression, including karyogamy, epigenetic reprogramming, embryonic genome activation (EGA) and cellular divisions [31, 32].

Figure 1.2 (page 6) summarizes some of the genome properties that are described in the following sections.

1.2.1 Chromatin structure and genome access

Within cells, DNA localizes to the nucleus (with exception of mitochondrial DNA) where it is densely packed and complexed with structural and regulatory proteins forming chromatin. Chromatin states and domains underlie genome function, namely throughout development and in response to diverse stimuli, as chromatin structure determines the physical access to genes, which ultimately dictates the transcriptional potential [35–37]. Spatial constraints and the highly dynamic chromatin structure (considerably varying throughout the cell cycle) temporally regulate gene expression based on genome accessibility.

Nucleosomes are the basic repeating elements of chromatin and their positioning and density determine chromatin structure. Actively transcribed less densely packed domains (euchromatin) are accessible to RNA polymerases and the transcriptional machinery, opposed to highly condensed largely inactive regions (heterochromatin) [38–48].

During cellular division, chromatin attains the highest packing density when chromosomes of cells in metaphase become visible under the microscope. Surprisingly, unwinding the DNA from one single cell would lead to a linear chain larger than 1 m[(1)], whereas the nucleus of a human cell has an average diameter

[(1)] The human genome has more than 3 billion base pairs ($3 \cdot 10^9$ bp) and one nucleotide unit length

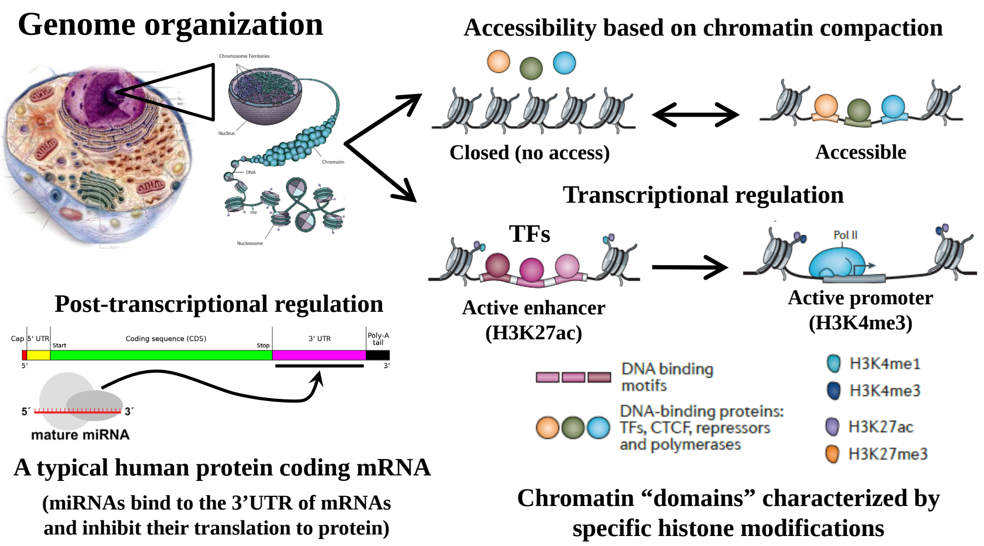

Genome organization
Chromosome Territories
Nucleus
Chromatin
DNA
Nucleosome
Accessibility based on chromatin compaction
Closed (no access)
Accessible
Transcriptional regulation
TFs
Pol II
Active enhancer (H3K27ac)
Active promoter (H3K4me3)
Post-transcriptional regulation
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'
mature miRNA
A typical human protein coding mRNA
(miRNAs bind to the 3'UTR of mRNAs and inhibit their translation to protein)
DNA binding motifs
DNA-binding proteins: TFs, CTCF, repressors and polymerases
H3K4me1
H3K4me3
H3K27ac
H3K27me3
Chromatin "domains" characterized by specific histone modifications

Figure 1.2: Summary of the organization and functioning of the genome. The genome is highly packed within the nucleus, with very dense inactive regions of heterochromatin and more active regions of euchromatin accessible for transcription. Active enhancers (marked by H3K27ac) associate with transcriptional activation and harbour binding sites for specific TFs, whose binding favours interactions with coregulators sunch as Mediator and looping events bridging to the transcriptional machinery on the promoters of target genes. Gene expression can be post-transcriptionally regulated by microRNAs which bind to the 3'UTR of target mRNAs and preclude their translation into protein. Additional regulation at the RNA and protein levels represent another strong component of the regulation of gene expression. Sources: [33, 34], `http://en.wikipedia.org/wiki/Messenger_RNA`.

of 10 µm ($1 \cdot 10^{-6}$ m), greatly exemplifying the huge packing degree of nuclear DNA. Among others, dynamic chromatin remodelling can be achieved through **nucleosome density and positioning** (section 1.2.1), **covalent histone post-translational modifications** (section 1.2.1), **DNA methylation** (section 1.2.1) and non-coding RNA regulation of chromatin structure [28, 49–54].

Nucleosome density and positioning

Allowing to achieve a huge DNA compaction in the nucleus are the nucleosomes, composed of the nucleosome core, ≈ 147 bp of DNA wrapped around a histone octamer[2] and a DNA stretch with variable length depending on chromatin compaction (linker DNA)[3] [55]. The DNA sequence itself contributes to the nucleosome positioning by dictating the affinity between the nucleotides and the core histones. Transcription factors (TF, section 1.2.2) and other proteins also influence nucleosome distribution through cooperative or competitive interactions, besides

is 0.33 nm.

[2] 2 copies of each core histones H2A, H2B, H3 and H4.

[3] Histone H1 links different nucleosomes through the linker DNA, contributing to higher order chromatin structures.

being involved in the recruitment of ATP-dependent remodeling complexes which can actively translocate, dissociate or restructure nucleosomes [56, 57].

Bargaje *et al.* [58] have shown that liver highly expressed genes were depleted of nucleosomes in their TSS, whereas in the brain, where they were lower expressed, their TSS was nucleosome masked. These results suggest a role for nucleosome positioning and possibly sliding on the control of tissue-specific gene expression, further requesting systematic elucidation across tissues and conditions (e.g on health and disease). Remarkably, despite nucleosomes, *in vivo* transcription occurs at comparable speeds to that of naked DNA templates [48, 59]. Nucleosome density and positioning in the genome is therefore an important means of regulating gene expression, being highly complex and dynamically re-arranged upon diverse stimuli [60, 61] and contributing to disease [62].

The histone code

Histone proteins compose the nucleosomes[4] and allow for DNA compaction and de-compaction as needed within the nucleus. Due to their highly basic nature with a net positive charge, histones have affinity to the negatively charged phosphate groups of the DNA. Histones are subject to reversible covalent post-translational modifications (PTMs) mainly on the amino (N)-terminal tails oriented to the exterior of nucleosomes, causing a local closing or opening of chromatin.

Some PTMs occurring in histones include methylation, acetylation, phosphorylation, ubiquitylation, propionylation, crotonylation, succinylation and glycosylation (and their reciprocal) via specific enzymes, namely histone acetyltransferases (HATs), deacetylases (HDACs), sirtuins, methyltransferases (HMTs), and kinases [41, 63–66]. For instance, histone lysine acetylation in general associates with chromatin opening, due to the neutralization of the lysine's positive charge with the addition of the acetyl group, thereby decreasing its affinity to DNA. These modifications influence inter-nucleosomal interactions, thus directly impacting the overall chromatin structure, and the interactions with many proteins, including effectors such as ATP-hydrolizing remodelling enzymes that reposition nucleosomes [67, 68], additional histone modification reader and writer proteins and coregulators, leading to effects on gene expression.

Several histone modification marks have been identified in association with chromatin states and components [69], defining a "histone code" that underlies specific cellular responses based on sequential and combinatorial histone modifications that can be interpreted and interact with a multitude of proteins and enzymes [51, 70–78], with $>$ 400 described histone modification types [79].

[4] See footnote (2) in page 6.

For instance, H3K4me3 is associated with promoters [80, 81] and H3K4me1 preferentially associates with enhancers [82], genomic regions proximal and distal to target gene TSSs associated with activation of gene transcription [83]. H3K27ac and H3K9ac associate with active enhancers [84, 85]. H3K36me3 and H4K20me1 associate with transcribed regions [81, 86]. Furthermore, H3K27me3 associates with Polycomb-repressed regions [86]. These and other modifications testify a transcriptional regulation system based on chromatin allostery effected through histone modifications [87] and shown to play important roles in several diseases [88–92] as well as beneficial mechanisms [93].

Histone modifications thereby confer a signalling dimension coupled to the genome, as part of the epigenome [94], propagating multiple latent signals through recruitment of particular proteins that mediate biological functions, such as for DNA methylation by unmethylated H3K4 through interaction with DNMT3L [95], or for hetero-chromatin formation through interactions between H3K9me3 and hetero-chromatin-associated protein 1 (HP1) [96, 97].

DNA methylation and de-methylation

Along with nucleosome density and positioning and histone modifications, DNA base modification is also involved in the chromatin packing. Cytosine methylation at the fifth carbon is the most common base modification, originating 5-methylcytosine (5mC). Methylation of cytosines adjacent to guanines (CpG islands) on gene promoters often associates with gene silencing with proteins binding to methylated DNA also forming complexes with HDACs, promoting an hyper-methylated and de-acetylated state that results in compact and silent chromatin [98]. Antagonically, DNA methylation within gene-bodies, more frequent than in promoters, has been positively correlated with gene expression [99–101]. This dynamic effect of DNA methylation reveals a complex context-specific role of DNA methylation on genome regulation, including exon usage [102, 103].

In particular, DNA methylation/de-methylation is an heritable epigenetic trait impacting several cellular processes, namely embryonic development, differentiation, chromatin structure, transcription, genomic imprinting and chromosome stability [98, 104, 105]. The enzymes involved include DNA methyltransferases (DNMTs), which use S-adenosylmethionine as a methyl group donor, and ten eleven translocation enzymes (TET1-3), which sequentially convert 5mC to the demethylation intermediates 5-hydroxymethyl-cytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [106–109].

Tissue-specific genes often have un-methylated promoter CpG islands which appear methylated in other tissues [28, 110]. DNA methylation effect on gene expression has been recognized for long [111, 112] and gene silencing described as its most prominent effect, with prime example on the silencing of one X chromosome in females [113]. However, more recently, a broader complexity and variety of DNA methylation mechanisms and effects has been revealed [28, 98, 110, 114–118], namely gene expression changes resulting from alteration in non-CpG methylation in human adult brain, T cells and skeletal muscle [119]. DNA hyper/hypo-methylation is also highly involved in diseases, namely the hyper-methylation of tumour suppressor promoters as a tumourigenesis driver [105, 120], and involvement in cardiopathies including atherosclerosis [118, 121–124], diabetes [125] and autoimmune diseases [126, 127].

The interplay between DNA and histone lysine methylation has also recently become apparent, with direct links between H3K4, H3K9 and H3K36 methylation and the targeting of DNA methylation, mechanistically supporting normal chromatin function [109, 128]. Additionally, methylation processes are also directly linked to metabolism and nutrition, relying on co-factors such as folate and the vitamins B6 and B12 as well as on the cellular methyl-donor pools [128–130].

1.2.2 Transcriptional regulation

Genome accessibility dictates which genes are physically able to be transcribed (section 1.2.1) and is highly coupled with **transcriptional regulation**, the level of gene expression control determining **when accessible genes are transcribed and how much RNA is thereby synthesized**. Factually, these processes are intrinsic to each other. Genome accessibility has its basis on protecting and keeping the genome's integrity while allowing important "parts" to be used when and where necessary. Transcription and its regulation further build upon this accessibility to operate functions through gene products.

Transcription refers to the process of producing an RNA strand complementary to a template DNA and it involves three sequential steps, all subject to regulation - **initiation, elongation and termination** - that occur in transcription factories harboring the transcriptional machinery (RNA polymerase(s) and many proteic and enzymatic complexes with affinities for multiple partners that overall catalize transcription). Chakalova and Fraser [131] nicely illustrate the 3D concept of the highly dynamic transcription of multiple genes on transcription factories, a complex process with a tight spatio-temporal control.

RNA polymerases, the transcriptional machinery core, are the master enzymes for RNA synthesis, being capable of binding to DNA on gene promoters and polymerize ribonucleotides into a nascent RNA transcript (nucleotidyl transferase activity). Distinct RNA polymerases transcribe different gene classes, namely RNA polymerase I transcribing genes coding for rRNA, RNA polymerase II transcribing protein coding genes into mRNA as well as microRNA (miRNA) genes and RNA polymerase III transcribing genes coding for tRNAs and some rRNAs. Following descriptions will refer mainly to RNA polymerase II.

Gene transcription can be regulated by direct binding to a gene, interaction with the transcriptional machinery or via an epigenetic mechanism and its "effectors" include trans-acting elements (e.g. genes encoding general TFs, activators, repressors, general cofactors and the resulting proteins themselves) and cis-acting DNA sequences (enhancers, insulators and silencers).

Transcription initiation is an ATP-dependent step during which nascent RNA transcripts are initialized from the 5' to the 3' direction, meaning the first ribonucleotide of the nascent transcript has a triphosphate group at the fifth carbon (5'end), while additional ribonucleotides are linked to the other extremity of the nascent transcript on the third carbon (3'end). A promoter-bound complex including the pre-initiation complex, the RNA polymerase II and additional general TFs is formed, eliciting DNA melting, strand separation and subsequent initiation of the synthesis of nascent RNA transcripts. Abortive cycles resulting in truncated RNAs occur until a transcript longer than 10 nucleotides is synthesized, which triggers the polymerase escape from the promoter and the elongation phase. During elongation, interaction with many factors from the initiation phase is lost and new interactions with elongation factors take place. The process is not continuous nor at a constant rate, with several pause periods, and involves DNA unwrapping, correct nucleoside triphosphate selection, phosphodiester bond synthesis and proof-reading through pyrophosphorolysis or phosphodiester bond hydrolysis for removal of non-complementary nucleotides. RNA processing steps, namely 5' capping(5) (during elongation) and splicing(6) (during or after transcription) give rise to mature mRNAs, which dissociate from the RNA polymerase II upon transcription termination, a step involving additional termination factors and RNA cleavage and poly-adenylation(7) in case of polymerase II transcripts (transcription of protein coding genes).

Basal transcription levels can be sustained through the promoter-bound complex formed during initiation, while fine-tuning of gene expression is exerted by often signal dependent specific TFs that are able to directly bind short specific regulatory DNA regions (response elements (REs)). These regulatory regions often

(5) Results in the addition of 7-methylguanosine with a 5' triphosphate bridge to the 5' end of the nascent RNA, confering protection to exonuclease degradation and involved in the regulation of nuclear export and promotion of 5' proximal intron excision and translation [132–134].

(6) Intron removal and exon joining involving trans-esterifications (except for tRNAs) which gives rise to mature transcripts, namely mRNA [135–138].

(7) Addition of $\approx$ 200 adenine residues to the 3' cleaved RNA [139–141].

locate within gene promoters. They can also be either proximal or distal (far up- or down-stream) to the target gene TSS and are defined **"enhancers"** or **"silencers"** if the effect on gene transcription is activation or inhibition, respectively.

Transcription factors directly bind DNA through their DNA binding domain (DBD), the majority composed of zinc-coordinating or helix-loop-helix domains, with 10-20 protein-DNA interactions most often on DNA major grooves where attractive forces between particular amino acids and the nucleotide pairs overall produce very strong and specific binding (hydrogen bonds, ionic bonds and hydrophobic interactions) [142, 143]. Additionally, many TFs bind DNA as dimers, which increases binding intensity and stability through doubling the DNA-protein contact area, resulting in higher effect consistency. Additively, this oligomerization capacity increases the binding combination possibilities ranging from monomers, homo- and hetero- dimers to higher order associations, contributing to the high regulatory complexity of the genome.

TF diversity and action spectrum directly relate with organism complexity and in humans they represent the largest protein family, with $\approx$ 2500 DBD-containing proteins and roughly 10% of the human genes coding for TFs [144]. The nucleotide sequence as well as the amino acids in the DNA binding domain of the TF determine which TFs bind which regulatory regions, ultimately defining which genes will be under the regulatory action of a certain TF. The large number of TFs and their many binding regions in the genome as well as the multiple regulators per gene result in a vast possible state space defining a complex gene regulatory network, responsible for the spatio-temporal control of gene expression and allowing for adaptation and versatile responses to numerous stimuli [145–147]. Consequent activation or repression of target gene transcription occurs either by a direct mechanism or in conjugation with other proteins that lack a DBD. Besides the DBD, TFs contain binding sites for other proteins, namely transcriptional cofactors, in a trans-activating-domain (TAD) responsible for the protein interactions that can induce or inhibit target gene transcription via interactions with the transcriptional machinery. In addition to the DBD and the TAD, some TFs harbour as well a signal-sensing-domain (SSD), that responds to external or internal stimuli, often through ligand binding such as with nuclear receptors (NRs) (ligand binding domain (LBD)) or PTMs, upon which effect on the transcription of target genes is exerted. PTMs, including phosphorylation, represent major mechanisms regulating protein activity, including TFs [145].

General cofactors such as Mediator [148, 149] physically bridge specific TFs with the basal transcriptional machinery, allowing for the integration of signals that results in gene expression adjustments. Protein interactions triggered by the specific TFs will result in the recruitment of particular interacting partners leading to an overall conformation that can induce or prevent gene transcription. Within this pool of multiple interacting partners, the presence of activators allows for higher transcription levels whereas their absence holds low transcription. Repressor TFs can prevent gene transcription by direct physical obstruction of the RNA polymerase to the gene promoter, via binding to silencers or via binding to mRNA with translation inhibition. Through 3D genomic looping and energetically favorable protein-protein interactions, distant regions and associated TFs come to proximity with gene promoters and the necessary cofactors and transcriptional machinery, thereby effecting their action on the control of gene expression. Long-range interactions with TSSs represent additional means for controlling gene expression, constituting yet another complex 3D interaction network at the level of chromatin, largely undisclosed. Attesting for this complexity is an average 3.9 distal elements interacting with a TSS and an average 2.5 TSSs interacting with distal elements, as derived by the ENCODE project [12], raising the possibility for a huge combinatorial network when considering all genes and regulatory regions.

In addition to spanning multiple TF-binding sites (TFBS) and highly enhancing gene transcription via the enhanceosome complex, enhancers relate with cell-type specific characteristics. Villar and colleagues compared liver active promoters and enhancers of 20 mammals including humans [150]. In contrast to promoters, enhancers present a rapid evolution and low conservation among mammals. Recently evolved enhancers often associate with genes under lineage-specific positive selection.

Specific transcription factors and the precise tuning of gene expression

The >2000 different TFs have been classified based on their properties (e.g. structural, namely according to their DBD, and functionally), with many available public and commercial databases on curated TF classification and details, including their consensus binding motifs [151–156]. TF activity is regulated by several different mechanisms, namely ligand binding, membrane release, PTMs including phosphorylation, subunit coupling, unmasking and nuclear transport [157].

Well known TF classes, based on the DBD, include “Immunoglobulin domains” often involved in cell cycle, apoptosis and immunity processes (e.g. p53 and STAT family), “Basic domains” with many diverse cellular functions, including cAMP response, steroid hormone synthesis and cancer development (e.g. C/EBP, CREB, Fos, Jun, Myc and SREBP factors) and “zync-coordinating domains” (e.g. GATA factors involved in hematopoietic cell development) including **“NRs”** (e.g. steroid and thyroid hormone receptors and RXR-related receptors).

In particular, the animal-exclusive (metazoans) family of nuclear receptors regulates the transcription of many genes, presenting a wide-range of actions, spanning from development and differentiation to homeostasis and metabolism, with 48 known family members in humans, of which 24 are ligand-dependent [158–164]. NR ligands include lipophilic compounds such as endogenous hormones and bile acids, fatty acids, sterols, vitamins (e.g. A and D) and xenobiotics, which upon binding to the SSD induce a conformational change that results in the activation of the NR. In the presence of a ligand, activated NRs bind their specific genomic response elements and through recruitment of cofactors and chromatin remodelers increase target gene transcription. Some NRs have affinity for corepressors in the absence of ligand, in which case they silence gene transcription [163]. Due to their role in ubiquitous functions, NRs are majorly involved in pathological processes, namely in many metabolic, immune and cancerous diseases [165–169], with >10% of US Food and Drug Administration (FDA) approved drugs targeting NRs [170].

Combinations of different inputs (e.g. ligand binding and/or PTMs) on a TF can have different effects on target genes, through variable interactions with coactivators and corepressors that lead to the differential activation/repression of target gene sets [171]. Such coactivators and corepressors are often involved in chromatin remodelling and histone modification, as earlier introduced, and the combinatorial nature of interactions further increases response diversity and adds to the fine control of gene expression.

Transcriptional cascades elicited by TFs in response to diverse stimuli amplify signals and result in the precise spatio-temporal control of gene expression. During development, they drive the embryo through decreasing pluripotency stages with concomitant increase in morphological complexity and functional determination, including TFs such as OCT4, SOX2, KLF4, NANOG, GATA6, CDX2 and SOX17 [31, 172]. The spatio-temporal coordination and regulation of the dosage and circuitry of developmental TFs is essential for the correct development of an organism [32]. Cellular differentiation is the prime example of a process in which specific TFs act in order to elicit a determined phenotype based on particular gene expression programs [173], with several feedback loops where pioneer factors and master regulators of stage transitions sustain their own expression while inducing or repressing key transcription factors of other stages as well as target genes [172, 174–177]. The combination of activating and repressing signals from specific TFs together with the interacting partners composing the multi-protein complexes involved in transcription are key for gene- and cell-type-specific transcription.

Besides presenting sequence specificity, TFs colocalize whithin **transcriptional hotspots**, characterized by a high density of TF and coregulator binding with cooperative regulation of gene expression and high transcriptional activity [178–181].

These genomic "islands" or "epicentres" locate particularly within **super-enhancers** [182–185], Mediator-rich [186] large enhancer clusters densely occupied by pioneer, pluripotency, master regulator and cell-type specific factors [187, 188]. Spatio-temporal differentiation, fate acquisition and response to stimuli involve disruption and re-establishment of new super-enhancers and associated hotspots, as shown for lineage progression [189] and inflammation [190].

1.2.3 Post-transcriptional regulation

As we have seen so far, the timing, quantity and tissue distribution of RNA is finely regulated by conjugations of diverse stimuli and different genome usage (e.g. on different tissues), evoked mainly through genome accessibility (section 1.2.1) and transcriptional regulation (section 1.2.2). Another regulatory level impacting the fate of RNA transcripts comes to action, including RNA processing, stability, sequestration, transport and degradation rate, named **Post-transcriptional regulation**, mainly exerted by RNA-binding-proteins (RBPs) and small RNA species themselves, such as microRNAs (miRNAs).

Many RBPs interact with RNA transcripts and are involved in their regulation, through their RNA-binding-domain (RBD) which contains particular amino acids able to bind short specific RNA sequences (RNA-recognition-motif (RRM)), somewhat similar to the DBD earlier described for TFs. These RBPs are involved in diverse processing steps such as alternative splicing and nuclear export [191–196], exosome- and P-body-mediated RNA processing, storage or degradation [197–202] and ultimately translation (section 1.2.4).

In addition, another class of short RNAs, named miRNAs, has a high impact in the post-transcriptional control of gene expression [203]. miRNAs are non-coding double stranded short RNAs of $\approx$ 22 nucleotides capable of binding target mRNA sequences through perfect matching over a “seed region” of $\approx$ 7 nucleotides. By allowing imperfect base pairing throughout the remaining, mainly causing mRNA destabilization, miRNAs preclude mRNA translation to protein, leading to gene silencing [204–208].

Since miRNAs discovery in humans in the early 2000s [209–212], their number has rapidly grown [213], with 1881 precursor miRNAs and 2588 mature miRNAs listed in miRBase[(8)] [214], a microRNA sequence and annotation database. This large number of miRNAs, comparable to that of known TFs, is in agreement with at least 60% of human genes harbouring miRNA binding sites in the 3’-untranslated-region (3’UTR) of their mRNA transcripts [215].

[(8)] `http://www.mirbase.org/`, as of June 2014.

The biological relevance of miRNAs is readily illustrated by the lethality, abnormal development or diseases caused by miRNA knockout in animal models [216–225].

Individual miRNAs can repress hundreds of target genes and likewise, individual target genes are often regulated by multiple miRNAs, overall imposing a complex miRNA-target network that further extends the gene regulatory network [226, 227]. Although the global impact of miRNAs on gene expression is not completely understood, they seem to act as "buffers" against abrupt expression changes while keeping cellular output to sensible levels, with a reported modest effect on protein levels of target genes [228–233].

On the other hand, "switch"-like behaviours during signal propagation or lineage commitment associated with feedback-loops and transient high miRNA levels have also been reported [234–236]. Expression of miRNAs itself is a highly regulated process that results in context-specific post-transcriptional gene expression control [237–242].

miRNAs are frequently encoded by families of genes with a common promoter and often identical seed sequences. They are either inter- or intragenic, the latter being co-regulated with the host mRNA, which supports their role as "buffers" to avoid the uncontrolled expression rise of certain genes [241, 243, 244].

miRNA gene transcription mainly by RNA Polymerase II [245] gives rise to primary-miRNA (pri-miRNA) transcripts that can have multiple hairpin structures of ≈ 70 nucleotides each. These double-stranded RNA hairpins undergo several cleavage steps mainly by the endoribonucleases Drosha and Dicer, in the nucleus and cytosol, respectively, with energy-dependent export into the cytosol in between, upon which generally one strand of the ≈ 22 nucleotide miRNA is incorporated within the RNA-induced silencing complex (RISC), containing one of the Argonaute protein family (Ago) members and many associated proteins including the GW182 family, where the miRNA and target mRNA interact [246–251]. Additionally, RNA editing steps such as methylation, uridylation and adenylation, Ago loading and RNA decay all further influence miRNA post-transcriptional control of gene expression [251–253].

miRNAs impact diverse cellular processes such as development, differentiation, proliferation, cell cycle, apoptosis and metabolism, affecting the expression of proteins with varied functions, such as TFs, signalling and metabolic enzymes, receptors, interleukins and growth factors [235, 254, 255]. As expected based on their numerous biological functions, miRNA involvement in disease has been shown in patologies such as cancer and immune, cardiovascular and metabolic diseases [256–262]. Owing to the progress of the past 15 years in the miRNA field, the complexity and dynamics of miRNA-mediated gene regulatory processes in conjugation with RBPs and including cellular storages such as P-bodies and exosomes, as well as their overall impact on cellular as well as systemic processes, are starting to be elucidated [263–268].

1.2.4 Translation and protein characteristics

Following gene expression control at the levels of mRNA synthesis and processing (sections 1.2.2 and 1.2.3), regulation of translation is the next control step, dictating which, where and how much mRNAs are translated to proteins. Protein translation is an energy dependent anabolic process that leads to the synthesis of proteins based on a template mRNA, involving **initiation** with ribosome recruitment, peptide **elongation, translocation** (ribosome displacement along the mRNA molecule) **and termination** (upon binding to a stop codon) steps, with general analogies to transcription (described in 1.2.2). Its main effectors are ribosomes and it occurs in the cytoplasm or across the endoplasmic reticulum membrane.

Ribosomes are protein-RNA complexes (ribonucleoprotein) capable of reading mRNA molecules and synthesizing a complementary polypeptide, through their ribosomal RNA catalytic peptidyl transferase activity (ribozyme). Ribosomes are composed of two subunits, a small one responsible for the interactions with the mRNA and a large one, which binds tRNAs and associated amino acids. Through mRNA codon pairing with aminoacyl-tRNA anti-codons(9), amino acids are polymerized based on the sequence dictated by the mRNA molecules, starting from their 5' end by an AUG start codon. Likewise in transcriptional regulation, all steps of protein translation are subject to regulation, namely by several translational protein complexes including kinases, and can thereby cause considerable differences in expression levels when comparing to the transcriptional and post-transcriptional outputs, being involved in many diseases [269–275].

Besides internal regulation via PTMs on interacting partners of the translational machinery, many different external stimuli elicit effects on protein translation, namely hormones such as insulin which leads to increased protein synthesis, prostetic groups such as heme which can restrict globin mRNA translation if at insufficient levels, viral-infection-derived interferon synthesis with translation inhibition via inactivating phosphorylation or mRNA degradation, as well as pathogen-derived toxins [272, 276, 277].

Despite all regulation upon protein translation, the actual protein effect, e.g. binding to specific DNA region in case of a TF, target phosphorylation in case of

(9) Codons of 3 nucleotides define the “genetic code”, a translation between nucleotides and amino acids.

a kinase or metabolic reaction catalysis in case of a metabolic enzyme, depends on many subsequent factors such as correct protein folding and targeting, stability, activity and turnover. Their final state arises from the many PTMs and the huge interaction network with multiple molecular classes so characteristic of the highly complex and promiscuous “protein world”, representing yet another regulatory layer impacting gene expression.

1.2.5 The combinatorial nature of genome regulation

The previous sections attempt to give an overview of the individual components and processes making up the genome and its dynamic and complex usage (1.2.1 through 1.2.4). In reality, all these processes and their components undissociably compose an intricate interplay in which they co-operate to elicit phenotypic changes, including irreversible state transitions and lineage commitment during cellular differentiation or immune responses.

As described in previous sections, the different layers regulating the genome usage and gene expression commonly share the same cellular targets and influence each other on multiple ways, which accounts for the complexity of biological systems and confers them robustness with increased survival chances.

Overall, despite a largely invariant hardly coded genetic information, genome usage and output are dynamically tuned in time and space, giving rise to a huge set of cellular responses, based on sequential and combinatorial processes comprising epigenetic mechanisms, such as chromatin remodelling with nucleosome re-positioning involving histone modifications and DNA methylation, mutually influenced [128], causing the opening of previously unaccessible chromatin to master regulators and specific TFs that upon binding within enhancers can trigger cell-type specific transcriptional cascades and gene programs [189, 278]. These processes are further controlled via feedback loops also involving miRNAs [279, 280], RBPs and signalling cascades operating on the cellular protein pool, subject to protein availability and activity, as largely modulated by external stimuli [281, 282].

Regulatory cross-talking on genome usage and gene expression control has been shown, for instance, for mRNA biogenesis and metabolism [283], on lineage-specific master regulators targeting Mediator for super-enhancer formation on cell-type specific genes, such as for PPARγ regarding adipocyte differentiation and for GATA1 on blood cell lineage development [284, 285] or for the miRNA-mediated DNA methylation control through targeting of transcriptional repressors [286].

In agreement and further illuminating Waddington’s "epigenetic landscape"

[68, 174, 175, 287–292], we are currently gathering more and more evidence of a very dynamic and plastic transcriptional landscape, in which chromatin state changes upon diverse stimuli [293–296], including through the remodelling of nucleosomes and super-enhancers which allow for variable interactions with key regulators. This dynamic re-arrangement underlies cell identity, lineage commitment and disease mechanisms [58, 61, 297, 298], as already shown for Friedreich ataxia [62], Huntington disease [299], inflammation and atherosclerosis [300], oncogenic drivers [301–306] and variants associated with cell-type relevant traits such as diabetes or immune-mediated disorders [307, 308].

The epigenetic landscape is characterized by peaks representing initial or transition phases between two cellular states, of lower stability, while valleys represent well defined or differentiated cellular states, of higher stability and from which it is harder to divert. With so many cell types, the complete epigenetic landscape is highly complex. **Gene regulatory networks** (section 1.2.6) acting on cells dictate their fate through the epigenetic landscape, via the different concentrations of key TFs, the effects of miRNAs, other ncRNA species as well as other interactions arising from the integration of external and internal signals among different cellular components and allowing to precisely control the expression of different gene sets in time and space.

1.2.6 Gene regulatory networks

As described in the previous sections, the genome is dynamically and differentially used upon external or internal stimuli, as the result of an intricate interplay within and between gene, signaling and metabolic networks, including signaling molecules, receptors, enzymes, TFs, genomic regulatory regions, miRNAs, RBPs, target genes and their products, which define regulatory networks associated with specific cellular processes, functions or development stages [144, 309]. Figure 1.3 (page 19) contains a schematic representation of the interdependencies between diverse cellular components which interact through regulatory networks.

While less focus is given here to signalling networks, these are intricate to gene regulatory and metabolic networks, making up one system of highly connected and inter-dependent components assuring the correct functioning of the organism [311], allowing to keep homeostasis and so diversely adapt to changes and respond to stimuli, including through the activation and inhibition of transcription factors and metabolic enzymes, with a huge degree of combinatorial interactions and cross-talk that provides higher diversity and robustness [312–315].

Recent technological advances and international consortia including the EN-

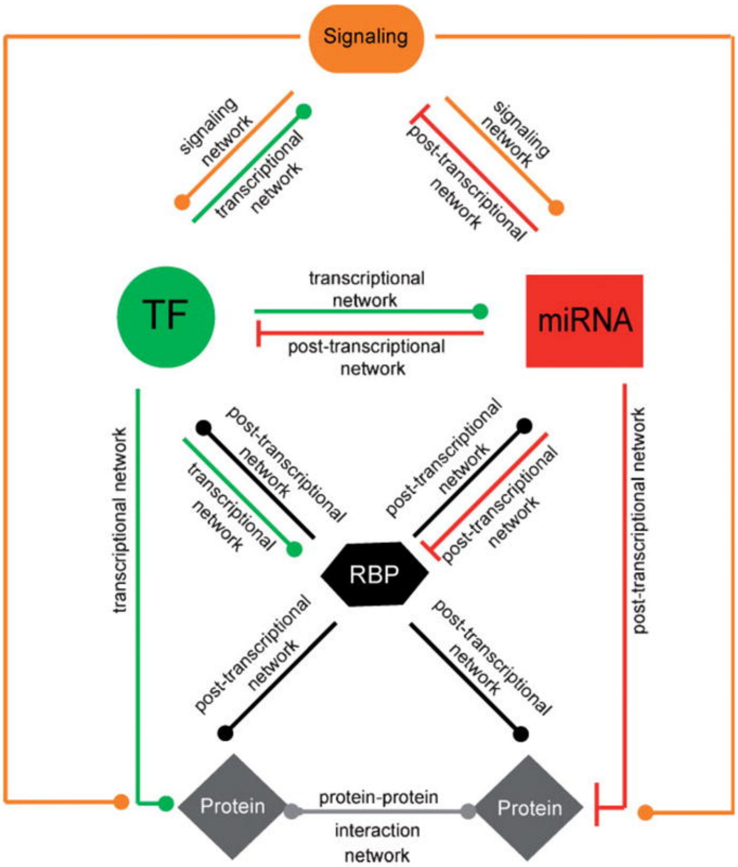

Signaling
signaling network
transcriptional network
signaling network
post-transcriptional network
TF
transcriptional network
post-transcriptional network
miRNA
transcriptional network
post-transcriptional network
transcriptional network
post-transcriptional network
post-transcriptional network
post-transcriptional network
RBP
post-transcriptional network
post-transcriptional network
Protein
protein-protein interaction network
Protein

Figure 1.3: Schematic representation of key players from regulatory networks and their relationships [310]. Regulatory networks operate within cells via signal transduction from the interactions between many partners. TFs and microRNAs closely relate and delineate the output from transcription.

CODE [12], the BLUEPRINT epigenome [316] and the NIH Roadmap Epigenomics [317] projects have started to provide novel insights into genome components, function, organization and regulatory network [318–325], which will help leveraging our understanding of pathological mechanisms and thereby our capacity to effectively revert or cure diseases. The ENCODE project is a large-scale public research initiative launched by the US National Human Genome Research Institute (NHGRI) in 2003 aiming to identify all functional elements in the human genome. The BLUEPRINT epigenome project is a large-scale EU research initiative launched in 2011 aiming at providing a blueprint of haematopoietic epigenomes and comparing healthy and disease states. The NIH Roadmap Epigenomics project was launched in 2010 with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research. By 2015, it has provided reference human epigenomes for more 100 tissues.

Conceptually, gene regulatory networks can be studied as a control system with inputs, intermediary components and outputs connected through feedback and feedforward loops (FFLs). For instance, TFs, miRNAs and target genes define cellular modules that build up large complex gene regulatory networks, varying on time and in response to different stimuli [235, 326]. The variety and complex architecture of FFLs within gene regulatory networks underlies binary switching or oscillatory behaviours accompanying critical cellular steady state transitions, built upon signal frequency, duration and amplification through the individual modules [173]. The difficulty relies on cataloguing all individual components and the motifs describing their interactions, knowing when they operate and which of the partners are in action in different contexts, as well as their interaction strength and direction, all of these contributing to the overall outcome [327–329]. Modelling is thereby particularly necessary for studying gene regulatory networks due to the large number of possible combinations of interactions between the elements, additionally providing the means to *in silico* test network properties and simulate network perturbations, such as the effect of a drug or of changing the expression of particular components. Multiple modelling frameworks have been developed specifically for the purpose of modelling complex gene regulatory networks [330].

1.3 Metabolism and the metabolic network

Metabolism defines the interconnected set of **biochemical reactions constantly occurring in an organism**. It underlies all processes sustaining and perpetuating life, allowing cells to extract energy from nutrients. Likewise, it is the realm of enzymes and metabolites which are transported and chemically transformed through series of connected reactions in which the products of certain are the substrates of others. Individual reactions are commonly grouped into reaction sets often with a defined function, named **metabolic pathways**(10). Metabolite availability partially represents the environment and can vastly influence cellular processes and viability. Metabolism thereby represents one layer of the complex genome-environment interplay that extends from the simpler unidirectional information flow (central dogma of Biology). Metabolic pathways make up one large inter-dependent reaction system, the **metabolic network**. Systemic or multi-organ-level metabolism includes processes such as digestion or an inflammatory response and intimately relates with homeostasis, while at the cellular level the metabolic network comprises all reactions taking place, including all metabolites, enzymes and cofactors, with $>$ 7000 reactions [331, 332] and $>$ 40000 metabolites [5] already known in humans. In this thesis, focus is given to cellular level metabolism, in particular to the metabolic changes occurring during adipocyte differentiation

(10) visit `http://biochemical-pathways.com/` for an overview on metabolic pathways.

(chapter 4.2).

In general, metabolism involves compounds from four main organic classes: **nucleotides, amino acids, lipids and carbohydrates**, together with many necessary coenzymes and cofactors such as vitamins, as well as xenobiotics and other inorganic compounds including iron, zync or phosphate groups. **Synthesis, degradation and transport reactions** involving compounds from these classes, most catalized by enzymes with respective cofactors, sustain energy transfer and primely distinguish lifeless matter from living organisms. Through metabolism, the energy from nutrients can be directly used on necessary cellular processes, via the catabolic break-down of macromolecules with energy release, or the anabolic assembly of simple molecules into macromolecules such as nucleotides and proteins, with energy consumption. Alternatively, energy can be stored into cellular reserves for use upon energetic demand, such as glycogen in the liver and triglycerides in adipocytes.

Enzymes are the master effectors of metabolism, catalizing the conversion of metabolites without being therein consumed and allowing to increase the reaction rates by decreasing their activation energy (Gibbs free energy), often speeding up many orders of magnitude in comparison to their spontaneous occurrence. A large proportion of the regulation of metabolic activity is achieved through enzymes, namely in response to environmental changes or signals from other cells. The notion of tighter control of only a few key enzymes within a pathway as been advocated since long [333], with recent work providing hints regarding the transcriptional control of metabolic pathways and its roles in evolution [334]. Energy metabolism involves mainly the catabolism of proteins, lipids and carbohydrates through the pathways of amino acid degradation, fatty acid oxidation and glycolysis, respectively, which converge in acetyl-coA, entry point into the tricarboxilic acid cycle (TCA) which bridges to the oxidative phosphorylation, a cellular pathway with high efficient production of ATP, the cellular energetic force. Besides its involvement in the TCA cycle, acetyl-coA is also a building block for fatty acid, ketone body and cholesterol synthesis, the latter via the condensation of two acetyl-coA molecules by acetoacetyl-coA transferase (ACAT), the first step of the mevalonate pathway. Additionally, it forms the neurotransmitter acetylcholine combined with choline in a reaction catalized by choline acetyltransferase (CHAT), linking metabolism and signalling.

Indeed, one large cellular network exists which encompasses and integrates all biochemical reactions from metabolism and signalling with the genome, its regulation and gene products and the environmental stimuli.

Since molecular conversions and interactions are the basis for all cellular processes, **metabolism is virtually linked with all cellular activities** and subject to fine regulation. Multiple regulatory levels are exerted on metabolism [335], including compartmentation which segregates enzymes and metabolites into pools and confers higher organization; self-regulation often via allosteric control, namely based on the levels of pathway intermediates, e.g. hexokinase inhibition by glucose-6-phosphate; post-translational modifications such as glycosylation, phosphorylation and acetylation, which can activate or repress enzyme activity; cooperativity in which the binding of a substrate induces conformational changes increasing or decreasing the affinity to additional substrate molecules, e.g. the increased affinity of hemoglobin upon binding of the first oxygen molecule; or extrinsic control such as that exerted by hormones, namely insulin, which leads to signalling cascades resulting in glucose uptake and a shift in metabolism towards glycogen and fatty acid synthesis. Regulation through feedback and feedforward loops modulates both the activity and concentration of enzymes, the first case occurring at the order of seconds or minutes and accounting for the fine tuning of metabolic activity and capacity, and the second taking minutes to hours and defining a more coarse control of metabolism, including transitions to different metabolic states.

As metabolism is so diverse, complex and inter-connected, with varying specialization degrees throughout the body according to the tissue or cell-type, with highly dynamic and fast adptation and responses to a large range of stimuli [336, 337], a big challenge remains still in characterizing all cellular responses and relating them with conditions or metabolic states and signatures that could be used as disease markers or predictors of systemic states [338, 339]. In this context, **metabolic models and modelling** (more details in 1.7.1) hold the promise to help us to better understand metabolic pathways in health and disease. Metabolic disorders such as diabetes, metabolic syndrome and dyslipidemia arise mainly from an intricate interplay between genetic, environmental, and nutritional factors impacting the cellular metabolic state to a greater or lesser extent.

Over 100 years of research in biochemistry provide a vast knowledge on metabolism, systematically initiated already in the 19th century with works on fermentation and related enzymes by A. Payen, L. Pasteur, W. Kühne and E. Buchner. Nowadays, metabolism is one of the best known cellular processes and spans a vast set of pathways and components, including the central carbon and energy metabolism, largely uniform across species.

Due to its vastness and complexity, researchers have manually compiled information on metabolic pathways since long and this knowledge has been available for instance in textbooks and journal articles. Since a few decades, electronic databases provide the resource of choice for storing, organizing and sharing information, including that related to metabolism. These databases are an important source of knowledge on metabolism going beyond what is described in textbooks which are often outdated. Unfortunately it has been very hard to curate and unify the multiple existing databases, which are both redundant and exclusive to certain degrees, with no unique choice giving a fully comprehensive result. Additionally, these databases often contain information on pathways and reactions including both the metabolic and regulatory levels, the degree to which they are separated depending on the author's criteria. Therefore, expert knowledge and curation is further required in order to focus on each of these cellular dimensions separately. Some of these databases include MetaCyc [340], including the HumanCyc (Encyclopedia of Human Genes and Metabolism), KEGG [341], BRENDA [342], REACTOME [343], WikiPathways [344], ConsensusPathDB [345], SMPDB [346] and HMDB [5].

1.4 The interplay between the genome and the metabolism

In the section **"Genome and gene regulatory networks"** (1.2), we saw that the genome represents the physical material existing in cells which contains information for cellular multiplication and life perpetuation, defining the identity of a species.

Diploid organisms such as humans have two copies of the genetic material, in a combination of gene pairs that defines each individual's **genotype**, manifested through the **phenotype** representing the observable characteristics of an organism, such as hair and eye color, which result from specific allele combination. In the cellular context, **the metabolic state is at the frontline of the observed phenotypic characteristics**: macromolecules are responsible for information storage (DNA), processing (RNA) and execution (proteins) and the set of reactions, metabolites and compounds present in a cell at a given time reflect their overall interaction and joint effect, representing the cellular state (details in **"Metabolism and the metabolic network"**, 1.3).

The genome and the metabolism are part of the same network specialized in maintaining life and responsible for the cellular response diversity, with extensive relationships and overlaps between the gene regulatory and metabolic networks [9, 347]. For instance, metabolic signaling through metabolite-sensing TFs including NRs readily modulates gene expression, namely by the altered molecular interactions with coregulators, often also metabolite-sensing, whose function alteration has been associated to diseases [348–353]. On the other hand, in response to stimuli or transcriptional cascades, **the expression of metabolic enzymes is modulated**, representing **control of metabolic capacity and activity via transcription**.

Furthermore, PTMs which regulate the stability and activity of many proteins are dependent on the metabolic availability of chemical groups such as phosphate, methyl and acetyl, the latter also implicated on genome accessibility through DNA methylation or histone modification. Acetyl-coA is a key metabolite linking metabolism, signaling, chromatin structure and transcription, namely as source of the acetyl group transferred in protein acetylation, including of histone lysine residues [354]. **Metabolism** thereby **exerts an effect on transcriptional control**, namely through the cellular pools of active chemical groups and their donors, required for epigenetic mechanisms.

The existence of multifaceted proteins, conferring evolutionary advantages, is yet another example of the intricate interplay within cells. For instance, some glycolytic enzymes present non-glycolytic activities as well, including the regulation of transcription, apoptosis and of cell motility [355–358].

Alterations in the interplay between gene regulatory and metabolic networks have a pivotal role in disease aetiology, as we will better understand in the following section.

1.5 Diseases as perturbations of biological networks

The leading causes of death worldwide are cardiovascular diseases, cancers, diabetes and chronic lung diseases, as reported by the World Health Organization (WHO) with data from 2012[11]. Since 1980, the number of obese people worldwide has more than doubled, with 600 million obese adults in 2014 and more deaths due to overweight and obesity than underweight. Obesity is a complex chronic disease with epidemic proportions worldwide that continues to increase regardless of age and gender, characterized by a body mass index (BMI) $\geqslant 30$ [12]. It is highly preventable and mainly caused by an excessively high calorie diet with low energy expenditure (e.g. physical exercise). While not a cause of death by itself, obesity is a major risk for metabolic syndrome, type 2 diabetes mellitus (T2DM), cardiovascular diseases and chronic inflammation, often resulting in premature death. Concerning Luxembourg, a study from LISER[13] states that, in 2008, 55% of the population older than 15 years was either overweight or obese [359]. Measures to stop obesity expansion across the globe and to decrease the current number of obese persons

[11] http://www.who.int/mediacentre/factsheets/fs310/en/, as of 01.06.2015.

[12] http://www.who.int/mediacentre/factsheets/fs311/en/, as of 01.06.2015.

[13] former CEPS/INSTEAD, http://www.ceps.lu/

are therefore essential.

Despite a yet unfullly disclosed aetiology and pathophysiology of many complex diseases, they all share a common overall characteristic: **a complex interplay between genetic and environmental factors with altered metabolism** [360, 361].

From about 30000 known human diseases, over 10000 are monogenic[14] [362], including sickle cell anemia, haemophilia and cystic fibrosis. Biological networks are the natural way how cells, tissues, organs and ultimately the whole organism communicate and operate changes. They comprise many interactions among many components [363], including genes, their products, metabolites and other compounds, and result from millions of years of evolution, presenting a high redundancy level that confers them robustness against perturbations, including environmental changes or pathogenic attacks, increasing the organism chances of survival [262, 364, 365]. Most diseases affect the information flow within these networks.

In diseased individuals, a combination of genetic and environmental factors leads to altered interactions between network partners, which can be suppressed, diminished, increased or additional in comparison to that of healthy individuals, with impact on protein folding, stability and molecular affinity that affects multiple cellular processes including protein-DNA, protein-protein and enzyme-substrate interactions [366–368]. Single gene mutations such as indels and SNPs can cause aberrant protein functioning, affecting both regulatory and metabolic tasks. An accumulation of such low impact mutations across multiple genes can jointly lead to a disease [367].

The study and understanding of biological networks in health and disease, including their comparison between related and unrelated disorders, is currently a major endeavour to increase our capacity to effectively diagnose and heal complex diseases. Thereby, network biology provides means to model inheritance traits and other genetic phenomena, besides shedding light into disease mechanisms through the identification of disease modules (Figure 1.4, page 26), aiding to prioritize diagnostic markers or therapeutic candidate genes [369–371].

1.5.1 Disease databases

In a data-rich era, gene-disease databases storing details on genetic association to diseases have emerged, namely the OMIM [373], UNIPROT [374], CTD [375], ClinVar [376] and GAD [377], which include Mendelian, complex and environmental

[14] http://www.who.int/genomics/public/geneticdiseases/en/, as of 01.06.2015.

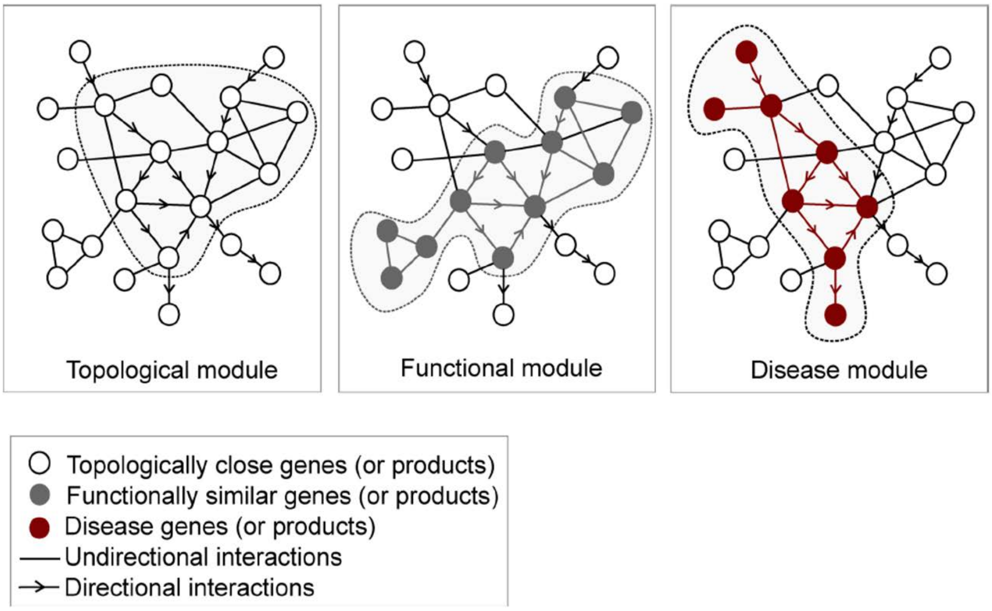

Topological module
Functional module
Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Undirectional interactions
→ Directional interactions

Figure 1.4: Example of a disease module [372]. "A disease module represents a group of nodes whose perturbation (mutations, deletions, copy number variations, or expression changes) can be linked to a particular disease phenotype, shown as red nodes. (...) a disease can be viewed as the breakdown of a functional module."

diseases. Each of these and other disease databases do not provide the complete set of genetic associations to diseases, in part because they focus on different aspects, such as links between chemicals and diseases (CTD), protein function based on sequence (UNIPROT) or candidate gene and genome-wide association studies (GWAS, GAD), becoming complementary and requiring efforts to integrate their heterogeneous information if attempting to work at the most complete level.

In this context, integrative databases such as DisGeNET[15] [378] are valuable resources providing more comprehensive overviews on current knowledge about gene-disease associations and details on the information sources. DisGeNET combines data from various expert curated databases and text-mining derived associations[16] and the current version (v3.0) contains 429111 associations between 17181 genes and 14619 diseases, ranked with a score based on the supporting evidence[17]. In this thesis, DisGeNET was used as source for gene-disease associations.

[15] http://www.disgenet.org/.

[16] http://www.disgenet.org/web/DisGeNET/menu/dbinfo#sources.

[17] http://www.disgenet.org/web/DisGeNET/menu/dbinfo, as of 03.06.2015

1.5.2 Disease networks

In the context of multigenic diseases which do not follow a Mendelian inheritance, disease networks emerge based on their multiple associated genes and their products, each with a relatively low impact in the overall pathology, but together defining a gene-disease network that characterizes each disease, with shared disease-associated genes among different diseases [366, 372].

In 2007, Goh *et al.* [366] published the first human disease network, systematically linking genetic disorders with the disease-associated genes known at the time. In a gene centric representation, genes (nodes) are connected to each other (edges) if they associate to a common disease, allowing clustering based on the genes link to disease, revealing network topology features such as disease modules as well as bridges among different diseases. From this representation it became apparent that genes shared between diseases contribute to comorbidity susceptibility. Additionally, using a disease centric representation, the authors generated a network of diseases (nodes) connected to each other (edges) if sharing at least one gene in which mutations are associated with both diseases. Based in this representation, a large disease cluster of multiple cancer types was prominent, owing to common tumour suppressor genes such as TP53 and PTEN. Metabolic diseases appeared poorly connected in that original disease network, leading Lee *et al.* [379] to investigate the link between disease and the metabolic network, through metabolic links such as shared mutated enzymes, reactions or metabolites, revealing a higher comorbidity susceptibility with significant comormidity for 31% of all metabolically linked diseases [380]. Similar disorders associate to genes with a higher likelihood of interacting [366, 381].

Later in 2014, Zhou *et al.* [382] derived a human symptoms-disease network from biomedical literature that quantifies the similarity between the symptoms of any two disease pairs.

1.5.3 Examples of diseases as network perturbations

The concept of diseases arising from network perturbations can be exemplified by the typical cancer abnormal methylation on promoters of tumour suppressor genes and frequent de-methylation on promoters of oncogenes, leading to a network rewiring that spans silencing tumour suppressors and activating oncogenes, in a disruption of the normal interactions that allows cancer cells to proliferate and escape apoptosis [383].

In this section, focus is given to T2DM, a chronic systemic disease with multifactorial causes impacting many different cell types and organs, including the liver, pancreas, vasculature and adipose tissue, with many implications in biological networks. As T2DM is a complex disease impacting systemic functions and impairing adipocyte differentiation (studied within my thesis) and functioning, more detail is given to its description. The link between T2DM and Alzheimer's disease (AD), or more generally, hyperglycemia and dementia, is also shortly addressed, exemplifying how studying cellular networks can help understanding the links between apparently unrelated diseases.

T2DM is a complex disease resulting from the body's ineffective use of insulin and characterized by hyperglycemia (high blood sugar levels), usually with adult onset and highly preventable by a healthy diet and frequent physical activity. T2DM accounts for 90% of all people with diabetes in the world [384], with an estimated 1.5 million deaths directly caused by diabetes in 2012(18). Obesity, high caloric diet and sedentary lifestyle are major risk factors to develop T2DM, in conjugation with genetic susceptibility and environmental factors, with several known metabolic changes, including of BCAA levels [385], and epigenetic alterations [386, 387]. As the pathology progresses, individuals with normal glucose tolerance degenerate into impaired glucose tolerance, with high blood sugar and insulin levels, associated with insulin resistance leading to a decreased glucose uptake from the blood by cells across the body in response to insulin, affecting muscles, then the liver and lastly adipose tissue [388]. Pancreatic β-cells initially respond to insulin resistance by increasing the production and secretion of insulin, leading to even higher insulin blood levels, in a less severe condition that can last for years. In this setting, vascular endothelial cells, which directly face increased glucose, fatty acid and inflammatory cytokine levels in the blood, are primely affected during the progression into T2DM, with vascular complications due to accelerated atherogenesis and endothelial dysfunction, namely through reduced endothelium-dependent vasodilator response to acetylcholine with impairment of the regulation of blood pressure, also contributing to exacerbate the inflammatory state [389–392]. For this reason, the study of the endothelium and its functions has received much attention, and human umbilical vein endothelial cells (HUVEC) represent one of the most widely used cellular model for the study of endothelial-relevant processes [393–396]. In context of this thesis, public data from the binding of 10 TFs on HUVEC cells was used to assess the enrichment for vascular-disease-associated genes among metabolic genes with varying number of associated TFs.

Together with life deteriorating conditions such as high fat and sugar diet and inactivity, impaired glucose tolerance declines into a more severe state in

(18) `http://www.who.int/mediacentre/factsheets/fs312/en/`, as of 02.06.2015.

which insulin no longer suppresses glucose release by the liver, with persistent hyperglycemia at advanced stages. In parallel, increased insulin levels often trigger weight gain via increased fat storage in adipose tissue, more insulin sensitive than muscle and liver, in particular in the abdominal cavity with concomitant local chronic inflammation and worsening of the obesity condition [397]. At this stage, gluco- and lipo-toxicity, endoplasmic reticulum (ER) and mitochondrial stress burden cells, ultimately leading to pancreatic β-cell apoptosis [398] and insulin deficiency, further increasing hyperglycemia and enhancing several T2DM complications and comorbidities. As sugar levels increase, the blood becomes more viscous and normal circulation and diffusion processes are particularly affected in small capillaries, in some cases resulting in diabetic retinopathy (causing blindness) and peripheral neuropathy (extremity numbness and pain). Poor wound healing often leading to amputation is another clinical complication from T2DM, together with comorbidities such as fatty liver disease, kidney disease and increased risk for cardiovascular diseases.

Many genes were found to contribute to the risk for T2DM including TCF7L2, PPARγ, FTO, KCNJ11 and HNF4A. Thiazolidinedione (TZD) drugs such as rosiglitazone and pioglitazone, PPARγ agonists, have been used for improving insulin resistance. These agonists activate PPARγ which associates with a coactivator complex including a histone acetylase, translating into opening of chromatin on PPARγ responsive elements, with consequent induction of the expression of target genes involved in glucose transport, insulin signalling and fatty acid metabolism, namely glucokinases, glucose transporter type 4, malic enzyme, lipoprotein lipase, fatty acyl-CoA synthase, adipocyte fatty acid binding protein and adiponectin [387]. This PPARγ-induced transcriptional cascade results in a systemic potentiation of insulin action with decreased liver glucose secretion and increased peripheral glucose uptake (e.g. by adipose tissue and muscle), overall attenuating hyperglycemia. See section **"Adipocytes as an example of disease relevant cell type"** (5.2) for details in the mechanism of action of PPARγ in context of adipocyte differentiation.

In addition to the mentioned effects, hyperglycemia has been shown to associate with increased susceptibility for dementia, in both diabetic[(19)] and non-diabetic persons [399], namely to AD [400], with insulin signalling modullating the phosphorylation of tau protein [401], reflecting the highly connected and inter-dependent nature of biological networks and organism functioning. AD is a neurodegenerative disease causing gradual loss of cognitive and functional body capacity with marked deposition of amyloid-β plaques and neurofibrillary tangles of hyperphosphorylated

[(19)] `http://www.alz.org/national/documents/topicsheet_diabetes.pdf`, as of 02.06.2015.

tau protein in the brain, accounting for 60% to 70% of dementia cases[20]. In 2013, Carvalho and colleagues showed that the simultaneous presence of high glucose and amyloid-β peptide reduced cell viability and membrane potential, with increased mitochondrial superoxide radical and peroxide production (oxidative stress indicators), in both rat and mice cells [402]. More recently in 2015, Macauley *et al.* showed that hyperglycemia increased the production of the amyloid-β protein in an AD mouse model, with a 38.8% increase in amyloid-β levels in older mice already with neuritic plaques upon doubling glucose blood levels [403]. These researchers showed that the high-blood-glucose-induced increase in amyloid-β level in the brain was mediated by ATP-sensitive potassium (KATP) channels, thereby coupling metabolism and neuronal activity. Interestingly, KATP channels are also the ones used by pacreatic β-cells to secrete insulin in response to high blood sugar levels, possibly revealing a higher order link between glucose levels, insulin secretion and amyloid-β levels in the brain, connecting insulin resistance and diabetes with dementia.

1.6 Adipocytes as a disease relevant cell type

Adipocytes are the cells responsible for storing fat in the organism, as energy source, and they compose the adipose tissue. The importance of adipose tissue for normal body functioning is illustrated by the disease spectrum arising with both extremely decreased and increased adipose tissue functioning, respectively in **congenital generalized lipodystrophy** (or Berardinelli-Seip syndrome) and **obesity**, for instance. Both associate with insulin resistance and metabolic syndrome due to nutrient overload in the blood and across vital organs. Persons with congenital generalized lipodystrophy, a very rare autosomal recessive disease with lack of adipose tissue and muscular hypertrophy, develop insulin resistance with hyperglycemia and in some cases T2DM, as well as hypertriglyceridemia with enlarged internal organs due to ectopic fat deposition. As fat accumulates in vital body organs, their correct functioning is impaired, namely with hepatomegaly and fatty liver, hypertrophic cardiomegaly, hypertension and splenomegaly, besides several other complications that overall can cause premature death. In agreement with the body impairment in the absence of adipose tissue observed in humans, transgenic mice without white adipose tissue present similar features such as diabetes, lipid-loaded liver, enlarged internal organs, reduced leptin and higher serum triglycerides, with premature death [404]. Likewise, sugar and fat overload typical of obesity impair adipocyte differentiation with an increase of incompletely differentiated adipocytes that are less capable

[20] `http://www.who.int/mediacentre/factsheets/fs362/en/`, as of 02.06.2015.

of taking up fat from the blood and activate the immune system, leading to inflammation and fat accumulation in vital organs. Overall, the excess of highly caloric and fatty nutrients increases the risk for comorbidities like **metabolic syndrome**, encompassing T2DM, cardiovascular diseases (e.g. hypertension, coronary heart disease, stroke) and inflammation [405], as well as higher susceptibility for certain cancer types, including from the digestive or female reproductive systems [406].

Two types of adipose tissue are commonly described, white adipose tissue (WAT), specialized in storing triglycerides and a major endocrine organ [407–409] that releases leptin, adiponectin, resistin and composes most body fat, and brown adipose tissue (BAT), capable of releasing heat from fatty acid oxidation and present in smaller amount in adults compared to newborns. Interestingly, BAT was firstly described in 1551 by the Renaissance naturalist Konrad Gessner from his observations of the anatomy of alpine marmots, in which he reports a tissue in the interscapular area as "neither fat, not flesh [*"nec pinguitudo, nec caro"*], but something in between" ([410], page 842) [411]. This tissue was afterwards recognized as BAT and more than 450 years later, the presence of active BAT in adult humans became consensual [412–415]. A distinct developmental origin between WAT and BAT has been demonstrated, the latter sharing a common precursor with muscle cells [416, 417]. Additionally, recent reports have shown the presence of "brite" (brown-in-white) or "beige" adipocytes within WAT depots, with brown-adipocyte-like characteristics, pointed with interest for the activation of energy expenditure pathways and potential intermediates for rescuing obesity [418–424]. Adipose tissue is dynamic and presents a high plasticity [425]. Food intake higher than energy expenditure leads to WAT hypertrophy, which can be reversed through adipocyte mobilization when energy expenditure is higher than intake; cold exposure and sensitizing compounds induce BAT activation and expansion to increase thermogenic capacity; and pregancy and lactation induce the appearance of "pink" adipocytes, mammary gland alveolar epithelial cells that produce and secrete milk from the transdifferentiation of white adipocytes, recently characterized in mouse [426].

White adipocytes are round cells containing a single large fat droplet that can occupy over 90% of the cell volume, with few mitochondria. In contrast, brown adipocytes are polygonal and contain numerous mitochondria and several small lipid droplets, being smaller in size than white adipocytes. Molecularly, the expression and induction of uncoupling protein 1 (UCP1) is a brown adipocyte mark, with much lower levels in white adipocytes. Both WAT and BAT are highly vascularized and innervated. Besides adipocytes, WAT contains a multitude of other cell types including pre-adipocytes, fibroblasts, endothelial cells, macrophages and leukocytes, which contribute for the large endocrine and paracrine activity of WAT. Pre-adipocytes are undifferentiated fibroblasts and adipocyte-precursor cells capable of differentiating into adipocytes upon diverse stimuli including nutrient abundance and hormones like insulin. Both pre-adipocytes and adipocytes are highly sensitive to insulin, which promotes adipose tissue expansion through increased triglyceride storage and decreased lipolysis, or counter-acting hormones like ACTH, glucagon and epinephrine, which promote fat mobilization and a decrease in adipose tissue from fatty acid oxidation.

Adipogenesis and the full differentiation of adipocytes are crucial in both physiological and pathological events related with adipose tissue, including obesity, being subject to tight regulation [427], as we will see in more detail in the following section (**“Adipogenesis and its regulation"**, 1.6).

Adipogenesis and its regulation

Adipogenesis is the process of differentiation of precursor cells (pre-adipocytes) into adipocytes, capable of storing fat, responsive to several stimuli and actively secreting hormones and citokines (adipokines) based on the energy load state [409]. Adipocytes originate during development from mesenchymal stem cells (MSCs) derived from the embryonic germ layer mesoderm, which also gives rise to muscle cells and chondrocytes [428]. BAT is formed earlier around week 20 and large WAT accumulation is already visible by week 25 of fetal development(21). Brown adipocytes and muscle cells have a common Myf5+ precursor not common to white adipocytes [416], with recent lineage tracing studies in mouse providing a complex picture for the cellular origins and precursors of adipocytes(22), including a small subset of white adipocytes with Myf5+ precursors [429–433]. The following descriptions apply mainly to white adipocytes, whose differentiation was studied in context of this thesis. Figure 1.5 (page 33) summarizes adipocyte origin and key players in their differentiation programme.

During development, some MSCs commit to pre-adipocytes upon extracellular signals and regulation from several pathways including the Wnt, Hedgehog, bone morphogenic protein (BMP) and insulin growth factor (IGF) signaling, which regulate the balance between the myo, adipo and osteo lineages through inhibitory and activating effects [435–437]. Proliferating pre-adipocytes are no longer capable to differentiate into cell types other than adipocytes. In the presence of adipogenic

(21) http://discovery.lifemapsc.com/library/review-of-medical-embryology/, as of 05.06.2015.

(22) http://discovery.lifemapsc.com/in-vivo-development/adipose, as of 05.06.2015.

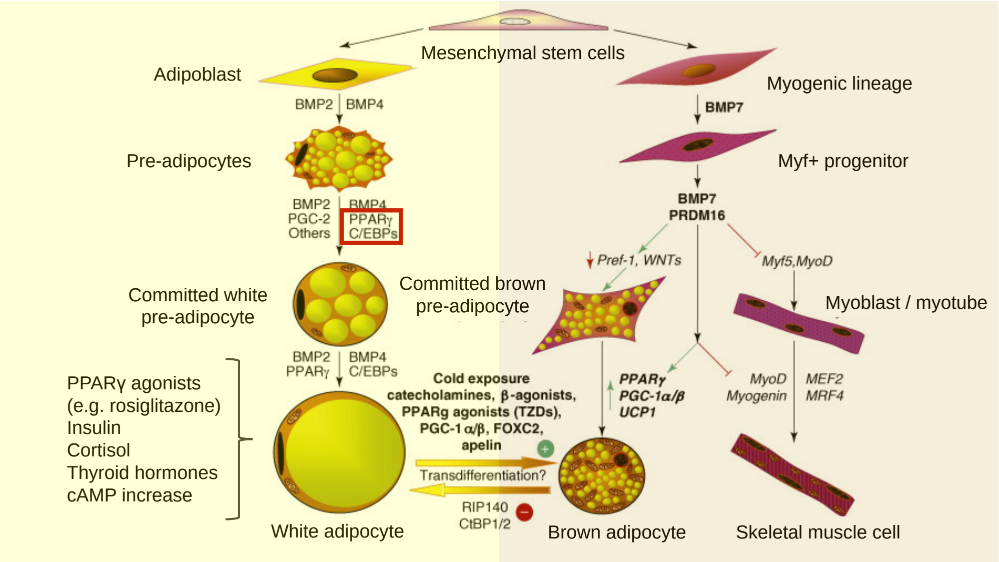


Figure 1.5: Adipocyte lineage and differentiation programme. Adipocytes arise from mesenchymal stem cells which give also rise to muscle cells. White and brown adipocytes present distinct progenitors, brown adipocytes being from the myogenic lineage together with myocytes. BMP2 and BMP4 favour the white adipocyte lineage whereas BMP7 promotes the myogenic lineage. At later stages, PPARγ and members of the C/EBP family regulate terminal white adipocyte differentiation. PPARγ is also involved in brown adipocyte differentiation, including via the stimulation of UCP1, a brown adipocyte marker. *In vitro* differentiation of adipocytes can be achieved through cellular stimulation with an "adipogenic" cocktail containing PPARγ agonists, insulin, cortisol, thyroid hormones and cAMP activators. Adapted from [434].

stimuli such as insulin, glucocorticoids, PPARγ agonists, thyroid hormones and elevation of cyclic adenosine monophosphate (cAMP) levels, pre-adipo-cytes cease proliferating and through an epigenomic transition state differentiate into mature adipocytes [438]. Adipogenesis exemplifies well a phenotypic change with precise spatio-temporal regulation eliciting chromatin re-arrangements and repositioning of genes in the nucleus. DNA methylation contributes to the silencing of pluripotency genes and those specific to other lineages. De-methylation and histone modifications are involved in the opening of adipocyte-related genes. Together with the concerted action of transcription factors, these events sequentially shape cells into the adipocyte phenotype. In this setting, PPARγ interactions with Mediator elicit the establishment of adipocyte-specific enhancers [180, 284, 436, 439–443].

Early adipogenic factors include ZNF423, TCF7L1, C/EBPβ and δ, GR, SREBF1, STAT5, AP1, KLF15, KLF4, KLF5 and EBF1 which operate changes leading to cell rounding, LPL expression and C/EBPα and PPARγ activation which induce a transcriptional cascade resulting into terminal adipocyte differentiation with expression of lipid droplet formation, many metabolic and adipokine genes, including glycerophosphate dehydrogenase, fatty acid synthase, acetyl-coA carboxylase, malic enzyme, glucose transporter type 4, insulin receptor and adipocyte protein 2 (the adipocyte-selective fatty acid binding protein) [444, 445]. While hundreds of TFs have been reported to be involved in adipogenesis [445], PPARγ appears as a master regulator, being necessary and sufficient to induce fibroblast differentiation into adipocytes [445, 446]. PPARγ is member of the ligand-dependent family of NRs and upon activation, namely through natural lipophilic compounds, dimerizes with another NR, RXR, being then capable of binding abundant peroxisome proliferator-activated receptor response elements (PPREs) across the genome, thereby regulating the expression of target genes and playing also a role in the synthesis of biologically active compounds in vascular endothelial and immune cells [447–451].

PPARγ activation by agonists such as TZD drugs has been clinically used on patients with T2DM, increasing glucose uptake and decreasing insulin resistance. However, these drugs can cause severe side effects including water retention, weight gain, hepatotoxicity and increased risk for heart failure and bone fracture [452–457], urging the discovery of new drugs able to elicit insulin-sensitization without such side effects. Resulting from alternative splicing, two PPARγ isoforms are produced, PPARγ1, largely ubiquitously expressed, and PPARγ2, more exclusive of adipose tissue and involved in lipid storage in WAT or energy dissipation in BAT [458, 459]. Recently, NFAT5 has been shown to inhibit PPARγ2 and associate with suppression of adipogenesis and insulin signaling. Thereby, reducing its expression has been pointed as a possible therapeutic target to replace PPARγ agonists [443]. Additionally, partial PPARγ agonists including natural compounds might hold promise to improve hyperglycemia with decreased side effects [460, 461]. Interestingly, the selective activation of PPARγ in adipocytes has been shown to be sufficient for systemic insulin sensitization in mice, with improved adipokine, inflammatory and lipid profiles and serum insulin levels without increased adipogenesis [462].

PPARγ, C/EBPα and LXRα were studied within this thesis in the context of Simpson-Golabi-Behmel syndrome (SGBS) adipocyte differentiation. C/EBPα is a basic leucine zipper domain (bZIP domain) TF that can bind to response elements in the genome as a homodimer or heterodimer with C/EBPβ or C/EBPδ and has been involved in cell cycle regulation, body weight homeostasis, energy metabolism and several cancer types [463]. C/EBPα is an important regulator of adipogenesis, inducing PPARγ and being required for acquiring insulin sensitivity, with differential roles in white and brown adipose tissue [445, 464, 465]. LXR is another ligand-dependent NR also dimerizing with RXR and closely related with PPARγ, being involved in the regulation of fatty acid, cholesterol and glucose homeostasis and implicated in neurodegenerative changes [466–469]. Two isoforms, LXRα and LXRβ have been identified, the latter ubiquitously expressed while LXRα is expressed in the liver, kidney, intestine, lung, spleen, adipose tissue and macrophages. Oxysterols, oxygenated cholesterol derivatives including 22(R)-hydroxycholesterol, 24(S)-hydroxy-cholesterol, 27-hydroxycholesterol and cholest-enoic acid, are their natural ligands, with two synthetic agonists (T091317 and GW3965) being widely used in research [470, 471].

More recently, miRNAs have also been shown to play important roles in adipocyte differentiation and lipid metabolism [472–475]. For instance, miRNA-143 up-regulation promotes adipogenesis [476, 477], let7 favours the osteocyte lineage [478, 479], miRNA-27 familly is down-regulated with adipogenesis and targets PPARγ and C/EBPα [480, 481], and many others [482]. Within this thesis, miRNA expression was profiled during SGBS adipogenesis in order to obtain a list of miRNAs therein involved, leading to the selection of three down-regulated miRNAs, miR-27a, miR-29a and miR-222, which were over-expressed in differentiating adipocytes in order to assess their target genes. All these miRNA families have been subject to extensive research in the recent years, with reported links in several processes including in the pathology of cancer, atherosclerosis and insulin resistance [483–485].

Due to the diverse mechanisms, dynamic behaviour and the systemic impact it can have, adipogenesis has been one of the most widely studied processes and model system for transcriptional, epigenetic and metabolic regulation of cell-type-specific gene expression and phenotypic determination. Several adipocyte cellular models have been developed [486], including the extensively used mouse 3T3-L1 pre-adipocytes, established already in 1975 [487], or the human SGBS pre-adipocyte cell line, established in 2001 from the stromal cell fraction of subcutaneous adipose tissue of an infant with Simpson-Golabi-Behmel syndrome [488], a rare X-linked congenital disorder characterized by pre- and post-natal overgrowth with features like macrosomia, renal and skeletal abnormalities as well as an increased risk of embryonic cancers. SGBS cells reliably recapitulate human adipogenesis compared with primary adipocytes, being neither transformed nor immortalized and providing an almost unlimited source due to their ability to proliferate for up to 50 generations with retained capacity for adipogenic differentiation [488–490]. These reasons motivated the use of SGBS cells for studying human adipogenesis within the work described in this thesis, being their differentiation protocol previously described [488–490].

1.7 Systems Biology

Systems biology represents the natural progression of the life sciences [8], which conceal vast amounts of knowledge on organism functioning, resulting from thousands of years during which mankind experienced and tried to understand the most diverse life related phenomena, in particular during the last few centuries, with so many discoveries.

Basic research through the life sciences has uncovered many of the mechanisms underlying a large number of common and less common observations and the current medical practices allow us to treat many diseases and conditions otherwise lethal. With the advent of biochemistry, cellular and molecular biology and their large expansion during the 20th century, details on so many individual cellular components, processes and to some extent on their relationships were generated.

As the sub-cellular complexity was being unraveled, with so many components and interactions, the technologies and methods to study them also evolved, going from single-throughput to high-throughput starting from the 1990's. Such technological advances span areas such as microscopy, cytometry, spectroscopy, amplification & hybridization techniques and sequencing. Together with computer science and bioinformatics they lead to a flood of ever increasing complex biological data, in particular in the post-genome era where the different *omics* techniques generate large-scale data on genomes, transcriptomes, proteomes, metabolomes, interactomes, etc. [491].

As knowledge increases and more discoveries and inventions are made, more complex and harder mysteries remain, which is the root for an ever-increasing complexity and demands for unraveling biological processes and relationships. In agreement with a shift from individual or few components to network approaches, attempting to simultaneously consider many interacting partners, more data has been generated, covering multiple aspects of the cellular or organism functioning.

These data require a higher capacity to process, catalog and store them as well as effective methods to translate information into reliable knowledge. In this context, online databases became indispensable resources housing huge amounts of information on the most varied topics. One of the current challenges for the scientific community is exactly on which information sources to rely on for obtaining the most correct evidences.

Parallel to the comprehension of a very complex picture of human functioning with so many components and dynamic interactions and with the increasing amounts of data necessary to address biological questions, modelling also became an invaluable resource for biology, in particular for the study of phenomena for which experimentation is limited. Figure 1.6 (page 37) presents an overview of systems approaches to study cellular processes and functioning.

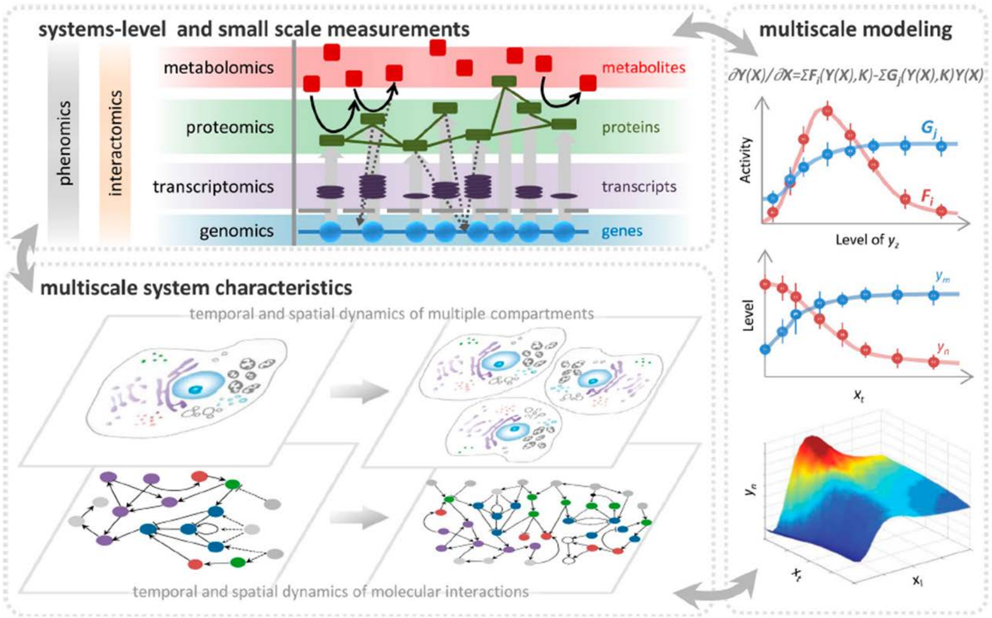

systems-level and small scale measurements
phenomics
interactomics
metabolomics
proteomics
transcriptomics
genomics
metabolites
proteins
transcripts
genes
multiscale modeling
∂Y(X)/∂X=ΣF_i(Y(X),K)-ΣG_j(Y(X),K)Y(X)
Activity
G_j
F_i
Level of y_z
Level
y_m
y_n
x_t
y_n
x_t
x_1
multiscale system characteristics
temporal and spatial dynamics of multiple compartments
temporal and spatial dynamics of molecular interactions

Figure 1.6: Overview of a systems approach to study cellular functioning [8]. Systems biology holds upon the *omics* techniques to acquire vast amounts of data from many different cellular aspects. Data integration is necessary to place together these different aspects and provide a unified view of the system or part of it. Modelling cellular processes is useful for better understanding their complexity, to predict unknown or missing information and to visualize properties of the system.

1.7.1 Modelling Biological Processes

Modelling in biology involves models representing a process or organism of interest, which is a defined system. Mathematical algorithms are used to perform specific tasks and the computational power is used to find possible solutions to problems defined through the algorithms, models and available data. Modelling is entangled with data integration and visualization of biological entities and their relationships, with different detail levels. Such models can be of biological networks, e.g. gene regulatory, signalling, metabolic; whole cell models, increasingly hard with the complexity of an organism; models of a specific process; multi-cellular or organ models; and whole organism models.

Due to the heterogeneity of cellular processes and interactions, different cellular dimensions are usually modelled through different mathematical formalisms, which relate with the nature of the processes in study and question to be answered as well as historical aspects [492–494].

Vast amounts of *omics* data from individual components and their interactions allow building models and iteratively simulate and curate them, making models of diverse cellular processes abundant in biology, with common model repositories such as BioModels [495].

Within this thesis, more focus was given to the gene regulatory and metabolic networks, with metabolic modelling of the adipocyte differentiation. Therefore, the following section addresses modelling in context of metabolism.

Metabolic models and modelling

Metabolic models are mathematical formulations of the knowledge about biochemical reactions and their properties. They represent the classical biochemical pathways and reaction details in a frame that can be simulated and challenged *in silico*. For that, the equations of biochemical reactions are stored in files with a specific structure and format, specifying the components of a metabolic network, including reactions, metabolites and their relationships, defining substrates and products and the stoichiometry of their conversions, often with additional information regarding their properties and usage. These are called metabolic models and several different formats exist with particular syntaxes.

One of the most widely used formats is the Systems Biology Markup Language (SBML) [496], "a machine-readable exchange format for computational models of biological processes"(23), based on XML, freely available and capable of representing many different classes of biological phenomena, including metabolic networks, cell signaling pathways, regulatory networks, infectious diseases, and many others. Within this thesis SBML metabolic models were used in context of human adipocyte differentiation.

Metabolic models embed the stoichiometric matrix (S), in which the rows represent biochemical compounds (metabolites), the columns biochemical reactions and whose entries contain the stoichiometric coefficients (integers) that link a reaction to a metabolite, which can be directly used as the basis on which to computationally infer network properties (e.g. null space analysis).

First models of human metabolism

The first genome-scale human metabolic network models appeared in 2007, Recon1 [497] and the Edinburgh Human Metabolic Network (EHMN) [498], by two independent groups. These were the result of spurious and methodical manual

(23) `http://sbml.org/Documents/FAQ#What_is_SBML.3F`, as of 24.07.2015.

curation of the literature together with genome annotation and biochemical evidence for reactions. They have since then been updated [499, 500] and several additional generic and context-specific human metabolic models were developed [501–506], together with the dissemination of a protocol for the high-quality generation of genome-scale metabolic reconstructions [507].

Besides containing details on biochemical reactions, their properties and involved species, metabolic models also describe the genes that encode the enzymes of the metabolic network. A metabolic pathway is included in a model through the connections between its metabolites and reactions, which represent enzymes, transporters, diffusion or exchanges. Gene-protein-reaction associations (GPRs) are then used to include genes in a metabolic model, based on the knowledge of which protein or protein subunit the gene encodes for, and of which reaction(s) the protein is involved in. The annotation for gene-reaction associations in Recon1 has been more careful, and it provides rules of which genes are necessary for a reaction to occur as well as their relationship, allowing for easy coupling with gene expression data. For this reason, within this thesis only Recon1 was used.

Recon1 is based on the genome annotation Build 35 (2004) and accounts for 1905 genes, 2004 proteins, 2766 metabolites, and 3742 metabolic and transport reactions, while Recon2 contains $\geqslant$ 7000 metabolic reactions [500]. Tissue/cell-specific models can be generated from the generic human models by introducing or removing reactions based on literature curation or various *omics* techniques, namely e.g. transcriptomic, proteomic, metabolomic and phenotypic data. The context-specific data is mapped to the metabolic reconstruction via GPRs that allow for excluding links contained in the general model for which no evidence is found (e.g. from inactive genes, proteins or absent metabolites) and also adding links based on specific data. These links consolidate experimental noise and allow to more accurately infer the metabolic activity reflected in the data. By specifying inputs, outputs and constraints derived from experimental data, one can achieve a highly specific model for the study system. The accuracy of such models depends on the knowledge available for the system, on how well transcription correlates with enzyme abundance and activity and on metabolite availability for a particular reaction (e.g. metabolite concentration, complexation state, intracellular location).

Constraint-based modelling of metabolism

Organisms are limited to live a certain time and attain a certain growth. Species evolution represents a prime example of how adaptation has been directing which characteristics are retained and which are lost throughout generations, with advantageous outfits being positively selected. Overall, from the simplest to the most complex characteristics of living organisms, constraints are everywhere.

Cells, the units of life, represent as well a constrain that delineates an internal or intrinsic space from an external environment. In regards to metabolism, its intrinsic chain organization where the products of upstream reactions are the substrates of the following with many transport reactions connecting different processes and compartments, many constraints apply. Therefore, modelling metabolism by imposing constraints to the known biochemical conversions follows a biological reasoning. The challenge relies on identifying the set of constraints characterizing different cellular processes, cell and tissue types and conditions [508–510].

Constraint-based modelling (CBM) allows to study large-scale metabolic networks relying only on simple physico-chemical and physiological constraints without accurate kinetic constant values or enzyme and metabolite intracellular concentrations, as required by kinetics-based models describing the change in metabolite concentrations over time [511–513].

One simplification of most CBM methods is the assumption of a steady state for the metabolic network, with an equal rate of production and consumption for each metabolite, meaning no accumulation or depletion of metabolites despite flux through reactions. In this case, exchange reactions account for replenishment and drainage of metabolites, keeping a flow through the system where the overall supply equals the drainage. Equation 1.1 generally describes the change in the concentration of metabolites ($\vec{C}$, vector of concentrations) in a network over time (d$\vec{C}$/dt), which corresponds to the product of the stoichiometric matrix (S) and the vector of fluxes through each reaction ($\vec{v}$).

$$\frac{d\vec{C}}{dt} = S * \vec{v} \tag{1.1}$$

Under the steady-state assumption, with no change in the concentration of metabolites, we have the relation described in Equation 1.2.

$$\frac{d\vec{C}}{dt} = 0 \tag{1.2}$$

And substituting in Equation 1.1, we get the relation:

$$S * \vec{v} = 0 \tag{1.3}$$

Equation 1.3 is generally used to compute the metabolic state of a network assuming the steady state.

Available information on protein localization, kinetic constants and intracellular concentrations of enzyme and metabolite further extend and allow various mathematical and *in silico* network analyses. Constraints are mathematically described as balances or bounds. Bounds constraining the values of individual variables are usually identified, leading often to inequalities (e.g. $v_{min} \leqslant v \leqslant v_{max}$). Several methods/algorithms for constraint-based analysis have been developed and validated [508, 514].

In general terms, there are four constraint types limiting cellular functions: physico-chemical (e.g. mass, energy and momentum conservation or enzyme turnover), spatial or topological (e.g. crowding of molecules inside cells), condition-dependent environmental constraints (e.g. nutrient availability, temperature, pH) and regulatory or self-imposed constraints (e.g. amounts of gene products [transcriptional and translational regulation] and their activity [enzyme regulation]; gene repression in response to external signals).

A summary of CBM methods and their applications can be found on `http://cobramethods.wikidot.com/start`, which contains descriptions for many of the CBM methods available, several implemented under the Constraint-based Reconstruction and Analysis toolbox (COBRA) [515] for Matlab.

CBM methods are thereby very useful to study metabolism at the genome-scale, providing a framework to simulate cellular growth, drug response, gene deletion, the effect of a treatment or differentiation in metabolism as well as a platform for integrating and conciliating multi-*omics* data for integrative analysis. The prediction of metabolic activity at the genome-scale is an important task in Systems Biology and can aid to understand how cells adapt to perturbations and evolve with time by providing a snapshot of the metabolism and cellular state on a particular condition or state.

In 2008, Shlomi *et al.* [516] introduced a constraint-based computational method allowing to systematically predict specific metabolic behaviour based on metabolic models and the integration of *omics* data. They exemplified the usage of the method integrating the genome-scale metabolic network Recon1 with tissue-specific gene- and protein-expression data, allowing them to predict tissue-specific metabolic activity in ten human tissues. The method of Shlomi *et al.*, updated in 2010 by Zur *et al.* [517], prompted the prediction of the metabolic activity during adipocyte differentiation done within this thesis, allowing to integrate transcription data and metabolism and acquire a better notion of how they relate. The authors consider genes to be post-transcriptionally up- or down-regulated based on the discrepancy between the measured levels and the predictions. For additional details on Shlomi's method, refer to the Material and Methods (Chapter 3, page 49).

Still in 2008, another method allowing to generate context-specific networks was presented, GIMME [518]. GIMME removes genes and respective reactions from a general metabolic network based on gene (in)activity, generating smaller specific networks that are able to produce a pre-defined output, set through an objective function. Since the definition of objective function is not straightforward for human cells, throughout the thesis, only Shlomi *et al.* method was used to perform metabolic predictions of adipocyte differentiation.

1.7.2 Data integration and visualization

Biomedical data are increasingly complex and large, having attained massive proportions in the past few decades with constant advancements in the high-throughput technologies. Such data flood has the potential to unravel further knowledge, but also brings us challenges not faced before, such as on data storage and safety, as well as processing and analysis. The different technical methods produce vast quantities of heterogeneous data, prone to different types of noise. Therefore, keeping track with data and revealing new insights is a hard task [519]. Data integration appears thereby very relevant, if not necessary, in order to deal and percept such vast amount of data, or extracting dependencies not observable from any component individually [520]. The collection of more diverse data sets spanning multiple cellular aspects already makes data integration a natural avenue, but having more data *per se* does not imply a gain in understanding from compiling those varied datasets. Gene expression can be mapped into metabolic networks allowing to visualize the distribution of expression levels through the network. Improving from this direct mapping, metabolic modelling can integrate the expression values, the network structure and the constraints therein contained to predict metabolic activity that obeys to those impositions. Direct mapping of expression data into a network fails to expose fluxes that can not occur due to upstream reagent absence or inactivity, for instance. A few of these individual cases might be easy to spot, but not attained by human perception at the level of metabolic or other biological networks. Therefore, data integration is not a straightforward endeavour, requiring much careful methodologies to not confound data meanings and expose often dispersed relationships [521–523], but is useful to improve our understanding of biological processes. Several methods and tools provide means to perform data integration with network analysis or visualization [524–530]. As no known routine provides easy means to automatically generate image metanodes depicting multi-*omics* data that can be overlaid on biological networks, we set out to develop such a tool for automated generation of metanode images of omics data and provide means to quick and easily map and visualize those image nodes on Cytoscape networks.

1.8 Outline

In light with a post-genome era in which the *omics* techniques are providing more and more means to profile and measure diverse cellular processes and components, areas such as bioinformatics and systems biology are constantly being utilized and developed in order to process, integrate, analyze and visualize biological data. Comprehensive and multi-faceted approaches are needed. Here we adopted integrative systems biology approaches to describe processes and relationships. In particular, we focussed on human adipogenesis as an experimental model for cellular differentiation, suitable for studying the regulatory and metabolic cellular architectures, and relevant for diseases including obesity, T2DM, metabolic syndrome and related comorbidities. We set out to integratively describe adipogenesis by collecting a diverse set of experimental data relative to the process and combining it with metabolic modelling and prediction of metabolic activity. Besides integratively analyzing and presenting adipocyte differentiation, we were interested in highlighting disease-associated genes in context with regulators and the metabolic network, as well as exploring the relationship between the regulatory load or convergence on genes and their association to diseases. Despite a limited breadth and scope, we believe our work is a pioneering example of an integrative effort to study and present regulatory and metabolic processes characterizing adipocyte differentiation. Furthermore, it reveals a general principle of higher regulatory load on disease-associated genes, which can be explored through the epigenomic mapping of active enhancers to prioritize novel candidate genes for disease association.

2 Scope and Aims

In the present post-genome and *omics* era, a major challenge and bottleneck for improving our understanding of biological processes in health and disease and translate it into actionable treatment routines is the ability to effectively integrate and consolidate huge amounts of heterogeneous data on various complex cellular components, related processes and mechanisms and ultimately provide a unified view of organism funtioning. Such endeavour is required because that is the setting of an organism, which functions as an integrated unit that can not be disentangled without functional loss. Additionally, the >100 years of biological and medical research and knowledge led to the accumulation of a large information repository on individual components and processes, spanning most avenues of human functioning. The biomedical data needed and produced nowadays is heavier and more complex than ever, imposing integrative approaches to conciliate the existing knowledge with the newly generated data in order to depict the missing links and pieces and further our comprehension on human functioning in health and disease [531] [(1)].

How to effectively integrate and visualize omics data in order to easily grasp biological meanings and relationships? is still a largely unanswered question urging us to develop strategies and methods fulfilling such need.

The ability to integratively describe biological processes would allow us to be more exact about conclusions, more comprehensive regarding mechanisms and interacting partners and more aware of possible off-target interactions, overall providing a better understanding of a process in context of the organism and reducing the burden from inadequate treatments.

Within this thesis, focus is given to adipogenesis as an experimental model for cellular differentiation, suitable for studying the regulatory and metabolic cellular architectures, and relevant for diseases including obesity, T2DM, metabolic syndrome and related comorbidities.

Based on the cellular setting with a signal responsive dynamic genome organization, one could speculate that genes with more control, or with a higher regulatory load, might be more relevant for cellular-specific functions and thereby also more likely to be implicated in diseases. My work addressed this question, in particular the relation between TF and enhancer load on genes and their association to disease (Results chapter 4.3), in particular regarding complex diseases, which could facilitate prioritizing novel candidate genes for disease-association.

Accordingly, the **aims** of the work described in this thesis were:

[(1)] `http://apps.who.int/iris/handle/10665/152819`, as of 10.06.2015.

1. **To study human adipocyte differentiation with focus on the interplay between the gene regulatory and metabolic networks by collecting multiple experimental data and predicting metabolic activity;**

2. **To present an integrated view of human adipogenesis based on the collected data and metabolic modelling, focussing on the expression dynamics and regulator incidence on known key lipid metabolism and dyslipidemia genes;**

3. **To test whether a general principle of higher regulation of disease genes can be observed across cell-types using public TF and active enhancer data and if that principle applies for cell-type related diseases;**

4. **To investigate properties of the high regulatory load genes that could segregate them from other genes**.

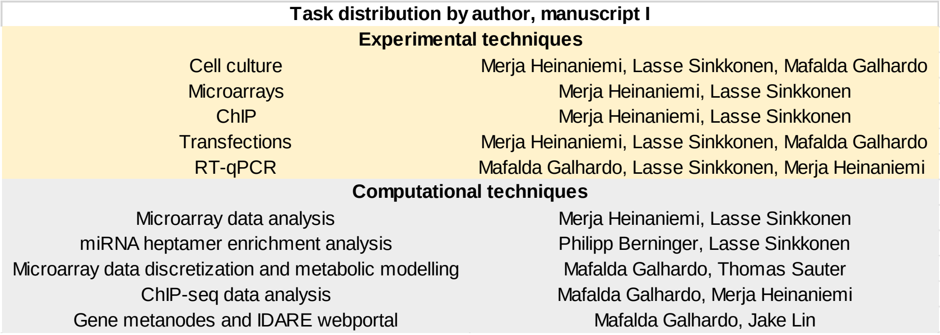
Resulting from the above mentioned aims, a total of 5 manuscripts have been produced.

The first, **Manuscript I** (page 59, section 4.2), fulfills aims 1 and 2 and is entitled “Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network”. It was published in *Nucleic Acids Research* on 2014 (PMID: 24198249).

The author contribution for the main tasks within **Manuscript I** is summarized within Figure 2.1 (page 47). Accordingly, I contributed to all figures except Figures 5 and S4-S7.

The second, **Manuscript II** (page 117, section 4.3), fulfills aims 3 and 4 and is entitled “Cell type-selective disease-association of genes under high regulatory load”. It has been accepted for publishing also in *Nucleic Acids Research* on 14.08.2015 and was published online on 03.09.2015 (PMID: 26338775).

As stated in the “AUTHOR CONTRIBUTIONS” section of Manuscript II, I performed all the analysis steps except the 3’UTR and miRNA analysis, done by Philipp Berninger, and the liver disease network extraction and betweeness centrality calculation for each gene in the network done by Thanh-Phuong Nguyen. Accordingly, I contributed to all figures except Figures 7, S4 and S5.



Task distribution by author, manuscript I	
Experimental techniques	
Cell culture	Merja Heinaniemi, Lasse Sinkkonen, Mafalda Galhardo
Microarrays	Merja Heinaniemi, Lasse Sinkkonen
ChIP	Merja Heinaniemi, Lasse Sinkkonen
Transfections	Merja Heinaniemi, Lasse Sinkkonen, Mafalda Galhardo
RT-qPCR	Mafalda Galhardo, Lasse Sinkkonen, Merja Heinaniemi
Computational techniques	
Microarray data analysis	Merja Heinaniemi, Lasse Sinkkonen
miRNA heptamer enrichment analysis	Philipp Berninger, Lasse Sinkkonen
Microarray data discretization and metabolic modelling	Mafalda Galhardo, Thomas Sauter
ChIP-seq data analysis	Mafalda Galhardo, Merja Heinaniemi
Gene metanodes and IDARE webportal	Mafalda Galhardo, Jake Lin

Figure 2.1: Summary of the distribution of main tasks from Manuscript I, per author.

Additionally, IDARE2, a tool developed for the automated generation of multi-*omics* image metanodes and mapping into Cytoscape networks, is described in **Manuscript III** (page 147, section 4.4), entitled “IDARE2 - Simultaneous visualization of multi-omics data in Cytoscape”, which is in preparation for a soon submission. The tool was mainly developed by Thomas Pfau (thomas.pfau@uni.lu), with support for setting up the web-server from Jake Lin (jake.lin@uta.fi). IDARE2 consists of an upgrade to IDARE, a web-portal designed in the context of **Manuscript I** presented in this thesis. My contribution was on the conceptual framework and extensive testing.

Still in context of **Manuscript I**, two technical summary reports were generated in order to describe in more detail the experimental and technical methods employed within the analysis, presented in appendix, respectively, **Manuscript IV** (page 224) and **Manuscript V** (page 228). These have been published in *Genomics Data*.

3 Materials and Methods

Detailed description of the materials and methods utilized within this thesis can be found in the three attached manuscripts - Results sections **4.2** (page 59), **4.3** (page 117) and **4.4** (page 147) - respectively, **Manuscript I** - "Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network", **Manuscript II** - "Cell type-selective disease association of genes under high regulatory load" and **Manuscript III** - "IDARE2 - Simultaneous visualization of multi-omics data in Cytoscape".

In the following pages, I summarize the methodology applied to achieve the aims outlined in Chapter **2** (page 45, **"Scope and Aims"**), mainly covering **Manuscript I** and **Manuscript II**, followed by a short overview on methods used in **Manuscript III**.

3.1 Overview of the materials and methods employed in Manuscript I

In order to study human adipogenesis comprehensively and integratively, we collected diverse experimental data sets from the differentiation of SGBS pre-adipocytes into lipid-loaded adipocytes, described in detail in **Manuscript I** (4.2), in its "MATERIALS AND METHODS" section starting from page 63.

The SGBS human pre-adipocyte cell line was kindly provided by Prof. Dr. M. Wabitsch (martin.wabitsch@uniklinik-ulm.de).

A time course of the gene expression during adipocyte differentiation using Illumina HT-12 v3 microarrays served to obtain a global view of the expression dynamics during adipogenesis, namely which genes changed the most and at what stage of differentiation.

Firstly, based on the gene expression profile of TF genes, the 3 highest differentiation-induced TFs were selected for profiling their genome-wide binding in adipocytes using ChIP-seq, a keystone of our work which served the purpose of linking gene expression changes with key regulators.

Secondly, based on the expression of metabolic genes, we employed the constraint-based method by Shlomi *et al.* [516] to predict the metabolic activity of adipocytes at different time points of the differentiation, a fairly innovative and little used approach allowing to consider metabolism at the genome scale. This CBM method takes the expression data together with a metabolic model and employs a mixed integer linear programming (MILP) problem that finds a metabolic activity distribution satisfying the stoichiometric and thermodynamic constraints embedded in the model and maximizing the number of reactions whose predicted activity is consistent with the expression of respective enzymes or transporters. Shlomi's method keeps an overall consistent network state, meaning that if an upstream enzyme is lowly expressed and a downstream enzyme is highly expressed, the reaction associated to the latter will still be predicted inactive as the reaction from a reagent step was inactive. Thereby, for each reaction in the network, the method runs two calculations for the overall network similarity with gene expression, one setting the current reaction active and the other inactive. The one giving the highest overall network similarity to expression data will be taken as predicted activity. In case both an active and inactive predictions give the same similarity, the activity of the reaction remains undetermined. The difference between Shlomi's method and the simple overlay of expression data on the metabolic network is the effect of the metabolic network topology and overall network activity distribution, which has to be consistent. Therefrom, the reaction activity predictions will deviate from the expression data according to the network properties. Owing to the MILP nature of the method, a conversion of continuous gene expression data into a discrete category (discretization step) is required. Low and highly expressed genes are taken into account for the predictions, which try to inactivate reactions associated to lowly expressed genes and to activate reactions linked to highly expressed genes. Therefore, the discretization method, for instance applying a threshold to segregate genes into three categories, has a significant impact on the activity predictions, which varies according to the GPR associations affected. One of the drawbacks of Shlomi's method is its considerable running time, which depends mainly on the number of reactions with non-null data.

Thirdly and supplementing the genome-wide binding profiles for 3 key adipogenesis regulators, we also profiled the H3K4me3 histone modification mark in pre-adipocytes in comparison to adipocytes, which could be indicative of changes in the chromatin state at the origin of a disparate gene program usage between pre-adipocytes and adipocytes. Additionally, we expanded our view on the regulation of adipogenesis by profiling as well the expression of miRNAs throughout differentiation using miChip (v.11.0) arrays and selecting 3 down-regulated miRNAs for further investigating their target genes, which we did by profiling the expression of genes with Illumina HT-12 v4 microarrays after over-expressing the 3 selected miRNAs and then collecting down-regulated genes containing in their sequence the complementary of the miRNAs seed sequence.

The previous paragraphs addressed the multiple experimental data types collected during SGBS adipocyte differentiation, with a focus on key regulators and prediction of metabolic activity based on gene expression, as described in the **first aim** of Chapter **2** (**"Scope and Aims"**, page **45**) .

The above mentioned diverse experimental datasets paved our way into an integrated analysis and concerted view of adipocyte differentiation, leading us to develop custom gene "metanodes" enclosing the gene centric data (multiple data types associated to one node). These gene metanodes were then embedded on familiar metabolic pathways, linked to reactions representing enzymes. In this representation, enzymes are connected by the metabolites involded in the respective reactions, through edges representing the reversibility of an enzyme with direct or reverse arrows. The edges of the metabolic pathways were colored based on the difference in predicted metabolic reaction activity comparing pre-adipocytes and adipocytes, depicting the general metabolic flow upon differentiation. In order to represent the diverse experimental datasets encompassing the gene expression dynamics during adipogenesis, the regulatory convergence from 3 TFs and 3 miRNAs and the H3K4me3 change between pre-adipocytes and adipocytes and the metabolic activity prediction, we set up a webportal for interactive data exploration, **IDARE**[(1)], with highlight for disease-associated genes, fulfilling the aim of presenting an integrated view of human adipogenesis based on the collected data and metabolic modelling, with focus on the expression dynamics and regulator incidence on known key lipid metabolism and dyslipidemia genes, the **second aim** described in Chapter **2** (**"Scope and Aims"**, page **45**).

Additionally at this step, we collected public ChIP-seq data from the genome-wide binding of 10 TFs in HUVEC cells (bed peak files), either unstimulated from the ENCODE project [12] (cMYC, GATA2, MAX, cJUN, cFOS) or from the SRA archive (ETS1 from VEGFA stimulation, MEF2C from statin stimulation, p65 from TNF stimulation, FLI1 representing an endothelial cell developmental TF, HIF1A from hypoxia). After peak calling and peak-to-gene association with the tool GREAT [532], a list of the number of TFs per gene was obtained. We further focussed on metabolic genes and used the hypergeometric distribution to test for the enrichment of vascular disease-associated genes among genes with the highest TF load (up to 10).

Detailed description of the materials and methods employed in **Manuscript I** can be found from page 63, in its "MATERIALS AND METHODS" section. In addition, two technical summary reports were generated in context of **Manuscript**

[(1)] `http://systemsbiology.uni.lu/idare.html`

I, which describe the experimental techniques and computational analysis therein employed, respectively, **Manuscript IV** and **Manuscript V**, in appendix.

Manuscript IV - "Transcriptomics profiling of human SGBS adipogenesis" (page 224), describes the experimental design, material and methods for the identification of differentially expressed genes during SGBS adipocyte differentiation using microarrays combined with the prediction of metabolic activity associated to the gene expression, using the method of Shlomi *et al.* (2008).

Manuscript V - "ChIP-seq profiling of the active chromatin marker H3K4me3 and PPARγ, CEBPα and LXR target genes in human SGBS adipocytes" (page 228) describes the experimental design, material and methods for identifying the most highly induced TFs and their putative targets during SGBS adipogenesis, using ChIP-seq, including deep-sequencing quality controls.

Given their technical nature, no further details are given to **Manuscript IV** and **Manuscript V**.

3.2 Overview of the materials and methods employed in Manuscript II

To address the link between regulatory load and disease association, we used public data from the ENCODE [12], the BLUEPRINT [316] and the NIH Roadmap Epigenomics [317] projects, in an exclusively computational analysis described in section 4.3, **Manuscript II**. Detailed description of all analysis steps can be found in the "MATERIALS AND METHODS" section within Manuscript II (page 120).

The tests for enrichment of disease-associated genes among the genes with a high regulatory load were done using the hypergeometric distribution [532–534], with gene-disease associations based on the DisGeNET database[(2)] [378]. We required a minimum association score of 0.08 to exclude associations supported by automated text-mining and a minimum of 15 genes associated to a disease in order to avoid significant results due to very small set sizes, resulting in 340 diseases from the DisGeNET tested per sample. The hypergeometric distribution is useful for calculating the significance of a known result from a two-by-two factor experiment, without replacement. In the present case, our two-by-two factors are the setting of high regulatory load (HRL) and non-HRL genes *versus* disease and non-disease genes, being the null hypothesis that the proportion of disease-associated genes

[(2)] `http://www.disgenet.org/`.

among HRL genes is not statistically significantly higher from that among non-HRL genes. Genes were grouped in bins of regulatory load, per sample, and each bin of genes was tested for enrichment of the genes associated to each of the 340 DisGeNET diseases, using the hypergeometric distribution.

To test whether genes under high regulatory load enrich for disease association in comparison to other genes in multiple cell types, we gathered ChIP-seq data from the genome-wide binding profiles of 93 TFs in 9 cell lines from the ENCODE project. As enhancers are known to be involved in gene transcription activation mechanisms, we equally took the ENCODE data from the same cell lines on the H3K27ac mark, known to characterize active enhancers. Furthermore, we realized the availability of the H3K27ac data for another 11 ENCODE cell lines, as well as from the BLUEPRINT and the NIH Roadmap Epigenomics projects, which we obtained totalizing a set of 139 samples covering 96 cell types and tissues (Supplementary file I). The GREAT tool [532] was used to perform associations between ChIP-seq peaks and genes, resulting in a list of genes ranked by the number of associated TFs or by the number of H3K27ac peaks falling within each genes regulatory domain as defined by the “Basal + extension” rule of the GREAT tool, respectively the TF load per gene and enhancer load per gene, or generally, the regulatory load per gene, for each sample. The H3K4me3 data of each sample was used to filter out genes lacking the mark within their TSS ± 1000 bp, considered to be in closed chromatin regions and thereby not transcribed.

The previous paragraphs addressed the different datasets and processing steps performed in order to test for a general principle of higher regulation among disease genes across cell-types, for both TFs and active enhancers, with a diverse set of diseases which could expose cell-type-related enrichments, thereby **addressing the third aim** described in Chapter **2** (**“Scope and Aims”**, page **45**).

Additionally, Kyoto Encyclopaedia of Genes and Genomes (KEGG) [341] pathways were used to assess the participation of the HRL genes in multiple pathways, in comparison to other genes, which could give a hint about their network relevance. To compare the average number of KEGG pathways per HRL genes to that of other genes, we performed 10000-fold permutation tests by randomly picking an equal number of genes as HRL and obtaining the average number of pathways they participated in. From here, we derived a p-value from the ratio between i) the number of times the average number of KEGG pathways per gene from the randomly selected genes was at least that of the HRL genes and ii) 10000 (the total number of random permutations). The described routine was separately done for each of the 139 samples. Remaining within network properties, we then focused on liver samples (primary liver and HepG2) and derived a liver gene-disease network based on MeSH terms including 137 liver-related disorders. Based in these diseases, liver disease genes (nodes) were extracted from the Comparative Toxicogenomics Database (CTD) [375], considering only curated disease-gene associations, and extending with interactions (links or edges) from the Human Protein Reference Database (HPRD) [535] (accessed on June 2015), containing manually curated protein interactions from literature and leading to a liver-disease network of 3775 genes and 8278 interactions. We then calculated the betweenness centrality for each gene and compared the average betweenness centrality of the high regulatory load and other genes in liver samples, to investigate the role of these genes in the network.

We further compared the average 3'UTR length of the HRL *versus* other genes, which could be indicative of the post-transcriptional control, based on annotation from Biomart (Ensembl Genes 78), obtaining as well predicted target sites for conserved miRNAs from TargetScan 6.2 [215] and summing up the target site count per 3'UTR, resulting in an average target site count per transcript, which we also compared between the HRL and other genes. In this way, we studied multiple properties of the HRL genes and tested if they were different from other genes, fulfilling **the fourth aim** described in Chapter **2** (**"Scope and Aims"**, page **45**).

3.3 Overview of the methods employed in Manuscript III

To address a generalized lack for automated generation of custom multi-omics image metanodes and easily mapping those images onto Cytoscape networks, **IDARE2** has been developed and is described in **Manuscript III** (section 4.4), starting from page 148.

Briefly, IDARE2 image metanode generation is implemented on Matlab and access to users is granted via the web-server `http://idare-server.uni.lu/` (currently still only available within the uni.lu network), which is an interface for data upload and to request jobs for automated metanode generation.

The tool takes as input one or more user-provided datasets, which can range from discrete to continuous values, currently supporting a maximum of 9 simultaneous input files and 20MB per file. Several "DataType" choices are available, including ItemData for a less stringent organization and repositioning of the data within the image node. "Heatmap" draws a boxed representation also free on the positioning. "ItemGrid" or "TimeSeries" data types are useful when the user would like to specify an order defining the positioning of the different elements within the node. IDARE2 user guide (page 236) provides detailed descriptions for additional features and properties.

In addition to the metanode generation frontend, IDARE2 also provides a Cytoscape App for an easy mapping and network visualization containing the generated metanodes. The app is composed by an image mapping tool and the IDARE visual style, based on matching keys between images and nodes; a COBRA specific SBML reader for importing gene-protein-reaction associations from COBRA models or metabolic models in general; and a network extractor tool, which collapses large networks into sub-systems, or pathways containing related reactions and metabolites, while keeping the links to the original network and between sub-networks, creating links from each subnetwork to the position of a connecting metabolite, thereby facilitating network navigation and inspection.

4 Results

4.1 Overview

The current **“Results”** chapter is integrally composed by sections containing the manuscripts generated based on the work performed within my PhD studies.

In **section 4.2**, page 59, I present **Manuscript I**, **“Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network”**, *Nucleic Acids Research*, 2014 (PMID: 24198249).

By collecting a diverse and comprehensive set of regulatory data on adipocyte differentiation we were able to show that enzymes crucial or rate limiting for lipid metabolism were often associated with miRNAs and high occupancy binding by multiple TFs, suggesting tight combinatorial effects of TF upregulation and miRNA downregulation driving metabolic changes in adipogenesis.

In **section 4.3**, page 117, I present **Manuscript II**, **“Cell type-selective disease-association of genes under high regulatory load”**, accepted for publishing also in *Nucleic Acids Research* on 14.08.2015 (PMID: 26338775). Herein, Manuscript II refers to the online published version from 03.09.2015 which has been included in the thesis during the correction phase after defence.

Through a computational analysis of public ChIP-seq data, we could observe a general principle of higher regulatory load on disease-associated genes, across multiple tissues and cell types, suggesting that genes under higher regulation and integrating multiple signals might incur a higher likelihood for disease. As a result, we showed that the epigenomic mapping of active enhancers can serve as tool for the unbiased prioritization of novel candidate genes for disease association.

In context of **Manuscript I**, we generated a web-portal for the interactive inspection of metabolic pathways overlaid with metanodes depicting the gene-related data colected for the SGBS and HUVEC cells, **IDARE** (`http://systemsbiology.uni.lu/idare.html`).

IDARE2 supplements the need for a tool to automatically generate user tailored multi-omics metanodes for mapping within (biological) networks and additionally provides easy means to visualize them in Cytoscape.

In section **4.4**, page 147, I present **Manuscript III**, **“IDARE2 - Simultaneous visualization of multi-omics data in Cytoscape”**, which is an upgrade of IDARE and will be submitted for publication soon, being first-authored by Thomas Pfau.

Finally, in **Appendix** (page 223), I include **Manuscript IV** and **Manuscript V**, summary reports of the experimental techniques and analysis employed within **Manuscript I**, which focussed on the integrated analysis and interpretation of biological findings with less details on the technical aspects. I shortly describe next **Manuscript IV** and **Manuscript V** with no further highlights as all biological descriptions and interpretations can be found in section **4.2**, page 59, **Manuscript I**.

Manuscript IV, **"Transcriptomics profiling of human SGBS adipogenesis"**, page 224, describes the experimental design and quality controls applied to profile the expression of SGBS cells during adipocyte differentiation using microarrays, with identification of differentially expressed genes and coupling with constraint-based modelling of metabolism to predict metabolic changes associated with the gene expression.

At last, in **Manuscript V**, **"ChIP-seq profiling of the active chromatin marker H3K4me3 and PPARγ, CEBPα and LXR target genes in human SGBS adipocytes"**, page 228, the experimental design and quality controls applied to identify the putative target genes of the 3 highest induced TFs during SGBS adipocyte differentiation using ChIP-seq are described. PPARγ, CEBPα and LXR showed the highest increased expression, respectively with $\geqslant$ 10000, $\geqslant$ 6000 and $\geqslant$ 2000 putative target genes in SGBS adipocytes. Additionally, it contains similar descriptions for the profiling of the H3K4me3 mark in pre-adipocytes and adipocytes by ChIP-seq.

4.2 Manuscript I - "Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant no-des of the human metabolic network"

Although the biomedical field has largely advanced in the recent decades, complex diseases are increasing among the population and account for a considerable fraction of deaths worldwide[1]. In regards to obesity, T2DM and metabolic syndrome, comorbid diseases affecting many millions of people worldwide, an impaired functioning of adipose tissue has been recognized for long [536–547], resulting in adipogenesis being one of the most studied biological processes in an attempt to understand the mechanisms underlying such impairment and finding ways to revert it. Thereby, by the time I began my PhD studies in September 2012, a vast literature body was already available on adipocyte biology, to cite only a few see [409, 548–556]. A Pubmed search with the key "adipose tissue function review" with a limit for papers until September 2012 retrieves 7890 articles (Supplementary file II). While not all of these will exactly describe adipose tissue function, the number reflects how the topic has been under appreciation throughout the years.

The TFs PPARγ and C/EBPα were recognized master regulators of adipogenesis, capable of triggering a transcriptional cascade leading to lipid-loaded fully differentiated adipocytes. TZD drugs, peroxisome proliferator-activated receptor (PPAR)γ agonists, were often used to improve insulin resistance on patients with T2DM, although having many reported side effects. By activating PPARγ, TZDs would improve adipocyte metabolism and differentiation, with activation of genes involved in glucose and fatty acid uptake and anabolism, leading to a systemic decrease in hyperglycemia and increasing insulin sensitivity, although often causing weight gain due to increasing triglyceride storage.

Despite such vast knowledge, facts were rather dispersed over the literature and therefore hard to realize, and an integrated view of adipogenesis combining the gene regulatory and metabolic networks and highlighting disease-associated genes was not available. Furthermore, the interactive exploration of links between these networks was also hardly possible. To meet this lack, we aimed at **integratively study human SGBS adipocyte differentiation**, in particular combining an experimental and computational approach to globally **depict adipogenic changes involving the gene regulatory and metabolic networks in concert**, through key regulators and prediction of metabolic activity, exposing their convergence on lipid disease-related genes.

In order to achieve such endeavour, we generated:

[1] `http://www.who.int/mediacentre/factsheets/fs310/en/`, as of 01.06.2015.

- a time series gene expression profile during differentiation (microarrays);
- a time series miRNA gene expression profile during differentiation (miRNA microarrays);
- candidate target genes of the miRNAs miRNA-27a, miRNA-29a and miRNA-222, downregulated with adipogenesis, based on seed match analysis upon repression in the array experiment from the over-expression of the 3 miRNAs;
- the genome-wide binding profiles of the three highest differentiation-induced TFs, PPARγ, C/EBPα and LXR in adipocytes (chIP-seq);
- the genome-wide profile of the H3K4me3 histone modification, a mark for actively transcribed genes.

Additionally, data analysis and integration resulted in the generation of:

- metabolic models with the predicted metabolic reaction activity during the SGBS cell transition from pre-adipocyte to day 12 differentiated adipocytes (based on gene expression data and the constraint-based modelling (CBM) method by Shlomi *et al.* [516, 517]);
- the IDARE webportal[2], containing Recon1 metabolic pathways integrated with metabolic predictions and metanodes of the gene regulatory data collected in order to facilitate inspection and interpretation of their relationships.

The main results obtained with such integrative analysis include:

1. a highly dynamic gene expression during SGBS adipocyte differentiation with up-regulation of lipid metabolism genes (Figures S2, S3, 4 (A), 6, 8 (A) and S8);
2. predicted activation of metabolic pathways involved in lipid metabolism (Figures 3, 6 (C), 7, 8 (A) and S8);
3. hundreds of miRNA significant putative targets, ranging from 6 to 12 for metabolic genes (Figures 4 and S4);
4. PPARγ, C/EBPα and LXRα as the highest induced TFs having a large set of putative target genes including on lipid metabolism pathways (Figures 5, 6, 7, 8, S6 and S8);
5. little changes on the H3K4me3 mark between SGBS pre-adipocytes and adipocytes (Figures 5 (D) and S5);

[2] `http://systemsbiology.uni.lu/idare.html`.

6. combinatorial TF and miRNA regulation of the triacylglyceride synthesis and the BCAA catabolism pathways and on several lipid and glucose metabolism genes such as ACADM, CYP1B1, GPAM, HK2, LPL, PISD, RPIA and SCD, several known dyslipidemia-associated genes (Figures 6, 7, 8, S1, S8).

The results described above suggest an extensive and combinatorial regulation on key genes for lipid metabolism, including those already known to associate with dyslipidemia, leading us to hypothesize that disease genes might be under tighter regulation than genes in general. To test this hypothesis in a larger dataset, we gathered chIP-seq data from the genome-wide binding of 10 TFs in HUVEC cells, publicly available, and derived a set of vascular-disease-associated genes from the DisGeNET, using the hypergeometric distribution to calculate the enrichment of disease genes among the genes with between 1 to 10 TFs, obtaining a $>$ **2-fold enrichment for vascular-disease-associated genes among genes with between seven and nine TFs**, namely the nitric oxide synthase 3 (endothelial cell, NOS3) putatively bound by eight out of ten TFs (Figure 2).

Manuscript I is integrally presented starting from page 62.

Nucleic Acids Research Advance Access published November 5, 2013

Nucleic Acids Research, 2013, ***1–23***
doi:10.1093/nar/gkt989

Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network

Mafalda Galhardo[1], Lasse Sinkkonen[1], Philipp Berninger[2], Jake Lin[3,4], Thomas Sauter[1,*] and Merja Heinäniemi[1,5,*]

[1]Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg, [2]Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland, [3]Institute for Systems Biology, 401 Terry Avenue North, 98109-5234, Seattle, Washington, USA, [4]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, House of Biomedicine, 7 Avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg and [5]Department of Biotechnology and Molecular Medicine, A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, FI-70211 Kuopio, Finland

Received July 8, 2013; Revised September 17, 2013; Accepted October 2, 2013

ABSTRACT

Metabolic diseases and comorbidities represent an ever-growing epidemic where multiple cell types impact tissue homeostasis. Here, the link between the metabolic and gene regulatory networks was studied through experimental and computational analysis. Integrating gene regulation data with a human metabolic network prompted the establishment of an open-sourced web portal, IDARE (Integrated Data Nodes of Regulation), for visualizing various gene-related data in context of metabolic pathways. Motivated by increasing availability of deep sequencing studies, we obtained ChIP-seq data from widely studied human umbilical vein endothelial cells. Interestingly, we found that association of metabolic genes with multiple transcription factors (TFs) enriched disease-associated genes. To demonstrate further extensions enabled by examining these networks together, constraint-based modeling was applied to data from human preadipocyte differentiation. In parallel, data on gene expression, genome-wide ChIP-seq profiles for peroxisome proliferator-activated receptor (PPAR) γ, CCAAT/enhancer binding protein (CEBP) α, liver X receptor (LXR) and H3K4me3 and microRNA target identification for miR-27a, miR-29a and miR-222 were collected. Disease-relevant key nodes, including *mitochondrial glycerol-3-phosphate acyltransferase* (GPAM), were exposed from metabolic pathways predicted to change activity by focusing on association with multiple regulators. In both cell types, our analysis reveals the convergence of microRNAs and TFs within the branched chain amino acid (BCAA) metabolic pathway, possibly providing an explanation for its downregulation in obese and diabetic conditions.

INTRODUCTION

Several diseases caused by dysfunction in metabolism have become prevalent in human populations worldwide. Among these, cardiovascular disease (CVD) represents the leading cause of death worldwide. Obesity is a major risk factor for CVD, in particular when accompanied with insulin resistance, hypertension and altered blood lipid profiles (1). These in combination are referred to as the metabolic syndrome that also confers risk to develop diabetes and cancer (1).

High-quality genome-scale metabolic reconstructions are now available that represent the entire network of metabolic reactions a given organism is known to exhibit (2,3). Metabolic fluxes within the network adapt according to enzyme activity, substrate, cofactor, energy, metabolite and product availability as well as posttranslational regulation (4,5). Current technologies allow the characterization of global phenotypes on the transcriptome level through deep sequencing of RNA and DNA

*To whom correspondence should be addressed. Tel: +358 40 3553049; Fax: +358 17 163751; Email: merja.heinaniemi@uef.fi
Correspondence may also be addressed to Thomas Sauter. Tel: +352 46 66446296; Fax: +352 46 66446435; Email: thomas.sauter@uni.lu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

molecules. However, global measurements of proteome activity or metabolic fluxes remain a bottleneck. To address the latter limitation, it is possible to leverage the ability of mathematical models to integrate various data types to reveal central changes in metabolism. These mathematical representations allow the computation of physiological states. For estimating reaction activities, a method was proposed (6) where the expression levels serve as a soft-constraint to favor consistent solutions in concordance with the mass conservation in the metabolic network.

Alterations in the expression status are an initial step for a metabolic shift and can serve as a predictor of the metabolic activity cells are able to sustain. For this reason, the regulator molecules actuating this shift represent candidate therapeutic targets. In adipocytes, two transcription factors (TFs), peroxisome proliferator-activated receptor γ (PPARγ) and CCAAT/enhancer binding protein α (CEBPα), have been shown to be the key regulators: they are required to initiate terminal differentiation and are sufficient to convert other cell types to adipocytes (7), manifested through their genome-wide binding profile that positions them as master regulators of the adipocyte expression profile (8–10). Several antidiabetic drugs have been developed that activate PPARγ (11). The widely used CVD drugs statins on the other hand impact cholesterol levels through genes regulated by the signal-responsive TFs sterol-regulatory element binding factors (SREBFs) and liver X receptors (LXRs) (12). It is highly likely that interactions among TFs could play a role in disease, yet less is known so far how their targets overlap. Recent studies have also placed attention on the role of noncoding RNA regulators known as microRNAs (miRNAs) during adipocyte differentiation of cell culture and *in vivo* models (13,14), identifying counteracting regulators such as the miR-27 family and let-7 (15–18). We have recently identified several miRNAs as primary PPARγ target genes in mouse adipocytes (19), yet it remains unclear to what extent these different regulators converge to control the metabolic phenotype and whether identifying their convergence points could improve therapeutic interventions.

The Encyclopedia of DNA Elements (ENCODE) project has built an extensive list of functional elements in the human genome, including regulatory elements bound by TFs that control gene activity (20). Human umbilical vein endothelial cells (HUVECs) belong to the panel of ENCODE cell types with most data available and are also widely used as a model cell line in CVD research. Here, we hypothesized that observing the regulation of metabolic genes via multiple regulators (epigenetic, transcriptional and posttranscriptional) could indicate a plausible high relevance in a disease context. Moreover, to delineate the metabolic activity shifts affected by these key nodes, such an integrative analysis could become informative coupled with mathematical modeling of reaction activities. To allow data sources of gene regulation [such as ENCODE (20)] and metabolic network models (2,3) to be explored in an integrative manner, we used a tripartite graph representation and developed an interactive web portal, Integrated Data Nodes of Regulation (IDARE, http://systemsbiology.uni.lu/idare.html, see User Guide in Supplementary Material), that can be used to visualize global or tissue-specific data. This integrative experimental and computational analysis allowed us to address the connectivity between the human regulatory and metabolic networks.

Using just the overlap of TF-gene associations and the metabolic network, we observed a strong enrichment of disease-associated nodes among genes that show TF binding in multiple HUVEC ChIP-seq studies, including the *nitric oxide synthase* (NOS) gene family. We collected further experimental data on TFs and miRNAs in adipocytes differentiated from Simpson–Golabi–Behmel syndrome (SGBS) preadipocyte cell line, an established model for human adipogenesis (21). Interestingly, each of the previously characterized dyslipidemia genes *LDLR* (*LDL receptor*) (22), *LPIN1* (*lipin 1*) (23) and *LPL* (*lipoprotein lipase*) (24) that belong to the triacylglycerol synthesis and release pathway are highlighted as shared TF- and miRNA-associated genes. Moreover, the cell fate determining TFs were observed to form a multi-TF feed-forward loop with binding sites nearby genes from the cholesterol synthesis and fatty acid activation pathways. Finally, the convergence of miRNAs and TFs highlight the branched-chain amino acid (BCAA) metabolism as a key nonlipid pathway for which altered regulation by the factors studied here may provide an explanation for its association with obesity and diabetes.

MATERIALS AND METHODS

Cell culture and differentiation

The human preadipocyte cell line isolated from a Simpson-Golabi-Behmel syndrome patient (SGBS) have previously been shown to be in many ways identical to differentiated primary adipocytes from healthy donors but maintain their differentiation capacity longer than other isolated cells (21), therefore representing an optimal model system for high-throughput analysis. The SGBS cells were cultured in Dulbecco's modified Eagle's medium (DMEM)/Nutrient Mix F12 (Gibco, Paisley, UK) containing 8 mg/l biotin, 4 mg/l pantothenate, 0.1 mg/ml streptomycin and 100 U/ml penicillin (OF medium) supplemented with 10% FBS in a humidified 95% air/5% CO_2 incubator. The cells were seeded into culture medium flasks or plates, which were coated with a solution of 10 μl/ml fibronectin and 0.05% gelatine in phosphate-buffered saline (PBS). Confluent cells were cultured in serum-free OF medium for 2 days followed by stimulation to differentiate with OF media supplemented with 0.01 mg/ml human transferrin, 200 nM T3, 100 nM cortisol, 20 nM insulin, 500 μM IBMX (all from Sigma-Aldrich) and 100 nM rosiglitazone (Cayman Chemical, Ann Arbor, USA). After day 4, the differentiating SGBS cells were kept in OF media supplemented with 0.01 mg/ml human transferrin, 100 nM cortisol and 20 nM insulin. SGBS cells differentiate within 10–12 days as determined by microscopic analysis (Oil red O staining). At this time point, the cells are filled with small-sized lipid droplets and are most responsive, whereas at later time points (20 days), the lipid droplets fuse and cells are less active (Dr Martin Wabitsch, personal communication). RNA samples were collected at 0, 4, 8 and 12 h and on days 1, 3 and 12 of differentiation and chromatin samples from day 0 (H3K4me3) and day 10 (TFs and H3K4me3). To find LXR-responsive genes, the day 10 differentiated SGBS cells were stimulated with 1 µM T0901317 for 4 h (synthetic agonist for LXRs), while control cells received DMSO (final concentration 0.1%).

miRNA transfection

MiRNA mimics for miR-27a, miR-29a and miR-222 (Thermo Scientific Dharmacon, Colorado, USA) or a scrambled double-stranded siRNA sequence as control (siCtrl) (Eurogentec, Liège, Belgium) were introduced into 4 days differentiated SGBS adipocytes using Lipofectamine RNAiMAX reagent (Invitrogen, Halle, Belgium) according to manufacturer's instructions. Shortly, miRNA mimics or siRNAs were mixed with Lipofectamine RNAiMAX reagent, incubated for 20 min and diluted with plain DMEM-F12 medium to a final concentration of 100 nM. The first differentiation medium was replaced by the transfection mixture and incubated for 2 h before changing to the second differentiation medium (see above). Twenty-four hours after transfection, the cells were collected for RNA extraction.

RNA extraction and real-time quantitative polymerase chain reaction

Total RNA was extracted using TriSure (Bioline, London, UK). One milliliter of TRIsure was added per a confluent six-well to lyse the cells. RNA was extracted with 200 µl of chloroform and precipitated from the aqueous phase with 400 µl of isopropanol by incubating at −20°C overnight. cDNA was synthesized by using 1 µg of total RNA, 0.5 mM dNTPs, 2.5 µM oligo-dT18 primer, 1 U/µl RiboLock RNase Inhibitor (Fermentas, Vilnius, Lithuania) and 10 U/µl M-MuLV Reverse Transcriptase (Fermentas) for 1 h at 37°C. The reaction was stopped by 10-min incubation at 70°C. Real-time quantitative polymerase chain reaction (RT-qPCR) was performed with Applied Biosystems 7500 Fast Real-Time PCR System using Absolute Blue qPCR SYBR Green Low ROX Mix reagent (Thermo Fisher Scientific, Surrey, UK). Five microliters of cDNA template was used with 1 µl of gene-specific primer pairs (10 µM) and 10 µl of the qPCR SYBR mixture in a final reaction volume of 20 µl. The PCR reaction started with 15 min at 95°C to activate the polymerase. The PCR cycling conditions were as follows: 40 cycles, of which each was composed of 15 s at 95°C, 15 s at 55°C and 30 s at 72°C. Fold inductions were calculated using the formula $2^{-(\Delta\Delta Ct)}$, where ΔΔCt is $[Ct_{(\text{target mRNA})} - Ct_{(RPL13A)}]_{\text{differentiated}} - [Ct_{(\text{target mRNA})} - Ct_{(RPL13A)}]$ and the Ct is the cycle at which the threshold is crossed. PCR product quality was monitored using post-PCR melt curve analysis. The primer sequences are provided in Supplementary Table S1.

miRNA assays

The miRNA detection was performed by using TaqMan MicroRNA Reverse Transcription Kit with TaqMan MicroRNA Assays (Applied Biosystems). The miRNA cDNA synthesis and miRNA real-time PCR were done following manufacturer's instructions and by using an Applied Biosystems 7500 Fast Real-Time PCR System. Relative expression levels in the undifferentiated and the 5-day differentiated adipocytes were calculated using the formula $2^{-(\Delta\Delta Ct)}$, where ΔΔCt is $[Ct_{(\text{target miRNA})} - Ct_{(U6)}]_{\text{differentiated}} - [Ct_{(\text{target miRNA})} - Ct_{(U6)}]_{\text{undifferentiated}}$ and the Ct is the cycle at which the threshold is crossed.

Microarray profiling

Total RNA in triplicates from the differentiation time series, LXR agonist stimulation and the miRNA transfections were processed according to the manufacturer instructions to prepare cDNA that was hybridized on microarrays (for the time series and ligand stimulation, the array hybridizations were performed on Illumina HT-12 v3 arrays at the Turku Centre for Biotechnology, Microarray and sequencing facility; for the miRNA transfections, on Illumina HT-12 v4 arrays at DNA Vision, Charleroi, Belgium). The raw data files were processed and quality controlled using the R/Bioconductor lumi package. Raw and normalized expression values are available via GEO (GSE41578). Genes that had a detection $P < 0.05$ were selected for statistical analysis performed using the limma package. The F-test was used to assess significance of overall dynamic response over the differentiation and a two-tailed *t*-test to compare specific time points to day 0 undifferentiated cells (Benjamini–Hochberg adjusted $P < 0.01$ was considered significant). For the miRNA transfections, statistical analysis was based on *t*-test significance comparing mean expression levels on miRNA transfection to a scrambled siRNA control transfection, and similarly the LXR agonist-treated cells were compared with solvent-treated cells. The expression profiles of metabolic genes or TFs were clustered for visualization using self-organizing maps [GEDI software (25)] and AutoSOME (26) as instructed in the tool documentation. Enriched pathways from the human metabolic reconstruction (2) were determined using a hypergeometric test testing for overrepresentation. Genes from Gene Ontology categories with similar gene numbers as Recon1 (1040) were obtained using the GO Online SQL Environment (http://www.berkeleybop.org/goose), as of 12 August 2013: cell projection (747), envelope (630), locomotion (775) and receptor activity (464). The number of probes detected in the array is indicated in brackets for each category from a total of 12 756 detected probes.

miRNA array profiling

Total RNA samples from time points day 0, day 1, day 3 and day 12 of the differentiation time series used for mRNA array analysis (see above) were also used to profile miRNAs using miChip arrays (v.11.0) arrays (27) at the EMBL Genomics Core facility at Heidelberg. The raw signal values from total RNA array hybridizations were median normalized and then further normalized to the respective signals from day 0 samples. Only probes corresponding to mature human miRNAs were included in the analysis.

Chromatin immunoprecipitation

Nuclear proteins were cross-linked to DNA by adding formaldehyde directly to the medium to a final concentration of 1% for 8 min at room temperature. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M and incubating for 5 min at room temperature on a rocking platform. The medium was removed and the cells were washed twice with ice-cold PBS. The cells were then collected in lysis buffer [1% sodium dodecyl sulphate (SDS), 10 mM EDTA, protease inhibitors, 50 mM Tris–HCl, pH 8.1] and the lysates were sonicated by a Bioruptor UCD-200 (Diagenode, Liege, Belgium) to result in DNA fragments of 200–500 bp in length. Cellular debris was removed by centrifugation and the lysates were diluted 1:10 in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, protease inhibitors, 16.7 mM Tris–HCl, pH 8.1). Chromatin solutions were incubated overnight at 4°C with rotation with antibodies against H3K4me3 (4 µl per immunoprecipitation (IP) of 17-614, Millipore, Billerica, MA, USA), PPARγ (mixture of 0.5 µl per IP of sc-7196x, Santa Cruz Biotechnologies, Santa Cruz, CA, USA and 5 µl per IP of 101700, Cayman, Ann Arbor, MI USA), CEBPα (5 µl per IP of sc-61, Santa Cruz Biotechnologies) and LXRα (5 µl per IP, kind gift from Eckardt Treuter, Karolinska Institute, Stockholm, Sweden). The LXR antibody recognizes also LXRβ that maintains a constant low level of expression during differentiation. The immuno complexes were collected with 20 µl of MagnaChIP protein A beads (Millipore) for 1 h at 4°C with rotation. Nonspecific background was removed by incubating the MagnaChIP protein A beads overnight at 4°C with rotation in the presence of bovine serum albumin (250 µg/ml). The beads were washed sequentially for 3 min by rotation with 1 ml of the following buffers: low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris–HCl, pH 8.1), high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl, 20 mM Tris–HCl, pH 8.1) and LiCl wash buffer (0.25 M LiCl, 1% Nonidet P-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris–HCl, pH 8.1). Finally, the beads were washed twice with 1 ml of TE buffer (1 mM EDTA, 10 mM Tris–HCl, pH 8.1). The immuno complexes were then eluted by adding 500 µl of elution buffer (25 mM Tris–HCl, pH 7.5, 10 mM EDTA, 0.5% SDS) and incubating for 30 min on rotation. The cross-linking was reversed and the remaining proteins were digested by adding 2.5 µl of proteinase K (Fermentas) to a final concentration of 80 µg/ml and incubating overnight at 65°C. The DNA was recovered by phenol/chloroform/isoamyl alcohol (25:24:1) extractions and precipitated with 0.1 volume of 3 M sodium acetate, pH 5.2, and 2 volumes of ethanol using glycogen as carrier. Immunoprecipitated chromatin DNA was then used as a template for real-time quantitative PCR or for library preparation and sequencing (performed at EMBL Core facility).

PCR of chromatin templates

Real-time quantitative PCR of ChIP templates was performed using chromatin-region–specific primers in a total volume of 20 µl with Applied Biosystems 7500 Fast Real-Time PCR System using Absolute Blue qPCR SYBR Green Low ROX Mix reagent (Thermo Fisher Scientific, Surrey, UK). The PCR cycling conditions were preincubation for 15 min at 95°C, 40 cycles of 15 s at 95°C, 15 s at 55°C and 30 s at 72°C and a final elongation for 10 min at 72°C. Relative association of chromatin-bound protein was calculated using the formula $2^{-(\Delta Ct)}$*100, where ΔCt is $Ct_{(output)} - Ct_{(IgG\ \ control)}$, output is the DNA immunoprecipitated with TF-specific antibodies and IgG control is the DNA from immunoprecipitations using nonspecific control antibody. The primer sequences are provided in Supplementary Table S1.

Discretization of array expression values and constraint-based model

Metabolic changes resulting from human SGBS preadipocyte cell differentiation were qualitatively predicted from gene expression data using an implementation of the constraint-based method by Shlomi *et al.* (6). Constraint-based modeling is a widely used mathematical approach for the description and analysis of metabolic networks. It relies on the stoichiometric structure and does not require detailed kinetic parameters. By assuming steady state for the intracellular metabolites, the respective dynamic balance equations can be simplified to easy to handle linear equations. Besides the law of mass conservation, other constraints might be included, e.g. enzyme capacities, irreversibility information or measured uptake and secretion rates, as well as optimality considerations, to further constrain the possible solution space, i.e. the possible flux values, which can be realized within the given network structure. Recent efforts, like the aforementioned applied method of Shlomi *et al.*, focus on the generation of context-specific and thus more predictive models via the additional integration of omics data. A consistent version of the generic human metabolic model Recon1 (2) served thereby as modeling platform on which own microarray data were overlaid as soft-constraints based on gene-protein-reaction associations to allow the prediction of network activity distributions. *LPIN1* was missing and owing to its central role in adipocytes, was added to the model and assigned to the triacylglycerol pathway. Continuous log2-normalized expression values were first discretized into three categories: lowly expressed (−1), moderately expressed (0) and highly expressed (1) genes, based on mean expression ± 0.5*standard deviation cutoffs. These values were mapped to the reactions contained in Recon1 and used as input for the metabolic reaction activity prediction.

Heptamer enrichment analysis and miRNA target identification

To identify heptamer motifs whose frequency is significantly different in the 3′-untranslated regions (3′-UTRs) of downregulated transcripts, relative to their frequency in the entire set of 3′-UTRs, we considered all RefSeq transcripts for which a corresponding probe set was significantly downregulated ($P < 0.01$, log2-fold change < -0.3) 24 h after miRNA mimic transfection. The 3′-UTR sequences of the RefSeq transcripts were downloaded from the UCSC genome browser (14.05.2012) and the human miRNA sequences were obtained from miRBase release 18 (http://www.mirbase.org/). The enrichment analysis was performed using a Bayesian model originally introduced for comparing miRNA frequencies between samples (28,29) that have been successfully applied to determine motif enrichments of small RNAs (30,31).

The transcripts for the list of putative target transcripts of miR-27a, miR-29a and miR-222 were selected based on two criteria: (i) at least one probe set corresponding to the transcript was significantly downregulated on the respective miRNA transfection and, (ii) the transcripts 3′-UTR contained at least one hit for any of the possible heptameric reverse complements for the corresponding seed sequences of miR-27a (CUGUGAA or ACUGUGA), miR-29a (GG UGCUA, AUGGUGC or UGGUGCU) or miR-222 (AU GUAGC), respectively.

ChIP-seq analysis

The HUVEC H3K4me3 peak data available from ENCODE (wgEncodeUwHistoneHuvecH3k4me3StdPkRep1) was overlapped with transcription start site (TSS) coordinates from Refseq to limit the analysis to active genes in HUVECs. Gene to disease associations were obtained from DisGeNET (32). A list of endothelial-relevant disease-associated genes was compiled by combining genes associated with CVDs, vascular diseases, coronary artery diseases, cerebrovascular disorders, peripheral arterial occlusive disease and pulmonary arterial hypertension. ENCODE data from untreated HUVEC cells was retrieved as peak coordinate files (UTA cMYC, SYDH GATA2, SYDH MAX, SYDH cJUN and SYDH cFOS). Other public data were obtained from the SRA database as .sra files (SRR576805 ETS1 from VEGFA stimulation, SRR351351 MEF2C from statin stimulation, SRR390745 p65 from TNF stimulation, SRX096362 FLI1 representing an endothelial cell developmental TF, SRR518265 HIF1A from hypoxia and PPARG samples SRX032890 and SRX019521, each with their respective control samples) that were converted to fastq files using sratools v.2.1.7. (data from own experiments were already in fastq format). Raw reads were first quality controlled using the FASTQ software v.0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). A deviation from the expected GC-content was observed in the input sample of SGBS cells and this sample was replaced in the downstream analysis by a new input obtained from similarly differentiated cells. All reads that were detected as read artifacts or that had low-quality base pair calling were removed and read stacks collapsed using the FASTX software v.0.0.13 (minimum quality score of phred 10 across the read length was required) (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The reads that passed the quality control steps were aligned to the hg19 human genome using the Bowtie (33) software v0.1.25 (one mismatch allowed, maximum three locations in the genome from which the highest quality match was reported). The mapping capability with these settings was tested by aligning all 36-mers of the hg19 fasta genome available via UCSC and was determined to be 0.88 and used in the subsequent peak-calling step.

SGBS histone data were analyzed using the EpiChIP software v.0.9.7 (34), where the H3K4me3 signal was quantified from −750 to +1250 region centered at Refseq TSS coordinates. This region was detected to have the highest signal by window analysis. TF peak detection from SGBS was performed using the Quest software (35) v.2.4. run in the advanced mode with default settings applied except for the mappable genome fraction (set to 0.88) and enrichment (ChIP enrichment set to 15 and ChIP to background enrichment to 2.5). Fastq files and signal tracks from SGBS cells can be accessed via NCBI GEO (GSE41578). The final peak lists were filtered to remove peaks with log10Qvalue <3. We chose to apply two cutoffs to detect both low-occupancy (enrichment >15) and high-occupancy (enrichment >30) binding sites. In the text, the complete list of low- and high-occupancy genomic regions (Supplementary Table S2) has been analyzed unless otherwise specified. The public HUVEC ChIP-seq data were processed with default settings (enrichment 30), which corresponds to settings used in their respective publications (information was available for three of five studies via GEO). To assess what biological pathways could be most affected by the given TF, a genomic region enrichment test was performed using the GREAT software (binomial P-value, false discovery rate (FDR) 1%) (36). The same software was used to obtain the peak to gene association files for analyzing TF convergence on shared targets. The complete gene association and enrichment term lists for the SGBS TFs can be found in Supplementary Table S3

Gene metanodes and IDARE web portal

Gene Metanodes (Figure 1) showing gene-related data were generated with Matlab®. Recon1 metabolic gene Entrez IDs were used for data mappings. Based on homogeneous (HUVEC TFs) or heterogeneous (SGBS) data we customized the metanode visual appearance. The open sourced IDARE web portal and its HUVEC and SGBS instances are built using HTML5 standards and javascript libraries jQuery, highchart.js, bootstrap and cytoscapeweb. The release contains a configuration-based python workflow responsible for building graph objects from Recon1 SIF and XGMML pathways into javascript object notation files (full description on the User Guide provided in the Supplementary Material). In addition, gene metanode image files, annotations from hg19 and time course expression data are integrated along with reaction and metabolite relationships. The images and graph object files are then deployed to the appropriate directories according to the

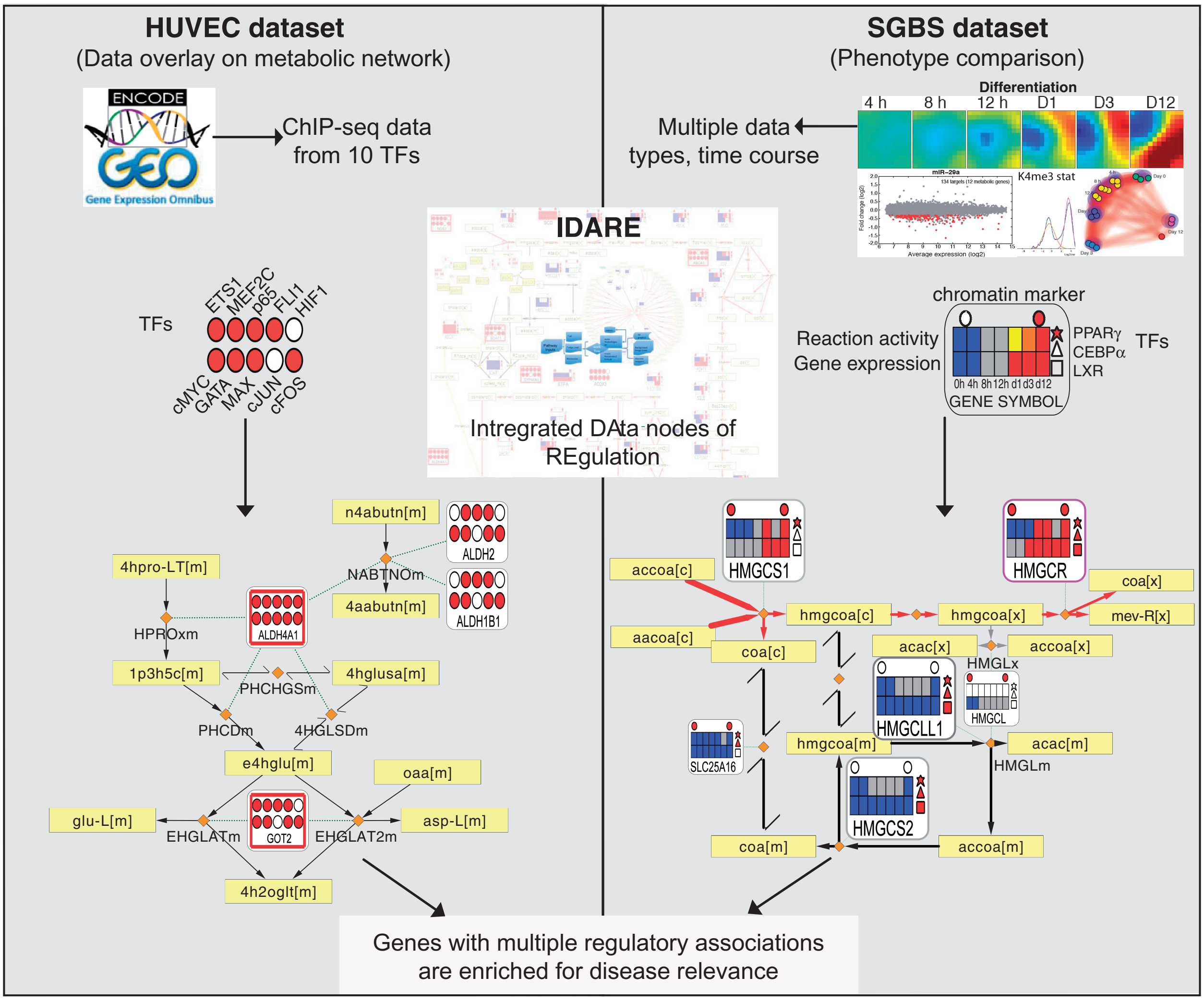


Figure 1. Conceptual overview of the analysis performed that links the regulatory and metabolic networks for exposing disease-relevant genes. A summary of the data sets and data integration approach is shown. By integrating regulatory network components with metabolic pathways, gene metanodes shown from the IDARE webportal provide a simple and intuitive means to analyze relationships between the metabolic and regulatory networks. HUVEC data set: Overlaying TF-binding data indicated in filled color circles with the metabolic network reaction backbone can be used to examine the co-occurrence of transcriptional regulators that is informative of likely disease association. Yellow rectangular nodes represent metabolites and small orange diamonds represent reactions in the metabolic network. SGBS data set: Often the biological question involves comparison of phenotypic states. Constraint-based modeling methods, established in context of metabolic reconstructions, can be used as exemplified with data from adipocyte differentiation. Those pathways that are predicted to shift to an active state in adipocytes compared with preadipocytes are indicated by the red edge color, yellow and black color correspond to active or inactive reactions in both states. The heatmaps show for each differentiation time point the gene expression (below) and predicted reaction activity (above) where blue indicates low/absent gene expression or inactive reaction, gray, moderate expression or undetermined reaction and red stands for highly expressed gene and active reaction, respectively. The regulators can be added as different shapes. Here, circles above represent the presence (red) or absence (white) of the H3K4me3 marker for active transcription and the three different polygons to the right represent PPARγ (star), CEBPα (triangle) and LXR (square) peak associations (red, present; white, not present).

instance workflow configurations. This architecture is extendable and allows for easy inclusion of other data sets like HUVEC as well as custom pathways. IDARE instances can be deployed locally or on web and cloud servers.

RESULTS

Integrating gene regulation data with a human metabolic reconstruction

A metabolic network typically consists of metabolite and reaction nodes. To make such models actionable in context of disease pathways or drug target identification, it becomes useful to integrate the regulation of genes that catalyze the reactions within the model. We included versatile nodes (referred to as gene metanodes) associated to reactions, which can represent any gene-centric data collected from different experiments (Figure 1). TF binding data indicated in filled circles in the metanode can be overlaid with the metabolic network that includes rectangular nodes representing metabolites and diamonds representing reactions. Exemplified using HUVEC data, the visualization can be used to examine the co-occurrence of transcriptional regulators. The display options are flexible including presenting the regulators as different shapes, as shown in SGBS. Moreover, the edges forming the network backbone can be colored to represent reaction activity obtained using constraint-based modeling methods. The metanode heatmaps show for each differentiation time point the gene expression (below) and predicted reaction activity (above). By integrating regulatory network components with metabolic pathways, gene metanodes shown from the IDARE web portal (see Supplementary Data—IDARE User Guide for further details) provide a simple and intuitive means to analyze relationships between the metabolic and regulatory networks.

To illustrate the concept and to test whether such data could be informative to highlight key parts of the large metabolic reconstruction networks, we collected ChIP-seq data from HUVECs that represent the most studied primary cell type among ENCODE data sets. Using the H3K4me3 chromatin mark to focus on active loci, we associated Recon1 genes to TF peaks from 10 studies, 5 collected from NCBI GEO database and 5 available from ENCODE (see 'Materials and Methods' section for details). Disease information was collected from the DisGeNET database (32), focusing on endothelial-relevant disease. Table 1 shows those genes that are associated with eight or more TFs while the complete TF and disease association result is shown in Supplementary Table S4. Interestingly, expressed genes associated with between seven and nine TFs are >2-fold enriched in vascular disease-relevant genes, a result that points to the prominent link between high TF-mediated gene expression regulation and disease. The hypergeometric *P*-values for the different number of TF associations to Recon1 genes are shown in Figure 2A, comparing vascular disease-relevant associations to all disease associations (the observed increasing trend in significance and enrichment apply also when analysis is not restricted to metabolic genes, data not shown).

The gene with most disease associations overall, the *endothelial nitric oxide synthase*, *NOS3* (also known as *ENOS*), is visualized in Figure 2B. The release of NO is a key paracrine signal in the vascular system that is essential for the regulation of blood flow and pressure (37). The two other nitric oxide synthases (*NOS1* and *NOS2*) can catalyze the same reaction to convert L-arginine to NO. Interestingly, each of these gene metanodes show multiple TF associations (Figure 2B). The respective genomic region around *NOS3* TSS with TF signal from the 10 ChIP-seq experiments is shown in Figure 2C. The other multi-TF–associated nodes in the proline-arginine metabolic pathway shown in Supplementary Figure S1 include *ALDH4A1* that participates in multiple amino acid pathways and is known to cause the autosomal recessive disorder known as type II hyperprolinemia (38) and *MTAP* from the coronary artery disease genome-wide association study (GWAS) reported locus on chromosome 9 (39). Encouraged by these findings, we next evaluated whether the same principle generalizes in human adipocytes that represent a key cell type in obesity and metabolic disease. However, a more limited set of available genome-wide regulator profiles motivated the collection of experimental data and in parallel using mathematical predictions based on the metabolic reconstruction to expose relevant pathways as described in more detail below.

Predicted metabolic activity changes during adipocyte differentiation

To outline plausible metabolic activity changes during adipocyte differentiation using the SGBS cell line, which represents an established human adipocyte cellular model isolated from a SGBS patient (21), we leveraged the constraint-based modeling approach (see 'Materials and Methods' for details) to predict the dynamic activity changes of metabolic reactions (6). Based on a time-course measurement of the transcriptome of differentiating SGBS preadipocytes a dynamic shift is evident, in particular, in the expression levels of metabolic genes among which 18%, 2-fold more when compared with other gene categories with similar numbers of genes or even with all detected genes (Supplementary Figures S2 and S3) (25,26), are differentially regulated. An overall trend of increasing levels from day 1 onward results in 219 upregulated metabolic genes by day 12 of differentiation, compared with 98 downregulated genes (Supplementary Table S5).

As gene expression levels alone are insufficient to describe the metabolic adaptation that occurs during terminal differentiation, we used them as soft-constraints to predict reaction activity for Recon1 (2,6). The predicted pathway activity changes are shown in Figure 3 and the complete prediction results for all seven differentiation time points and the 2469 consistent reactions contained in Recon1 are provided in Supplementary Table S6 ('Materials and Methods' section). Consistent with the mRNA level changes, a much higher number of reactions were predicted active in adipocytes than in preadipocytes (556 compared with 290, respectively) with 259 reactions predicted to become active during differentiation. Five pathways with highest predicted activation between preadipocytes and adipocytes were cholesterol metabolism (76% reactions predicted to change), fatty acid activation (64%) and oxidation (93%), triacylglycerol synthesis (60%) and branched chain amino acid (BCAA: valine, leucine, isoleucine) metabolism (50%) (for metabolites and enzymes involved refer to Supplementary Table S7), highly involved in lipid metabolism and metabolic diseases, suggesting the ability of the approach to recapitulate adipocyte characteristics. On a metabolite level, these pathways converge at acetyl-CoA, which can produce intermediates to be converted to fatty acids or to be consumed in the energy-producing mitochondrial oxidation. Pathways excluded from further analysis contained reactions predicted undetermined in one of the two phenotypic states, concretely, heme biosynthesis with 10 reactions, all predicted active in adipocytes, but nine of them undetermined in preadipocytes; heme degradation with only two reactions, both predicted active in adipocytes but undetermined in preadipocytes and the biosynthesis of tyrosine, phenylalanine and tryptophan, which contains only one

Table 1. Endothelial disease relevant genes exposed by association to multiple TFs in HUVEC data

Symbol	Pathway	Entrez GeneID	Number of disease	Number of TFs
NOS3	Arginine and proline metabolism	4846	140	8
PTGS2	Eicosanoid metabolism	5743	97	8
HMOX1	Heme degradation	3162	56	8
ABCA1	Transport, extracellular; transport, golgi apparatus	19	20	9
PTGS1	Eicosanoid metabolism	5742	12	8
LIPG	Triacylglycerol synthesis	9388	9	10
ADA	Nucleotides; purine catabolism	100	9	9
PDE4D	Nucleotides	5144	9	8
PDE4B	Nucleotides	5142	8	9
ABCC4	Transport, extracellular	10 257	7	9
PDE3A	Nucleotides	5139	7	8
PIK3CG	Inositol phosphate metabolism	5294	7	8
SLC12A2	Transport, extracellular	6558	6	9
PAFAH2	Glycerophospholipid metabolism	5051	4	9
GCLM	Glutathione metabolism	2730	3	10
MTHFD1L	Folate metabolism	25 902	3	10
GALNT2	*O*-glycan biosynthesis	2590	3	9
PPAP2B	Triacylglycerol synthesis	8613	1	10
NNMT	NAD metabolism	4837	1	10

The 19 Recon1 metabolic genes annotated with endothelial-relevant disease terms in the DisGeNET database (32) and having an active TSS mark (H3K4me3) with putative binding of at least 8 TFs from ChIP-Seq HUVEC studies (see ‘Materials and Methods’ section for details) are presented.

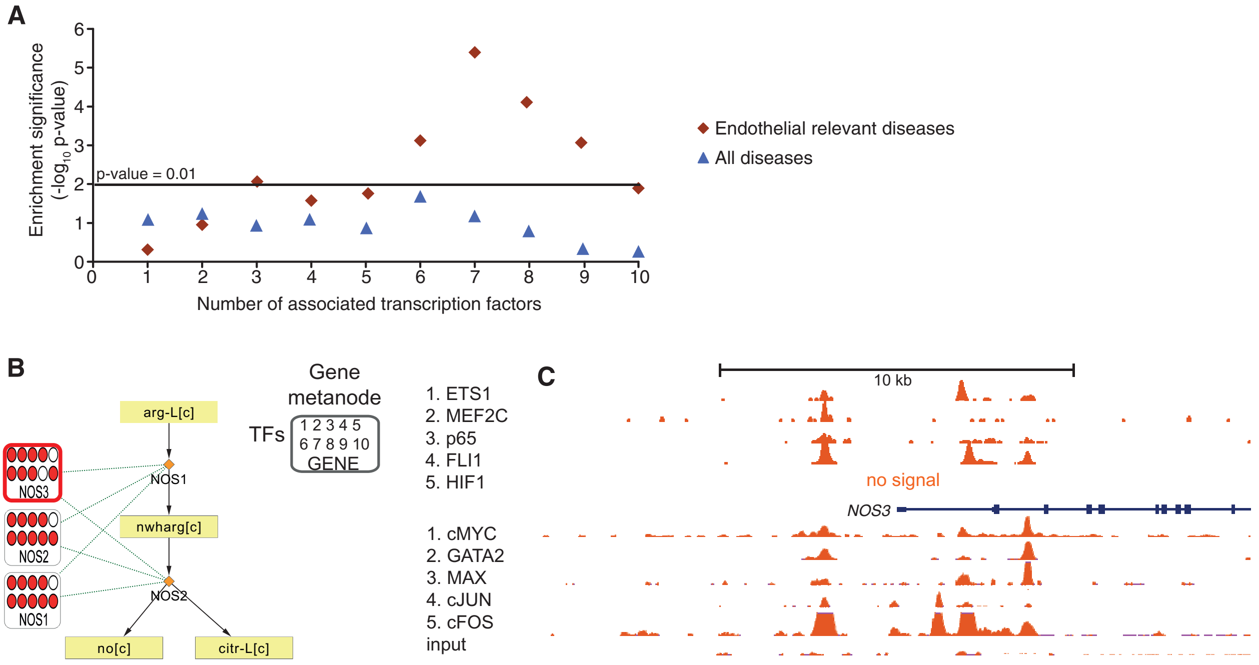


Figure 2. Gene metanodes reveal frequent TF association to disease-relevant genes exemplified by *nitric oxide synthases* in HUVEC data. The result of disease enrichment tests for genes associated with 1–10 TFs are shown (**A**), where the horizontal line corresponds to hypergeometric *P* <0.01. The endothelial relevant diseases (diamonds) is compared with all diseases (triangles). (**B**) The three enzymes encoding nitric oxide synthases (NOS1, NOS2 and NOS3) and the two reactions that convert L-arginine to NO for regulating vascular dilation are shown. Data related to genes are displayed in gene metanodes superimposed on the metabolite-reaction network. Peak associations from 10 ChIP-seq studies in HUVECs (ETS1, MEF2C, p65, FLI1, HIF1α available via NCBI SRA and cMYC, GATA2, MAX, cJUN, cFOS and input available via ENCODE) are displayed in the indicated order where color indicates TF association and the respective TF signal tracks are shown in C. The value range 1–100 is used in the first five tracks and 1–78 in the ENCODE tracks.

reaction (O2 + L-Phenylalanine + Tetrahydrobiopterin -> Tetrahydrobiopterin-4a-carbinolamine + L-Tyrosine), predicted active in adipocytes and inactive in preadipocytes.

More than 300 metabolic reactions are predicted to change from preadipocytes to adipocytes (Supplementary Table S6). In agreement with an increased gene expression for the majority of metabolic genes, most reaction changes

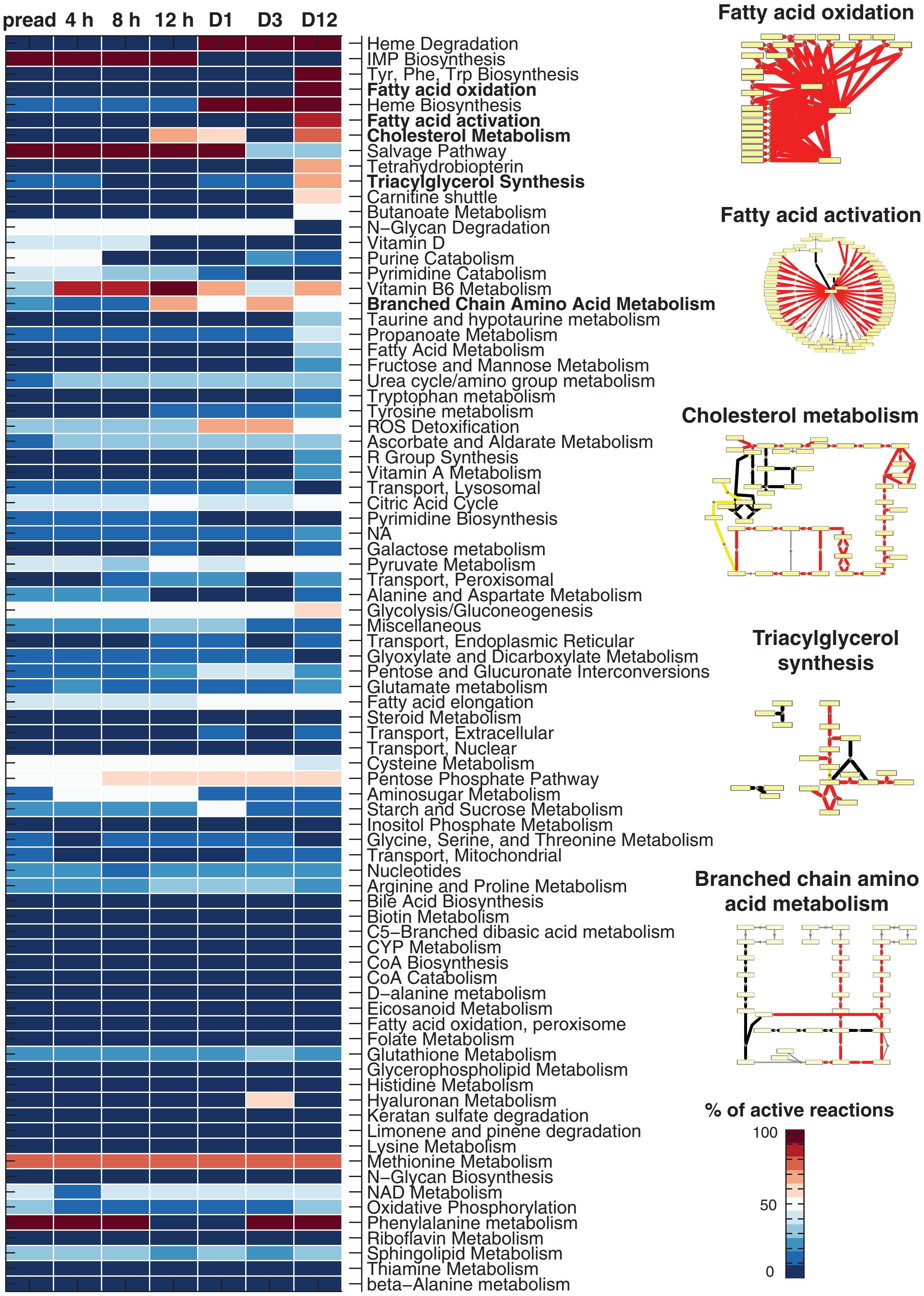


Figure 3. Predicted metabolic activity changes during adipocyte differentiation and reaction backbones of selected pathways. A heatmap representing the predicted percentage of active reactions for each differentiation time point (columns) and metabolic pathway (rows) is shown for the Recon1 pathways (2). Blue color corresponds to no active reactions and red to complete activation. Notice that reactions can change between active and either inactive or undetermined. The pathways are shown in a descending order of activity considering the difference between preadipocytes ('preadip') and adipocytes ('day 12') and secondly alphabetically. Pathways highlighted (cholesterol metabolism, fatty acid activation and oxidation, triacylglycerol synthesis and BCAA metabolism) represent those with highest predicted activation with high prediction confidence score. The metabolic pathway backbones of these pathways are shown based on the stoichiometric information in the generic human metabolic model Recon1 (2). These pathways are predicted to shift to an active state in adipocytes compared with preadipocytes as indicated by the red edge color, yellow and black colors correspond to active or inactive reactions in both states, respectively. Yellow rectangular nodes represent metabolites and small orange diamonds represent reactions.

predicted are activations, especially those in lipid metabolism pathways. The BCAA metabolism pathway appeared as the most affected nonlipid pathway based on the predictions.

Genome-wide measurements of transcriptional, posttranscriptional and chromatin-level regulation in adipocytes

We performed a diverse set of genome-wide measurements from SGBS adipocytes, summarized in Figure 4A. We first focused on identifying the most highly induced TFs and their putative targets. In agreement with metabolic gene expression changes, a highly dynamic TF expression profile is also observed (Figure 4A). Among the highest upregulated TFs (see Supplementary Table S8), *PPARG* and *CEBPA* have been previously associated with a key role in adipocyte differentiation and global regulation of metabolic genes (8,40,41). The most upregulated TF gene was the signal responsive *LXRA* (also known as *NR1H3*), an established regulator of cholesterol reverse transport (42) for which the genome-wide occupancy has not yet been studied in adipocytes. We obtained the genome-wide binding profiles of PPARγ, CEBPα and LXRα to reveal their possible interplay in SGBS adipocytes, by collecting ChIP samples for sequencing (ChIP-seq). We also included published primary adipocyte (day 9) (10) and SGBS PPARγ (day 20) (41) data sets retrieved from the Sequence Read Archive database for comparison. The ChIP-seq reads that satisfied quality criteria were aligned to the hg19 human genome (see 'Materials and Methods' section for details). Statistics about initial and final read numbers are indicated in Supplementary Table S2.

In addition to regulation of the transcriptional output, the measured expression changes of metabolic genes could also be controlled at the posttranscriptional level. In particular, miRNAs have emerged as important regulatory molecules and regulation of a number of miRNAs have been described as important for successful adipogenesis and lipid accumulation (43). We hypothesized that different miRNAs that become down-regulated during adipogenesis might contribute to allow the observed upregulation of metabolic genes to take place, serving as gatekeepers in preadipocytes. To investigate this possibility, we performed microarray profiling to detect differentially regulated miRNAs on days 0, 1, 3 and 12 of differentiating SGBS cells, revealing several candidate miRNAs for further analysis (Supplementary Figure S4). We focused on miRNA clusters with several members repressed already at early stages of differentiation, which lead to the selection of miR-27a that has previously been studied in mouse models, and two miRNAs whose role in adipocytes has not been characterized, miR-29a and miR-222 for further experiments. Their downregulation was validated by RT-qPCR to occur already by day 5 of differentiation (Figure 4B). To identify candidate target genes, we performed a miRNA mimic transfection (corresponding to a specific overexpression of each miRNA) at day 4 of differentiation and analyzed by microarrays the mRNA profiles at 24 h posttransfection and compared these with the cells similarly transfected with a scrambled control siRNA. An analysis for enriched heptamer motifs in 3′-UTRs of the downregulated mRNAs from the microarray analysis at day 5 reveals enrichment of motifs complementary to respective miRNA seed sequences (Figure 4C and D), suggesting that the observed mRNA downregulation could be due to their direct targeting by the miRNA.

Finally, we used the H3K4me3 chromatin modification, indicative of the transcriptional potential of the associated TSS, to evaluate changes in chromatin state of metabolic genes. We collected ChIP samples from undifferentiated and differentiated SGBS cells for sequencing and analyzed the ChIP-seq data obtained and public data from primary preadipocytes and adipocytes (10,41).

Predicted target gene profiles and their overlap

The comparison of all genes (in A) or metabolic genes (in B) associated to each TF is shown in Figure 5. Figure 5C shows the intersection of metabolic genes associated with PPARγ in the three independent data sets (1278). This first genome-wide mapping of LXR binding in human adipocytes revealed 2117 associated putative target genes. For CEBPα, we obtained 6880 putative target genes, while for PPARγ, the 11 078 putative target associations kept reflect peak associations that were found in our data set and observed in at least one of the two public data sets (10,41). From these genes, 1691 were associated with peaks from all three TFs (Figure 5A, Supplementary Table S3), with 147 common metabolic putative target genes.

To evaluate TSS activity changes, we analyzed the H3K4me3 ChIP-seq data using a mixture model method (34) that separates histone marker labeled gene TSS from those lacking the marker (Supplementary Figure S5) based on their read count distribution estimates. According to this analysis, most genes did not completely switch their TSS activation state during differentiation in SGBS cells (15 263 active; 19 311 inactive; 470 unclassified), while among the transcripts with altered TSS activity, the H3K4me3 marker mostly decreased: 1223 metabolic gene TSS are labeled with H3K4me3 in both preadipocytes and adipocytes, 1125 are inactive, TSS activity decreased for 37 transcripts (corresponding to 25 gene loci) while only 2 gained the activity marker (Figure 5D and Supplementary Table S9). As indicated, there was considerable agreement between primary adipocyte (10) and SGBS data (Supplementary Table S9).

The large intersection of TF-associated genes with the metabolic genes from Recon1 (1069 out of 1496 genes, Figure 5B) suggests a high contribution of these 3 TFs to the metabolic changes observed on terminal adipocyte differentiation. Interestingly, the metabolic pathways with most changes (those highlighted in Figure 3) show extensive TF binding. The *acyl-CoA synthetase long-chain family member 1* (*ACSL1*) gene of the fatty acid activation pathway was among the top genes associated with high-occupancy binding sites for all three TFs (Figure 5E). The co-localized binding seen in Figure 5E was rather exceptional; genome-wide overlap in peak regions by all three TFs occurred only at 223 peak locations. At gene ontology

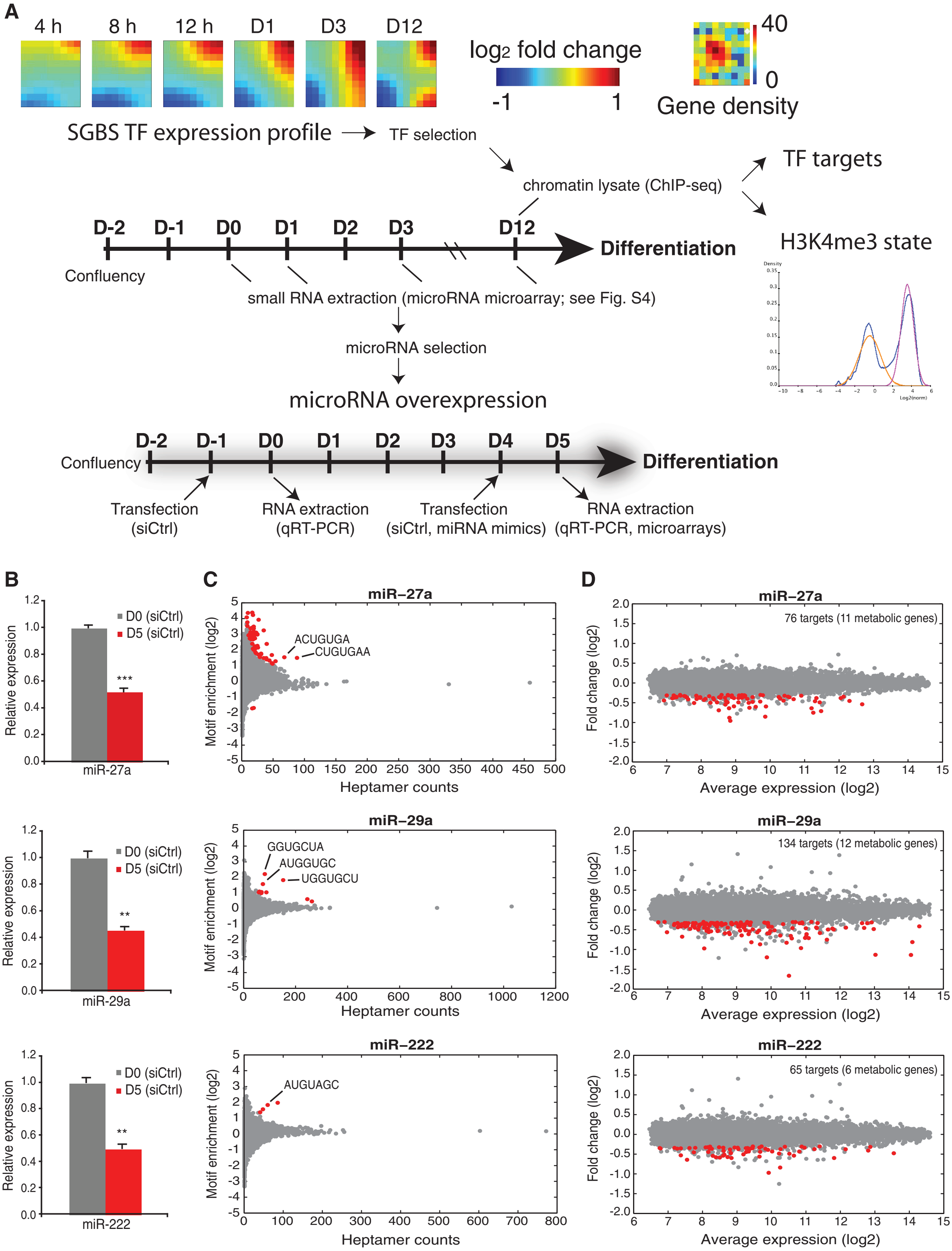


Figure 4. Selection of regulatory molecules for further study. (**A**) The TF expression profile used to select the highest upregulated TFs during adipogenesis is visualized using GEDI maps (25) that cluster TF genes with similar expression profiles. The number of genes in each cluster is displayed in the Gene density panel beside. For the identification of their genome-wide targets, ChIP-seq samples were collected as indicated on day 12, including lysates used to assess the H3K4me3 chromatin marker status using TSS activity status mixture modeling. The schematic representation of the experimental procedure indicates the sample collection intervals for microarrays, miRNA microarrays and ChIP-seq. For the identification of primary target mRNAs of the selected miRNAs, a similar differentiation procedure was used, accompanied by transfections with miRNA mimics for miR-27a, miR-29a or miR-222 or scrambled siRNA as a control at day 4. Samples were collected 24 h after transfection for microarray and

(continued)

term level (Supplementary Table S3), the TFs overlap in regulation of cholesterol efflux. Several genomic regions each bound by one or several of the TFs studied concentrated in the vicinity of *ATP-binding cassette, subfamily A member 1 (ABCA1)* a key gene in cholesterol reverse transport (Figure 5G). Interestingly, *ACSL1* is required for oleate- and linoleate-mediated inhibition of cholesterol efflux through *ABCA1* in macrophages indicating cross-talk between the pathways (45). ChIP-qPCR validation from independent immunoprecipitations is shown for the two regions indicated (Figure 5F and H), and the *CDH1* and *FABP4* loci that served as a negative and positive control region, respectively, are shown in Supplementary Figure S6.

Next we identified target genes for each miRNA based on repression in the array experiment and a seed match analysis (putative targets are indicated in red in Figure 4 and listed in Supplementary Table S10), ranging from 65 (miR-222) to 134 (miR-29a) significant target calls per miRNA. Figure 5I shows the metabolic target genes of each miRNA overlapped with TF target association lists (those genes that are significantly regulated during differentiation are indicated with a star).

Interestingly, those putative miR-27a and miR-222 targets that are overlapping with more than one TF are in fact mainly associated with high-occupancy binding, namely *PISD*, *CYP1B1*, the *mitochondrial glycerol-3-phosphate acyltransferase* (*GPAM*), *hexokinase 2* (*HK2*), *LPL*, *RPIA*, *the C-4 to C-12 straight chain acyl-CoA dehydrogenase* (*ACADM*) and *stearoyl-CoA delta-9-desaturase* (*SCD*), suggesting that key metabolic genes are under tight combinatorial transcriptional and posttranscriptional regulation. Interestingly, the DisGeNET resource reports a disease association for each of these genes, only *HK2* is not supported by this data source but in light of recent literature is implicated in cancer (46,47). All but two putative miRNA targets were also associated with TF-mediated regulation. Moreover, as previously reported (17,16), we could confirm *PPARG* among the miR-27a targets in adipocytes. There is also overlap between the miRNAs: according to our analysis miR-27a and miR-222 both target *CYP1B1*, miR-27a and miR-29a target the amino acid transporter *solute carrier family 7 member 5 (SLC7A5)* while miR-29a and miR-222 target *dihydrolipoamide branched chain transacylase E2* (*DBT*).

In conclusion, the putative shared TF and miRNA target genes that our data integration revealed were *ACADM*, *DBT*, *GPAM*, *HK2*, *LPL* and *SLC7A5*. Taken together, by collecting a diverse and comprehensive set of regulatory data on adipocyte differentiation we were able to show that enzymes crucial or rate limiting for lipid metabolism were often associated with miRNAs and high occupancy binding by multiple TFs, suggesting tight combinatorial effects of TF upregulation and miRNA downregulation driving metabolic changes in adipogenesis.

Cell fate determining TFs engage signal-dependent TFs in a feed-forward motif

Defects in adipocyte differentiation represent an important early event in obesity and related metabolic dysfunction. To address the role of cell fate master regulators interfacing the metabolic and regulatory networks, heatmaps showing the top-ranked upregulated TFs according to differential expression between day 12 and day 0 are shown in Figure 6A. Focusing on target gene associations to high occupancy binding by the three TFs (lines connecting the studied TFs to upregulated genes), the prominent role for PPARγ in regulating other TFs in adipogenesis becomes apparent. Notably, each TF studied here binds its own regulatory region, while CEBPα and LXR show few high-occupancy interactions to the other upregulated TFs, in contrast to PPARγ. In fact, the only other TF association is the LXR binding to *SREBF1*.

We confirmed the binding of LXR to the prominent LXR peaks in the *SREBF1* locus by ChIP-qPCR in adipocytes (enrichment was also observed for CEBPα), while the prominent PPARγ peak further upstream is supported by all three ChIP-seq studies (Figure 6B). According to our data, the cell fate regulating TFs form two closely connected feed-forward motifs to these two signal responsive TFs, both known to play key roles in cholesterol metabolism, with PPARγ associated to *SREBF1* through both LXR and CEBPα (Figure 6A and Supplementary Figure S7).

As a representative of a ligand-responsive candidate drug target TF, we examined the high-occupancy LXR binding sites and confirmed binding to 11 previously reported LXR targets (Supplementary Table S3 in bold). Notably, all these genes were also upregulated during differentiation. To test the ligand-responsiveness of genes in the loci occupying the LXR binding sites (<500 kb from the TSS), we performed a microarray with the LXR agonist T0901317 (Supplementary Table S11) from a 4-h ligand stimulation of differentiated SGBS cells and could

Figure 4. Continued
RT-qPCR analysis. (**B**) RT-qPCR analysis of the relative expression values of the endogenous miR-27a, miR-29a and miR-222 from undifferentiated and 5-day differentiated SGBS cells. The measured expression values were normalized to U47 snRNA and are shown relative to undifferentiated cells, value of which was set to 1 (gray bars). Data points indicate the mean expression values of triplicate experiments and the error bars represent SD. Student's *t*-test was performed to determine the significance of downregulation on differentiation (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). (**C**) Enrichment analysis of heptamer motifs in the 3′-UTR sequences of significantly downregulated transcripts. The count of all possible heptamer motifs (each represented by a circle) and their log2-enrichment within the 3′-UTRs of downregulated transcripts are depicted on the x-axis and y-axis, respectively. The significantly enriched heptamers are marked in red. The most enriched abundant heptamers are corresponding to the reverse complement sequences of the overexpressed miRNA seeds as indicated (see 'Materials and Methods' section for details). (**D**) MA-plot depicting the log2-expression levels (x-axis) of all transcripts in cells transfected with indicated miRNA mimics and the log2-fold change relative to cells transfected with siCtrl. The significantly downregulated transcripts (unadjusted $P < 0.01$, log2-fold change < -0.3) containing at least one putative binding site for the respective miRNA are marked in red. The total number of putative miRNA targets is indicated (with metabolic target genes in brackets).

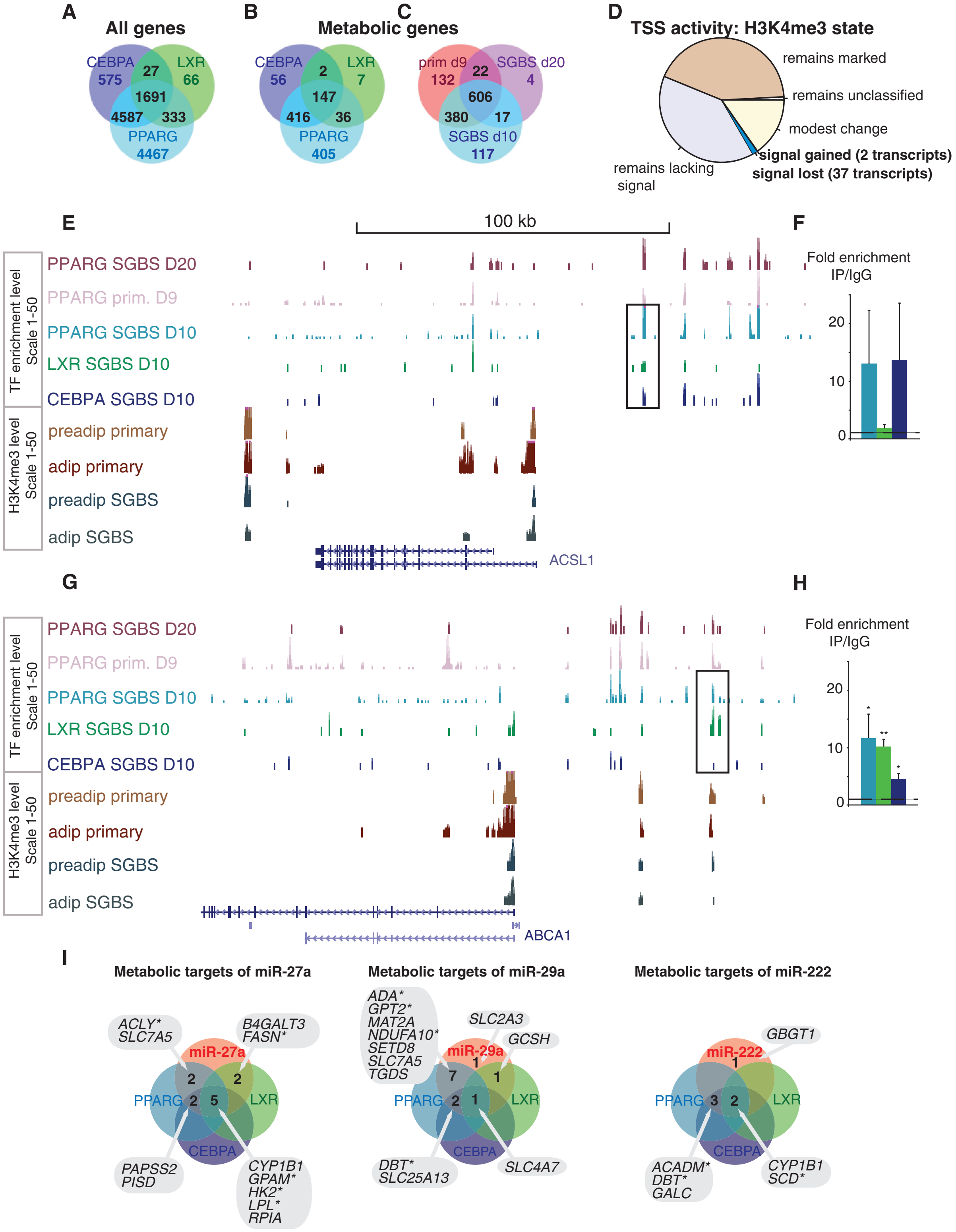


Figure 5. Overlaps in the genome-wide target profiles of PPARγ, CEBPα, LXR and the miRNAs −27a, −29a and −222 from SGBS adipocytes. Venn diagrams comparing putative target genes for PPARγ, CEBPα and LXR (in **A** and **B**) or between different studies that profile PPARγ binding (B) as obtained using the GREAT tool (36) are shown. Among all genes (A), 1691 genes are associated with all TFs, representing a large fraction from each individual TF peak-gene association list. Metabolic genes (**C**) show a similar highly overlapping target association profile. (C) In total, 606 PPARγ target associations to metabolic genes are supported by ChIP-seq data from day 10 and day 20 differentiated SGBS cells (41) and day 9 differentiated primary adipocytes (44), with an additional 419 metabolic genes supported by at least two data sets. The pie chart (**D**) illustrates how many genes gain or lose the H3K4me3 signal, showing that most genes retain their activity marker status. (**E**) The ChIP-seq signal tracks from

(continued)

validate 39 additionally ligand-responsive genes at this early time point (Supplementary Table S12). Among known target genes, *MYLIP1*, *SREBF1*, *ABCA1* and *ABCG1* were significantly ligand responsive at this time point, while the upregulation of *stearoyl-CoA delta-9-desaturase* (*SCD*) was modest and did not pass the multiple testing correction (unadjusted $P < 0.05$). New putative target associations to *ADH1B*, *TMEM17* and *WSB1* were supported by the array data and high-occupancy binding sites. However, the majority of responsive genes were associated with less-prominent LXR binding under the unstimulated condition represented by the ChIP-seq experiment, including *ELOVL6*, *LIPA*, *SCARB1*, *HSD17B7* and *LPIN1* that function in lipid metabolism. Therefore, ligand stimulation greatly impacts LXR-mediated regulation of the metabolic network, in agreement with observations from mouse liver (49).

The interaction of PPARγ, LXR and SREBF1 has been studied before on selected target genes (50), and therefore our focus here was how they interact on pathway level. First, to confirm that SREBF1-mediated regulation converges with the pathways observed to change most during adipogenesis (Figure 3 and Supplementary Figure S2), we obtained a list of SREBF1-regulated genes observed in human myocytes (48). Cholesterol metabolism was the most significant pathway (hypergeometric P = 3.15e-06, Supplementary Table S13). The TF feed-forward circuits additionally engage the acyl-CoA synthetase enzyme genes, namely *ACSL3*, *−4*, *−5* (Supplementary Figure S8) and the previously highlighted *ACSL1* (Figure 5), in fatty acid activation that fuels triglyceride synthesis.

The integrative pathway map of cholesterol metabolism is shown in Figure 6C. Red edges represent predicted reaction activation from the preadipocyte state and black edges represent reactions predicted inactive. A detailed description of all the components can be found in the figure legend. In short, filled shapes indicate the putative binding of a TF (polygons on the right) or the presence of the active TSS marker (circles above) from ChIP-seq data. The central heatmaps display the discretized gene expression (below) and the reaction activity during differentiation as predicted by the mathematical model (above).

Based on the microarray data, all genes participating in the active branch of the cholesterol synthesis pathway, except *CYP51A1*, show an increased expression on differentiation. RT-qPCR validation for eight induced genes is displayed in Figure 6D. Genes involved in the inactive branch, namely *3-hydroxy-3-methylglutaryl-CoA* (*HMGC*)*-lyase-like-1*, *HMGC-synthase-2*, *SLC25A16* and *SOAT1*, show a constant low level of expression. To place the candidate regulatory motifs including SREBF1 (Supplementary Figure S7) in context of the ChIP-seq data, we used these integrative pathway maps to check which expression profiles potentially reflect more complex dynamics that can be achieved through feed-forward motifs by identifying those that could not be explained by simple direct binding by the three most up-regulated TFs selected for ChIP-seq.

Binding sites for at least one of the three TFs were associated to most genes in the pathway (81%). PPARγ has high-occupancy binding sites in the vicinity of *HMGC-synthase 1* and *mevalonate kinase (MVK)* indicating that it may have a predominant role in regulating these enzymes at key upstream reactions of the synthesis pathway (Figure 6C, upper left corner). All TFs bound nearby the *HMGC-reductase* (*HMGCR*) gene locus encoding the rate limiting enzyme and gene loci encoding enzymes of the initial steps of the alternative ketogenesis pathway (mitochondrial) that start by producing HMGC from acetyl-CoA and acetoacetyl-CoA, namely *HMGC-synthase 2* and *HMGC-lyase-like-1*. Concordant regulation by SREBF1 observed in myocytes concentrates on *HMGCR*, and, in particular, on the central and terminal parts of the pathway, including all genes starting from *MVD*. Among these, five were not bound by the other TFs in the ChIP-seq data, suggesting that their upregulation occurs indirectly via the regulation of *SREBF1*. These include *farnesyl diphosphate synthase (FDPS)* that has been reported to synthesize isoprenoid natural ligands for PPARγ (51), constituting a putative metabolite positive feedback loop. Interestingly, nodes associated to multiple TFs reappear at the end of the synthesis pathway (lower left corner) at the *dehydrocholesterol reductases* (*DHCR*)*-7* and *−24* loci, the latter being a known LXR target gene (52) in addition to *ABCA1*. However, only *ABCA1* is significantly up-regulated in the microarray after 4 h of ligand stimulation (Supplementary Table S11 and S12), which suggests that more complex dynamics may be used to control the terminal step of the pathway at the *DHCR24* locus.

In summary, a tight regulatory circuit between TFs necessary for adipocyte differentiation and those implicated in proper cholesterol homeostasis was observed and associated with the major lipid synthesis pathways. More generally, analyzing TF binding in context of the metabolic network allows formulating testable hypothesis about the regulatory mechanism.

Figure 5. Continued
PPARγ studies in SGBS cells (41) and primary adipocytes (44), CEBPα and LXR from SGBS adipocytes and H3K4me3 from primary and SGBS cells are displayed at the *ACSL1* locus that shows high-occupancy binding of PPARγ, CEBPα and LXR. The ChIP-qPCR validation comparing enrichment with specific antibody to IgG unspecific control for the PPARγ and CEBPα occupied region indicated is shown in (**F**). (**G**) Similarly as above, the TF signal tracks show multiple peaks at the *ABCA1* locus including the LXR response elements that show significant enrichment also with PPARγ and CEBPα antibodies as validated using ChIP-qPCR in (**H**). The enrichment values are shown relative to the enrichment of IgG and indicate the mean enrichment values of triplicate experiments and the error bars represent SEM. One sample *t*-test was performed to determine the significance of TF enrichment compared with IgG (*$P < 0.05$; **$P < 0.01$). (**I**) Venn analysis of metabolic target genes of the tested miRNAs and their targeting by TFs. The lists of metabolic genes targeted by the individual miRNAs and by the TFs PPARγ, CEBPα or LXR are overlapped to identify the metabolic genes under combinatorial multilevel regulation. The genes significantly changing during SGBS differentiation are indicated with a star.

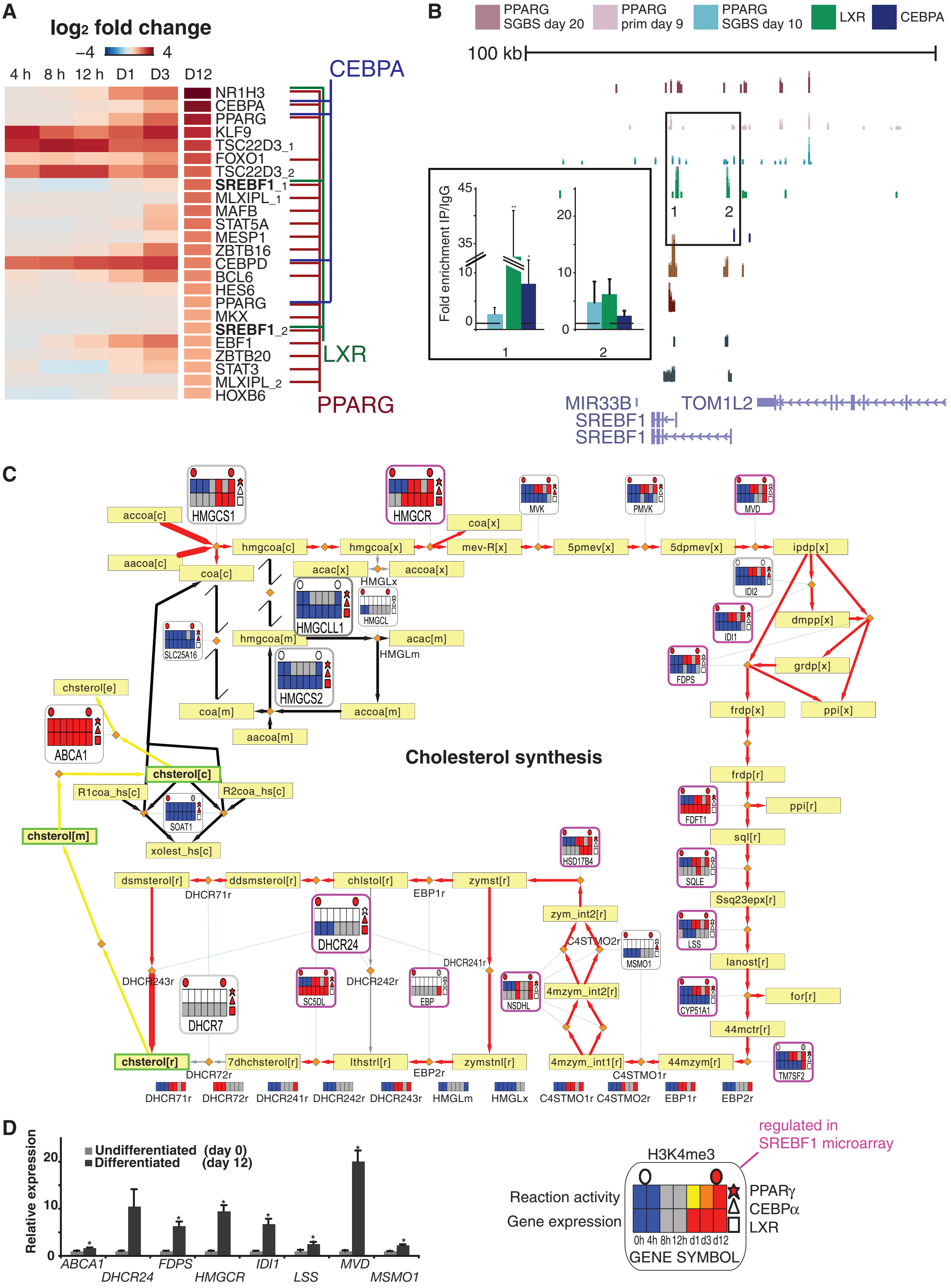


Figure 6. Integrated analysis of the regulation of the cholesterol synthesis pathway. (**A**) The average logarithmic fold change values from 4, 8 and 12 h and days 1, 3 and 12 of differentiation displayed as a heatmap and sorted based on day 12 values identify *LXRA* (NR1H3), *CEBPA* and *PPARG* as the most upregulated TF genes. Association to high-occupancy ChIP-seq peaks of PPARγ, CEBPα and LXR in SGBS cells are indicated with colored lines identifying autoregulation of each TF, regulation of *SREBF1* by LXR and PPARγ, of *CEBPD* and *PPARG* by CEBPα and regulation of majority of TF genes shown by PPARγ. (**B**) The ChIP-seq signal tracks as in Figure 4 are shown at the *SREBF1* locus. Regions with

(continued)

Convergence of miRNAs and TFs exposes further disease-relevant nodes

Our data also revealed several convergence points of miRNAs and TFs on metabolic genes (Supplementary Figure 5I). Figure 7 shows the BCAA catabolism pathway that represents the main nonlipid pathway highlighted here. The early steps include several genes that are upregulated during differentiation, two of which are potentially targeted by all three TFs studied, the *branched chain keto acid dehydrogenase E1 beta* (*BCKDHB*) and the *dihydrolipoamide dehydrogenase* (*DLD*), while *DBT* is a putative target of PPARγ, CEBPα, miR-222 and miR-29a. Together with *BCKDHA*, these genes in fact encode for the large protein complex called branched-chain alpha-keto acid dehydrogenase. According to our results, this enzyme complex is the key regulatory point under the control of multiple regulators and an interesting target for further analysis. Moreover, the *cytosolic branched chain aminotransferase 1* (*BCAT1*) that catalyzes the first steps of the BCAA turnover in cytosol is associated with all 3 TFs. The upregulated genes downstream include *propionyl CoA carboxylase beta (PCCB)* and *methylmalonyl CoA epimerase (MCEE)* whose genomic loci are occupied by both PPARγ and CEBPα and *PCCA* that is associated to all three TFs. Thus, similarly to cholesterol synthesis pathway, the TFs studied converge especially at the initial and terminal steps of the pathway, with the added complexity of posttranscriptional regulators at the large main enzyme complex. As described earlier, the *SLC7A5* gene that functions in BCAA transport is targeted by two miRNAs (miR-27a and miR-29a). The BCAA pathway from HUVECs is shown in Supplementary Figure S9. Among the miRNA target profiles available from HUVECs, a study of miR-663 targets reported regulation of *SLC7A5* (54), and moreover, 9 out of 10 TFs in HUVECs were associated with this gene. As in adipocytes, the *BCAT1* gene is associated with multiple TFs in HUVECs revealing key similarities between the regulator profiles and multi-regulator nodes of these cell types.

To visually assess the convergence of the studied TFs and miRNAs on the other highly activated metabolic pathways, we extended gene metanode metabolic maps to two additional pathways (fatty acid oxidation in Supplementary Figure S8 and triglyceride synthesis in Figure 8). All five maps, as well as the remaining 94 pathways in Recon1, and associated data can be interactively explored in our IDARE web portal (http://systemsbiology.uni.lu/idare.html).

MiR-27a is known to engage in the main TF circuitry through the inhibition of PPARγ (15–17). The triglyceride pathway was identified to contain multiple shared target associations, as shown in Figure 8. The regulatory associations from TFs and miRNAs converge along the pathway at three key enzymes: (i) GPAM that catalyzes the initial and committing step in glycerolipid biosynthesis, playing a pivotal role in the regulation of cellular triacylglycerol and phospholipid levels (57), (ii) LPIN1 that catalyzes the penultimate step in triglyceride synthesis including the dephosphorylation of phosphatidic acid to yield diacylglycerol (23) and (iii) LPL that catalyzes the release of fatty acids from triglycerides (24) [extracellular LPL facilitates fatty acid import, whereas also intracellular activity has been observed that could serve in fatty acid export (58)]. The enzymes from reactions directly connected to these highly regulated gene nodes are also upregulated and associated with PPARγ (and some CEBPα) binding, including the *AGPAT* gene family members (directly downstream GPAM), the triacylglycerol synthesizing *DGAT1* and *DGAT2*, and the lipase *MGLL*, supporting a tight transcriptional regulation of triglyceride synthesis spread across the pathway. We selected the *GPAM* locus for further validation experiments, as it was the first enzyme targeted by both TFs and miR-27a. We could confirm binding to several prominent peaks upstream of the *GPAM* locus (Figure 8B and C). Furthermore, miRNA motif analysis of the 3′UTR revealed two miR-27a binding sites, one of which corresponds to a conserved site that has been shown functional in mice (56). In agreement, transfection with miR-27a mimic, but not that of mir-29a or mir-222, significantly decreased *GPAM* mRNA levels (Figure 8D).

Finally, we also examined the H3K4me3 data in context of the pathways identified to change most. A reciprocal change in the TSS activity affecting carbohydrate and lipid metabolism was observed for two genes encoding enzymes involved in glycerol metabolism: the *glycerol-3-phosphate dehydrogenases GPD1* and *GPD2*. Glycerol-3-phosphate (G3P) can be synthesized from glucose via an intermediate step that forms dihydroxyacetone phosphate

Figure 6. Continued
high enrichment for one or several TFs were selected for validation by ChIP-qPCR (numbered in the figure). The enrichment values using antibodies against all three TFs are shown relative to the enrichment of IgG and indicate the mean enrichment values of triplicate experiments and the error bars represent SEM. One sample *t*-test was performed to determine the significance of TF enrichment compared with IgG (*$P < 0.05$; **$P < 0.01$). (**C**) The metabolic pathway of cholesterol synthesis is shown with several omics data overlayed extending the metanode features presented in Figure 2. The pathway starts with the condensation of acetyl-CoA (accoa[c]) and acetoacetyl-CoA (aacoa[c]) to form 3-hydroxy-3-methylglutaryl-CoA (hmgcoa[c]) catalyzed by HMG-CoA synthase encoded by the gene *HMGCS1*. The end point metabolite is cholesterol (chsterol[r][m][c][e], r—endoplasmic reticulum, m—mitochondria, c—cytosol, e—extracellular). The start and end reactions are indicated by a thicker arrow and genes discussed further in the text are shown as larger metanodes for clarity. For details of heatmap, node and edge descriptions, see Figure 1, and for a complete list of metabolite names, Supplementary Table S7. Here, regulation in the SREBF1 microarray (48) is indicated by purple lining of the node. The reaction activity heatmap is blank if the gene is associated to multiple reactions with different predicted activity and in those cases the respective reaction activity heatmaps can be found below the pathway with the reaction naming matching those shown on the pathway. (**D**) RT-qPCR validation of the relative expression values of selected genes from the cholesterol synthesis pathway during SGBS differentiation. The measured expression values are shown normalized to *RPL13A* mRNA and relative to undifferentiated cells (set to 1). Data points indicate mean expression values of triplicate experiments and the error bars represent SEM. Student's *t*-test was performed to determine the significance of upregulation (*$P < 0.05$).

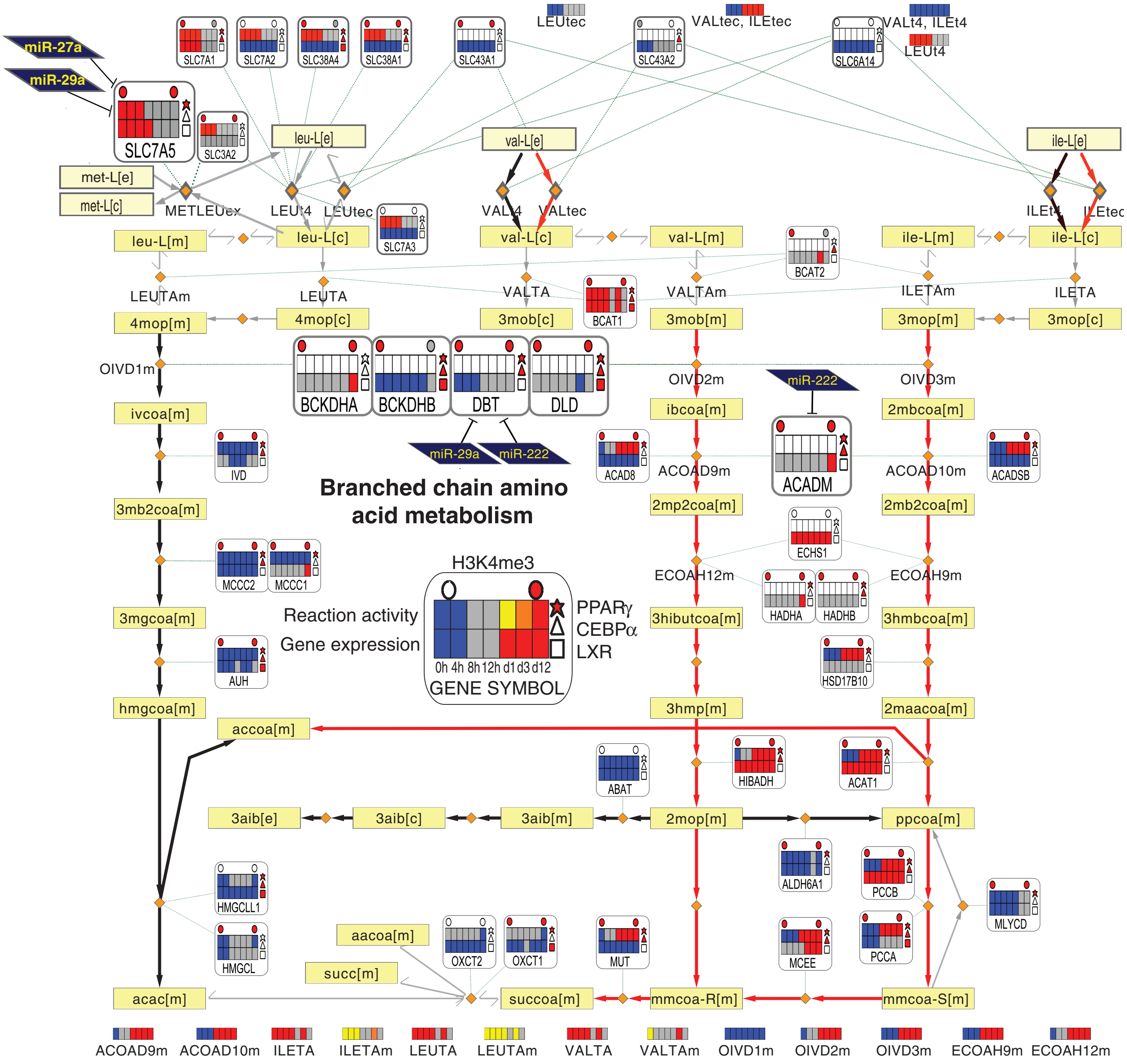


Figure 7. Integrated metabolic pathway of BCAA metabolism. The BCAA (valine, leucine and isoleucine) metabolism pathway is shown. The metanode and edge representation is identical to Figure 6. While leucine degradation is predicted inactive in both preadipocytes and adipocytes (left part), the degradation of both valine and isoleucine is predicted to become active in adipocytes (red edges). The respective end products acetyl-CoA and succinyl-CoA can be fed into the TCA cycle. The important intermediates malonyl-CoA and acetoacetate link to lipogenesis or ketone body formation, respectively. This pathway has similarities with FAO, sharing the enzyme *ACADM* and the metabolite propionyl-CoA (ppcoa[m]). The metanodes indicate that two nodes are associated with both TFs and miRNAs: The component of the large multienzyme complex, *DBT*, is associated with PPARγ, CEBPα, miR-29a and miR-222, while among BCAA transporters contained in Recon1, *SLC7A5* is associated to PPARγ, miR-27a and miR-29a. The genes discussed further in the text are shown as larger metanodes for clarity.

(DHAP). This metabolite and NADH are converted to G3P by GPD1, and G3P can subsequently be converted to lipids (55). *GPD1* increased H3K4me3 levels in differentiated cells, also in primary adipocytes (Figure 8E), matching its expression profile with a 7.9-fold increase in transcription (Supplementary Table S5). GPD2 in turn can convert G3P to quinone to fuel mitochondrial oxidation; in agreement with a shift to lipogenic metabolism its TSS activity was repressed (Figure 8E) and a low level of expression maintained.

Altogether, we obtained four novel genome-wide target gene profiles associating the TF LXRα and the miRNAs −27a, −29a and −222 with their likely target genes in human adipocytes to support the analysis performed using public data of TF binding for additional 12 TFs from adipocyte and endothelial cells. Such data on

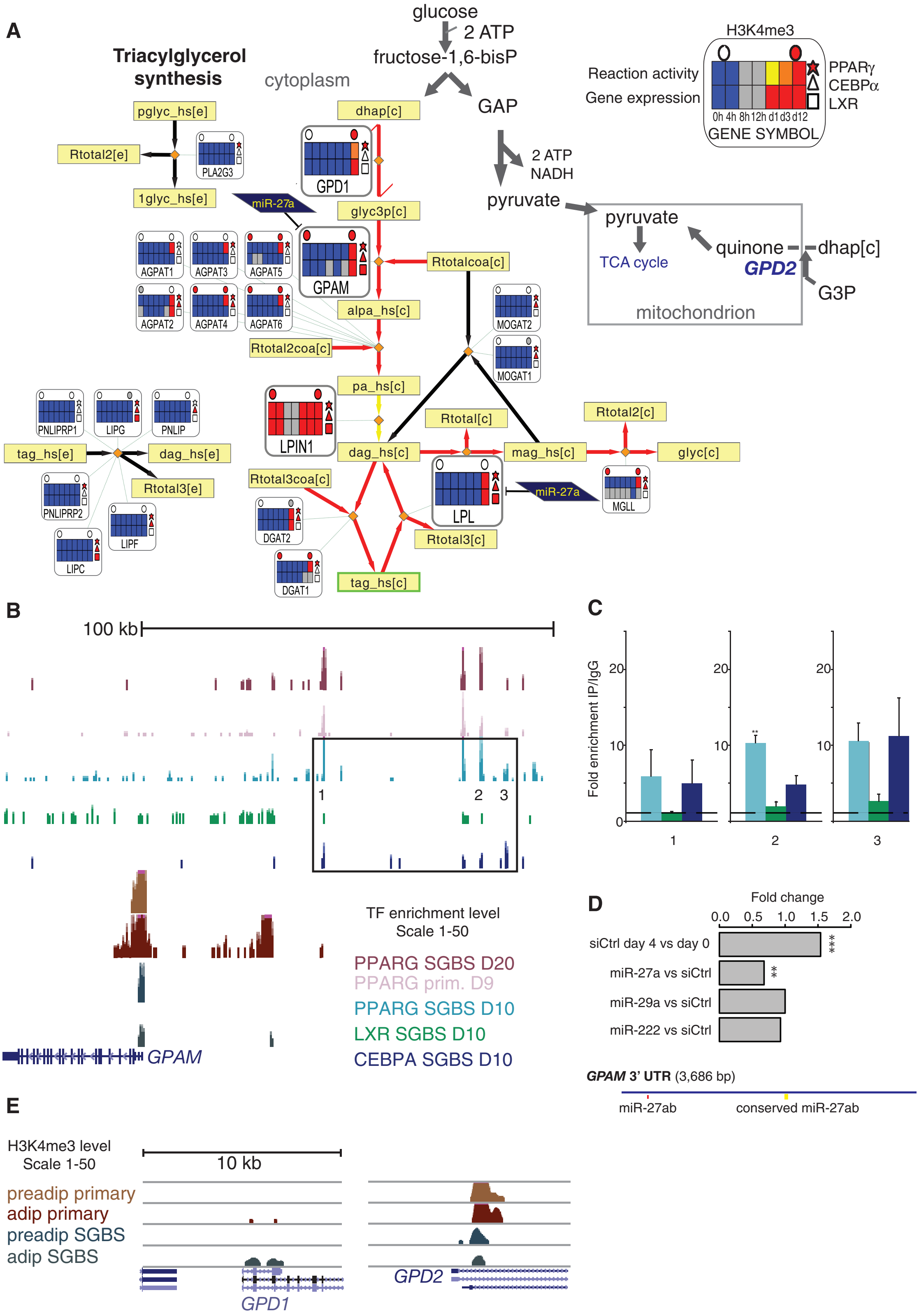


Figure 8. Integrated metabolic pathway of triacylglycerol synthesis. (**A**) The triacylglycerol synthesis pathway from Recon1 (2) is shown. The metanode and edge representation is identical to Figure 6. This pathway represents the synthesis of triacylglycerol (tag_hs[c]) from glycerol-3-phosphate (glyc3p[c]), which can be generated from the reduction of dihydroxyacetone phosphate (dhap[c]) by Glycerol-3-phosphate dehydrogenase (GPD1) (55). The *GPD1* and *GPD2* genes encode enzymes that function in an opposing manner in the conversion between DHAP and G3P. The mitochondrial GPD2 functions in a catabolic pathway that fuels the TCA cycle, whereas GPD1 plays a role in triglyceride synthesis. The initial and

(continued)

regulatory factors are highly relevant to describe their role in different cell types and potentially in driving disease. However, often the volume of data from genome-wide assays hinders tangible conclusions to be drawn. We provide here an easily accessible tool to analyze links between the metabolic and regulatory networks, to identify different regulatory mechanisms at the pathway level and for the discovery of key nodes as demonstrated here using the convergence of regulators to highlight such genes in the two cell types.

DISCUSSION

Despite the growing amount of data collected from gene expression studies, a common framework or a model to capture the key systems properties is often lacking. Here, we collect a comprehensive data set from own experiments and public data focusing on two key cell types implicated in the pathogenesis of metabolic syndrome, namely endothelial cells and adipocytes. To study in an integrated manner how components of transcriptional and posttranscriptional regulation impact the expression of metabolic genes, we introduce gene metanodes and the web portal IDARE (Integrated Data nodes of Regulation) for interactive data exploration of various data types within the metabolic network context.

The endothelium-derived relaxing factor NO and the catalyzing enzymes nitric oxide synthases that generate NO from the amino acid L-arginine represent a key discovery in CVD research. Together with genes implicated in hereditary monogenic disease (*ALDH4A1*) (38) or in recent GWAS studies (*MTAP*) (39), these enzymes from the arginine-proline metabolic pathway represent those associated with most TF binding in our analysis of 10 ChIP-seq studies from HUVEC cells.

TFs represent key factors to establish a cellular phenotype; however, they do not function in isolation. For a more comprehensive view, the analysis was extended from TFs to include the evaluation of chromatin marker levels and miRNAs. Three most upregulated TFs (PPARγ, CEBPα and LXRα) and prominent members of miRNA families (miR-27a, miR-29a and miR-222) downregulated on adipocyte differentiation were selected for genome-wide analysis. We validated the combinatorial binding in ChIP-qPCR and repression of mRNA using miRNA mimics for *GPAM*, representing a gene associated with TF and miRNA binding at a committing step in glycerolipid biosynthesis. Supporting the relevance of maintaining tight regulation of its expression level, Brockmöller *et al.* (59) reported a significant association between lowered *GPAM* mRNA levels and poor survival in breast cancer, a misregulation that in context of our results could be linked to the increased expression levels of miR-27a reported in invasive breast cancers (60). These results are highly supportive of the relevance of regulatory associations following the approach used here. However, mammalian regulatory regions may overlap and span across gene boundaries resulting inevitably in some false target gene associations. As further possible caveats, target mRNA dynamics impact detection of the miRNA regulatory effect and ChIP-seq signal only implies TF binding that constitutes a necessary, but not sufficient, condition to alter mRNA synthesis.

Despite possible inaccuracies in representing true regulatory interactions, the integrated analysis on metabolic pathway regulation clearly implicated the dyslipidemia loci *LPL* (24) and *LPIN1* (23) as well as *LDLR* (22) (the latter is missing from Recon1), each associated with multiple TFs and miRNAs. Furthermore, the *GPD1* locus that we identify among genes with increased TSS marker levels has recently been described to cause infantile hypertriglyceridemia (61). Thus, our analysis holds promise to identify key regulatory nodes that are important in different diseases by using microarray and sequencing data that are readily available from multiple tissues.

Our data extends data collected on PPARγ and CEBPα in human adipocytes (10,40,41,44), while for LXR and miR-27a, our genome-wide data on target genes are the first reported in human adipocytes and can be compared with data obtained from liver (49,56) and foam cells (62). PPARγ and CEBPα represent cell fate determining TFs widely studied in context of adipocytes. However, their interplay with signal-dependent TFs is less well understood, including LXRα that increased at expression level most during differentiation. The first glimpse to the LXR genome-wide binding profile through ChIP-seq showed binding in a few hundred to few thousand regions (high- versus low-occupancy cutoff), in agreement with a similar number of binding sites reported from unstimulated mouse liver cells (49). Most strikingly, among all upregulated TFs, only *SREBF1* was associated with TFs other than PPARγ, being bound by LXR and CEBPα. The cholesterol synthesis and fatty acid activation

Figure 8. Continued
committing step in glycerolipid biosynthesis is catalyzed by GPAM. *GPAM, LPIN1* and *LPL* are all associated with all three TFs and in addition *GPAM* and *LPL* are targeted by miR-27a. The synthesis of both triacylglycerol and glycerol is predicted to shift to active in adipocytes. The genes discussed further in the text are shown as larger metanodes for clarity. (**B**) The ChIP-seq signal tracks as in Figure 5 are shown at the *GPAM* locus. Regions with high enrichment for one or several TFs were selected for validation by ChIP-qPCR (numbered in the figure). Each region was tested for enrichment using antibodies against all three TFs and IgG as a control as is shown in adjacent plots (**C**). The enrichment values are shown relative to the enrichment of IgG and indicate the mean enrichment values of triplicate experiments and the error bars represent SEM. One sample *t*-test was performed to determine the significance of TF enrichment compared with IgG (*$P < 0.05$; **$P < 0.01$). (**D**) The GPAM 3′UTR is shown with miRNA target predictions from TargetScan. Two binding sites for miR-27a can be seen, one of which is conserved and previously validated in mouse liver (56). Fold change values from miRNA mimic transfections are displayed from the microarray data. Two sample *t*-test was performed to determine the significance of silencing compared with siCtrl transfection (**adjusted $P < 0.01$; ***adjusted $P < 0.001$). (**E**) Signal tracks at the vicinity of their TSS regions of *GPD1* and *GPD2* show the H3K4me3 ChIP-seq signal from primary preadipocytes and adipocytes (10) compared with SGBS preadipocytes and adipocytes. At the *GPD1* locus, the H3K4me3 signal increases in SGBS and primary adipocytes, whereas a decrease in signal is observed at the *GPD2* TSS in SGBS cells.

pathways were each associated with this putative multi-TF feed-forward circuit. It is intriguing that precisely these pathways have been reported to contribute to generation of endogenous PPARγ ligands (51,63), potentially providing a metabolite positive feedback to substantiate transcriptional autoactivation required for cell differentiation.

On activating signal, our microarray using an LXR agonist revealed regulation of several genes that in our ChIP-seq data were initially associated with low occupancy binding. Similar study in mouse liver (49) showed a dramatic increase in peak height on agonist activation, suggesting that the ligand-bound receptor may be more efficiently recruited to its genomic target loci. It will therefore be of interest to test the ligand-dependent binding profile also in adipocytes. LXRs have been shown to play critical roles in the regulation of overall cholesterol catabolism, absorption and transport in the intestine, macrophages and liver (42). Furthermore, the LXR target gene *MYLIP* that inhibits the LDLR pathway by targeting LDLR to proteasomal degradation (64) is a likely target gene of miR-222 according to our microarray and seed analysis, exposing a CVD-relevant novel regulatory factor in the cholesterol intercellular trafficking pathway.

The miRNA with most interaction with TFs (including regulation of *PPARG*) is miR-27a. The regulation of *GPAM* by miR-27b was recently described in mouse liver (56) and is supported by our microarray and heptamer analysis in human adipocytes. Moreover, we observe multiple other genes that are posttranscriptionally regulated along the pathway. These target associations include the *LDLR*, *LPL* and *LRP5* that function in lipid transport. It is worth pointing out that also *LPIN1* is among genes that have a modest downregulation on miR-27a transfection, in agreement with existence of a well-conserved binding site in its 3′-UTR. However, further validation is required to ascertain its regulation. Participation of carbohydrate metabolism in fueling the triacylglycerol synthesis is supported by the switch in regulation of *GPD* genes. The upstream enzyme *hexokinase-2* that phosphorylates and thereby activates glucose, similar to *GPAM*, *LPIN1* or *LPL*, was identified among the list of target genes associated with all TFs and miR-27a. Based on the earlier reports and our combined microarray and 3′-UTR heptamer analysis, miR-27 family is establishing itself as a key miRNA regulator of the triacylglycerol metabolism.

Interestingly, both miR-222 and miR-29a regulate the BCKD complex in the BCAA catabolism. Recently, increased levels of the BCAAs were shown to play an important role in diabetes (65). In a model proposed by Newgard (66), and supported by experiments applying *in vivo* mouse models (67), an obesity-related decline in BCAA catabolism in adipose tissue drives the rise of circulating levels of these amino acids. The model suggests that readily usable glucose and lipid substrates may obviate the need for amino acid catabolism in adipose tissue. However, the mechanism by which increased supply of these substrates causes downregulation of the BCAA catabolic enzymes is unknown. Drugs that activate PPARγ (thiazolidindiones or TZDs) can restore expression of the catabolic genes to normal (68), already suggesting a role of suppressed PPAR signaling in this metabolic adaptation. The miR-29 family is implicated in diabetes based on studies of hepatic gluconeogenesis in diabetic rat models (69). Based on our data, it will be relevant not only to study PPARγ, but to include CEBPα, miR-27a, miR-29a and miR-222 as other key regulators of this pathway, and by that potentially further elucidate the novel link of the BCAA pathway to diabetes.

Both miRNAs targeting the DBT subunit of the BCKD complex according to our data (miR-29a and miR-222) are upregulated in the adipose tissue of diabetic rats and are induced by increased glucose levels in mouse adipocytes (70,71). Moreover, the targeting of *DBT* by miR-29 family has already been validated in other cell types (72), making the combinatorial repression of *DBT* by the miRNAs one likely explanation for lowered catabolism of BCAAs in diabetic adipose tissue. In addition to the initial steps of BCAA catabolism, also the BCAA transport step mediated by SLC7A5 appears to be a highly regulated node possibly contributing to the diabetic phenotype. On top of being associated as a PPARγ target in our analysis, *SLC7A5* appears to be targeted by miR-27a and miR-29a, both glucose responsive and induced in diabetic condition (71). The complexity of *SLC7A5* regulation is further increased when looking at HUVEC data that reveal it as a potential target of as many as 9 TFs and an additional miRNA (miR-663) in the endothelial cells and in context of recent literature implicating it in key metabolic changes required for T-cell differentiation (73).

As an initial means to discover key pathways, we used the gene expression levels as soft constraints to obtain predictions for metabolic activity in Recon1, a generic model of human metabolism (2,6). Several other methods for the integration of expression data on genome scale metabolic networks have been and are currently being developed (74) and will be important to benchmark and consolidate the prediction results in future studies. Transcript-level measurements address the space of available network states that translational control and posttranslational modifications further fine tune [for HMGCR this is well established (53)]. The metanodes enable mapping and visualization of further data onto metabolic pathways, facilitating data exchange and hypothesis-driven research in context of the metabolic network. Here, two trends in transcriptional regulation were observed: (i) shared and high-occupancy binding nearby gene loci of the initial and terminal steps of a pathway (the cholesterol synthesis and the BCAA pathways), a type of transcriptional regulation that has been reported advantageous for fast responses to environmental conditions in pathways with low protein synthesis cost (75), and (ii) tight regulation spread along the entire pathway (triacylglycerol synthesis), which might link to the tight transcriptional regulation on pathways spanning high cost enzymes (75).

In conclusion, the analysis of genomic and transcriptomic data linked with a metabolic network model is useful as a means to explore high-throughput data in a global manner, revealing genes implicated in disease as convergence points of regulation. To focus on metabolic pathways that differ in activity comparing two phenotypes, constraint-based modeling to predict active metabolic pathways can be included. The putative shared TF and miRNA target genes from pathways activated during human adipocyte differentiation that our new data sets revealed were *ACADM*, *DBT*, *GPAM*, *HK2*, *LPL* and *SLC7A5*. Genes associated with all adipocyte TFs studied further include *ABCA1*, *ACSL1*, the *acetyl-CoA acetyltransferase 2* (*ACAT2*), the *BCAT1*; two more genes from the branched-chain alpha-keto acid dehydrogenase complex, namely *BCKDHB* and the *DLD*; four genes from the cholesterol synthesis pathway *DHCR7*, *HMGCS2*, *HMGCLL1*, *HMGCR*; and three other lipase genes from triglyceride metabolism, namely *lipase C*, *lipase G* and *LPIN1*. These data can now be compared with published data sets such as those from HUVECs using the IDARE tool. Our workflow extends from current routines in which these disparate but complementary types of cellular information are kept apart and further motivates study of biological processes from an integrative point-of-view.

ACCESSION NUMBERS

The microarray and deep sequencing data from this publication have been submitted to the NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/) and assigned the identifier GSE41578 and can be explored in context of the pathways described using the web resource at http://systemsbiology.uni.lu/idare.html, including a track hub for UCSC Genome Browser that allows fast visualization of ChIP-seq signal tracks at any gene locus.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr Anna-Liisa Levonen, Prof. Ilya Shmulevich and Prof. Carsten Carlberg for comments and suggestions to this manuscript, Prof. Martin Wabitsch for kindly providing the SGBS cell line, Dr Echardt Treuter for kindly providing the LXR antibody, Martine Schmitz for testing the antibodies used for ChIP-seq, Dr Patrick May, Fotis Georgatos, Christian Hutter and Dr Reinhard Schneider for providing access to the LCSB-Portal server and the high-throughput data service providers at Microarray and sequencing facility of Turku Centre for Biotechnology, at DNA Vision, Charleroi, Belgium and at EMBL Genomics Core at Heidelberg. M.H. and L.S. designed and performed the microarray time series and ChIP-seq experiments. L.S. designed and performed the miRNA array and miRNA mimic transfection experiments and ChIP-qPCR experiments. M.G. performed the constraint-based modeling jointly implemented together with T.S. M.H. performed the statistical analysis of the microarray data and P.B. designed and performed the miRNA motif enrichment and target association analysis. M.G. and L.S. designed and performed the RT-qPCR validations. M.H. and M.G. analyzed the ChIP-seq data. J.L., M.G., M.H. and T.S. designed the online web tool that J.L. and M.G. implemented. All authors participated in designing the integrative analysis and writing of the manuscript.

FUNDING

Fonds National de la Recherche Luxembourg [AFR, AM2c and C08/BM/01]; funding from the University of Eastern Finland and the Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg. Funding for open access charge: University of Luxembourg.

Conflict of interest statement. None declared.

REFERENCES

1. World Health Organization. (2000) *Global status report on noncommunicable diseases 2010*, World Health Organization, Geneva, Switzerland. http://www.who.int/nmh/publications/ncd_report2010/en/.
2. Duarte,N.C., Becker,S.A., Jamshidi,N., Thiele,I., Mo,M.L., Vo,T.D., Srivas,R. and Palsson,B.Ø. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.
3. Thiele,I., Swainston,N., Fleming,R.M.T., Hoppe,A., Sahoo,S., Aurich,M.K., Haraldsdottir,H., Mo,M.L., Rolfsson,O., Stobbe,M.D. *et al.* (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **31**, 419–425.
4. Rolfsson,O., Palsson,B.Ø. and Thiele,I. (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.*, **5**, 155.
5. Suarez,R.K. and Moyes,C.D. (2012) Metabolism in the age of "omes". *J. Exp. Biol.*, **215**, 2351–2357.
6. Shlomi,T., Cabili,M.N., Herrgård,M.J., Palsson,B.Ø. and Ruppin,E. (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.
7. Farmer,S.R. (2006) Transcriptional control of adipocyte formation. *Cell Metab.*, **4**, 263–273.
8. Nielsen,R., Pedersen,T.A., Hagenbeek,D., Moulos,P., Siersbaek,R., Megens,E., Denissov,S., Børgesen,M., Francoijs,K.-J., Mandrup,S. *et al.* (2008) Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.*, **22**, 2953–2967.
9. Lefterova,M.I., Steger,D.J., Zhuo,D., Qatanani,M., Mullican,S.E., Tuteja,G., Manduchi,E., Grant,G.R. and Lazar,M.A. (2010) Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. *Mol. Cell. Biol.*, **30**, 2078–2089.
10. Mikkelsen,T.S., Xu,Z., Zhang,X., Wang,L., Gimble,J.M., Lander,E.S. and Rosen,E.D. (2010) Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, **143**, 156–169.
11. Cho,N. and Momose,Y. (2008) Peroxisome proliferator-activated receptor gamma agonists as insulin sensitizers: from the discovery to recent progress. *Curr Medic Chem.*, **8**, 1483–1507.
12. Ono,K. (2012) Current concept of reverse cholesterol transport and novel strategy for atheroprotection. *J Cardiol.*, **60**, 339–343.
13. Mudhasani,R., Imbalzano,A.N. and Jones,S.N. (2010) An essential role for Dicer in adipocyte differentiation. *J. Cell. Biochem.*, **110**, 812–816.

14. Wang,Q., Li,Y.C., Wang,J., Kong,J., Qi,Y., Quigg,R.J. and Li,X. (2008) miR-17-92 cluster accelerates adipocyte differentiation by negatively regulating tumor-suppressor Rb2/p130. *Proc. Natl Acad. Sci. USA*, **105**, 2889–2894.
15. Karbiener,M., Fischer,C., Nowitsch,S., Opriessnig,P., Papak,C., Ailhaud,G., Dani,C., Amri,E.-Z. and Scheideler,M. (2009) microRNA miR-27b impairs human adipocyte differentiation and targets PPARgamma. *Biochem. Biophys. Res. Commun.*, **390**, 247–251.
16. Kim,S.Y., Kim,A.Y., Lee,H.W., Son,Y.H., Lee,G.Y., Lee,J.-W., Lee,Y.S. and Kim,J.B. (2010) miR-27a is a negative regulator of adipocyte differentiation via suppressing PPARgamma expression. *Biochem. Biophys. Res. Commun.*, **392**, 323–328.
17. Lin,Q., Gao,Z., Alarcon,R.M., Ye,J. and Yun,Z. (2009) A role of miR-27 in the regulation of adipogenesis. *FEBS J.*, **276**, 2348–2358.
18. Sun,T., Fu,M., Bookout,A.L., Kliewer,S.A. and Mangelsdorf,D.J. (2009) MicroRNA let-7 regulates 3T3-L1 adipogenesis. *Mol. Endocrinol.*, **23**, 925–931.
19. John,E., Wienecke-Baldacchino,A., Liivrand,M., Heinäniemi,M., Carlberg,C. and Sinkkonen,L. (2012) Dataset integration identifies transcriptional regulation of microRNA genes by PPARγ in differentiating mouse 3T3-L1 adipocytes. *Nucleic Acids Res.*, **40**, 1–15.
20. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
21. Wabitsch,M., Brenner,R.E., Melzner,I., Braun,M., Möller,P., Heinze,E., Debatin,K.M. and Hauner,H. (2001) Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *Int. J. Obes. Relat. Metab. Disord.*, **25**, 8–15.
22. Francke,U., Brown,M.S. and Goldstein,J.L. (1984) Assignment of the human gene for the low density lipoprotein receptor to chromosome 19: synteny of a receptor, a ligand, and a genetic disease. *Proc. Natl Acad. Sci. USA*, **81**, 2826–2830.
23. Péterfy,M., Phan,J., Xu,P. and Reue,K. (2001) Lipodystrophy in the fld mouse results from mutation of a new gene encoding a nuclear protein, lipin. *Nat. Genet.*, **27**, 121–124.
24. Harlan,W.R., Winesett,P.S. and Wasserman,A.J. (1967) Tissue lipoprotein lipase in normal individuals and in individuals with exogenous hypertriglyceridemia and the relationship of this enzyme to assimilation of fat. *J. Clin. Invest.*, **46**, 239–247.
25. Eichler,G.S., Huang,S. and Ingber,D.E. (2003) Gene expression dynamics inspector (GEDI): for integrative analysis of expression profiles. *Bioinformatics*, **19**, 2321–2322.
26. Newman,A.M. and Cooper,J.B. (2010) AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinf.*, **11**, 117.
27. Castoldi,M., Schmidt,S., Benes,V., Hentze,M.W. and Muckenthaler,M.U. (2008) miChip: an array-based method for microRNA expression profiling using locked nucleic acid capture probes. *Nat. Protoc.*, **3**, 321–329.
28. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
29. Berninger,P., Gaidatzis,D., Van Nimwegen,E. and Zavolan,M. (2008) Computational analysis of small RNA cloning data. *Methods*, **44**, 13–21.
30. Ma,J., Flemr,M., Stein,P., Berninger,P., Malik,R., Zavolan,M., Svoboda,P. and Schultz,R.M. (2010) MicroRNA activity is suppressed in mouse oocytes. *Curr. Biol.*, **20**, 265–270.
31. Sinkkonen,L., Hugenschmidt,T., Berninger,P., Gaidatzis,D., Mohn,F., Artus-Revel,C.G., Zavolan,M., Svoboda,P. and Filipowicz,W. (2008) MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.*, **15**, 259–267.
32. Bauer-Mehren,A., Rautschka,M., Sanz,F. and Furlong,L.I. (2010) DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, **26**, 2924–2926.
33. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
34. Hebenstreit,D., Gu,M., Haider,S., Turner,D.J., Liò,P. and Teichmann,S.A. (2011) EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic acids Res.*, **39**, e27.
35. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
36. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
37. Ignarro,L.J. (1989) Endothelium-derived nitric oxide: pharmacology and relationship to the actions of organic nitrate esters. *Pharm. Res.*, **6**, 651–659.
38. Geraghty,M.T., Vaughn,D., Nicholson,A.J., Lin,W.W., Jimenez-Sanchez,G., Obie,C., Flynn,M.P., Valle,D. and Hu,C.A. (1998) Mutations in the Delta1-pyrroline 5-carboxylate dehydrogenase gene cause type II hyperprolinemia. *Hum. Mol. Genet.*, **7**, 1411–1415.
39. McPherson,R., Pertsemlidis,A., Kavaslar,N., Stewart,A., Roberts,R., Cox,D.R., Hinds,D.A., Pennacchio,L.A., Tybjaerg-Hansen,A., Folsom,A.R. *et al.* (2007) A common allele on chromosome 9 associated with coronary heart disease. *Science*, **316**, 1488–1491.
40. Schmidt,S.F., Jørgensen,M., Chen,Y., Nielsen,R., Sandelin,A. and Mandrup,S. (2011) Cross species comparison of C/EBPα and PPARγ profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics*, **12**, 152.
41. Soccio,R.E., Tuteja,G., Everett,L.J., Li,Z., Lazar,M.A. and Kaestner,K.H. (2011) Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol. Endocrinol.*, **25**, 694–706.
42. Calkin,A.C. and Tontonoz,P. (2012) Transcriptional integration of metabolism by the nuclear sterol-activated receptors LXR and FXR. *Nat. Rev. Mol. Cell Biol.*, **13**, 213–224.
43. Rottiers,V. and Näär,A.M. (2012) MicroRNAs in metabolism and metabolic disorders. *Nat. Rev. Mol. Cell Biol.*, **13**, 239–250.
44. Siersbæk,M.S., Loft,A., Aagaard,M.M., Nielsen,R., Schmidt,S.F., Petrovic,N., Nedergaard,J. and Mandrup,S. (2012) Genome-wide profiling of peroxisome proliferator-activated receptor γ in primary epididymal, inguinal, and brown adipocytes reveals depot-selective binding correlated with gene expression. *Mol. Cell. Biol.*, **32**, 3452–3463.
45. Kanter,J.E., Tang,C., Oram,J.F. and Bornfeldt,K.E. (2012) Acyl-CoA synthetase 1 is required for oleate and linoleate mediated inhibition of cholesterol efflux through ATP-binding cassette transporter A1 in macrophages. *Biochim Biophys Acta*, **1821**, 358–364.
46. Gregersen,L.H., Jacobsen,A., Frankel,L.B., Wen,J., Krogh,A. and Lund,A.H. (2012) MicroRNA-143 down-regulates Hexokinase 2 in colon cancer cells. *BMC cancer*, **12**, 232.
47. Kwee,S.A., Hernandez,B., Chan,O. and Wong,L. (2012) Choline kinase alpha and hexokinase-2 protein expression in hepatocellular carcinoma: association with survival. *PloS One*, **7**, e46591.
48. Rome,S., Lecomte,V., Meugnier,E., Rieusset,J., Debard,C., Euthine,V., Vidal,H. and Lefai,E. (2008) Microarray analyses of SREBP-1a and SREBP-1c target genes identify new regulatory pathways in muscle. *Physiol. Genomics*, **34**, 327–337.
49. Boergesen,M., Pedersen,T.Å., Gross,B., Van Heeringen,S.J., Hagenbeek,D., Bindesbøll,C., Caron,S., Lalloyer,F., Steffensen,K.R., Nebb,H.I. *et al.* (2012) Genome-wide profiling of liver X receptor, retinoid X receptor, and peroxisome proliferator-activated receptor α in mouse liver reveals extensive sharing of binding sites. *Mol. Cell. Biol.*, **32**, 852–867.
50. Qin,Y., Dalen,K.T., Gustafsson,J.-A. and Nebb,H.I. (2009) Regulation of hepatic fatty acid elongase 5 by LXRalpha-SREBP-1c. *Biochim. Biophys. Acta*, **1791**, 140–147.
51. Goto,T., Nagai,H., Egawa,K., Kim,Y.-I., Kato,S., Taimatsu,A., Sakamoto,T., Ebisu,S., Hohsaka,T., Miyagawa,H. *et al.* (2011)

Farnesyl pyrophosphate regulates adipocyte functions as an endogenous PPARγ agonist. *Biochem. J.*, **438**, 111–119.
52. Wang,Y., Rogers,P.M., Stayrook,K.R., Su,C., Varga,G., Shen,Q., Nagpal,S. and Burris,T.P. (2008) The selective Alzheimer's disease indicator-1 gene (Seladin-1/DHCR24) is a liver X receptor target gene. *Mol. Pharmacol.*, **74**, 1716–1721.
53. Gibson,D.M., Parker,R.A., Stewart,C.S. and Evenson,K.J. (1982) Short-term regulation of hydroxymethylglutaryl coenzyme A reductase by reversible phosphorylation: modulation of reductase phosphatase in rat hepatocytes. *Adv. Enzyme Regul.*, **20**, 263–283.
54. Ni,C.-W., Qiu,H. and Jo,H. (2011) MicroRNA-663 upregulated by oscillatory shear stress plays a role in inflammatory response of endothelial cells. *Am. J. Physiol.: Heart Circ. Physiol.*, **300**, H1762–H1769.
55. Coleman,R.A. and Lee,D.P. (2004) Enzymes of triacylglycerol synthesis and their regulation. *Prog. Lipid Res.*, **43**, 134–176.
56. Vickers,K.C., Shoucri,B.M., Levin,M.G., Wu,H., Pearson,D.S., Osei-Hwedieh,D., Collins,F.S., Remaley,A.T. and Sethupathy,P. (2012) MicroRNA-27b is a regulatory hub in lipid metabolism and is altered in dyslipidemia. *Hepatology*, **01**, 1–10.
57. Dircks,L.K. and Sul,H.S. (1997) Mammalian mitochondrial glycerol-3-phosphate acyltransferase. *Biochim. Biophys. Acta*, **1348**, 17–26.
58. McNamara,J.P., Azain,M., Kasser,T.R. and Martin,R.J. (1982) Lipoprotein lipase and lipid metabolism in muscle and adipose tissues of Zucker rats. *Am. J. Physiol.*, **243**, R258–R264.
59. Brockmöller,S.F., Bucher,E., Müller,B.M., Budczies,J., Hilvo,M., Griffin,J.L., Orešič,M., Kallioniemi,O., Iljin,K., Loibl,S. *et al.* (2011) Integration of metabolomics and expression of glycerol-3-phosphate acyltransferase (GPAM) in breast cancer–link to patient survival, hormone receptor status and metabolic profiling. *J. Proteome Res.*, **11**, 850–860.
60. Tang,W., Zhu,J., Su,S., Wu,W., Liu,Q., Su,F. and Yu,F. (2012) MiR-27 as a prognostic marker for breast cancer progression and patient survival. *PLoS One*, **7**, e51702.
61. Basel-Vanagaite,L., Zevit,N., Zahav,A.H., Guo,L., Parathath,S., Pasmanik-Chor,M., McIntyre,A.D., Wang,J., Albin-Kaplanski,A., Hartman,C. *et al.* (2012) Transient infantile hypertriglyceridemia, fatty liver, and hepatic fibrosis caused by mutated GPD1, encoding glycerol-3-phosphate dehydrogenase 1. *Am. J. Hum. Genet.*, **90**, 49–60.
62. Feldmann,R., Fischer,C., Kodelja,V., Behrens,S., Haas,S., Vingron,M., Timmermann,B., Geikowski,A. and Sauer,S. (2013) Genome-wide analysis of LXRα activation reveals new transcriptional networks in human atherosclerotic foam cells. *Nucleic Acids Res.*, **41**, 3518–3531.
63. Shiraki,T., Kamiya,N., Shiki,S., Kodama,T.S., Kakizuka,A. and Jingami,H. (2005) Alpha,beta-unsaturated ketone is a core moiety of natural ligands for covalent binding to peroxisome proliferator-activated receptor gamma. *J. Biol. Chem.*, **280**, 14145–14153.
64. Zelcer,N., Hong,C., Boyadjian,R. and Tontonoz,P. (2009) LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor. *Science*, **325**, 100–104.
65. McCormack,S.E., Shaham,O., McCarthy,M.A., Deik,A.A., Wang,T.J., Gerszten,R.E., Clish,C.B., Mootha,V.K., Grinspoon,S.K. and Fleischman,A. (2012) Circulating branched-chain amino acid concentrations are associated with obesity and future insulin resistance in children and adolescents. *Pediatr. Obes.*, **8**, 52–61.
66. Newgard,C.B. (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance. *Cell Metab.*, **15**, 606–614.
67. Herman,M.A., She,P., Peroni,O.D., Lynch,C.J. and Kahn,B.B. (2010) Adipose tissue branched chain amino acid (BCAA) metabolism modulates circulating BCAA levels. *J. Biol. Chem.*, **285**, 11348–11356.
68. Hsiao,G., Chapman,J., Ofrecio,J.M., Wilkes,J., Resnik,J.L., Thapar,D., Subramaniam,S. and Sears,D.D. (2011) Multi-tissue, selective PPARγ modulation of insulin sensitivity and metabolic pathways in obese rats. *Am. J. Physiol.: Endocrinol. Metab.*, **300**, E164–E174.
69. Liang,J., Liu,C., Qiao,A., Cui,Y., Zhang,H., Cui,A., Zhang,S., Yang,Y., Xiao,X., Chen,Y. *et al.* (2012) MicroRNA-29a-c decrease fasting blood glucose levels by negatively regulating hepatic gluconeogenesis. *J. Hepatol.*, **58**, 535–542.
70. He,A., Zhu,L., Gupta,N., Chang,Y. and Fang,F. (2007) Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes. *Mol. Endocrinol.*, **21**, 2785–2794.
71. Herrera,B.M., Lockstone,H.E., Taylor,J.M., Ria,M., Barrett,A., Collins,S., Kaisaki,P., Argoud,K., Fernandez,C., Travers,M.E. *et al.* (2010) Global microRNA expression profiles in insulin target tissues in a spontaneous rat model of type 2 diabetes. *Diabetologia*, **53**, 1099–1109.
72. Mersey,B.D., Jin,P. and Danner,D.J. (2005) Human microRNA (miR29b) expression controls the amount of branched chain alpha-ketoacid dehydrogenase complex in a cell. *Hum. Mol. Genet.*, **14**, 3371–3377.
73. Sinclair,L.V., Rolf,J., Emslie,E., Shi,Y.-B., Taylor,P.M. and Cantrell,D.A. (2013) Control of amino-acid transport by antigen receptors coordinates the metabolic reprogramming essential for T cell differentiation. *Nat. Immunol.*, **14**, 500–508.
74. Blazier,A.S. and Papin,J.A. (2012) Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.*, **3**, 299.
75. Wessely,F., Bartl,M., Guthke,R., Li,P., Schuster,S. and Kaleta,C. (2011) Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol. Syst. Biol.*, **7**, 515.

Supplementary File I

Integrated analysis of transcript level regulation of metabolism reveals disease relevant nodes of the human metabolic network

Mafalda Galhardo[1*], Lasse Sinkkonen[1*], Philipp Berninger[2], Jake Lin[3,4], Thomas Sauter[1#] and Merja Heinäniemi[1,5#]

[1]Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg

[2]Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

[3]Institute for Systems Biology, 401 Terry Avenue North, 98109-5234, Seattle, Washington, USA

[4]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, House of Biomedicine, 7 Avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg

[5]A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, FI-7120 Kuopio, Finland

[*]These authors contributed equally

[#]To whom correspondence should be addressed

Table of Content

Supplementary Table Legends

Table S1: Primer sequences for RT-qPCR and ChIP-qPCR. The primer sequences used in PCR reactions of the validation experiments are listed here.

Table S2: ChIP-seq peaks identified for PPARγ, CEBPα and LXR in day 10 differentiated SGBS cells and in a comparable analysis of SRX032890 and SRX019521data sets. The number of reads that passed through each processing step is indicated for the SGBS data obtained here and the public raw read data processed from SRX032890 and SRX019521. The ChIP-seq peaks together with enrichment quantification and statistical significance values as identified using QuEST tool (35) are presented for each data set. The same analysis settings were applied to generate highly comparable data. The peaks that pass the enrichment threshold >30, chosen here to distinguish high-occupancy binding, are highlighted in bold. Notice that a separate sheet exists in the xls file for each data set.

Table S3: Peak to gene association and ontology term enrichment analysis for ChIP-seq data sets. The ChIP-seq peak coordinates from Table S2 were used as input for the GREAT tool (36) that first associates the peaks to putative target genes listed. Significant ontology terms were collected and highlight the role of these TFs in lipid and carbohydrate metabolism.

Table S4: HUVEC TF and disease association result. Recon1 metabolic genes are shown in context of the number of associated diseases and TFs (from 10 ChIP-Seq studies on HUVEC), detailed analysis description is found on Materials and Methods. Data were sorted by gene relevance for endothelial disease, by the number of associated TFs and by the H3K4me3 active transcription mark. The number of diseases the gene is associated to is based on DisGeNET database (32).

Table S5: Differentially expressed metabolic genes during SGBS differentiation. The average logarithmic fold change values and statistical analysis including t-test for individual time points and F-test results across all time points is presented for differentially expressed metabolic genes (based on Recon1 (2)).

Table S6: Comparison on reaction activity predictions for pre-adipocytes and adipocytes. Metabolic changes resulting from human SGBS pre-adipocyte cell differentiation were qualitatively predicted from gene expression data using an implementation of the constraint-based method from (6). The 323 reactions with predicted reaction activity change are highlighted.

Table S7: Naming of metabolic genes and enzymes from selected pathways. The complete names for the metabolites and enzymes included in the pathway figures are presented.

Table S8: Differentially expressed TF genes during SGBS differentiation. The average logarithmic fold change values and statistical analysis including t-test for individual time points and F-test results across all time points is presented for differentially expressed TF genes.

Table S9: Genes with altered H3K4me3 status during SGBS differentiation. The H3K4me3 histone marker quantification from -1250 to +750 bp around gene TSS is presented for metabolic genes that changed their H3K4me3 status in SGBS cells, including the respective data from primary adipocytes (10).

Table S10: Target gene associations for miR-27a, miR-29a and miR-222 based on combined microarray target profiling and heptamer motif analysis. Genes identified to be responsive to miRNA mediated regulation from SGBS array profiling experiments following miRNA over-expression and 3’-UTR motif analysis are listed (see Methods for details). Notice the separate data sheets for each miRNA.

Table S11: Differentially expressed genes in 4 h LXR agonist T0901317 stimulated SGBS adipocytes. The average logarithmic fold change values and statistical analysis is presented for differentially expressed genes upon ligand activation of LXRs.

Table S12: Additional data supporting LXR peak to gene associations. Data in support of LXR mediated regulation of the genes associated with LXR peaks is presented collected from own microarrays (see Table S11) and GSE35262.

Table S13: Hypergeometric test for enriched pathway terms among genes regulated by SREBF1. Recon1 pathway enrichment results from a hypergeometric test on genes reported as SREBF1 targets on muscle. Cholesterol pathway ranked first, followed by oxidative phosphorylation and fatty acid activation and elongation.

Supplementary Figure Legends

Fig. S1: Integrated metabolic pathways of arginine and proline metabolism in HUVECs. The complete argine-proline metabolism pathway that contains the top disease associated gene *NOS3* is shown. Regulatory associations from ten ChIP-seq studies (as in Fig. 1) are displayed in the gene metanodes. Among genes involved the initial steps of the pathway, *ALDH4A1, ALDH2* and *MTAP*, represent genes associated with endothelial relevant disease and with multiple TFs. TF association is indicated with filled circles. The genes discussed further in the text are shown as larger metanodes for clarity.

Fig. S2: Time series expression profile of metabolic genes during SGBS differentiation. The average logarithmic fold change values from 4, 8 and 12 h and days 1, 3 and 12 are displayed in color using GEDI maps (25) to cluster metabolic genes with similar expression profiles. Initially, the responses seen are modest shifting to more prominent up- and down-regulation by day 3 with the largest changes observed at day 12. The number of genes in each cluster is displayed in the Gene density panel below. To distinguish pathway dynamics, overrepresented metabolic pathways among significantly regulated genes are listed beside each map. The clustering of sample replicates and the separation between the time points using AutoSOME (26) is illustrated in the figure inset, in agreement the day 12 samples separate most from the other time points.

Fig. S3: Time series expression profile of selected GO categories during SGBS differentiation. The average logarithmic fold change values from 4, 8 and 12 h and days 1, 3 and 12 are displayed in color using GEDI maps (25) as in Fig. S2 from other functionally related genes for comparison. All genes in the HT12 Illumina array (**A**), or genes from the GO categories cell projection, envelope, locomotion and receptor activity (respectively **B**, **C**, **D** and **E**) having similar number of genes as Recon1 were selected to show gene expression changes. Focusing on day 12, several up- and downregulated clusters relative to 4 h can be observed, however not as prominent as observed for metabolic genes based on color intensity or the percentage of significantly differentially expressed genes (adjusted F-test p-value <0.01, absolute log2 fold change >1) indicated below the panels.

Fig. S4: MiRNA expression profiling by microarrays reveals down-regulation of several miRNA clusters during adipocyte differentiation. Total RNA samples from time points day 0, day 1, day 3 and day 12 of SGBS differentiation time series were used to profile miRNAs using miChip arrays (v.11.0) arrays (27) containing probes for all miRNAs from miRBase version 11.0. In order to identify miRNAs that could contribute to prevailing upregulation of mRNAs during adipogenesis, the analysis is focused on down-regulated miRNAs. Bar graph depicts all miRNAs that have a normalized expression signal of > 50 at time point day 0, that become early down-regulated > 1.25-fold on day 1 and day 3 of differentiation and that remain down-regulated > 1.5-fold on day 12 of differentiation. The measured expression values were median normalized and are shown relative to undifferentiated cells, value of which was set to 1 (light grey bars). Data points indicate the mean expression values of triplicate experiments and the error bars represent SD. No statistical analysis was applied due to large variation between separate array hybridizations following the median normalization. The miRNA clusters with multiple downregulated mature miRNAs and early downregulation profile were selected for further analysis and are indicated with a black bar.

Fig. S5: To illustrate the separation of the measured values between the signal and noise distributions the model fits are shown for SGBS pre-adipocytes (**A**) and adipocytes (**B**). At the overlapping region, genes remain unassigned.

Fig. S6: Negative and positive control regions for ChIP-seq validation experiments. The ChIP-seq signal tracks from PPARγ studies in SGBS cells and primary adipocytes (10, 41), CEBPα and LXR from SGBS adipocytes and H3K4me3 from primary and SGBS cells comparing pre-adipocytes and adipocytes are shown from a 200 kb region centered at TSS regions of *CDH1* (negative control) (**A**) and *FABP4* (positive control) (**B**) regions. Regions with high enrichment for one or several TFs from each locus were selected for validation by ChIP-qPCR (indicated by a lined box and numbered for each locus). Each region was tested for enrichment using antibodies against all three TFs and IgG as a control as is shown in adjacent plots for the regions indicated on the ChIP-seq tracks. The enrichment values are shown relative to the enrichment of IgG and indicate the mean enrichment values of triplicate experiments and the error bars represent SEM. One sample t-test was performed to determine the significance of TF enrichment compared to IgG (*, $p < 0.05$; **, $p < 0.01$).

Fig. S7: Feed-forward loops based on ChIP-seq data. The regulatory connections to the SREBF1 locus as identified from high-occupancy ChIP-seq regions and qPCR validation experiments are shown. The arrows represent the directionality of regulation (regulated by), and the sign of regulation is indicated if inferred from data (shown here only for LXR based on the microarray data).

Fig. S8: Integrated metabolic pathways of fatty acid oxidation and activation. The fatty acid oxidation (A) and activation (B) pathways from Recon1 (2) are shown. The metanodes composition and edge color are identical to those in Figure 4. **A**) Fatty acids are broken down in the mitochondria to acetyl-CoA and a two-carbons shorter acyl-CoA, through β-oxidation. The figure represents the fatty acid oxidation in a simplified manner, where each fatty acyl-CoA is directly or via octanoyl-CoA (occoa[m]) oxidized to acetyl-CoA (accoa[m]). The pathway is largely predicted to shift to active in adipocytes (red edges). Two genes are controlling these reactions, *ACADS* and *ACADM*, that encode acyl-CoA dehydrogenases for short and medium chain fatty acids, respectively. *ACADM* is associated to PPARγ, CEBPα and miR-222. **B**) The fatty acid activation is shown. Fatty acids need to be esterified to coenzyme A (CoA) in order to be metabolically processed (oxidative degradation, elongation into complex lipids or attached to proteins as lipid anchors), catalyzed by fatty acyl CoA synthetases (ACSs). The pathway is

largely predicted to be activated in adipocytes (red edges) and *ACSL1* is among the top genes associated with high-occupancy binding sites for all three TFs.

Fig. S9: Integrated metabolic pathway of branched chain amino acid metabolism with HUVEC data. The BCAA metabolism pathway as in Figure 6 is shown for HUVEC data. Association with 8 or more TFs is highlighted and these nodes include five of the transporters including *SLC7A5* and genes from upstream reactions catalyzed by *BCAT1* and the branched-chain α-keto acid dehydrogenase complex that overlap highly regulated nodes in SGBS.

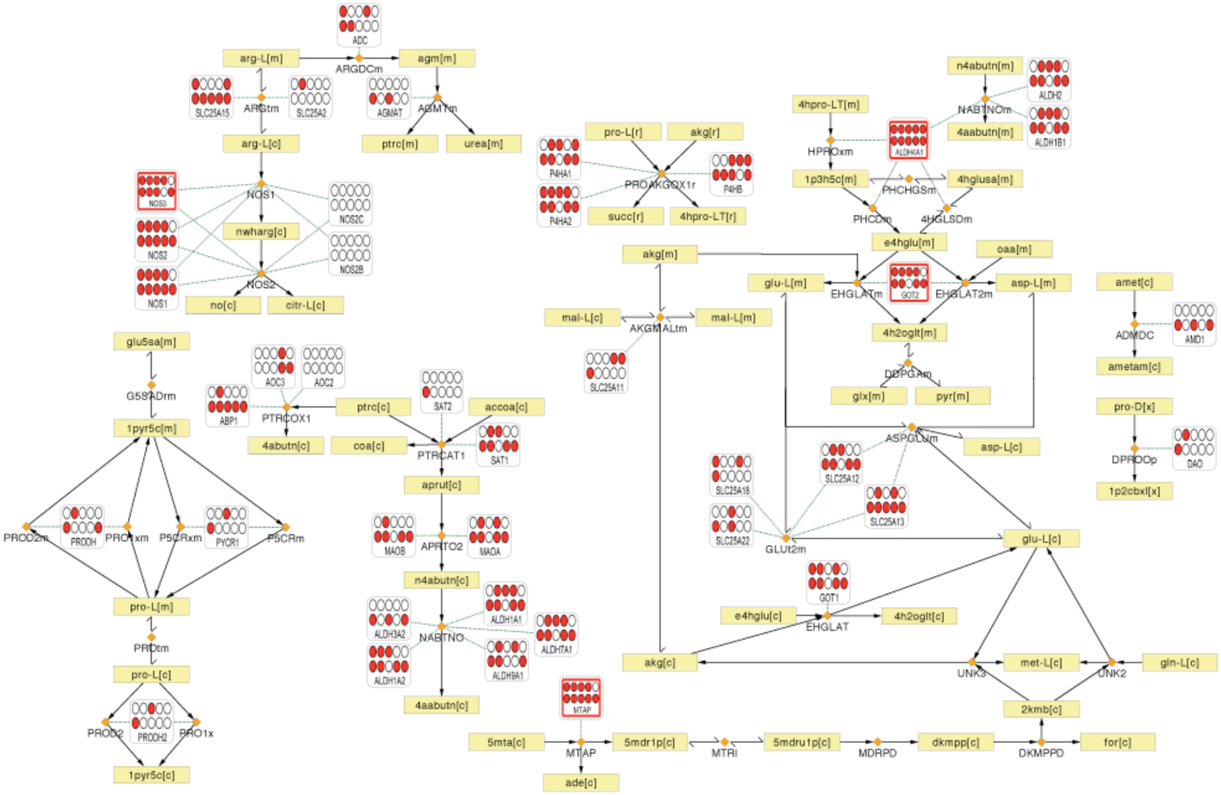
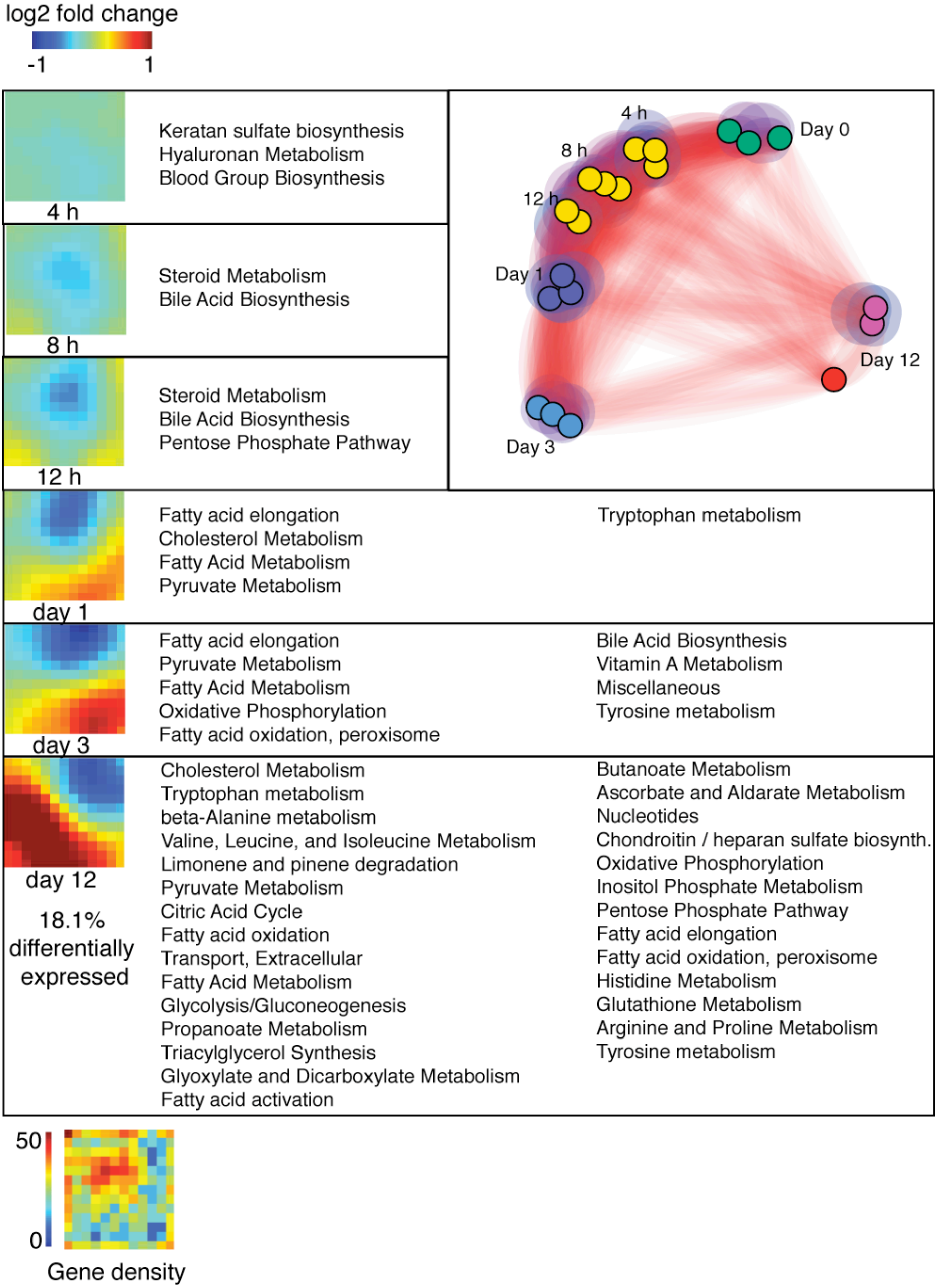

ADC
arg-L[m]
ARGDCm
agm[m]
SLC25A15
ARGtm
SLC25A2
AGMAT
AGMTm
arg-L[c]
ptrc[m]
urea[m]
NOS3
NOS1
NOS2C
NOS2
nwharg[c]
NOS2B
NOS1
NOS2
no[c]
citr-L[c]
glu5sa[m]
G5SADrm
1pyr5c[m]
AOC3
AOC2
ABP1
PTRCOX1
ptrc[c]
SAT2
accoa[c]
4abutn[c]
coa[c]
PTRCAT1
SAT1
aprut[c]
PROD2m
PRODH
PRO1xm
P5CRxm
PYCR1
P5CRm
MAOB
APRTO2
MAOA
pro-L[m]
n4abutn[c]
ALDH3A2
ALDH1A1
ALDH7A1
NABTNO
ALDH1A2
ALDH9A1
PROtm
pro-L[c]
4aabutn[c]
PROD2
PRODH2
PRO1x
1pyr5c[c]
P4HA1
pro-L[r]
akg[r]
P4HB
PROAKGOX1r
P4HA2
succ[r]
4hpro-LT[r]
n4abutn[m]
ALDH2
4hpro-LT[m]
NABTNOm
ALDH4A1
HPROxm
4aabutn[m]
ALDH1B1
1p3h5c[m]
PHCHGSm
4hglusa[m]
PHCDm
4HGLSDm
e4hglu[m]
akg[m]
oaa[m]
glu-L[m]
EHGLATm
GOT2
EHGLAT2m
asp-L[m]
amet[c]
mal-L[c]
AKGMALtm
mal-L[m]
ADMDC
AMD1
SLC25A11
4h2oglt[m]
ametam[c]
DDPGAm
glx[m]
pyr[m]
pro-D[x]
ASPGLUm
asp-L[c]
DPROOp
DAO
SLC25A18
SLC25A12
1p2cbxl[x]
SLC25A13
SLC25A22
GLUt2m
glu-L[c]
GOT1
e4hglu[c]
EHGLAT
4h2oglt[c]
akg[c]
UNK3
met-L[c]
UNK2
gln-L[c]
MTAP
2kmb[c]
5mta[c]
MTAP
5mdr1p[c]
MTRI
5mdru1p[c]
MDRPD
dkmpp[c]
DKMPPD
for[c]
ade[c]

Fig. S1



4 h	Keratan sulfate biosynthesis Hyaluronan Metabolism Blood Group Biosynthesis	
8 h	Steroid Metabolism Bile Acid Biosynthesis	
12 h	Steroid Metabolism Bile Acid Biosynthesis Pentose Phosphate Pathway	
day 1	Fatty acid elongation Cholesterol Metabolism Fatty Acid Metabolism Pyruvate Metabolism	Tryptophan metabolism
day 3	Fatty acid elongation Pyruvate Metabolism Fatty Acid Metabolism Oxidative Phosphorylation Fatty acid oxidation, peroxisome	Bile Acid Biosynthesis Vitamin A Metabolism Miscellaneous Tyrosine metabolism
day 12 18.1% differentially expressed	Cholesterol Metabolism Tryptophan metabolism beta-Alanine metabolism Valine, Leucine, and Isoleucine Metabolism Limonene and pinene degradation Pyruvate Metabolism Citric Acid Cycle Fatty acid oxidation Transport, Extracellular Fatty Acid Metabolism Glycolysis/Gluconeogenesis Propanoate Metabolism Triacylglycerol Synthesis Glyoxylate and Dicarboxylate Metabolism Fatty acid activation	Butanoate Metabolism Ascorbate and Aldarate Metabolism Nucleotides Chondroitin / heparan sulfate biosynth. Oxidative Phosphorylation Inositol Phosphate Metabolism Pentose Phosphate Pathway Fatty acid elongation Fatty acid oxidation, peroxisome Histidine Metabolism Glutathione Metabolism Arginine and Proline Metabolism Tyrosine metabolism

Fig. S2

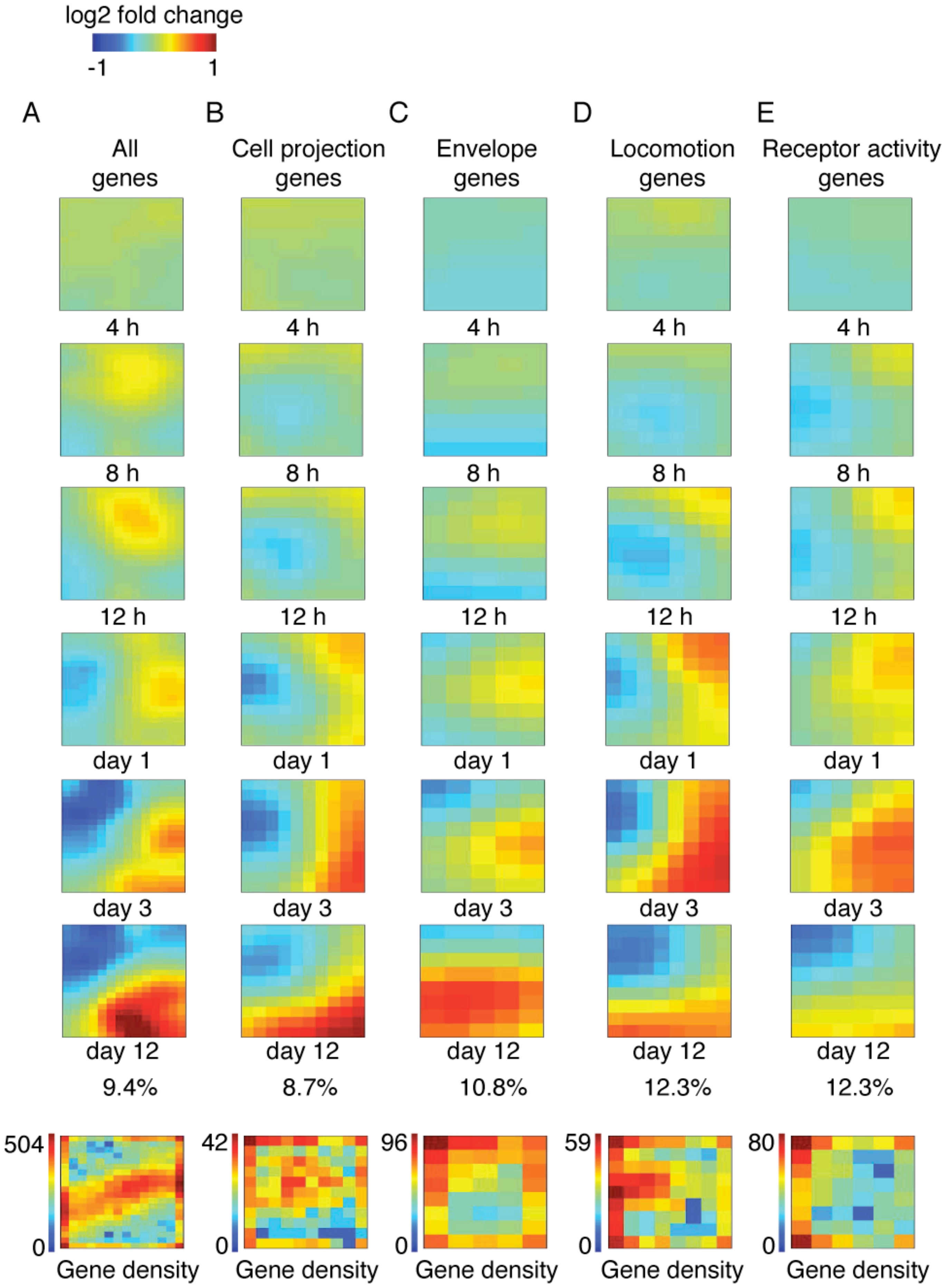

log2 fold change
-1
1
A
B
C
D
E
All genes
Cell projection genes
Envelope genes
Locomotion genes
Receptor activity genes
4 h
8 h
12 h
day 1
day 3
day 12
9.4%
8.7%
10.8%
12.3%
12.3%
504
0
42
0
96
0
59
0
80
0
Gene density

Fig. S3

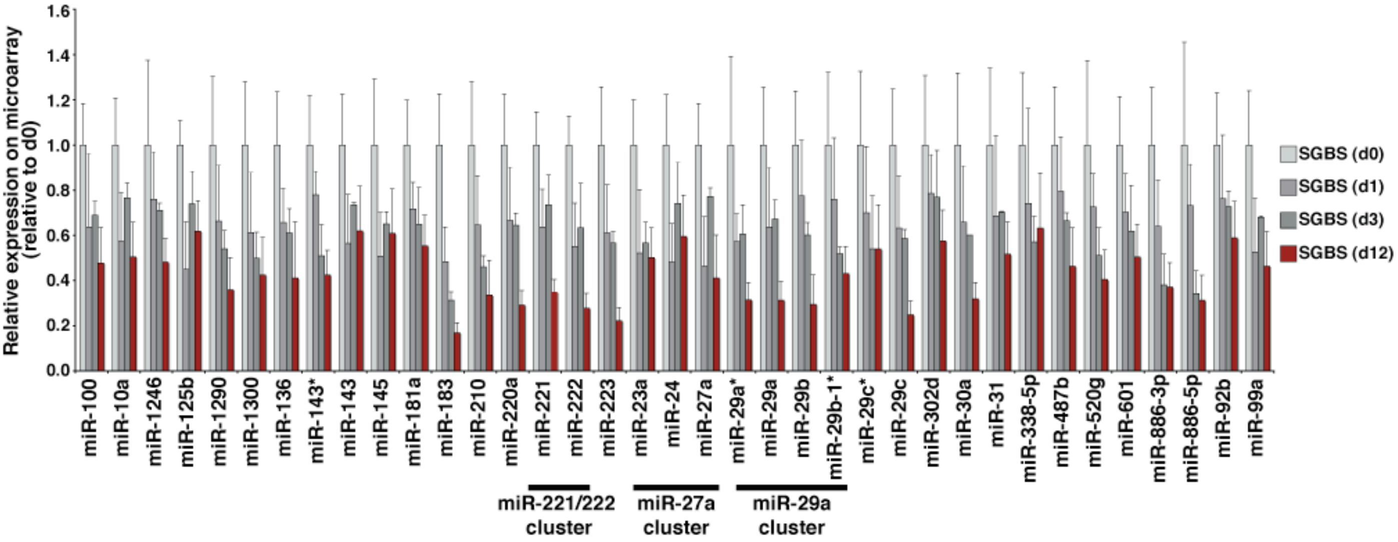

Relative expression on microarray
(relative to d0)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
miR-100
miR-10a
miR-1246
miR-125b
miR-1290
miR-1300
miR-136
miR-143*
miR-143
miR-145
miR-181a
miR-183
miR-210
miR-220a
miR-221
miR-222
miR-223
miR-23a
miR-24
miR-27a
miR-29a*
miR-29a
miR-29b
miR-29b-1*
miR-29c*
miR-29c
miR-302d
miR-30a
miR-31
miR-338-5p
miR-487b
miR-520g
miR-601
miR-886-3p
miR-886-5p
miR-92b
miR-99a
miR-221/222 cluster
miR-27a cluster
miR-29a cluster
SGBS (d0)
SGBS (d1)
SGBS (d3)
SGBS (d12)

Fig. S4

Mixture model distributions for log2 score from H3K4me3 ChIP-seq

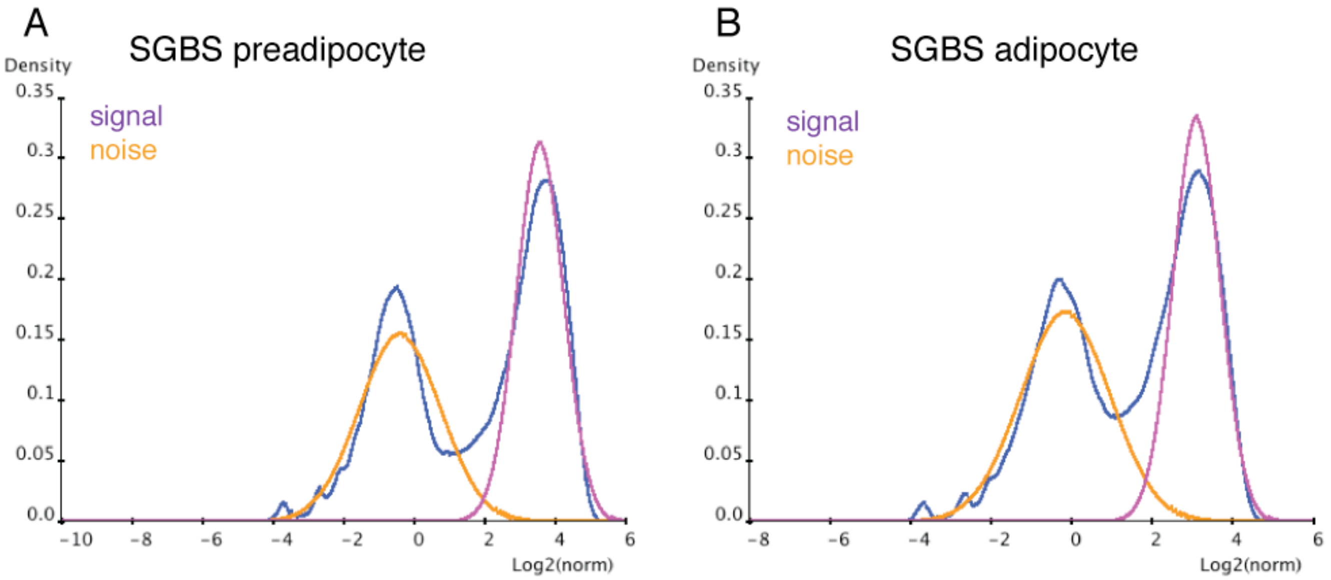


Fig. S5

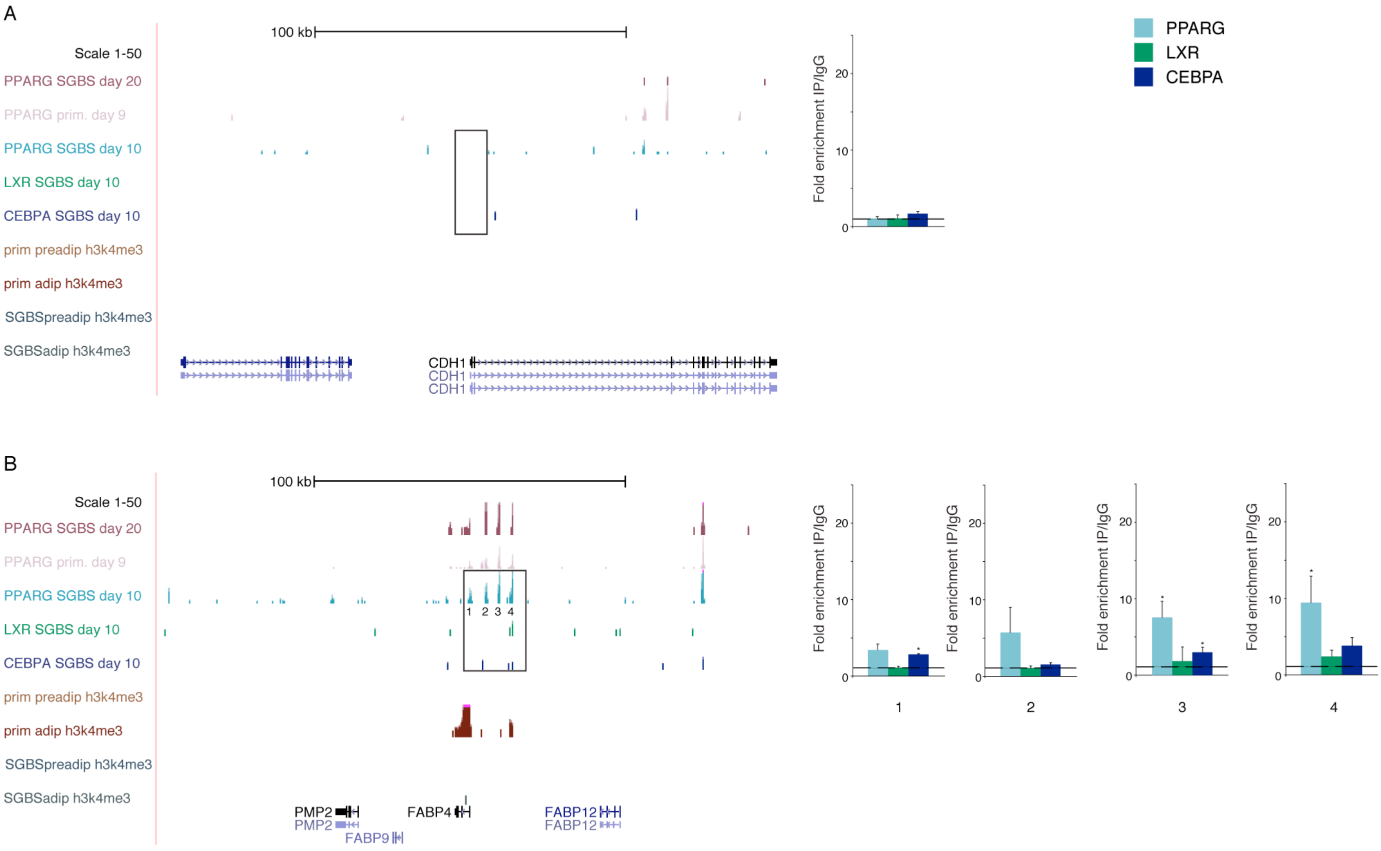

A
B
100 kb
Scale 1-50
PPARG SGBS day 20
PPARG prim. day 9
PPARG SGBS day 10
LXR SGBS day 10
CEBPA SGBS day 10
prim preadip h3k4me3
prim adip h3k4me3
SGBSpreadip h3k4me3
SGBSadip h3k4me3
CDH1
CDH1
CDH1
PMP2
PMP2
FABP9
FABP4
FABP12
FABP12
1
2
3
4
Fold enrichment IP/IgG
0
10
20
PPARG
LXR
CEBPA

Fig. S6

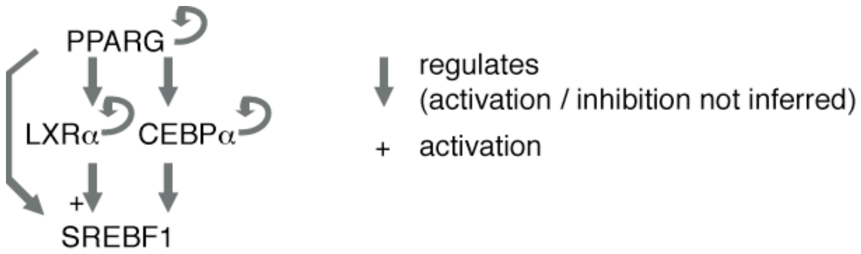

PPARG
LXRα
CEBPα
+
SREBF1
regulates
(activation / inhibition not inferred)
+
activation

Fig. S7

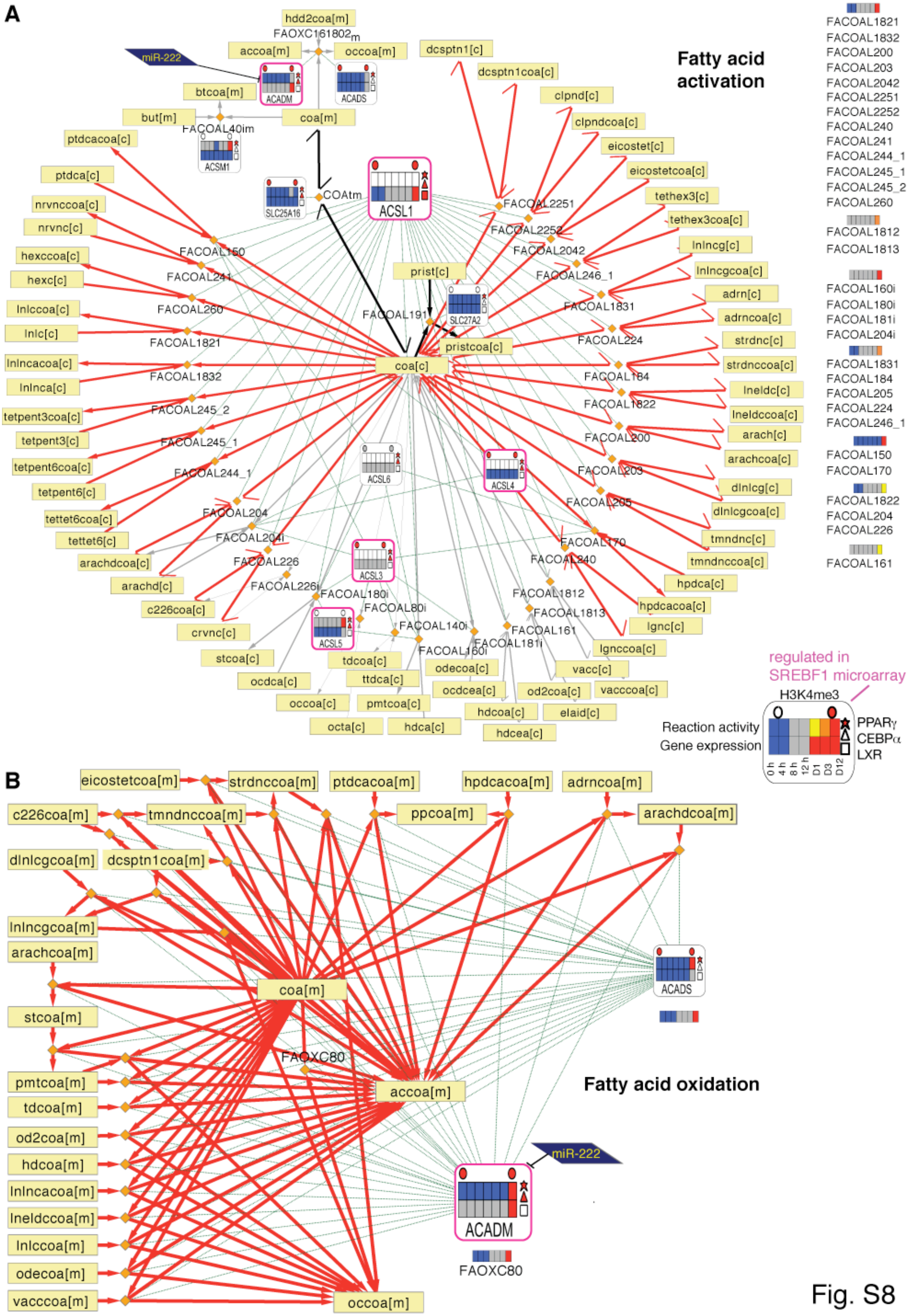

A
Fatty acid activation
hdd2coa[m]
FAOXC161802m
accoa[m]
occoa[m]
miR-222
ACADM
ACADS
btcoa[m]
but[m]
coa[m]
FACOAL40im
ACSM1
ptdcacoa[c]
ptdca[c]
nrvnccoa[c]
nrvnc[c]
hexccoa[c]
hexc[c]
lnlccoa[c]
lnlc[c]
lnlncacoa[c]
lnlnca[c]
tetpent3coa[c]
tetpent3[c]
tetpent6coa[c]
tetpent6[c]
tettet6coa[c]
tettet6[c]
arachdcoa[c]
arachd[c]
c226coa[c]
crvnc[c]
stcoa[c]
ocdca[c]
occoa[c]
octa[c]
SLC25A16
COAtm
ACSL1
FACOAL150
FACOAL241
FACOAL260
FACOAL1821
FACOAL1832
FACOAL245_2
FACOAL245_1
FACOAL244_1
FACOAL204
FACOAL204i
FACOAL226
FACOAL226i
FACOAL180i
FACOAL80i
FACOAL140i
FACOAL160i
FACOAL181i
FACOAL161
FACOAL1812
FACOAL1813
FACOAL240
FACOAL170
FACOAL205
FACOAL203
FACOAL200
FACOAL1822
FACOAL184
FACOAL224
FACOAL1831
FACOAL246_1
FACOAL2042
FACOAL2252
FACOAL2251
FACOAL191
prist[c]
pristcoa[c]
SLC27A2
coa[c]
ACSL6
ACSL4
ACSL3
ACSL5
dcsptn1[c]
dcsptn1coa[c]
clpnd[c]
clpndcoa[c]
eicostet[c]
eicostetcoa[c]
tethex3[c]
tethex3coa[c]
lnlncg[c]
lnlncgcoa[c]
adrn[c]
adrncoa[c]
strdnc[c]
strdnccoa[c]
lneldc[c]
lneldccoa[c]
arach[c]
arachcoa[c]
dlnlcg[c]
dlnlcgcoa[c]
tmndnc[c]
tmndnccoa[c]
hpdca[c]
hpdcacoa[c]
lgnc[c]
lgnccoa[c]
vacc[c]
vacccoa[c]
elaid[c]
od2coa[c]
hdceа[c]
hdcoa[c]
ocdcea[c]
odecoa[c]
ttdca[c]
tdcoa[c]
pmtcoa[c]
hdca[c]
FACOAL1821
FACOAL1832
FACOAL200
FACOAL203
FACOAL2042
FACOAL2251
FACOAL2252
FACOAL240
FACOAL241
FACOAL244_1
FACOAL245_1
FACOAL245_2
FACOAL260
FACOAL1812
FACOAL1813
FACOAL160i
FACOAL180i
FACOAL181i
FACOAL204i
FACOAL1831
FACOAL184
FACOAL205
FACOAL224
FACOAL246_1
FACOAL150
FACOAL170
FACOAL1822
FACOAL204
FACOAL226
FACOAL161
regulated in
SREBF1 microarray
H3K4me3
Reaction activity
Gene expression
PPARγ
CEBPα
LXR
0 h
4 h
8 h
12 h
D1
D3
D12
B
eicostetcoa[m]
strdnccoa[m]
ptdcacoa[m]
hpdcacoa[m]
adrncoa[m]
c226coa[m]
tmndnccoa[m]
ppcoa[m]
arachdcoa[m]
dlnlcgcoa[m]
dcsptn1coa[m]
lnlncgcoa[m]
arachcoa[m]
coa[m]
ACADS
stcoa[m]
FAOXC80
pmtcoa[m]
accoa[m]
Fatty acid oxidation
tdcoa[m]
od2coa[m]
hdcoa[m]
miR-222
lnlncacoa[m]
lneldccoa[m]
ACADM
lnlccoa[m]
odecoa[m]
FAOXC80
vacccoa[m]
occoa[m]

Fig. S8

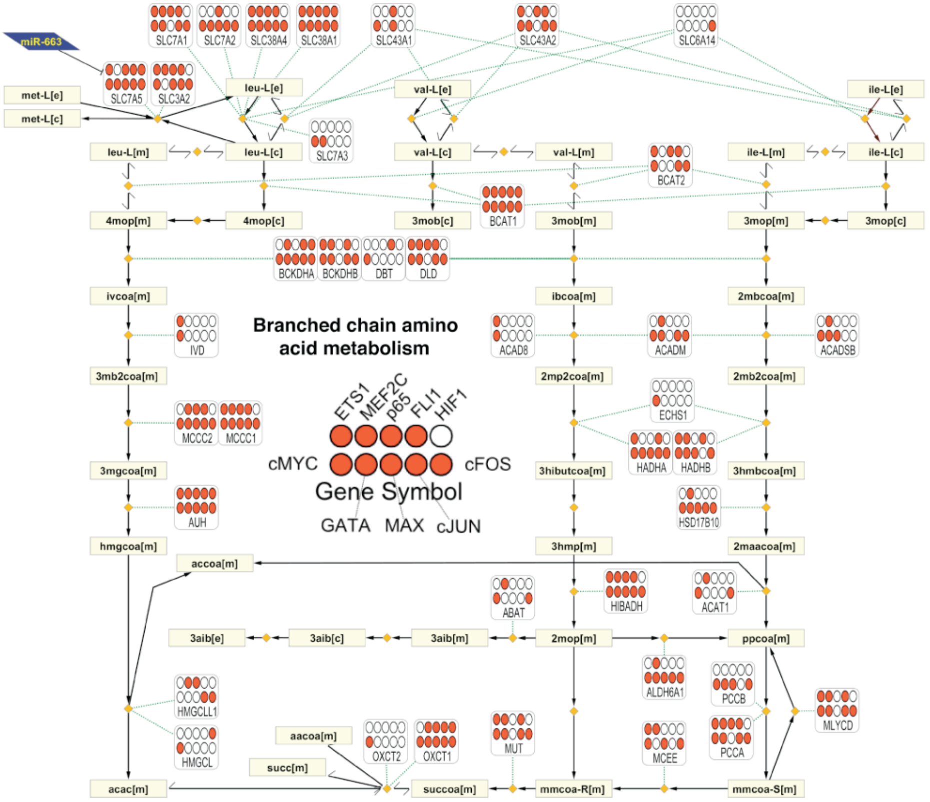
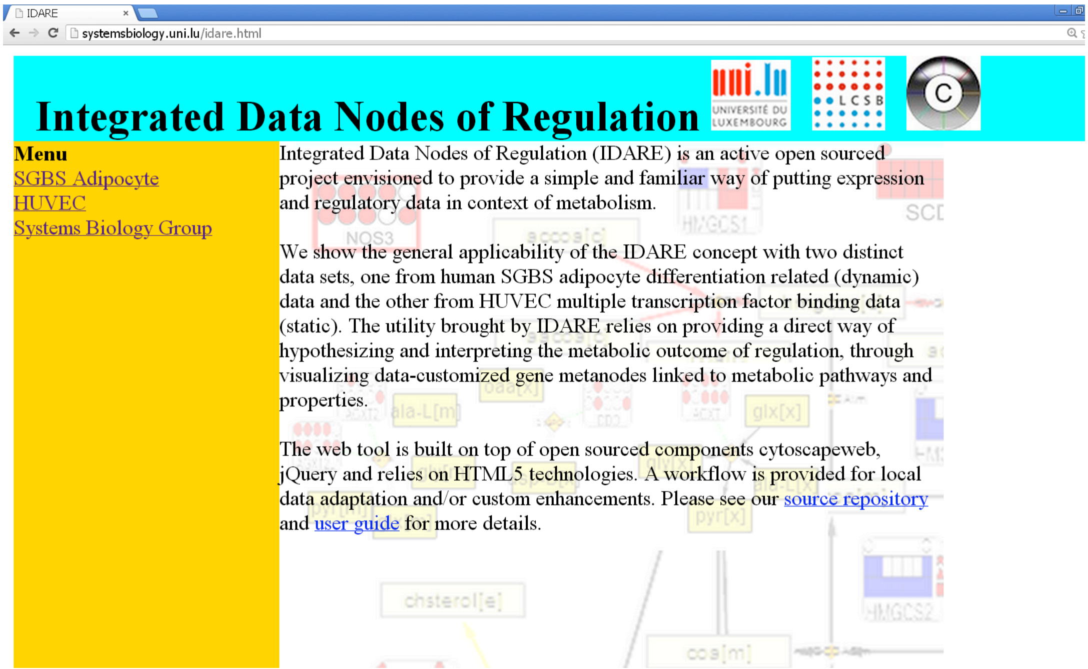

miR-663
SLC7A1
SLC7A2
SLC38A4
SLC38A1
SLC43A1
SLC43A2
SLC6A14
SLC7A5
SLC3A2
SLC7A3
met-L[e]
met-L[c]
leu-L[e]
leu-L[m]
leu-L[c]
val-L[e]
val-L[c]
val-L[m]
ile-L[e]
ile-L[m]
ile-L[c]
BCAT2
BCAT1
4mop[m]
4mop[c]
3mob[c]
3mob[m]
3mop[m]
3mop[c]
BCKDHA
BCKDHB
DBT
DLD
ivcoa[m]
ibcoa[m]
2mbcoa[m]
IVD
Branched chain amino acid metabolism
ACAD8
ACADM
ACADSB
3mb2coa[m]
2mp2coa[m]
2mb2coa[m]
ETS1
MEF2C
p65
FLI1
HIF1
cMYC
cFOS
Gene Symbol
GATA
MAX
cJUN
ECHS1
MCCC2
MCCC1
HADHA
HADHB
3mgcoa[m]
3hibutcoa[m]
3hmbcoa[m]
AUH
HSD17B10
hmgcoa[m]
3hmp[m]
2maacoa[m]
accoa[m]
HIBADH
ABAT
ACAT1
3aib[e]
3aib[c]
3aib[m]
2mop[m]
ppcoa[m]
HMGCLL1
ALDH6A1
PCCB
MLYCD
HMGCL
aacoa[m]
OXCT2
OXCT1
MUT
MCEE
PCCA
succ[m]
acac[m]
succoa[m]
mmcoa-R[m]
mmcoa-S[m]

Fig. S9

IDARE (Integrated DAta nodes or REgulation)

Visualizing regulatory data in context of metabolic pathways

http://systemsbiology.uni.lu/idare.html
http://systemsbiology.uni.lu/adipoflux.html
http://systemsbiology.uni.lu/huvec.html



Integrated Data Nodes of Regulation

Menu
SGBS Adipocyte
HUVEC
Systems Biology Group

Integrated Data Nodes of Regulation (IDARE) is an active open sourced project envisioned to provide a simple and familiar way of putting expression and regulatory data in context of metabolism.

We show the general applicability of the IDARE concept with two distinct data sets, one from human SGBS adipocyte differentiation related (dynamic) data and the other from HUVEC multiple transcription factor binding data (static). The utility brought by IDARE relies on providing a direct way of hypothesizing and interpreting the metabolic outcome of regulation, through visualizing data-customized gene metanodes linked to metabolic pathways and properties.

The web tool is built on top of open sourced components cytoscapeweb, jQuery and relies on HTML5 technologies. A workflow is provided for local data adaptation and/or custom enhancements. Please see our source repository and user guide for more details.

User guide

IDARE relies in two separate components:
1) The generation of gene ‘metanode’ image files on Matlab®;
2) Web interactivity through open sourced Html and CytoscapeWeb.

Web address: http://systemsbiology.uni.lu/IDARE
Contact: mafalda.galhardo@uni.lu and jlin@systemsbiology.org
Updated: June 4th, 2013

Table of content:

1. General purpose

IDARE was envisioned to provide a simple and familiar way of showing expression and regulatory data in context of metabolism. Using metabolic maps it provides easy links to biochemical knowledge and extends from current representations by introducing gene metanodes in association to the metabolic pathways.

We show the general applicability of the **IDARE** concept with two distinct data sets, one from HUVEC multiple transcription factor binding data (static) and the other from human SGBS adipocyte differentiation related (dynamic) data.

The utility brought by **IDARE** relies on providing a direct way of hypothesizing and interpreting the metabolic outcome of regulation, through visualizing data-customized gene metanodes linked to metabolic pathways and properties.

2. Available metabolic pathways and datasets

Currently, two datasets are available for exploring with **IDARE**:

1) **AdipoFlux** : human SGBS adipocyte differentiation dynamic data;
2) **Huvec**: human endothelial cell static multiple transcription factor binding data.

The basis for **IDARE** pathway representations is Recon1 (Duarte et al., 2007), a general human metabolic network reconstruction containing metabolic reactions (3742) and associated enzymes, genes and metabolites.

All Recon1 metabolic pathways are available, of which 5 were manually laid out due to their relevance in context of the adipocyte dataset (http://systemsbiology.uni.lu/adipoflux.html) we first analyzed:

- Cholesterol metabolism;
- Fatty acid activation;
- Fatty acid oxidation;
- Triacylglycerol synthesis;
- Valine, leucine and isoleucine metabolism.

These pathways were initially selected based on highest predicted metabolic activity difference between pre-adipocyte and adipocyte stages, as supported by our data and analysis. The networks were manually arranged on Cytoscape and saved as xgmml files containing xy node coordinates. The Cobra toolbox for Matlab® was used to extract network files from the Recon1 model that were imported into Cytoscape (sif).

Additionally, the Arginine and Proline metabolism pathway has been manually laid out in context of the HUVEC dataset (http://systemsbiology.uni.lu/huvec.html).

3. IDARE high level components

IDARE relies in two separate components:

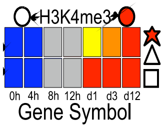
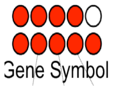
1) Generation of gene ‘metanode’ image files on Matlab®;

2) Web interactivity (Html and Cytoscape Web).

1) Generation of gene 'metanode' image files on Matlab®

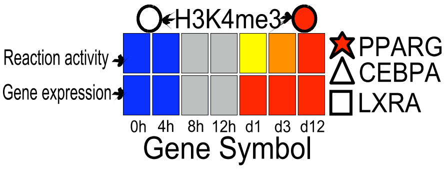
IDARE is currently supporting 2 metanode types, customized in context of SGBS adipocyte differentiation data (CASE 1) and HUVEC transcription factor data (CASE 2).

Summary table of metanode types and properties:

Metanode class	Dataset	Type	Input data	Icon
Heterogeneous	SGBS	dynamic (7 time points)	discrete gene expression (-1, 0, 1)	
			reaction activity prediction (-1, 0, 1, 2, 3)	
			TF putative binding of 3 TFs (gene list)	
			H3K4me3 presence or absence (0 or 1) on 2 time points	
Homogenous	HUVEC	static	10 TFs putative binding (0 or 1)	

CASE 1: SGBS data – dynamic gene expression, reaction activity predictions and few regulators.

We exemplify this metanode type with the SGBS adipocyte differentiation dataset.
Based in our dataset properties, we defined a gene metanode as containing:



- one lower line with discrete gene expression;
- one upper line with reaction activity prediction based on the gene expression;

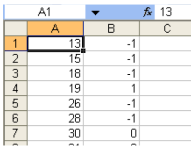
(both lines contain slots representing seven adipocyte differentiation time points)

- three polygons on the right side of the lines, representing the putative binding of three transcription factors (TFs: PPARγ, CEBPα, LXR);
- two circles on top of the lines, aligned according to the time point they belong to, which represent the presence of a histone modification mark associated with active transcription start sites (TSSs) – H3K4me3.

Matlab®: using mainly Entrez gene IDs for mappings, reads in data files, collects arrays for each data type and associates them to each metanode component. For each gene, colors the metanode based on those component arrays.

The following input data files were used (tabular text or excel files):

i. Discrete gene expression data for coloring the bottom line:
   - 1st column – Recon1 Entrez gene IDs;
   - Remaining columns – discrete gene expression values:
     - -1 – lowly expressed gene;



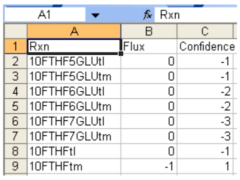
A1 | 13

	A	B	C
1	13	-1	
2	15	-1	
3	18	-1	
4	19	1	
5	26	-1	
6	28	-1	
7	30	0	

- 0 – moderately expressed gene;
- 1 – highly expressed gene.

Matlab® : first step is to build empty (white) rectangles, as many as time points (7), on defined positions (x,y) that are colored based on the data from correspondent time point.

ii. Reaction activity prediction for coloring the top line:

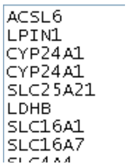


	A	B	C
1	Rxn	Flux	Confidence
2	10FTHF5GLUtl	0	-1
3	10FTHF5GLUtm	0	-1
4	10FTHF6GLUtl	0	-2
5	10FTHF6GLUtm	0	-2
6	10FTHF7GLUtl	0	-3
7	10FTHF7GLUtm	0	-3
8	10FTHFtl	0	-1
9	10FTHFtm	-1	1

- ○ 1st column – Recon1 reaction abbreviation;
- ○ Remaining columns – reaction prediction results (including confidence):
- 0 – inactive;
- 1 – active (direct way);
- -1 – active (reverse way);
- 2 – active (unknown direction);
- 3 – undetermined.

Matlab® : second step is to build empty (white) rectangles, as many as time points (7), on defined positions (x,y) on top of the gene expression line; each rectangle is colored based on the data from correspondent time point.

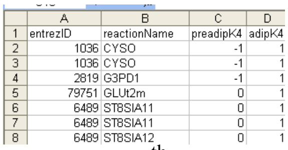
iii. Transcription factor Recon1 associated genes (list of gene symbols):



ACSL6
LPIN1
CYP24A1
CYP24A1
SLC25A21
LDHB
SLC16A1
SLC16A7
SLC4A4

Matlab® : third step is to build empty (white) polygons, one for each TF (3), on defined positions (x,y) on the right side of the expression line; from a discrete array (0 or 1) colors polygons in red (1) when gene is associated with a TF.

iv. H3K4me3 data for each gene in Recon1:



	A	B	C	D
1	entrezID	reactionName	preadipK4	adipK4
2	1036	CYSO	-1	1
3	1036	CYSO	-1	1
4	2819	G3PD1	-1	1
5	79751	GLUt2m	0	1
6	6489	ST8SIA11	0	1
7	6489	ST8SIA11	0	1
8	6489	ST8SIA12	0	1

- ○ 1st and 2nd columns: IDs for mapping (Recon1 Entrez gene IDs and reaction abbreviations);
- ○ Remaining columns: discrete values for the presence or absence of the histone mark.

Matlab® : 4th step is to build empty (white) circles, on defined positions (x,y) on top of the reaction activity prediction line, aligned accordingly to the time point they represent; from reading the H3K4me3 data file, forms a discrete array (-1, 0, 1) and colors circles in red (1) or grey (0) accordingly.

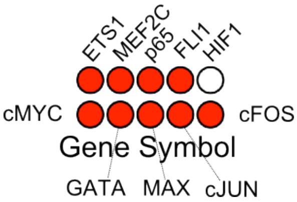
Metanodes are generated per pathway and within each pathway, per gene.
In cases when one gene is associated with multiple reactions, the following occurs:

- generate one metanode with white reaction activity prediction line (defines multiple reactions associated to that gene);
- for each reaction, plots that individual reaction prediction line which is shown on the left side panel when clicking on the gene metanode.

<u>CASE 2: HUVEC data – regulator data only (TFs, static metanode).</u>

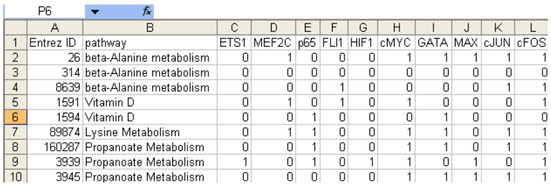
We exemplify this metanode type with the HUVEC transcription factor dataset.

Based in the HUVEC dataset properties, we defined a gene metanode as:



- two lines of circles (5 each) representing the putative binding of a total of 10 TFs:
    - Bottom line – cMYC, GATA, MAX, cJUN and cFOS (ENCODE data);
    - Top line – ETS1, MEF2C, p65, FLI1 and HIF1 (own data).
- Circles are filled in red when data supports the binding of correspondent TF to current gene.

Example input data file (tabular text or excel):



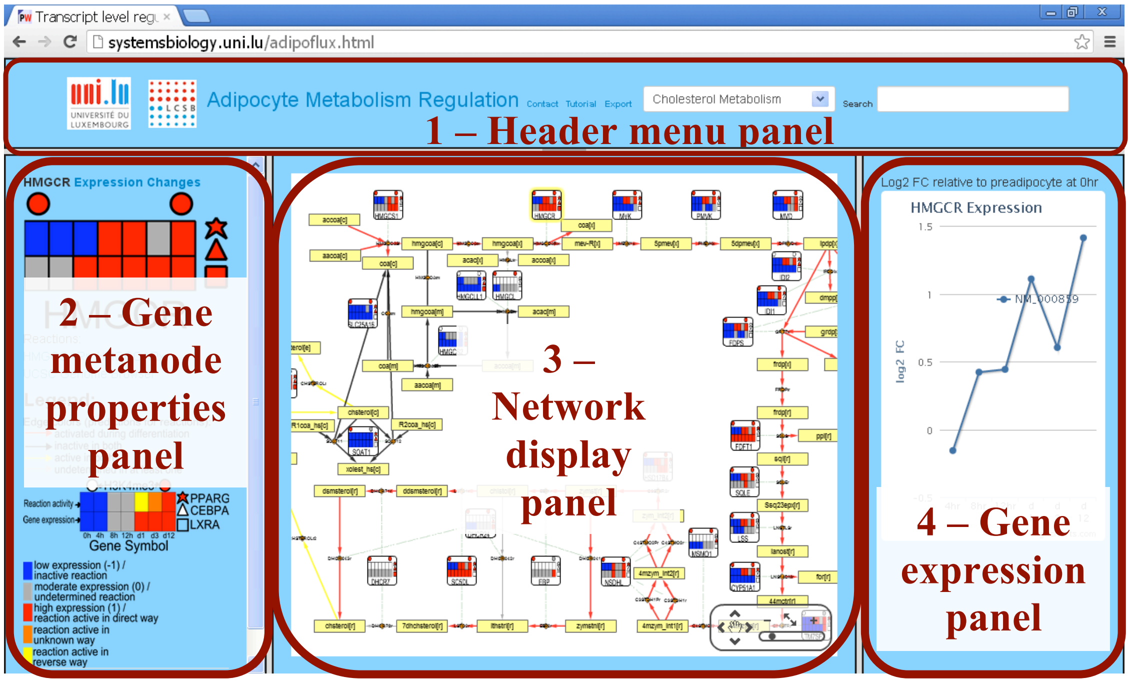
P6

	A	B	C	D	E	F	G	H	I	J	K	L
1	Entrez ID	pathway	ETS1	MEF2C	p65	FLI1	HIF1	cMYC	GATA	MAX	cJUN	cFOS
2	26	beta-Alanine metabolism	0	1	0	0	0	1	1	1	1	1
3	314	beta-Alanine metabolism	0	0	0	0	0	0	0	0	0	0
4	8639	beta-Alanine metabolism	0	0	0	1	0	0	0	0	1	1
5	1591	Vitamin D	0	1	0	1	0	1	0	0	0	1
6	1594	Vitamin D	0	0	1	0	0	0	1	0	0	0
7	89874	Lysine Metabolism	0	1	1	0	0	1	1	0	1	1
8	160287	Propanoate Metabolism	0	0	1	0	0	1	1	0	1	1
9	3939	Propanoate Metabolism	1	0	1	0	1	1	0	1	0	1
10	3945	Propanoate Metabolism	0	0	0	0	0	1	1	1	1	1

Matlab®: using Entrez gene IDs and pathways for mappings, reads in data file with discrete values for the presence or absence (1, 0) of the putative binding of a TF on a gene; draws ten white circles on defined coordinates (x,y), and colors them red when finding a data point "1".

4. IDARE display

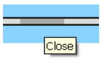
On the following, we exemplify **IDARE** details using as example the adipocyte dataset first analyzed (AdipoFlux instance). The same general characteristics apply to the Huvec dataset.


Transcript level regu ×
systemsbiology.uni.lu/adipoflux.html
uni.lu UNIVERSITÉ DU LUXEMBOURG
LCSB
Adipocyte Metabolism Regulation
Contact
Tutorial
Export
Cholesterol Metabolism
Search
1 – Header menu panel
HMGCR Expression Changes
2 – Gene metanode properties panel
Reaction activity
Gene expression
PPARG
CEBPA
LXRA
0h 4h 8h 12h d1 d3 d12
Gene Symbol
low expression (-1) / inactive reaction
moderate expression (0) / undetermined reaction
high expression (1) / reaction active in direct way
reaction active in unknown way
reaction active in reverse way
3 – Network display panel
Log2 FC relative to preadipocyte at 0hr
HMGCR Expression
1.5
1
0.5
0
log2 FC
NM_000859
4 – Gene expression panel

General view of a **IDARE** instance (AdipoFlux):

Panel overview

1 – **Header menu panel**: here the user can select which metabolic pathway to display as well as access the search function. Links to the associated publication, network export, this user guide and our group's homepage are also available.
2 – **Left side panel** – Gene metanode properties: opens by clicking on a gene metanode and displays details.
The first line on the panel shows the selected gene symbol and a link ('Expression Changes') to the right side panel where gene expression values are plotted.
On the second line, 'UCSC Genome browser' link opens a pop-up 'Adipocyte ChipSeq UCSC Genome Tracks' with a link to a UCSC Genome Browser Adipoflux hub containing ChIP-Seq tracks associated to this work (TFs and H3K4me3 modification) and the selected gene's position that the user should copy and paste to the genome browser in order to visualize the tracks in the selected gene location. Please refer to Chapter 6, section 2 for more details.
A large gene metanode is shown below the two first lines followed by the 'Reactions' associated to the gene. Clicking on a reaction name opens a pop-up with reaction static details.
On the bottom of the left side panel, a legend for edge colors and metanode is provided, so that the user can keep track of what is being represented. Please refer to 'section 3' for more details.
3 – **Central panel** – CytoscapeWeb metabolic network display: this panel contains the metabolic network which can be re-arranged in accordance to user's preference and embeds click-on functions for the nodes' additional details.
4 – **Right side panel** – gene expression: dynamically plots for a selected gene the log2 FC values of each differentiation time point relative to control pre-adipocyte values (microarray data from SGBS cell differentiation time course).



Panels 1, 2 and 4 can be expanded or collapsed by clicking on the center of the grey bar next to them.

5. IDARE interactive elements and functions

This chapter is a walk through **IDARE** interactive elements and functions, using as example 'Cholesterol metabolism' pathway. The metanodes exemplified are in context of the SGBS adipocyte differentiation data. All descriptions apply too for the Huvec data, except that the metanodes have a different visual display, as previously described on Chapter 4.1, case 2.
Below we exemplify how to 'read' the cholesterol synthesis metabolic pathway, which starts with the condensation of acetyl-CoA (accoa[c]) and acetoacetyl-CoA (aacoa[c]) to form 3-hydroxy-3-methylglutaryl-CoA (hmgcoa[c]) catalyzed by the enzyme HMG-CoA synthase (gene *HMGCS1*). The end-point metabolite is cholesterol (chsterol[r][m][c][e], r – endoplasmic reticulum, m – mitochondria, c – cytosol, e - extracellular).

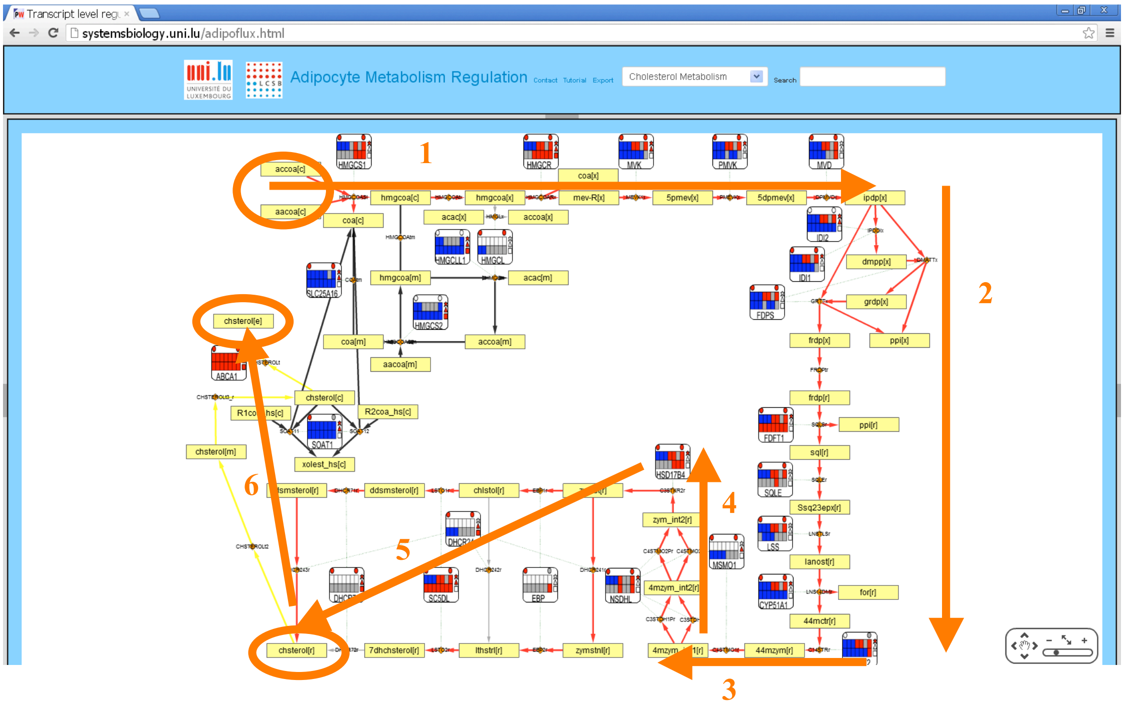


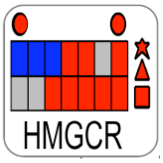
Figure 1: view of main adipoflux webpage, showing cholesterol synthesis pathway.

1. Pathway components:

a) Nodes:

accoa[c] - Metabolites (yellow boxes).

◆ - Reactions (orange diamonds), appear with reaction abbreviation (as of Recon1) on top.



- Gene metanodes representing 4 data levels: gene expression (bottom line rectangles), predicted reaction activity (upper line rectangles), TF association (right side polygons) and marker for active TSS (upper circles) – detailed metanode legend on the left side panel of the webtool and below. Gene metanodes link to reaction nodes (orange diamonds) representing gene-protein-reaction associations contained in Recon1.

miR-222 - miRNA nodes that link to target genes (gene metanodes). Data from miRs -27a, -29a and -222 are included, all the three consistently down-regulated during adipocyte differentiation.

b) Edges:

- Metabolic edges (solid lines): link substrate and product metabolites (yellow nodes) via reactions (orange nodes) that are catalyzed by enzymes.

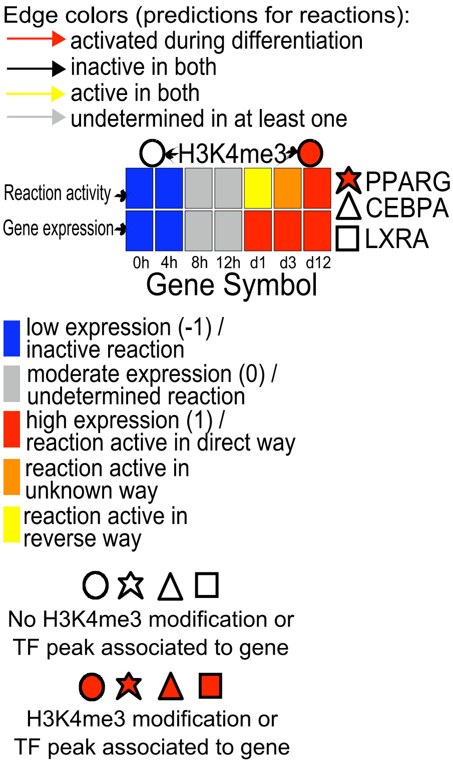
Edge color represents predicted reaction activity based on a constraint-based method (Shlomi et al., 2008), the general human metabolic model Recon1 (Duarte et al., 2007) and gene expression data from a differentiation time course experiment on human SGBS cells. See description below for each edge color.

Edge width represents prediction confidence, with thinner lines for reactions undetermined in the pre-adipocyte and/or the adipocyte stage (colored in grey) and thicker lines for confident reaction activity prediction on both stages (please refer to Shlomi's method for the concept underlying prediction confidence).

- Gene-protein-reaction (GPR) edges (dashed green lines): link metabolic reactions back to the gene(s) encoding the enzymes that catalize them; this info is contained in Recon1.

- miRNA target-inhibition edges (black solid lines in 'T' shape on target interface): link miRNAs with target genes, based on own experimental data.

c) Edge color and gene metanode legend: can be found on the left side panel from within the webtool (click to show/hide)



Edge colors (predictions for reactions):
activated during differentiation
inactive in both
active in both
undetermined in at least one
H3K4me3
Reaction activity
Gene expression
PPARG
CEBPA
LXRA
0h
4h
8h
12h
d1
d3
d12
Gene Symbol
low expression (-1) / inactive reaction
moderate expression (0) / undetermined reaction
high expression (1) / reaction active in direct way
reaction active in unknown way
reaction active in reverse way
No H3K4me3 modification or TF peak associated to gene
H3K4me3 modification or TF peak associated to gene

Red – reactions predicted inactive in pre-adipocytes and active in adipocytes.
Black – reactions predicted inactive in both pre-adipocyte and adipocyte stages.
Yellow - reactions predicted active in both pre-adipocyte and adipocyte stages.
Grey – reactions undetermined in at least one of pre-adipocyte or adipocyte stage.

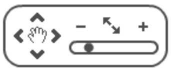
Metanode H3K4me3 panel – represents whether a tri-methylated Lisine-4 residue of Histone 3 was associated to the specific gene (red) or not (white). Left circle represents data on pre-adipocytes and right circle on adipocytes.

Metanode TF panel – represents whether at least one peak from PPARG (star), CEBPA (triangle) or LXR (square) was associated with the specific gene (red) or not (white). Peak-gene associations were obtained from the GREAT tool by providing a list of TF-peak genomic coordinates.

Metanode gene expression (bottom-line rectangles) and predicted reaction activity (upper-line rectangles) per differentiation time-point are represented by color with legend below the gene metanode icon.

2. Data interactivity and integration

Our web tool provides interactive access to the discussed five metabolic pathways combined with several *omics* data, future releases will incorporate many more pathways.



The network itself can be re-arranged by moving nodes as preferred by the user. This can be done by clicking on the 'hand' icon on the panel to

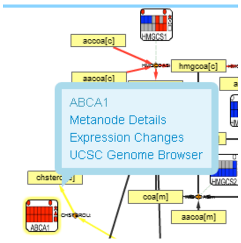
the lower right corner of the screen.
Once the ‘hand’ is selected, one can move each node to a desired position and the edges.

The network can be exported as an svg file, by hitting the link ‘export network’ on the ‘header menu panel’.

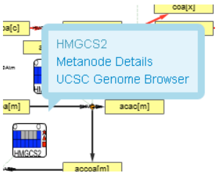
Data interactivity is provided through the network nodes, via click-in functions that open the left side panel or pop-up tool tips on mouse over events.

Mousing-over metabolite (yellow boxes), reaction (orange diamonds) or gene (white) nodes displays a callout with additional info/links.

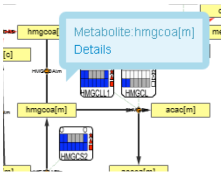
Gene mouse-over (blue callout):

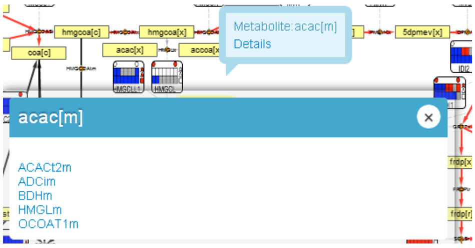


Clicking on ‘Metanode Details’ opens the left side panel, while clicking on ‘Expression changes’ opens the right side panel and dynamically plots the gene expression log2 FC of each differentiation time point relative to control pre-adipocytes. The ‘UCSC Genome Browser’ link opens a pop-up that can re-direct to UCSC GB to visualize ChIP-Seq data tracks on the gene’s position. Genes whose expression was too low (not detected on the microarray) do not contain the ‘Expression changes’ link to the right side panel (e.g. HMGCS2 on the right).
Direct click on a gene metanode opens both side panels (when expression values are available).



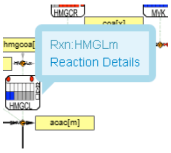
Metabolite mouse-over:

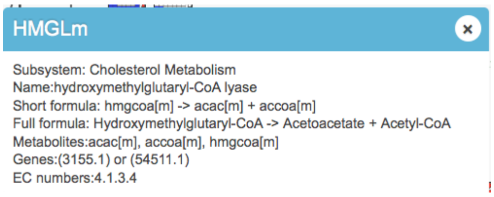




Clicking on ‘Details’ opens a pop-up containing the reactions the metabolite is associated to (either being consumed or produced). In the case of the reaction name containing ‘:’, that has been replaced with ‘_’ to conform to HTML5 Java Script Object Notation syntaxes.
Further clicking on the reaction link opens a new pop-up with reaction details (see below for reaction details description). Direct click on metabolite node opens same pop-up showing which reactions the metabolite associates with.

Reaction mouse-over:



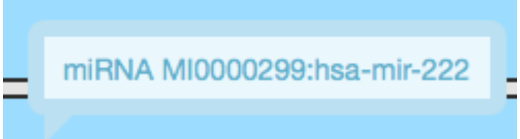

HMGLm

Subsystem: Cholesterol Metabolism
Name:hydroxymethylglutaryl-CoA lyase
Short formula: hmgcoa[m] -> acac[m] + accoa[m]
Full formula: Hydroxymethylglutaryl-CoA -> Acetoacetate + Acetyl-CoA
Metabolites:acac[m], accoa[m], hmgcoa[m]
Genes:(3155.1) or (54511.1)
EC numbers:4.1.3.4

Clicking on ‘Reaction Details’ opens a pop-up with static info on the current reaction. This info is contained in Recon1 and includes:

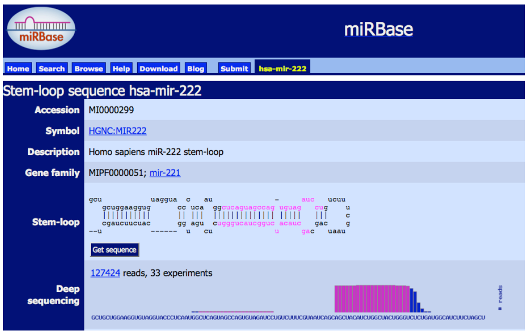
- Subsystem (Recon1 pathway the reaction belongs to);

- Name (full name of selected reaction);
- Short formula (biochemical reaction equation with metabolite abbreviations);
- Full formula (biochemical reaction equation with metabolite full names);
- Metabolites (abbreviations of the metabolites involved in the reactions);
- Genes (Recon1 gene-reaction rule – if multiple genes are associated with the reaction, shows the Boolean rules 'and'/'or' that characterize it; the notation was kept as of Recon1 and it represents as outdated version of Entrez Gene IDs; 'undefined' stands for reactions without associated genes as of Recon1 (e.g. transport reactions);
- EC numbers: Enzyme commission numbers for the enzyme(s) catalyzing the reaction, as of Recon1; 'undefined' is shown for reactions for which no EC number was available.


miRNA MI0000299:hsa-mir-222

MiR node mouse over shows the miRBase accession number and ID for the selected human microRNA (hsa-id).

On click of the miRNA node redirects to miRBase page for the specific human miRNA.



Stem-loop sequence hsa-mir-222

Accession	MI0000299
Symbol	HGNC:MIR222
Description	Homo sapiens miR-222 stem-loop
Gene family	MIPF0000051; mir-221

Stem-loop

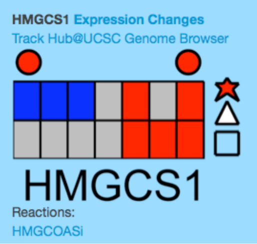
```
gcu            uaggua  c   au               -      auc    ucuu
   gcuggaaggug       cc uca  ggcucaguagccag uguag   cug    u
   |||||||||||       || |||  |||||||||||||| |||||   |||    c
   cgaucuucuac       gg agu  cugggucaucgguc acauc   gac    g
--u            ------  u   cu               u      gac    uaau
```

Deep sequencing

127424 reads, 33 experiments

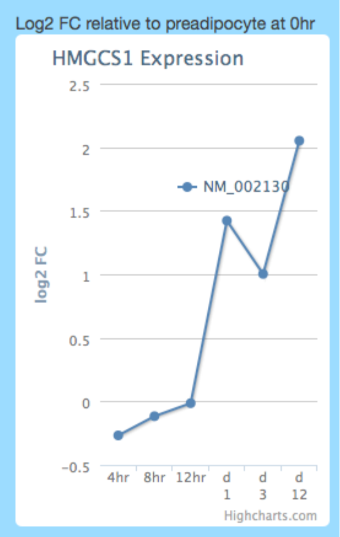
GCUGCUGGAAGGUGUAGGUACCCUCAAUGGCUCAGUAGCCAGUGUAGAUCCUGUCUUUCGUAAUCAGCAGCUACAUCUGGCUACUGGGUCUCUGAUGGCAUCUUCUAGCU

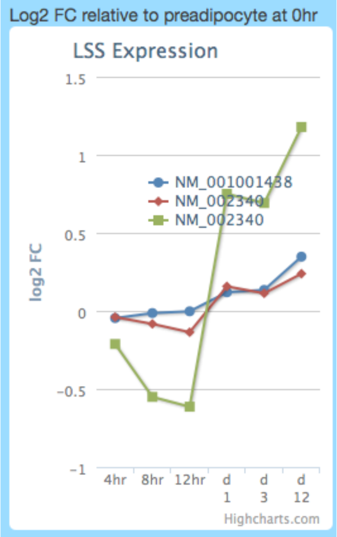
Plotting Gene Expression Changes:



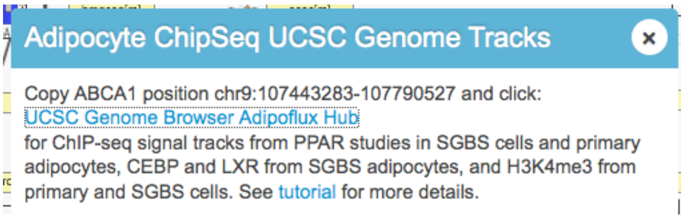
A key feature built in to adipoflux viz is the ability to plot gene expression changes collected over time; these values can serve as confirmatory metrics for the 'Gene expression' bottom line of the metanode graphs. Gene expression data is available for a large majority of genes in the pathways presented. Clicking on gene metanode automatically renders these dynamic plots when the data is available. For more convenience, clicking the hovering tooltip 'Expression Changes' or within the Metanode Details (left panel) 'Expression Change' will dynamically plot the selected gene expression values.

Shown below are HMGCS1 and LSS expression over a differentiation time course (cholesterol metabolism pathway). Notice that LSS gene has multiple microarray probes and therefore expression sets (different lines in the plot). The probe ID and fold change value relative to control pre-adipocytes are displayed on mouse over. Further, these charts are exportable.





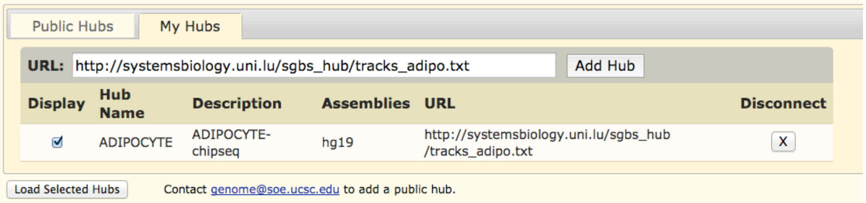
Launch Adipocyte Track Hub@UCSC Genome Browser:


Adipocyte ChipSeq UCSC Genome Tracks

Copy ABCA1 position chr9:107443283-107790527 and click:
UCSC Genome Browser Adipoflux Hub
for ChIP-seq signal tracks from PPAR studies in SGBS cells and primary adipocytes, CEBP and LXR from SGBS adipocytes, and H3K4me3 from primary and SGBS cells. See tutorial for more details.

On clicking of the UCSC Genome Browser link within the Metanode details section, a url to connect to the UCSC Adipocyte Track Hub is served, using ABCA1 as an example.

The URL launches to the following screen, then click Load Selected Hubs:

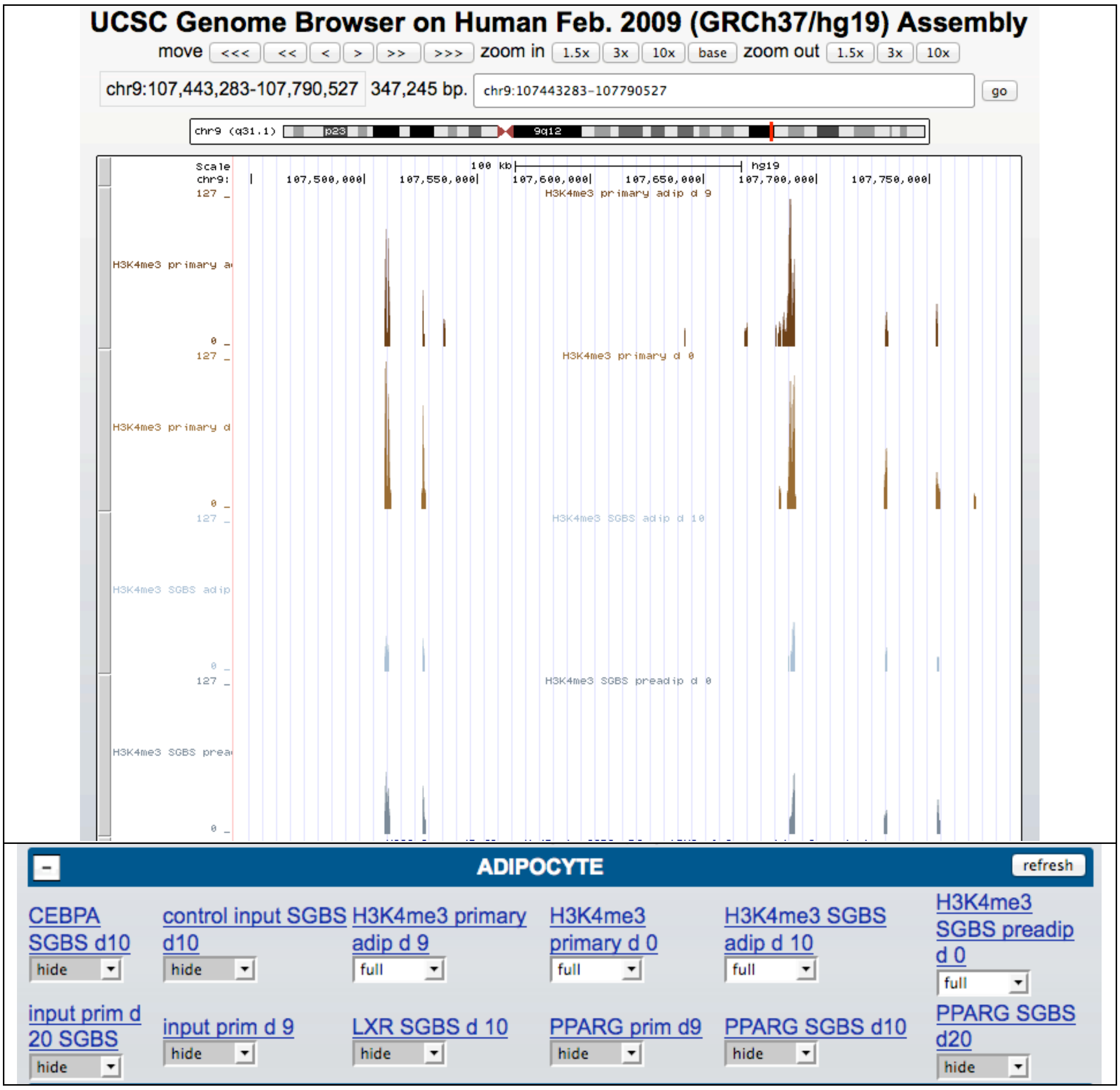


URL: http://systemsbiology.uni.lu/sgbs_hub/tracks_adipo.txt Add Hub

Display	Hub Name	Description	Assemblies	URL	Disconnect
☑	ADIPOCYTE	ADIPOCYTE-chipseq	hg19	http://systemsbiology.uni.lu/sgbs_hub/tracks_adipo.txt	X

Load Selected Hubs   Contact genome@soe.ucsc.edu to add a public hub.

UCSC Genome Browser with 4 H3K4me3 tracks visible, screen below indicates visibility control of Adipocyte Hub custom tracks:



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

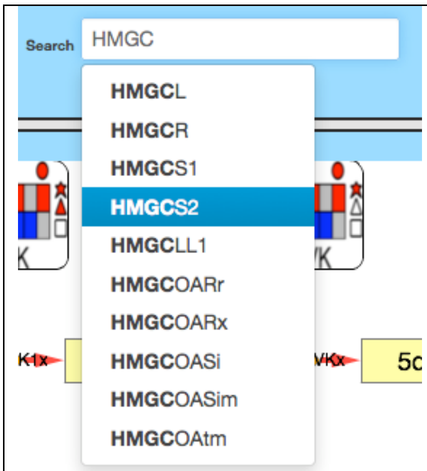
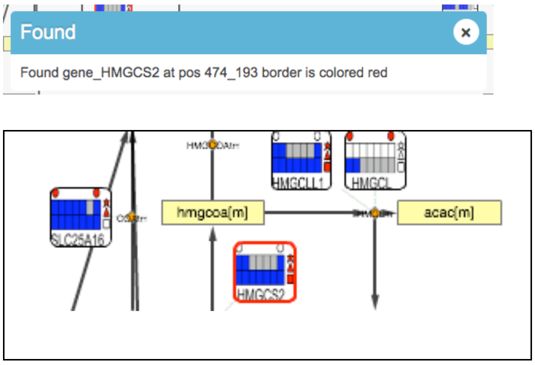
chr9:107,443,283-107,790,527 347,245 bp. chr9:107443283-107790527

CEBPA SGBS d10 hide	control input SGBS d10 hide	H3K4me3 primary adip d 9 full	H3K4me3 primary d 0 full	H3K4me3 SGBS adip d 10 full	H3K4me3 SGBS preadip d 0 full
input prim d 20 SGBS hide	input prim d 9 hide	LXR SGBS d 10 hide	PPARG prim d9 hide	PPARG SGBS d10 hide	PPARG SGBS d20 hide

3. Search function

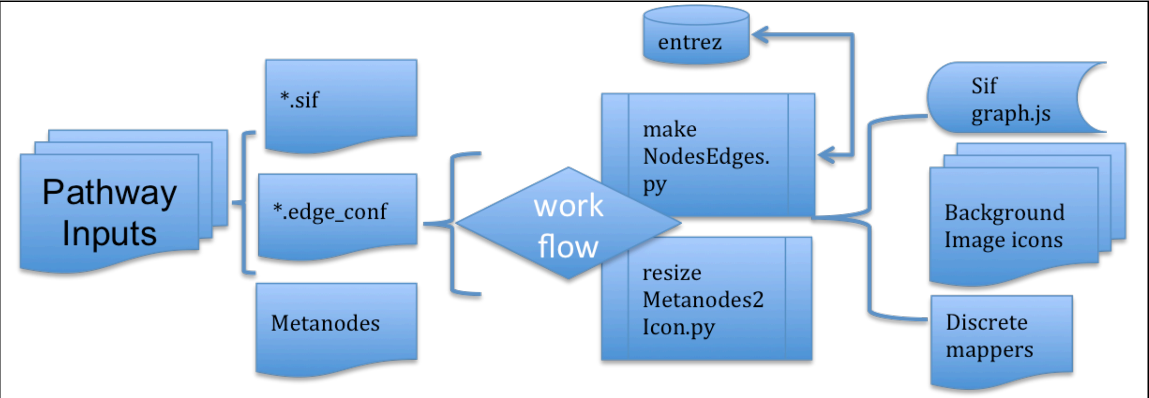
The search function has built in type ahead. A potential list of matches is displayed as a selectable list. This feature augments the power of search and is available for genes, reactions and metabolites; a message is shown on whether the term used is part of the active pathway, along with highlighting of the searched node.

An example scenario of searching in the cholesterol pathway for HMGCS2 is as follows:

 Search HMGC HMGCL HMGCR HMGCS1 HMGCS2 HMGCLL1 HMGCOARr HMGCOARx HMGCOASi HMGCOASim HMGCOAtm	 Found Found gene_HMGCS2 at pos 474_193 border is colored red SLC25A16 hmgcoa[m] HMGCLL1 HMGCL acac[m] HMGCS2
Type ahead automatically shows potential search terms.	A message dialog shows search status; and the found node is highlighted in red.

6. Web graph object automation

We defined a pathway as a set of nodes (genes, reactions, metabolites) connected by edges (tripartite graph). The genesis of AdipoFlux uses xgmml files that contains custom graphs, where the node positions (x,y) and graph attributes (node shapes, colors, edge arrow types…) read in on page load. While this layout scheme is informative and controlled, it is quite manual and only 5 pathways were included. Soon we realized the need for an automated workflow to integrate new pathways. Using the pathway definition – our solution is based on simple interaction (add link) and edge configuration files; in addition, the generated metanodes are also read in for creation of background image icons. All the required scripts are available as part of release.



Workflow inputs are set in a conf file and on workflow invocation, graph and image icons are created and placed into web container paths. Custom discrete mappers are coded as needed.

Because our graph object automation workflow is built using python and bash scripts and designed around principled graph theory, being extensible for other pathway datasets.

7. Web interactivity (HTML and CytoscapeWeb).

IDARE is an open sourced Web 2.0 application, built based on modern HTML specs. The source code is available online on https://code.google.com/p/adipoflux/, including a Matlab® file that can be used for producing the metanode image files (png).

We use Cytoscape Web to embed metabolic pathways as interactive networks.

We are grateful to the following open sourced projects that **IDARE** references:

- JQuery and JQuery plugins
  - Messi
  - JQuery.layout
  - JQuery.tipsy
- Bootstrap.js
- Highchart.js
- UCSC Genome Browser
- ImageMagick

The above tools and libraries were used to render data integration and interactivity and apply to all datasets on **IDARE**. A detailed description of the available functions can be found on the next chapter.

4.3 Manuscript II - “Cell type-selective disease-association of genes under high regulatory load”

Based on the observed association between lipid disease-related genes and combinatorial regulation in SGBS adipocytes, and the enrichment of vascular disease-associated genes among genes with $\geqslant 6$ TFs in HUVEC, we hypothesized that disease-related genes are under tight regulatory control as a mechanism to decrease errors as well as providing robustness to network perturbations.

The prioritization for novel disease-assotiation genes remains a rather complex task with few methods leading to plausible results. Here we investigated the relationship between regulatory load and disease, observing that HRL genes enrich for disease in a cell type-selective manner.

To test the hypothesis that **disease-related genes are under higher regulatory control** in a general setting, we used public chIP-seq data from the binding of a total of 93 TFs across 9 cell lines and a dataset on active enhancers (H3K27ac) from 139 samples comprising 96 tissue and cell types to rank genes based on their regulatory load and test the enrichment for disease association across multiple diseases, revealing a **cell type selective disease association enrichment for the high regulatory load genes**, as described in **Manuscript II**. The link between disease association and high regulatory load had not been shown previously to the presented extent.

Data analysis and integration resulted in the generation of:

– ranked lists of the TF and enhancer loads on protein coding genes for 9 cell lines or 139 samples, respectively;

– disease association enrichment test results for the TF and enhancer data, based in DisGeNET gene-disease associations and the hypergeometric distribution;

– the overlap of genes with the top 10% enhancer load across 139 samples;

– the average number of KEGG pathways per HRL gene and for an equal number of randomly selected genes, for each of the 139 samples;

– the average 3’UTR lengths of the HRL genes and other genes, per sample.

– the average betweenness centrality for all genes in a liver disease network *versus* the high regulatory load genes from two liver samples (E066 and HepG2).

The main results obtained were as follows:

1. higher proportion of disease genes among genes with more associated TFs and enhancers (Figures 2 and 3 (B,C));
2. positive correlation for the TF and enhancer loads across 9 ENCODE cell lines (Figures 3 and S1);
3. enrichment for disease association on the genes with highest TF and enhancer load across samples (Figures 2 (B), 3 (C), 4, 5 and S3);
4. low overlap between the genes with highest enhancer load per sample (average Jaccard index similarity $<$ 30%, Figure S2);
5. cell-type selective enrichment for disease association of the high enhancer load genes across samples (Figure 4 and S3);
6. enrichment for disease association for the HRL-non-super-enhancer associated genes, with little overlap between the top enhancer load genes and top expressed genes (Figures 5 and S2);
7. higher betweenness centrality for the HRL genes than other genes from a liver disease network (>2 FC greater, Figures 6 and S4);
8. HRL genes participate on average in more KEGG patwhays than random and show longer 3'UTRs with more miRNA binding sites (Figures 6, 7 and S5).

Manuscript II is integrally presented starting from page 119.

Nucleic Acids Research Advance Access published September 3, 2015

doi: 10.1093/nar/gkv863

Cell type-selective disease-association of genes under high regulatory load

Mafalda Galhardo[1], Philipp Berninger[2], Thanh-Phuong Nguyen[1], Thomas Sauter[1] and Lasse Sinkkonen[1,*]

[1]Life Sciences Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg and [2]Biozentrum, University of Basel and Swiss Institute of Bioinformatics, 4056 Basel, Switzerland

Received April 13, 2015; Revised July 27, 2015; Accepted August 14, 2015

ABSTRACT

We previously showed that disease-linked metabolic genes are often under combinatorial regulation. Using the genome-wide ChIP-Seq binding profiles for 93 transcription factors in nine different cell lines, we show that genes under high regulatory load are significantly enriched for disease-association across cell types. We find that transcription factor load correlates with the enhancer load of the genes and thereby allows the identification of genes under high regulatory load by epigenomic mapping of active enhancers. Identification of the high enhancer load genes across 139 samples from 96 different cell and tissue types reveals a consistent enrichment for disease-associated genes in a cell type-selective manner. The underlying genes are not limited to super-enhancer genes and show several types of disease-association evidence beyond genetic variation (such as biomarkers). Interestingly, the high regulatory load genes are involved in more KEGG pathways than expected by chance, exhibit increased betweenness centrality in the interaction network of liver disease genes, and carry longer 3′ UTRs with more microRNA (miRNA) binding sites than genes on average, suggesting a role as hubs integrating signals within regulatory networks. In summary, epigenetic mapping of active enhancers presents a promising and unbiased approach for identification of novel disease genes in a cell type-selective manner.

INTRODUCTION

Identification of disease-relevant genes and gene products as biomarkers and drug targets is one of the key tasks of biomedical research. Great progress has been made in diagnosing and treating various diseases over the past decades. Still, a great majority of research is focused on a small minority of genes while over a third of genes remain unstudied (1). Unbiased prioritization within these ignored genes would be important to harvest the full potential of genomics in understanding diseases.

Many databases to catalog disease-associated genes and the nature of their association, such as the Comparative Toxicogenomics Database (CTD) or the Online Mendelian Inheritance in Man (OMIM), have been created (2,3). One of the more comprehensive databases, DisGeNET (4,5), draws from multiple sources as well as text-mining approaches to generate gene-disease networks where genes are associated to diseases by various evidence ranging from altered expression and genetic variation to existing therapeutic association. DisGeNET already links many of the human genes to at least one disease, highlights the multigenetic background of most diseases and how many genes can be associated to multiple diseases (4,5).

Interestingly, as much as 90% of the human disease-associated genetic variants are located outside of the coding sequences of protein coding genes, suggesting that they affect the regulation of these genes instead (6,7). The active regulatory regions of the genome can be identified in a cell type-specific manner through chromatin immunoprecipitation coupled with deep sequencing (ChIP-Seq) analysis of selected covalent histone modifications such as histone H3 lysine 27 acetylation (H3K27ac; marking active enhancers) and histone H3 lysine 4 trimethylation (H3K4me3; marking open transcription start sites), among others. Indeed, by taking advantage of such epigenomic data produced by the Roadmap Epigenomics Mapping Consortium, Farh *et al.* (8) recently showed that up to 60% of human autoimmune variants are located within active enhancers of immune cells. In particular, the genetic variants seem to coincide with so called super-enhancers or stretch-enhancers, large enhancer regions often associated with key genes and master regulators of cellular identity (9–11). These enhancers function as hotspots with binding sites for multiple transcription factors (TFs) (12) and, within the enhancers, single nucleotide polymorphisms (SNPs) often disrupt these binding sites as shown, for example, for type 2 diabetes vari-

[*]To whom correspondence should be addressed. Tel: +352 4666446839; Fax: +352 4666446435; Email: lasse.sinkkonen@uni.lu

ants within islet enhancers (13). However, it remains unclear whether the genes controlled by multiple enhancers and TFs are associated to disease also beyond the genetic variation in their regulatory regions, as could be assumed from their role as regulators of cellular identity.

We have previously shown that metabolic genes regulated by multiple TFs in human umbilical vein endothelial cells (HUVEC) are enriched for genes associated to endothelial relevant diseases in DisGeNET (14). Here we set out to test whether this increased disease-association of genes under high regulatory load (HRL) is a general observation that holds across cell types and genes, and independent of the type of disease-association evidence. Analysis of ChIP-Seq data for 93 TFs across 9 ENCODE cell lines confirms an enrichment for disease-association among the highest regulated genes in all cell types. We find that the TF load of the genes correlates with their enhancer load in the respective cell types and thereby allows the identification of genes under high regulatory load by epigenomic mapping of active enhancers using H3K27ac. Consistently, genes associated with most enhancers are also most enriched for disease-association in all 9 cell lines. To elucidate the power of this approach and to analyze the cell type selectivity of the disease-associations, we perform disease-association enrichment analysis for high enhancer load genes from 139 ChIP-Seq samples of H3K27ac corresponding to 96 different cell types and tissues, with many diseases showing high level of cell type selectivity. Finally, we show that genes under high enhancer load are involved in more Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways and exhibit higher betweenness centrality in a liver disease gene network than other genes on average, suggesting a central role in integrating multiple signals in biological networks. Consistently, the genes under high regulatory load at the transcriptional level have longer 3′ untranslated regions (3′ UTRs) and contain more microRNA (miRNA) binding sites than other genes, suggesting that they could be under higher regulatory load also at the post-transcriptional level.

Taken together, these results paint a picture of high regulatory load genes as central nodes in biological networks, that are more likely to be associated with human disease, and identifies epigenomic analysis of active enhancers as a tool for cell type-selective prioritization of previously unstudied genes.

MATERIALS AND METHODS

Disease-associated genes

Gene-disease association data were retrieved from the DisGeNET Database (GRIB/IMIM/UPF Integrative Biomedical Informatics Group, Barcelona http://www.disgenet.org/ version 2.1, 5th of May 2014). DisGeNET provides gene-disease associations from several public data sources and literature text-mining, with a score ranking associations based on the supporting evidence. A minimum association score of 0.08 was used to select gene-disease associations supported by multiple data sources and to exclude associations that are based solely on text-mining results, resulting in 7428 disease-associated genes, of which 6167 were contained in our background set of 19 238 protein coding genes (Supplementary File 1). Alternatively, a minimum association score of 0.2 characterizes curated disease-associated genes (7110 genes, of which 5853 were in the background set) (gene-disease associations from UNIPROT, ClinVar and CTD human data set, see http://www.disgenet.org/web/DisGeNET/menu/dbinfo). Additionally, as a separate set of high confidence disease genes we used the OMIM database (downloaded from ftp://ftp.omim.org/OMIM/, as of June 2015) (4557 genes of which 3483 were in the background set). For gene set enrichment testing we selected only diseases with at least 15 associated genes, to avoid significant results only due to a very small set size, resulting in 340 diseases (Supplementary File 1) (15). Details about the gene-disease-association types defined in the DisGeNET for each disease are also found in the Supplementary File 1, and they include 'altered expression', 'biomarker', 'genetic variation', 'post-translational modification' and 'therapeutic'. To test whether 'genetic variation' was predominantly accounting for disease-association enrichment, we defined the group 'not genetic variation' by pooling all disease-associated genes with association evidence other than 'genetic variation'.

Background set of protein coding genes and their 'regulatory domain'

We focused on protein coding genes in the analysis. The NCBI Entrez Gene annotations for 'protein coding' genes (Homo_sapiens.gene_info file, ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/, downloaded on the 13th of May 2014) were used to derive a set of genes serving as 'background' for gene set enrichment testing. Their TSS was extracted by intersecting with the RefSeq genes file taken from the UCSC Table Browser (16) (http://genome.ucsc.edu/cgi-bin/hgTables, RefSeq genes, assembly: February 2009 (GRCh37/hg19), on the 13th of May 2014), resulting in 19 238 protein coding genes. In order to associate ChIP-seq peaks to the 19 238 genes, we used the Genomic Regions Enrichment of Annotations Tool (GREAT) (17) to derive a 'regulatory domain' for each gene, using the script 'createRegulatoryDomains' and the rule 'BasalPlusExtension' with default settings (source code from http://bejerano.stanford.edu/help/display/GREAT/Download, May 2014). Chromosome sizes of the human genome assembly hg19 were obtained using the script 'fetchChromSizes' from the UCSC BigWig and BigBed tools (18). Supplementary File 1 contains details on the 19 238 protein coding genes used for analysis, including their regulatory domains derived by the GREAT tool as start and end coordinates.

Data sources and processing

Public ChIP-seq data produced by the ENCODE project (19), the BLUEPRINT Epigenome project (20) and the NIH Epigenomic Roadmap project (21) were downloaded from the ENCODE Data Coordination Center (http://genomebrowser.wustl.edu/encode/) on May 2014, the BLUEPRINT consortium website (http://www.blueprint-epigenome.eu) on July 2014, and

NIH Epigenomic Roadmap supplementary website (http://compbio.mit.edu/roadmap) on January 2015, respectively. These data span 93 TFs, the H3K4me3 and the H3K27ac modification marks across 139 samples that comprise 96 tissues or cell types (Supplementary File 2). The ENCODE data were no further processed, while the BLUEPRINT and NIH Epigenomic Roadmap data were filtered to keep only peaks with a minimum fold change and ($-\log_{10}$ q-value) of 3.

The H3K4me3 data were used to filter out genes embedded in closed chromatin. A file containing genes with at least one H3K4me3 peak within their transcription start sites (TSS) ±1000 bp was obtained per sample, using the IntersectBed tool from the BEDTools suite (22) to intersect each sample's H3K4me3 data with a file containing RefSeq genes and their TSS ±1000 bp as start and end coordinates. In case of multiple H3K4me3 data files per sample, we considered evidence from one single file sufficient to call the mark present. The H3K27ac data served to map active enhancers. The ENCODE project was the only source of TF data. To select only TFs known to directly bind DNA, we used a list of manually curated TFs (23), 111 of which were included in the ENCODE TFs (Supplementary File 2) and 93 had been assayed in the used cell lines, resulting in the presented numbers of unique TFs assayed per cell line. In case of multiple files of the same TF in a cell line (e.g. different ENCODE data producing labs), a filtering step for keeping only peaks overlapping by at least 1 bp in two thirds of the 'replicates' was applied. The intersectBed tool was used to intersect TF or H3K27ac data with the file containing regulatory domains for each gene (see above), requiring a peak to completely fall within the genes regulatory domain in order to assign it to that gene. For each TF, we obtained a list of associated genes and derived the TF load per gene from the total number of associated TFs, across nine ENCODE cell lines (A549, GM12878, H1hESC, HCT116, HeLaS3, HepG2, HUVEC, K562 and MCF7). To obtain the enhancer load per gene, we used the count option of the IntersectBed tool to count the number of H3K27ac peaks falling within the genes regulatory domain. For both TFs and enhancers, we ranked genes based on the regulatory load and subsequently considered only genes with the H3K4me3 mark within ±1000 bp of the TSS. Following the above settings, on average 96% of peaks could be associated to a target gene.

Gene binning and hypergeometric enrichment tests

In order to group genes based on their regulatory load, we started binning ranked genes by deciles (bins containing 10% of the genes), with a separate group for genes with no associated TFs or enhancers (11 starting bins). Bins were then extended by inclusion of all genes with the same regulatory load as the last gene falling in a bin, excluding cases of genes with equal regulatory load falling in different bins (fewer bins depending on the sample). Top bin genes for each sample can be found in Supplementary File 3. We then performed hypergeometric distribution tests for the enrichment of disease genes among the different regulatory load bins per sample and the 340 DisGeNET diseases with at least 15 genes. For each sample, the 'population size' corresponded to the number of genes with the H3K4me3 mark (varying per sample), the 'number of successes' being the number of disease genes with the H3K4me3 mark (varying per sample) and the 'number of draws' the number of genes having the regulatory load of the bin in case (number of TFs or enhancers). Hypergeometric *P*-values were obtained using the Matlab® hypergeometric cumulative distribution function (hygecdf) and were adjusted for multiple testing with the Benjamini and Hochberg methodology as implemented in the Bioconductor's qvalue package (http://www.bioconductor.org/packages/release/bioc/html/qvalue.html).

Bicluster of hypergeometric enrichment statistical significance

In order to simultaneously cluster samples and diseases into homogeneous blocks based on the hypergeometric enrichment significance (adjusted $-\log_{10}$ *P*-values), the R package 'blockcluster' (24) was applied to the matrix (of adjusted $-\log_{10}$ *P*-values) from the 139 samples and 174 diseases, after binarization ('zero' for $-\log_{10}$ *P*-value < 1.301, 'one' otherwise) and exclusion of diseases or samples only containing 'zero'. Shortly, block clustering methods estimate a mixture model from permutations of objects and variables in order to draw a correspondence structure (thereby with certain order variability with repetition). 'Blockcluster' requires a predefined number of clusters for the rows and columns, which we fixed at 9 and 7, respectively (here, diseases and samples), in order to minimize redundant clusters. Supplementary File 4 contains the ordering for diseases and samples and their clusters (color shades), as obtained with the 'blockcluster' package.

Identification of super-enhancer genes

NIH Roadmap epigenomics raw data were downloaded from the GEO ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/by_experiment/) on May 2015, selecting data for all three from H3K4me3, H3K27ac and Input, resulting in 35 samples. These included bed files of reads aligned onto the hg19 human genome assembly using Pash 3.0 read mapper (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html). As the raw data contained sample names and the processed data used for the high regulatory load genes analysis contained sample IDs, mappings between the two were manually obtained based on descriptions from the original data sources (http://egg2.wustl.edu/roadmap/web_portal/meta.html). Next, the software HOMER (version 4.7, 25th of August 2014) (25) was used for super-enhancer calling on the H3K27ac bed files from each sample, pooling samples from the same origin, with default setting except the local fold change option (-L) which was set to 0 as recommended by the authors for super-enhancer analysis, resulting in the obtainment of the chromosome, start and end coordinates of super-enhancer peaks. We then used the IntersectBed tool from the BEDTools suite (22) and the genes 'regulatory domain' file obtained with GREAT (see previous descriptions) to derive a set of super-enhancer-associated genes per sample. These genes were subsequently used for testing the enrichment for disease-association using the hypergeometric distribution, as previously described.

Analysis of the RNA-seq data

Data were downloaded from http://egg2.wustl.edu/roadmap/web_portal/processed_data.html#RNAseq_uni_proc on June 2015, taking the file '57epigenomes.RPKM.pc' containing the RPKM (reads per kilobase per million mapped reads) for 57 samples, 38 of which were also in the set of 139 samples used for the analysis of high regulatory load genes. Conversion of ENSEMBL IDs to ENTREZ GENE IDs was done using the Bioconductor package 'biomaRt' (26), resulting in expression data for 18 220 ENTREZ GENE IDs across samples, out of which 18 181 were included in our background set of 19 238 protein coding genes. For each of the 38 samples, genes were ranked based on expression. Since the set of genes for high regulatory load and expression is not the same, we defined the top bin of highly expressed genes to contain the same number of high regulatory load genes in each sample, triplicating this number for the 30% top bins of expression. The 50% and 90% top expression bins were obtained relative to the total number of genes for which there was expression data (18 181).

KEGG pathway enrichment testing

KEGG (27) pathways were used to test whether high regulatory load genes appear in more pathways than expected. The list of KEGG pathways was obtained through the REST-style KEGG API from http://rest.kegg.jp/list/pathway/hsa, resulting in 282 pathways with at least one gene. KEGG pathways were downloaded and gene info per pathway was obtained using the R/Bioconductor package 'KEGGprofile'. The average number of pathways per KEGG gene (total of 6822 genes in all KEGG pathways), per high regulatory load gene (differing from sample to sample) or based on a random selection of an equal number of genes as the high regulatory load genes for each sample (10 000-fold) was calculated. Supplementary File 4 contains the results obtained for each sample. A *P*-value ($\leq$0.05 was considered significant) was calculated from this re-sampling test based on the probability to get at least the same average number of pathways per KEGG gene in random selections as obtained for the high regulatory load genes.

Constructing a liver disease gene network

The list of liver diseases was curated from the Medical Subject Headings (MeSH) database (http://www.ncbi.nlm.nih.gov/mesh/). The MeSH database is the National Library of Medicine's controlled vocabulary thesauruses consisting of sets of terms structured in a hierarchical form that facilitates searching at different levels. We curated 137 liver diseases. Based on the obtained list of liver diseases, 847 genes related to liver diseases (liver disease genes in short) were extracted from the Comparative Toxicogenomics Database (CTD) database (28). We considered only curated disease-gene associations to increase the reliability of the liver disease gene data. The construction of liver disease gene network was carried out by extracting human protein interactions published in the Human Protein Reference Database (HPRD) (29). The HPRD database contains manually curated protein interactions from literature and is one of the most well-known human protein interaction databases.

The final liver disease gene network of interest consisted of the liver disease genes and their neighbors (nodes), and their direct interactions (edges). In this study, we took into account one-step neighbors. The network was undirected and unweighted because we considered binary interactions. We obtained a network of 3775 genes and 8278 interactions. To unravel the role of genes in the network, we calculated betweenness centrality for each gene and compared the average betweenness centrality of the high regulatory load genes to that of all genes or all genes except those under high regulatory load. Betweenness shows the bridge role of a gene for other genes in the network (30). For each node *v* in the network, we computed the total number of shortest paths from node *s* to node *t*, called $d(s,t)$ and the number of those paths that pass through *v*, called $d(s,v\ t)$, and then ratio $d(s,v,\ t)/d(s,t)$ was calculated. These steps were repeated for all pairs of node *s* and node *t* in the network. The overall betweenness centrality of a node *v* is obtained by summing up those ratios. Betweeness $B(v)$ of a node *v* is defined as following:

$$B(v) = \sum_{s \neq v \neq t} \frac{d(s, v, t)}{d(s, t)} \qquad (1)$$

3′ UTR length and miRNA binding site analysis

Annotation data on 5′ UTR, CDS, 3′ UTR, spliced as well as unspliced transcript length for human mRNA genes was obtained from Biomart (Ensembl Genes 78). Transcripts lacking proper UTR annotation were filtered out. In cases where multiple transcripts correspond to one gene ID, a representative member was randomly chosen. A background set, consisting of 16 307 genes, was used for all comparisons. In order to test the hypothesis, that highly regulated genes tend to have longer 3′ UTRs, we compared in all 139 samples the length of 3′ UTRs, CDS, spliced as well as unspliced transcript length of the high enhancer load genes with the background set with the Kolmogorov-Smirnov test, testing if the background set is smaller than the test set. In order to correct for multiple testing, Bonferroni correction was used, with a significance level $\leq$0.0003597122 (0.05/139).

Predicted target sites for conserved miRNAs were obtained from TargetScan 6.2 (31). The target site count per 3′ UTR were summed up, resulting in an average site count per transcript. In cases where a site in the 3′ UTR was assigned to multiple miRNAs, it was counted only once.

RESULTS

Genes under high regulatory load from multiple transcription factors are enriched for disease-association across cell types

Our previous work on regulation of metabolic genes in human adipocytes and human primary macrophages has uncovered that combinatorial control by multiple regulators is in particular occurring at genes associated to key nodes such as entry points of the metabolic networks and at genes that are often disease-related (14; Pires Pacheco *et al.*, under revision). Moreover, analysis of metabolic genes controlled by multiple TFs in HUVEC cells revealed consistent enrichment for endothelial disease-relevant genes among the genes under the highest regulatory load (14).

To investigate whether this is a general finding across different cell types and gene categories, we took advantage of the numerous ChIP-Seq data sets of TF binding produced by the ENCODE project in a number of cell types (19). In detail, we used the existing TF binding data for a total of 93 different previously manually curated TFs (23) from nine ENCODE cell lines, representing different cell and tissue types (Figure 1, Supplementary File 2; see Methods for details). The number of assayed TFs per cell line varied from 6 TFs in HUVEC cells to 65 TFs in GM12878 cells. In addition we used ChIP-Seq data for H3K4me3 from each cell line to identify the putative active genes and associated all TF binding events to their proximal protein-coding genes marked by H3K4me3 following the 'BasalPlusExtension' rule of the GREAT tool (17). The number of unique associated TFs per each gene was then calculated and the genes were ranked according to the number of associated TFs, i.e. their regulatory load in each cell type (Figure 2A). The number of associated TFs ranged from 0 TFs per gene to as many as 57 TFs per gene for the genes with highest load in the GM12878 lymphoblastoid cell line (Supplementary File 2). Finally, all genes were classified either as disease-associated or non-disease-associated based on the evidence in the DisGeNET database (using a cut-off score of 0.08 for disease-association to exclude associations based only on text mining) (Figure 1; see methods for details) (4,5).

When focusing on the genes ranked according to their TF load, a similar pattern emerges in each cell line, independent of the number of assayed TFs: the proportion of disease-associated genes is usually close to or above 40% for the genes with highest TF load while for the majority of genes this proportion remains at 10–35% (Figure 2A). To test whether the observed enrichment is statistically significant, we ranked the H3K4me3 marked genes in each cell line into 10 bins of comparable size according to their TF load (6 bins in case of HUVEC cells) with an additional 11th bin in case the gene was not associated with any TF (Figure 2B). Next, the enrichment of disease-associated genes within each bin was tested using the hypergeometric distribution (see Methods for details). As shown in Figure 2B, only the bins of genes with highest TF load show a significant enrichment of 1.301 or higher (adjusted $-\log_{10}$ *P*-value corresponding to 0.05) for disease-association with bins based on top 10% genes always showing the most significant enrichment. Importantly, similar results were also obtained when using Gene Set Enrichment Analysis instead of hypergeometric distribution (15).

In conclusion, genes under combinatorial control from multiple TFs are enriched for disease association across multiple cell types, suggesting high regulatory load as a common feature of genes implicated in human diseases.

High transcription factor load correlates with high abundance of active enhancers

While presence of a TF binding event in proximity of a target gene could be indicative of either activation, repression or even no regulation by the TF, the presence of enhancer markers such as H3K27ac are indicative of active enhancers engaged in transcriptional activation via chromatin looping (32,33). To see whether the observed disease-association enrichment of genes under high TF load could be more easily observed by analyzing only few chromatin modifications, we used the H3K27ac ChIP-Seq data for active enhancers produced by the ENCODE project from the corresponding cell lines. Comparison of the average TF load and corresponding number of enhancer peaks at each open gene across the cell lines revealed a clear positive correlation, arguing that most genes under high TF load are also identifiable by a high active enhancer load (Figure 3A). Similar conclusion can be made when the genes are binned in comparable sized groups according to their TF or enhancer loads and analyzed for enrichment of genes within each bin (Supplementary Figure S1, Supplementary File 5). For example, the bins containing genes with highest TF load are significantly enriched for genes with highest enhancer load, and vice versa, the bins of genes with no associated TFs are also enriched for genes with no enhancers.

Based on the obtained correlations, we asked whether ranking of genes according to their enhancer peak abundance would also reveal higher proportion of disease-associated genes among the top ranking genes, similarly to high occupancy by multiple TFs. Indeed, the top ranking genes with highest enhancer load showed higher proportion of disease genes while genes associated with less than 10 enhancer regions rarely show a disease-gene proportion higher than 40% (Figure 3B). Again, the enrichments are also highly significant for the genes under the highest enhancer load in each cell line when tested with the hypergeometric distribution after grouping genes in comparable size bins, with top bins showing the most significant enrichments (Figure 3C). And yet again, similar results were also obtained when using Gene Set Enrichment Analysis instead of hypergeometric distribution (15). Moreover, similar enrichment patterns are also visible when more stringent groups of disease genes (DisGeNET score cut-off 0.2 or genes of monogenic diseases from OMIM database) are used (Supplementary Figure S2). Importantly, the disease-gene proportion profiles obtained using the enhancer load data appear more comparable between the different cell lines than in the TF load analysis that is highly dependent on the number and identity of the assayed TFs.

Taken together, the TF load of accessible (H3K4me3 marked) genes is positively correlated with the number of associated active enhancer peaks and the genes with highest enhancer load are enriched for known disease-relevant genes. This could allow the identification of novel disease genes through ChIP-Seq analysis of enhancer load using histone marks such as H3K27ac.

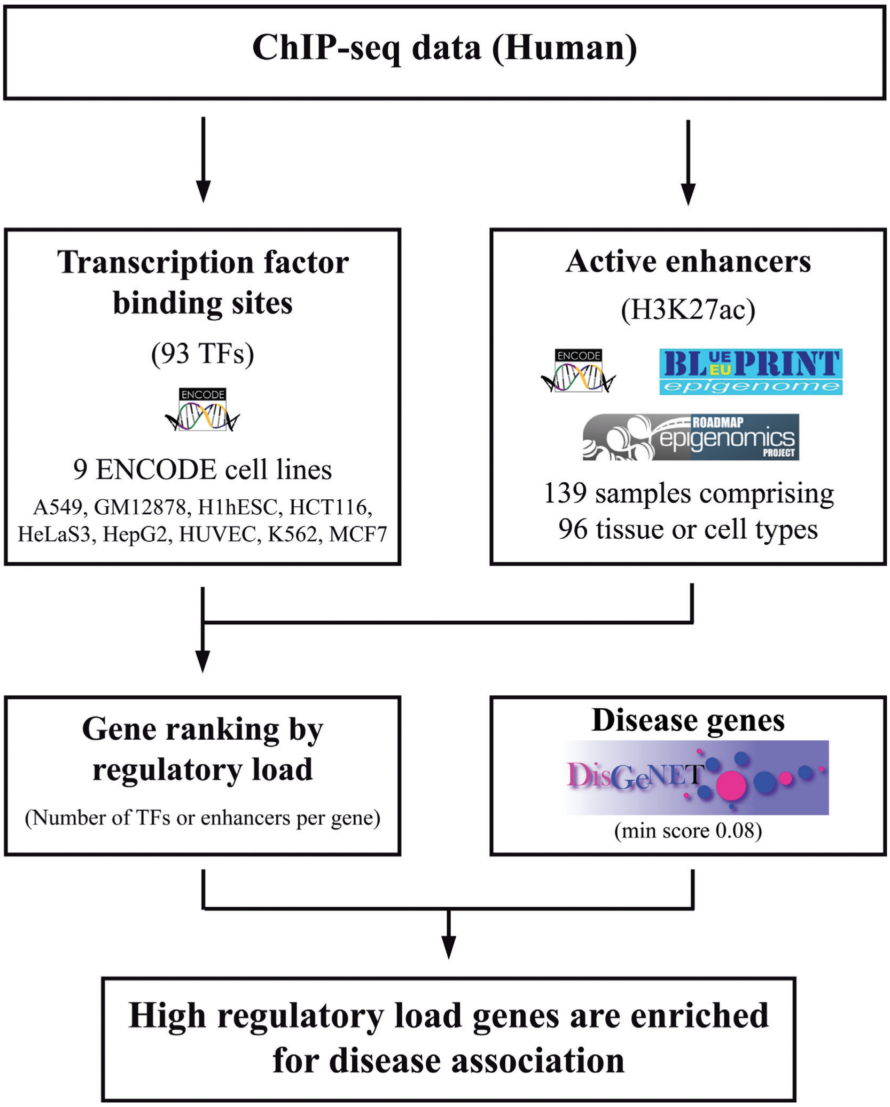

ChIP-seq data (Human)
Transcription factor binding sites
(93 TFs)
ENCODE
9 ENCODE cell lines
A549, GM12878, H1hESC, HCT116, HeLaS3, HepG2, HUVEC, K562, MCF7
Active enhancers
(H3K27ac)
ENCODE
BLUEPRINT epigenome
ROADMAP epigenomics PROJECT
139 samples comprising 96 tissue or cell types
Gene ranking by regulatory load
(Number of TFs or enhancers per gene)
Disease genes
DisGeNET
(min score 0.08)
High regulatory load genes are enriched for disease association

Figure 1. The workflow of the disease-gene enrichment analysis. Processed ChIP-seq data (bed files) from 93 transcription factors (TFs), H3K27ac and H3K4me3 across 139 sample sets were downloaded from the ENCODE (19), NIH Epigenomic Roadmap (21), and BLUEPRINT Epigenome (20) projects (see Supplementary File 2 for additional details). The H3K27ac was used as a mark for active enhancers. The GREAT tool (17) was used to derive a 'regulatory domain' for each protein coding gene ('BasalPlusExtension' rule with default settings) and the regulatory load per gene was obtained from the number of TF or enhancer peaks falling within the genes regulatory domain. Genes within closed chromatin regions (without the H3K4me3 mark within ±1000 bp from the TSS) were ignored. Gene-disease associations for 340 diseases with at least 15 genes were based on the DisGeNET database (requiring a minimum association score of 0.08), for a total of 7428 disease genes (4,5). The 19 238 protein coding genes (including 6167 of the disease genes) in our background set were grouped into comparable sized bins based on the regulatory load per sample. These 'regulatory load' bins were used for testing disease association enrichment (hypergeometric distribution) across 139 samples on the 340 diseases. The enrichment significance (adjusted $-\log_{10}P$-value) for each disease across samples was used to infer cell type and function related associations.

Cell type-selective disease-association of genes controlled by multiple active enhancers

H3K4me3 and H3K27ac profiles have already been mapped in numerous different tissue and cell types, allowing us to extend our analysis beyond the nine cell lines from the ENCODE project. To this end, we collected additional preprocessed ChIP-Seq data mapping both modifications from the ENCODE project (19), NIH Epigenomic Roadmap Consortium (21) and BLUEPRINT Epigenome project (20), obtaining a total of 139 sample sets corresponding to 96 different cell types and tissues (Figure 1, Supplementary File 2). For each sample set we performed the enhancer-to-gene association as described in Methods and binned the H3K4me3 marked genes according to their enhancer load to identify the genes under high regulatory load (in the top bin) in each sample set (Supplementary File 3). To compare the top bins, the Jaccard similarity index was calculated for the pair-wise combinations of the 139 samples (Supplementary File 4). Interestingly, the genes with high regulatory load varied a lot between the different cell types and tissues, with most cell types showing lower than 30% similarity when compared with the Jaccard similarity index (Supple-

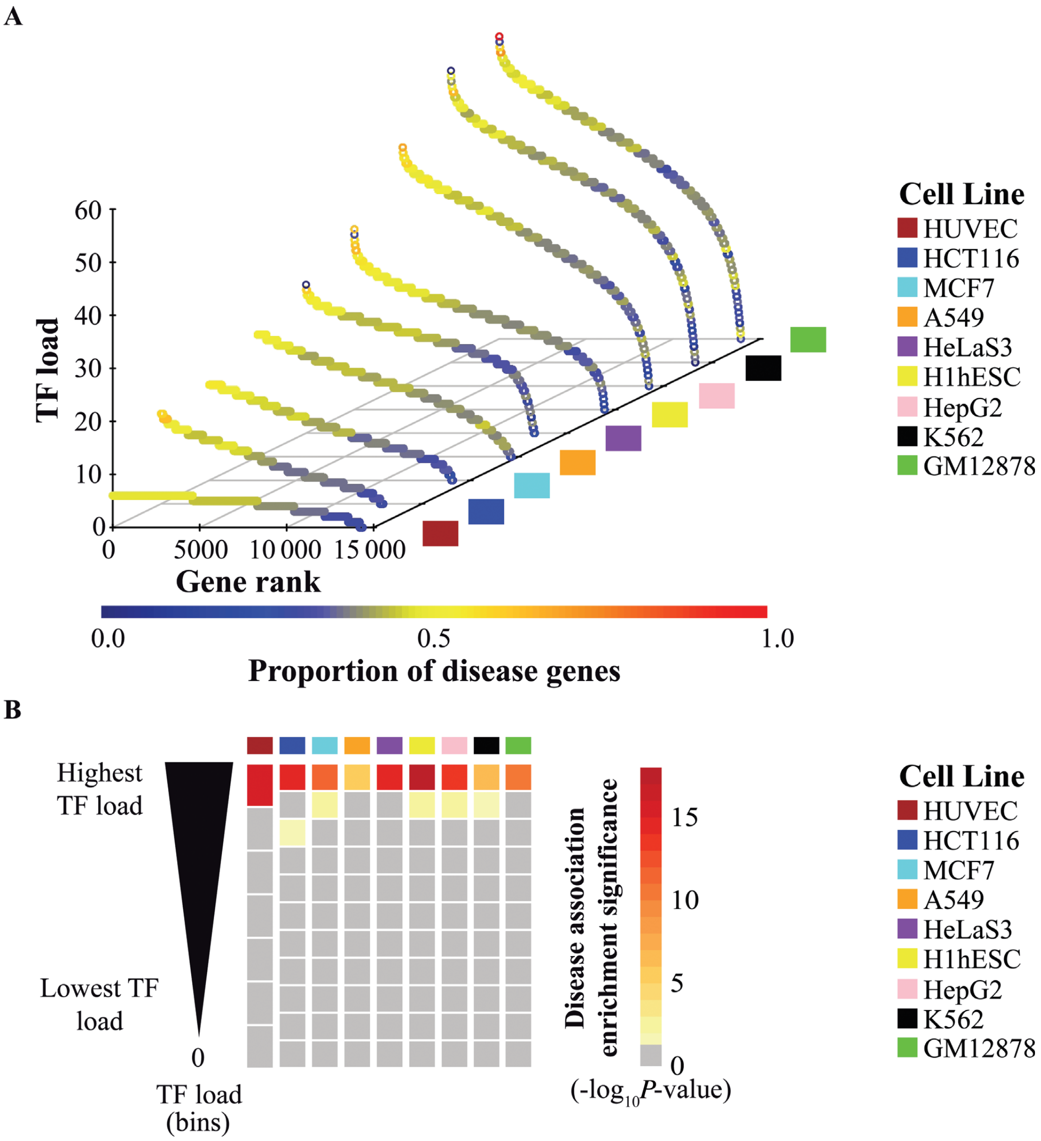


Figure 2. TF load enriches for disease association. Based on ENCODE data from 93 TFs across nine cell lines, the proportion of disease genes is higher among genes with high TF load. (**A**) 3D scatter plot of the TF load per gene and proportion of disease associated genes across nine cell lines. The proportion of disease genes is higher among genes with more TFs. Genes were ranked based on the number of TFs falling within their regulatory region (as defined in Methods). The TF load is depicted on the *z*-axis and the gene rank based on the TF load on the *x*-axis. Data from genes without the H3K4me3 mark within ±1000 bp of the transcription start site are not shown. The nine ENCODE cell lines are shown across the *y*-axis. 6167 disease genes were considered based on the DisGeNET version 2.1 associations (minimum association score of 0.08). The proportion of disease genes among all genes with each unique observed TF load is represented by the color gradient on the TF load for each cell line. (**B**) Heatmap depicting the statistical significance of the enrichment for disease associated genes in all TF load bins (adjusted $-\log_{10} P$-value), across nine cell lines. For each cell line, genes were grouped based on the number of TFs into deciles, and genes without TFs grouped separately. To avoid different bins having genes with the same TF load, the deciles were adjusted to contain all genes with the same number of TFs as the last gene in the decile. Using the set of 6167 disease associated genes derived from the DisGeNET, hypergeometric tests for each bin were performed. The statistical significance is indicated by the color gradient. Values below 1.301 (i.e. adjusted *P*-values larger than 0.05) are shown in gray and not considered significant. The enrichment significance is highest in the top bin of each cell line.

mentary Figure S3). Consistent with previous reports, the similarity was highest between the cell types from the same tissue, function or developmental origin.

Based on this cell-type selectivity of the high regulatory load genes, we hypothesized that high regulatory load would also enrich for diseases in a cell type-selective manner, and possibly allow informative links between different diseases and cell types. To test this, we collected all 340 diseases from DisGeNET database that had at least 15 associated genes with a minimum score of 0.08 (Supplementary File 1). Next, the enrichment of genes associated to each of these diseases was tested separately in all 139 sets of high regulatory load genes derived above based on the number of associated H3K27ac peaks (Supplementary File 3) to obtain a matrix of cell type- and disease-selective significant enrichments (Figure 4, Supplementary Figure S4,

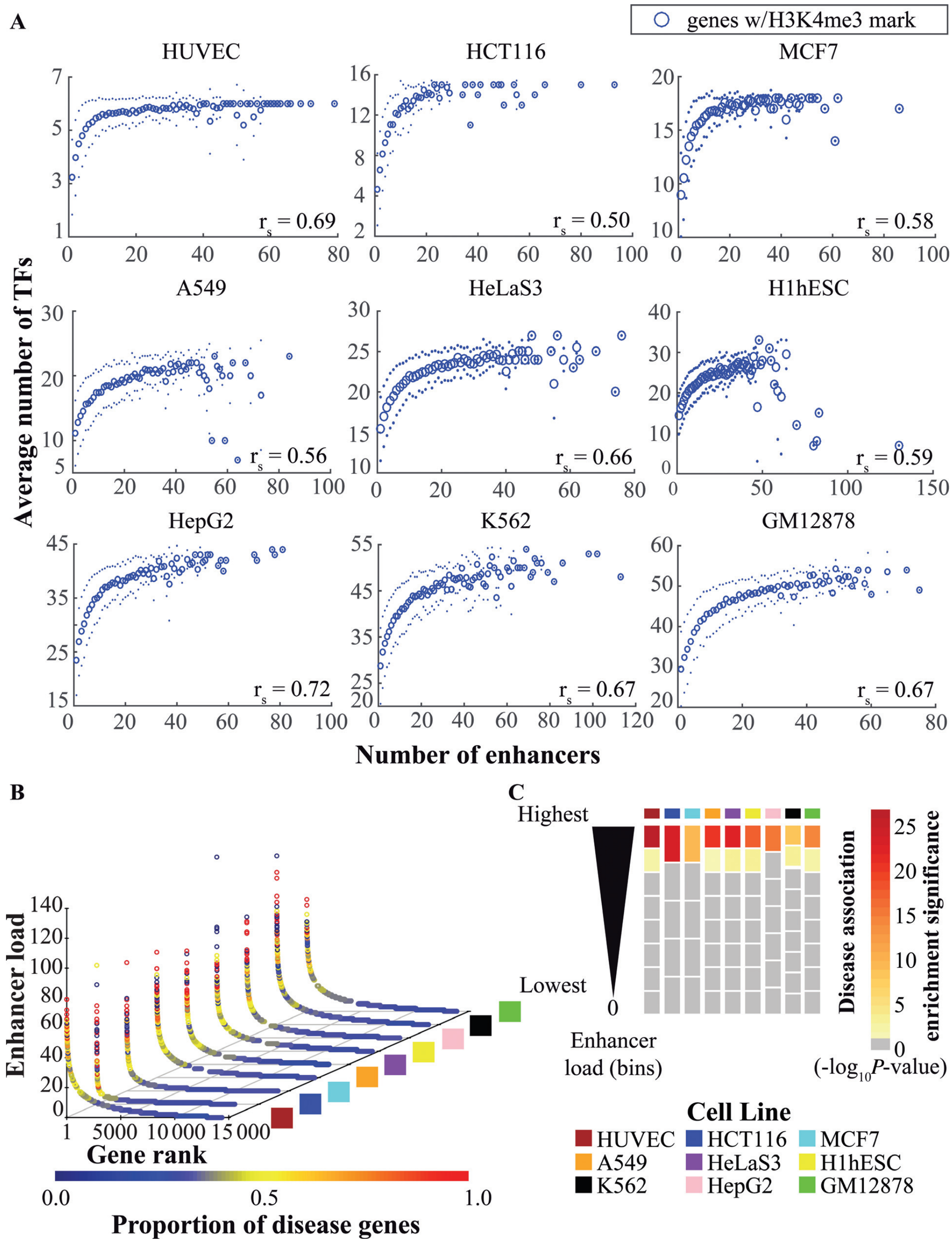


Figure 3. Enhancer load enriches for disease association. (**A**) Plots of the average number of TFs (*y*-axis) for each unique number of enhancer peaks per gene (*x*-axis) for nine cell lines. A positive correlation between the two is observed and the Spearman's rank correlation coefficient (r) is shown on the lower right corner of each plot varying from 0.5 (HCT116) to 0.72 (HepG2). (**B**) 3D scatter plot of the enhancer load per gene and proportion of disease associated genes across nine cell lines. The proportion of disease genes is higher among genes with more enhancers. Genes were ranked based on the number of enhancer peaks falling within their regulatory region (as defined in Methods). The enhancer load is depicted on the *z*-axis and the gene rank based on the enhancer load on the *x*-axis. Data from genes without the H3K4me3 mark within ±1000 bp of the transcription start site are not shown. The nine ENCODE cell lines are shown across the *y*-axis. A set of 6167 disease genes were considered based on the DisGeNET version 2.1 associations (minimum association score of 0.08). For each cell line, the proportion of disease genes among all genes with each unique enhancer load observed was calculated. This proportion is represented by the color gradient on the enhancer load for each cell line. (**C**) Heatmap depicting the statistical significance of the enrichment for disease associated genes on all the different bins of genes based on their enhancer load (adjusted $-\log_{10}P$-value), across nine cell lines. For each cell line, genes were grouped based on the number of enhancers into deciles, and genes without enhancers grouped separately. To avoid different bins having genes with the same enhancer load, the deciles were adjusted to contain all genes with the same number of enhancers as the last gene in the decile. Using the set of 6167 disease associated genes derived from the DisGeNET, hypergeometric tests for each bin were performed. The statistical significance is indicated by the color gradient. Values below 1.301 (i.e. adjusted *P*-values larger than 0.05) are shown in gray and not considered significant. The enrichment significance is the highest in the top enhancer load bin in all nine cell lines.

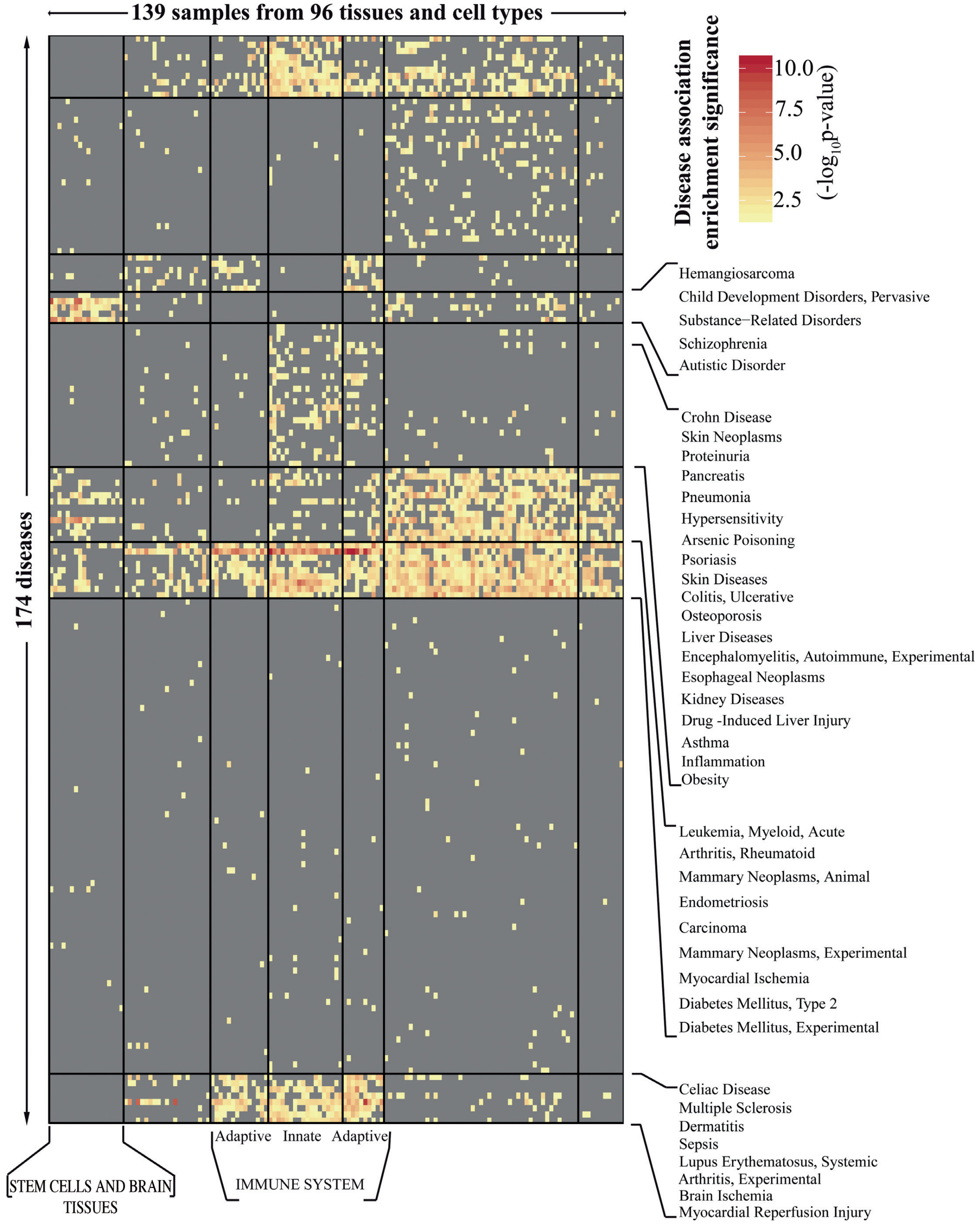


Figure 4. Cell type-selective disease-association of genes under high regulatory load. Heatmap showing the statistical significance (adjusted $-\log_{10}P$-value) of the disease association enrichment of the high enhancer peak load genes across 139 samples. For each of 139 samples, the set of genes with highest enhancer load (top 10% bin) was taken to perform hypergeometric enrichment tests for disease association on 340 diseases (disease associated genes from DisGeNET version 2.1, minimum 15 genes with a minimum score of 0.08 per disease). The significance of each test is represented as adjusted $-\log_{10}P$-value for the 139 samples (columns) across 174 diseases (rows), as indicated by the color gradient. Values below 1.301 (i.e. adjusted P-values larger than 0.05) are shown in gray and not considered significant. 166 diseases did not have an adjusted $-\log_{10}P$-value of at least 1.301 in any of the 139 samples. The R package 'blockcluster' (24) was used to perform the clustering for samples and diseases resulting in the observed pattern. Supplementary Figure S4 shows the same heatmap with names of all samples and diseases included and Supplementary File 4 contains the details of the diseases and samples as ordered in the heatmap.

Supplementary File 4). A total of 174 diseases showed significant enrichment (adjusted $-\log_{10}$ P-value $\geq$ 1.301) in the high regulatory load genes of at least one cell type. Figure 4 shows bi-clustering of the diseases and cell types or tissues according to the enrichment profiles. As expected, cell types are clustered together largely according to their function or developmental origin. For the different diseases the clustering patterns are not as obvious but still interesting clusters emerge. The largest cluster (second from the bottom) consists of 76 various diseases that are fairly weakly enriched in only one or a few different cell types or tissues. On the contrary, only very few diseases (mostly in the third disease cluster from bottom) showed enrichment in almost all cell types. These include many systemic diseases or syndromes such as type 2 diabetes and rheumatoid arthritis or broad categories related to cancer such as carcinoma and leukemia. Among the different cell types, the enrichments in the high regulatory load genes of the immune cells included many different diseases. Diseases like multiple sclerosis and systemic lupus erythematosus were particularly enriched for cells of both innate and adaptive immune systems while other autoimmune and inflammatory diseases, including Crohn's disease and asthma as well as acute inflammations, induced for example by pneumonia and drug-induced liver injuries, were preferably enriched in high regulatory load genes of the innate immune cells. Finally, the most selective disease enrichments were observed for the high regulatory load genes of the different brain regions and the closely clustering pluripotent stem cells. Most of these showed enrichments mainly for the disease groups such as pervasive child development disorders, substance-related disorders, schizophrenia and autistic disorder. Finally, among the cell types with a particularly low number of disease-associations, pancreatic islet was associated to only seven different diseases, with the most significant disease-association to type 2 diabetes. Such selective disease associations might reflect the highly specialized functions of the cell types like islet cells and stem cells, but might also reflect the fact that relatively little is still known about the disease mechanisms in tissues like brain.

In summary, the genes under high regulatory load vary between different cell and tissue types and, consistently, are enriched for different diseases in different cell types, often in accordance with known involvement of those cell types in the respective diseases. Therefore, identification of genes under high regulatory load using epigenomic data for active enhancers could guide identification of novel disease-associated genes in a cell-type-selective manner.

Identification of novel putative disease genes in human monocytes

Among the 139 samples of enhancer data the cell type with most samples are the monocytes that are innate immune cells involved in a wide range of diseases. To test the prediction of novel disease genes based on their regulatory load, we combined the high regulatory load genes from 10 monocyte samples to obtain an extensive list of 3131 monocyte high regulatory load genes. Next we compared this list to high regulatory load genes in all other samples in order to obtain a unique list of 82 monocyte-specific high regulatory load genes (Supplementary File 6). From these genes 25 were already included as disease-associated genes in DisGeNET version 2.1 (from 5th of May 2014) used in our analysis above, and 15 of them were associated to diseases with known involvement of monocytes or cell types derived from them (e.g. arthritis, pycnodysostosis, myeloid leukemia and properdin deficiency). This leaves 57 monocyte-specific high regulatory load genes that we expect to have higher probability of being associated with disease, especially in monocytes (Supplementary File 6).

After the initial submission of the manuscript a new version of DisGeNET (version 3.0, May 2015) was released, including 767 novel high confidence disease genes (cut-off score of 0.2 including only strong evidence associations), 710 of which are included in the gene background set used for our epigenomic analysis (Supplementary File 1). Searching for the 57 predicted monocyte disease genes described above among the 710 newly associated disease genes showed that as many as 14 of them had now been included as high confidence disease genes during the year between the two releases. These include genes such as *NUSAP1* and *MS4A6A* that are associated to glomerulonephritis, an IgA nephropathy (34); *GPBAR1* that is highly expressed in intestinal monocytes of patients with inflamed Crohn's disease (35), and; *SYNJ1* and *PLD3* that are both associated to Parkinson's disease and Alzheimer's disease (36–39). While the latter two genes have been studied mainly in the context of neurons, both associated neurodegenerative diseases have also a well-established neuroinflammatory component. And interestingly, *PLD3* shows the highest expression across all cell types in monocytes and related cell types, similarly to another non-classical phospholipase D family member, *PLD4*, that is known to be involved in microglial phagocytosis in the brain (40,41).

Finally, to perform a more robust test of the prediction power of high regulatory load for disease-gene association, we tested whether more of the newly associated 710 disease genes from DiGeNET version 3.0 could be found among the high regulatory load genes across all 139 samples used in our analysis. Notably, 469 or 66% of the new disease genes could indeed be found among the high regulatory load genes across the analyzed cell types, a significantly higher fraction than expected by chance (hypergeometric test, P-value = 1.6880e-12). Thus, arguing that high regulatory can guide identification of novel disease-associated genes.

Comparison of high regulatory load and super-enhancer genes

High regulatory load from multiple active enhancer peaks is conceptually very similar to previously described super-enhancers or stretch-enhancers that have also been associated to disease through high occurrence of disease-associated genetic variants within them (10,11). To compare high regulatory load genes with super-enhancer genes we used the 35 Epigenomics Roadmap samples for which mapped reads of H3K27ac ChIP-Seq data were available to call super-enhancer peaks in those samples (see Methods for details). This yielded between 300 and 900 super-enhancer genes per sample. Overlapping these genes with previously identified high regulatory load genes from the same samples showed that in all cases the majority (on average 67.9%) of the super-enhancer genes belong also to the group of high regulatory load genes (Figure 5A). However, these make up only 13–37% of all high regulatory load genes.

As expected, also super-enhancer genes were enriched for disease-association in all tested cell types (Figure 5A). This led us to wonder if the observed disease-associations for high regulatory load genes are simply due to the included super-enhancer genes. To address this possibility we generated separate lists of high regulatory load genes that exclude super-enhancer genes in all 35 samples and tested these genes for their disease-association enrichment. Importantly, in each case the remaining high regulatory load genes enriched for disease-association also when the super-enhancer genes were excluded from the analysis (Figure 5A).

Given that high regulatory load genes are associated with high number of active enhancers it could be assumed that they are also higher expressed than other genes on average. Consistently, this has already been shown to be the case for super-enhancer genes (9). To test this for high regulatory load genes we obtained normalized RNA-seq data for 38 cell types and tissues for which they were available from the Epigenomics Roadmap consortium. In keeping with the hypothesis, the high regulatory load genes showed approximately 2.1-fold higher expression levels than all genes on average (Supplementary File 7). And looking at all known disease genes, they too exhibited approximately 1.65-fold higher expression levels. This was mainly based on the two largest disease categories called ‘Biomarkers’ and ‘Genetic Variation’ which both showed the same average expression levels while the other smaller categories all showed even further elevated levels of expression between 2.65- to 3.2-fold above the average of all genes.

Based on these results we asked whether the high regulatory load genes could be obtained simply by focusing on the highest expressed genes in each cell type. To do this we grouped the genes in each sample according to their expression depending whether they were in the top 10%, top 30% or top 50% of highest expressed genes or in top 90% group containing most genes. Next we asked how large proportion of the high regulatory load genes in each cell type could be found in each group. As shown in Figure 5B, on average across the cell types, only 16.6% of high regulatory load genes could be found among the comparably sized top 10% of highest expressed genes. And only when considering the higher expressed half of all genes (top 50%) could 76.7% majority of high regulatory load genes be obtained.

Taken together, the majority of super-enhancer genes can be identified among the genes with high regulatory load, but they do not alone explain the observed disease-association enrichment of high regulatory load genes. Similarly to super-enhancer genes, both high regulatory load and disease genes show above average expression levels but expression level alone serves as a poor predictor of high regulatory load.

High regulatory load genes are not associated to disease only by genetic variation

As much as 90% of disease-associated genetic variants are located outside of coding genic sequences in humans and recent work integrating epigenomic analysis with GWAS has showed that around 60% of the variants are coinciding with active enhancers (6–8,19). This is particularly true for super-enhancers that serve as binding platforms for combinations of multitude of TFs (10,11). While the observed enrichment of disease genes among the genes under high regulatory load is not only due to super-enhancer genes, it might still be due to increased likelihood of these genes being associated to genetic variants.

In order to assess whether this is sufficient to explain our findings, we divided all protein coding genes into three categories: (i) genes not associated to any disease with a score above 0.08 according to DisGeNET database (13 071 genes); (ii) genes associated to diseases based on evidence for genetic variation (score $\geq$ 0.08; 2832 genes), and; (iii) genes associated to diseases based on other evidence than genetic variation (score $\geq$ 0.08; 4596 genes). Subsequently, enrichment of each of these gene sets in the high regulatory load genes of all 139 samples was tested and the boxplots of the adjusted enrichment *P*-values are depicted in Figure 5C. Importantly, the genes not associated to any disease also did not show any enrichment in any of the samples while genes associated to diseases through genetic variation showed significant enrichment in all samples with a median adjusted $-\log_{10}$ *P*-value of 6.1. However, also the other disease-associated genes, without evidence for genetic variation, showed a highly significant enrichment among all 139 sets of high regulatory load genes with a median adjusted $-\log_{10}$ *P*-value of 9.0. Thus, suggesting that there could be also other explanations for the frequent disease-association of the high regulatory load genes besides their higher likelihood of being affected by a genetic variation.

High regulatory load genes are involved in multiple pathways

The positioning of disease genes as central hubs in gene-regulatory or protein-protein interaction networks has been suggested to make the genes more likely to cause or be affected by perturbations than what would be the case for more peripheral genes (42). Indeed, one of the putative explanations for the higher occurrence of disease association among the genes under high regulatory load could lie within their role as central network nodes and as integration points within and between pathways. To see whether this hypothesis is supported by the current pathway knowledge we obtained the node information for all KEGG pathways (27) and calculated in how many pathways the high regulatory load genes occur on average in each of the 139 samples (Figure 6A). This was compared to the average pathway occurrence of an equal number of randomly selected H3K4me3 marked genes from each sample. Interestingly, in 135 of the 139 samples the average pathway occurrence was significantly higher (4.66 pathways per HRL gene on average) than for the randomly selected genes (3.52 pathways per gene on average) based on a re-sampling test (see Methods for details) with a large variation up to almost 6 pathways per gene in some cell types. Consequently, the

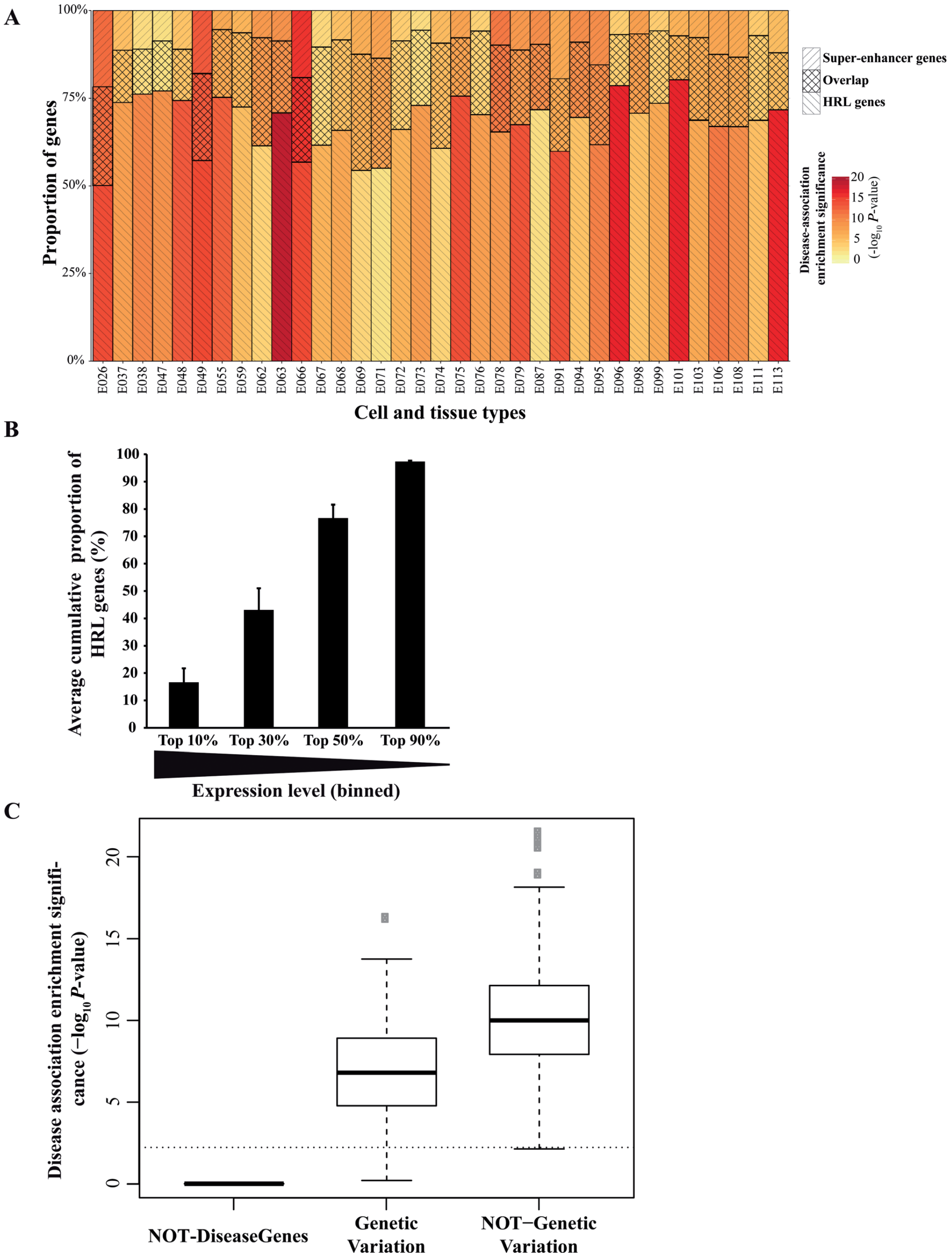


Figure 5. Features of the disease association of high regulatory load genes and comparison to super-enhancer genes. (**A**) Proportions of high regulatory load genes (descending diagonal stripes), super-enhancer genes (asscending diagonal stripes) and their overlap (crossing diagonal stripes) from their combined count in 35 samples from Epigenomics Roadmap consortium are indicated as cumulative bars. Code name for the each sample corresponds to those described in (21) and can be found in Supplementary File 2. Heatmap shows the statistical significance (adjusted $-\log_{10}P$-value) of the disease association enrichment of either the super-enhancer genes (upper part of the bar) or high regulatory load genes without super-enhancer genes (lower part of the bar). (**B**) Average cumulative proportion (±SD) of high regulatory load genes within the top 10%, top 30%, top 50% and top 90% of highest expressed genes across 38 RNA-Seq samples from Roadmap Epigenomics consortium corresponding to samples detailed in Supplementary Files 2 and 7 (see Methods for details). (**C**) Boxplot showing the statistical significance (adjusted $-\log_{10}P$-value) of the enrichment for disease association of the top enhancer load bin from each sample (n = 139) obtained considering the 2832 genes for which the association to a disease is defined as 'genetic variation' (based on the DisGeNET) *versus* the 4596 genes for which the association type is other than 'genetic variation'. Adjusted $-\log_{10}P$-value values below 1.301 (gray dashed line), i.e. P-values larger than 0.05 are considered non-significant. While disease-associated genes based on genetic variation enrich on the top enhancer load bin, this enrichment is not lost when excluding those genes and keeping disease-associated genes based on other types of association evidence based on DisGeNET ('altered expression', 'biomarker', 'post-translational modification' and 'therapeutic'). The set of 13 071 genes in our background set that are not disease associated was used as a control, showing no significant enrichment of non-disease genes among the genes with more enhancer peaks across all 139 samples.

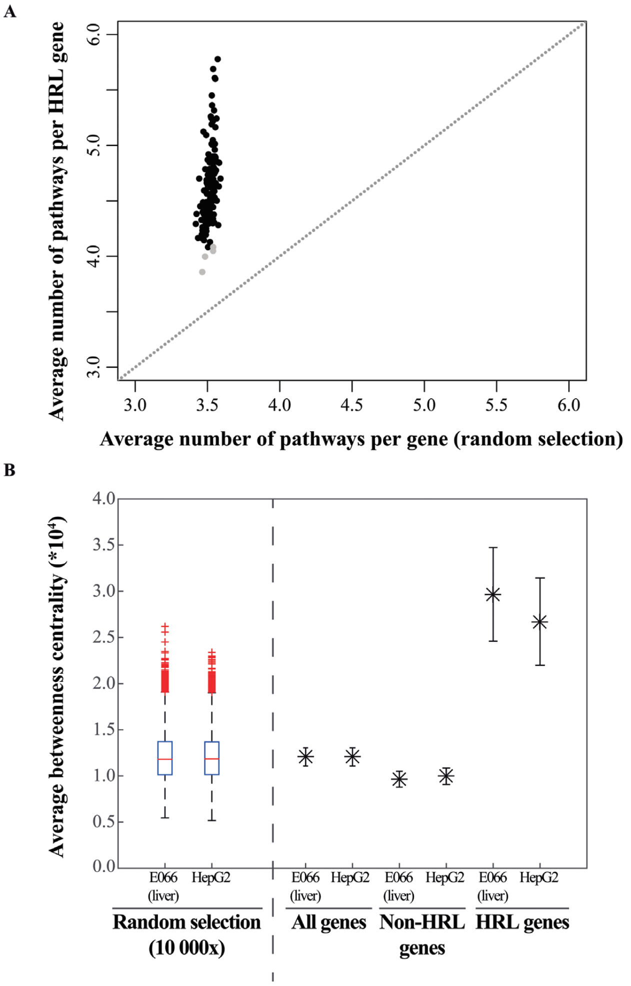


Figure 6. High regulatory load genes appear on average in more pathways and exhibit higher betweenness centrality than randomly observed. (**A**) Plot of the average number of pathways per gene for the high regulatory load genes across 139 samples (one dot per sample) as a function of the average number of pathways per randomly selected equal number of genes. While randomly selected genes appear on average in 3.52 pathways (constant number of pathways, the variation over the *x*-axis is very low), high regulatory load genes appear on average in 4.66 pathways and present a much higher variation (min. 3.86, max. 5.78), suggesting their importance as network nodes. 282 KEGG pathways were considered. For each set of highest enhancer load genes from the 139 samples, the average number of KEGG pathways they belong to was calculated (*y*-axis). Equal numbers of randomly selected genes were taken in a 10 000 fold re-sampling and the average number of pathways they belonged to are depicted on the *x*-axis. The statistical significance was determined by a re-sampling test and the significant ($P \leq 0.05$) and non-significant samples are shown as black and gray dots, respectively. (**B**) Betweenness centrality of genes under high regulatory load in liver disease gene network (for illustration see Supplementary Figure S5). A liver disease gene network was constructed as described in Methods and the betweenness centrality was calculated for each gene present in the network and potentially expressed in either liver sample based on the H3K4me3 mark. Boxplots (left side of the dashed line) represent the distribution of average betweenness centralities for 10 000 sets of equal numbers of randomly selected network genes. Asterisks (right side of the dashed line) represent the average betweenness centralities (±SEM) of all genes, all genes except those under HRL in primary liver tissue (E066) or HepG2 cell line and, high regulatory load genes in the two different samples as indicated. The average betweenness centrality of high regulatory load genes is significantly higher than for other genes as determined by a re-sampling test.

high regulatory load genes occur in more known pathways than other genes on average, suggesting that the identified disease-association enrichment could be due to central role of these genes within biological networks.

Genes under high regulatory load in liver exhibit high betweenness centrality in liver disease gene network

To directly address the positioning of high regulatory load genes in biological and disease networks, we constructed a liver disease-specific network covering 137 liver diseases that comprises of 3775 genes (nodes) and 8278 interactions (edges) based on human protein interactions from the Human Protein Reference Database (HPRD) (29) (see Methods for details of the network construction). An illustration of the network with positioning of high regulatory load genes can be found in Supplementary Figure S5. Next we obtained the lists of all H3K4me3 marked and high regulatory load genes in two liver samples, primary liver tissue and HepG2 hepatocarcinoma cell line, that were also present in the newly constructed liver disease gene network. As additional control, we created 10 000 lists of random selection of genes of equal numbers from both samples. Finally, to analyze the positioning of the high regulatory load genes we calculated the betweenness centrality for each gene in the network and compared the average betweenness centralities of the high regulatory load genes to the different control gene lists. Notably, while randomly selected genes showed similar mean betweenness as all genes, the high regulatory load genes showed in both samples almost 3 times higher betweenness than either of these control groups (Figure 6B). This is consistent with the somewhat lower betweenness centrality of the gene group where high regulatory load genes have been excluded. Accordingly, high regulatory load genes occupy the more central nodes within the liver disease gene network.

Genes under high regulatory load at transcriptional level have longer 3′ UTRs and contain more miRNA binding sites

Since high regulatory load genes appear to function as important nodes in biological pathways and integrate multiple signals at the transcriptional regulation level, we asked whether a similar finding could be made also at the other regulatory levels. More specifically, we assumed that high enhancer load genes might be under higher regulatory load also at the post-transcriptional level. Post-transcriptional regulation of mRNA stability and translation takes place mainly via the binding of miRNAs and various RNA-binding proteins to their regulatory regions in the mRNAs 3′ UTR with longer 3′ UTRs allowing higher number of regulatory regions (43,44). To test whether the 3′ UTRs of genes under high regulatory load from multiple enhancers in different cell types could in principle occupy more regulatory regions than all genes on average, we collected the 3′ UTR lengths for all genes and compared the 3′ UTR lengths in the different gene sets (Figure 7A, see Methods for details). Curiously, in 138 of the 139 samples the 3′ UTR length distribution was significantly longer for the top bin of highest enhancer load genes than for all genes (Kolmogorov-Smirnov test). The average 3′ UTR length for

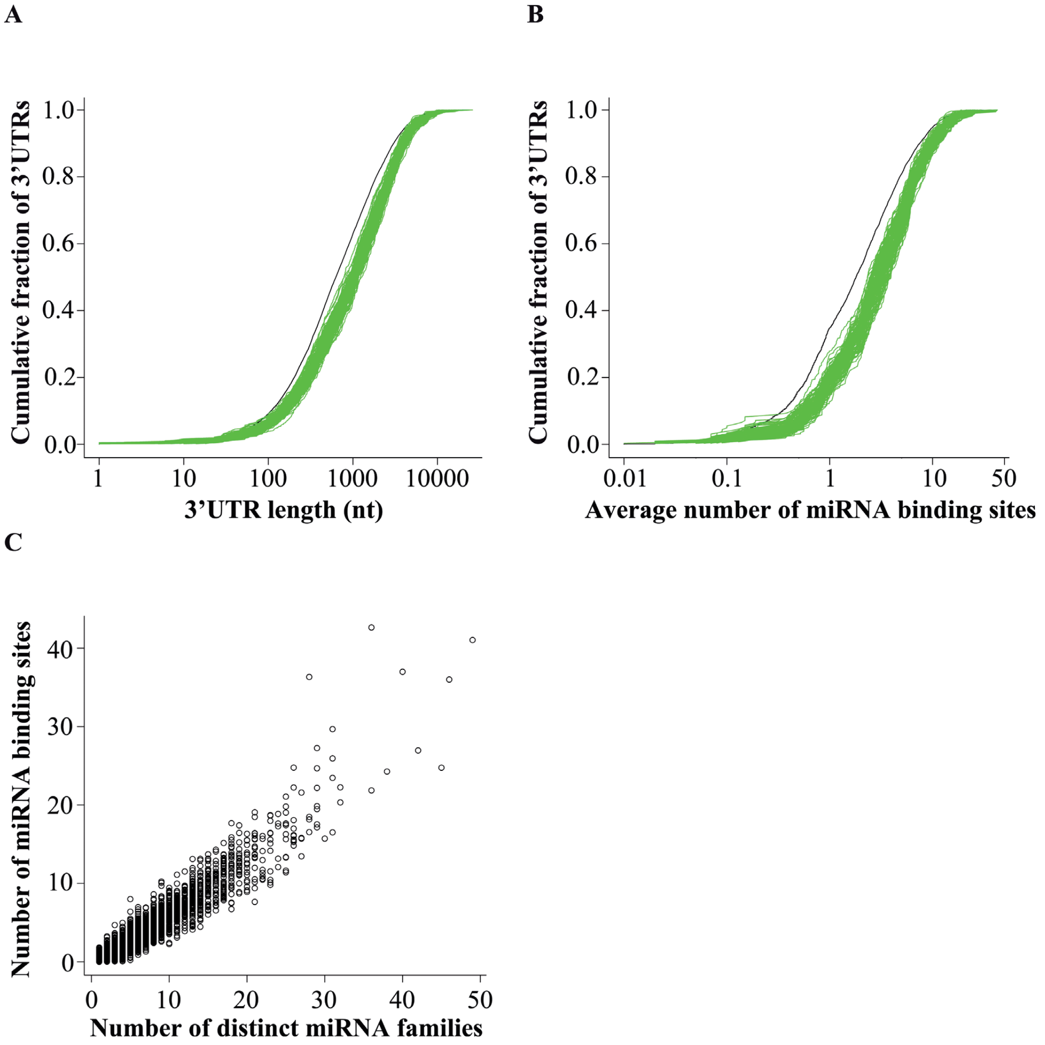


Figure 7. Genes under high regulatory load at transcriptional level have longer 3′ UTRs. (**A**) Distributions of 3′ UTR lengths in 139 sets of high enhancer load genes from different samples (each depicted by a green line) and in a background set of 16 307 3′ UTRs (depicted by the black line). The average 3′ UTR length of all mean lengths of the high enhancer load genes was 1695 nt, 39% longer than the average length of 1213 nt for the background set genes. For 138 samples the length was significantly longer than for the background set (Kolmogorov-Smirnov test, see methods). (**B**) Distributions of counts of predicted conserved miRNA binding sites (TargetScan 6.2) in 139 sets of high enhancer load gene 3′ UTRs from different samples (each depicted by a green line) and in a background set of 16 307 3′ UTRs (depicted by the black line). (**C**) The total number of predicted miRNA binding sites per 3′ UTR (*y*-axis) is positively correlated with the number of distinct miRNA families targeting the 3′ UTR (*x*-axis) across all genes.

all genes was 1213 nt while mean of all high regulatory load genes means was 1695 nt, i.e. 482 nt (or 39%) longer. To further see whether these longer 3′ UTRs indeed contain more regulatory regions, we analyzed the distribution of conserved miRNA binding sites predicted by the TargetScan software (31) within the 3′ UTRs (Figure 7B). In keeping with the longer length, the 3′ UTRs of the high enhancer load genes from each of the 139 sample sets contain significantly more miRNA binding sites than other genes on average, making them more prone to post-transcriptional regulation. In general, across all genes, the increased number of miRNA binding sites strongly correlates with the number of miRNAs from distinct miRNA families (Figure 7C), suggesting that the observed high number of miRNA binding sites also reflects targeting by multiple different miRNA families. Thus, the high regulatory load genes appear to be under combinatorial regulation by distinct regulators mediating multiple signals both at transcriptional and post-transcriptional level.

DISCUSSION

Much of the research is focused on elucidating the roles of selected few genes in human health and disease although this emphasis is not warranted by their connectivity, conservation or other features when compared to the less studied genes (1). The advent of different genome-wide approaches has allowed an improved ‘equality’ among genes and unbiased approaches to prioritize the previously uncharacterized genes based on these vast data sets will be increasingly important. Here we show that genes regulated by a high TF load are more likely to be disease-associated genes and can be identified across cell types through epigenomic mapping of active enhancers. The sets of high regulatory load genes vary between cell types, thereby allowing identification of putative disease-associated genes in a cell type-selective manner. Disease-association of these genes appears to rely on multiple different categories of association evidence and we propose central role within biological networks as one of the likely explanations for the observed enrichment. In keeping with the putative role as integrators of multiple signals between pathways, the high regulatory load genes appear also to be targeted by more post-transcriptional regulators such as miRNAs. This is consistent with earlier findings for positive correlation between numbers of TF and miRNA binding sites (45), and provides an additional feature that could be shared by the most relevant genes.

High load of active enhancers often assumes high expression levels of the target genes, a concept already suggested by many studies (9,10; Supplementary File 7). Therefore it is somewhat paradoxical why these genes would also be targeted by higher number of post-transcriptional regulators, such as miRNAs, that are mainly repressing their target genes. One possibility is that the miRNA regulation serves as a buffer to keep the abundant expression of the target genes within certain threshold in a robust manner (46). On the other hand, it is known that miRNAs and their target mRNAs are expressed in a mutually exclusive manner, suggesting that the high regulatory load genes could be under strong miRNA-mediated repression in other cell types where they are not occupied by high enhancer load, thus further enforcing their selective expression profiles (47). Consistently, multiple different miRNA binding sites might be needed to allow the repression of the genes by different miRNAs in different cell types.

While analyzing the 3′ UTR lengths we observed that also the coding sequences (CDS) of the high regulatory load genes are longer than the mean of all genes, albeit with smaller (24%) and less significant increase (Supplementary Figure S6A). And importantly, the unspliced primary transcripts are as much as 94% longer (Supplementary Figure S6B). This raises the possibility that these are simply longer genes occupying larger genomic regions, with the higher regulatory association at transcriptional level stemming from this feature. However, the 3′ UTR and CDS lengths of the different genes show no correlation and similar results for 3′ UTR lengths can be obtained when focusing only on enhancers or TF binding sites located upstream of the target genes (Supplementary Figure S6C and data not shown). Therefore, the longer 3′ UTR and indeed an overall longer gene length appear to be inherent features of the high regulatory load genes. This is particularly interesting in the light of the recent observation that human orthologs of mouse essential genes are significantly longer than all other genes on average (48). Indeed, 77% of the 2472 known essential genes with human orthologs are also identified as high regulatory load genes in our analysis and significantly enriched in the top regulatory load bins across all 139 samples (data not shown).

Our data suggest that epigenomic mapping of active enhancers could be used to predict disease-associated genes and thereby prioritize the analysis of previously unknown genes. Current analysis presented in Figure 4 provides an interesting starting point. More detailed analysis of the individual cell types and associated disease enrichments might provide novel insights into relationship of cell types and diseases in question, and in particular, how do the previously unassociated high regulatory load genes within different cell types fit into the network of the already known disease genes. To take the first step we already performed an analysis to identify monocyte-specific high regulatory load genes that could be novel disease genes and show this to be the case for 14 of them. Moreover, genes like *PLD3,* that has been linked to neurodegenerative diseases and studied in the context of neurons, is identified as monocyte-specific high regulatory load gene in our analysis. This suggests that PLD3's association to neurodegenerative diseases might be related to neuroinflammatory component of these diseases, similarly as has been shown for many Alzheimer's disease-associated genetic variants that are enriched in enhancer regions active in inflammatory cells (49).

On the other hand, the enrichment of disease genes associated to many systemic diseases across the high regulatory load genes of most cell types further highlights the need to find interventions to these diseases at whole-body level. Moreover, obtaining epigenomic data from diseased cell types or cells responding to different external signals could provide further interesting target genes for future analysis. In particular, the profiling of previously uncharacterized disease related cell types such as dopaminergic neurons in context of Parkinson's disease could reveal entirely new insights into the underlying epigenetic mechanisms of the disease development (50).

Our comparison of high regulatory load and super-enhancer genes (Figure 5) suggests these features to be two sides of the same coin and a less exclusive definition of these key genes might be beneficial for future analysis. The high enhancer load of the selected genes is largely a reflection of binding of multiple TFs in the regulatory regions of these genes as indicated by the correlations in Figure 2A. Similarly, Joshi has found TF hotspots to be enriched for enhancers and consistently, Siersbak *et al.* have shown super-enhancers to be enriched for TF hotspots (12,51). These findings together with the increased occurrence of these high regulatory load genes in more pathways than expected by chance and with increased betweenness centrality within liver disease gene network (Figure 6) lead us to propose the central role of these genes in regulatory networks as a possible explanation for their increased likelihood for disease association. The high regulatory load genes appear to serve as integration points within and between pathways, possibly also at the post-transcriptional level (Figure 7). Indeed, recent work by Hnisz *et al.* showed embryonic stem cell super-enhancers to consist from several constituents that together serve as binding platforms for a number of TFs to merge signals from multiple signaling pathways (52).

In conclusion, the central role of high regulatory load genes as signal integrators comes with an inherent feature of high enhancer load that can be taken advantage of to identify the genes through epigenomic profiling in a cell type-selective manner. In the future, an integrative approach using high regulatory load together with other features such as network centrality, post-transcriptional regulation, and expression data could be used to prioritize the previously unstudied genes in terms of their relevance for disease.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr. Aurélien Ginolhac, Dr. Nicolas Beaume and Dr. Petr Nazarov for support with bioinformatic analysis and statistics and Prof. Rudi Balling for inspiring discussions. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg (http://hpc.uni.lu). This study makes use of data generated by the ENCODE project (https://www.encodeproject.org/), the BLUEPRINT Epigenome project (www.blueprint-epigenome.eu) and the NIH Epigenomics Roadmap Consortium (http://nihroadmap.nih.gov/epigenomics/). A full list of the investigators who contributed to the generation of the data from BLUEPRINT project is available from www.blueprint-epigenome.eu. Funding for the BLUEPRINT project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 BLUEPRINT.

Author Contributions: M.G., T.S. and L.S. conceived the project and designed the analysis. M.G. performed all the analysis except for the 3′ UTR and miRNA analysis (performed by PB), and liver disease gene network construction and betweenness centrality calculation (performed by T.P.N.). All authors commented on the manuscript.

FUNDING

University of Luxembourg and Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg, National Research Foundation of Luxembourg (FNR) [AFR 1276195 and AFR 9139104 to M.G. and T.P.N.]. Funding for open access charge: University of Luxembourg.

Conflict of interest statement. None declared.

REFERENCES

1. Pandey,A.K., Lu,L., Wang,X., Homayouni,R. and Williams,R.W. (2014) Functionally enigmatic genes: A case study of the brain ignorome. *PLoS One*, **9**, e88889.
2. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
3. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2009) Comparative Toxicogenomics Database: A knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.*, **37**, D786–D792.
4. Bauer-Mehren,A., Rautschka,M., Sanz,F. and Furlong,L.I. (2010) DisGeNET: A Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, **26**, 2924–2926.
5. Bauer-Mehren,A., Bundschus,M., Rautschka,M., Mayer,M., Sanz,F. and Furlong,L. (2011) Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases. *PLoS One*, **6**, e20284.
6. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, **337**, 1190–1195.
7. Vernot,B., Stergachis,A.B., Maurano,M.T., Vierstra,J., Neph,S., Thurman,R.E., Stamatoyannopoulos,J.A. and Akey,J.M. (2012) Personal and population genomics of human regulatory variation. *Genome Res.*, **22**, 1689–1697.
8. Farh,K.K.-H., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shoresh,N., Whitton,H., Ryan,R.J.H., Shishkin,A.A. *et al.* (2014) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
9. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
10. Parker,S.C.J., Stitzel,M.L., Taylor,D.L., Orozco,J.M., Erdos,M.R., Akiyama,J.A., van Bueren,K.L., Chines,P.S., Narisu,N., Black,B.L. *et al.* (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17921–17926.
11. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
12. Siersbæk,R., Rabiee,A., Nielsen,R., Sidoli,S., Traynor,S., Loft,A., Poulsen,L.L.C., Rogowska-Wrzesinska,A., Jensen,O.N. and Mandrup,S. (2014) Transcription Factor Cooperativity in Early Adipogenic Hotspots and Super-Enhancers. *Cell Rep.*, **7**, 1443–1455.
13. Pasquali,L., Gaulton,K.J., Rodríguez-Seguí,S.A., Mularoni,L., Miguel-Escalada,I., Akerman,I., Tena,J.J., Morán,I., Gómez-Marín,C., van de Bunt,M. *et al.* (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.*, **46**, 136–143.
14. Galhardo,M., Sinkkonen,L., Berninger,P., Lin,J., Sauter,T. and Heinäniemi,M. (2014) Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. *Nucleic Acids Res.*, **42**, 1474–1496.
15. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
16. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
17. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
18. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
19. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
20. Martens,J.H.A. and Stunnenberg,H.G. (2013) BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, **98**, 1487–1489.
21. Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
22. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
23. Heinäniemi,M., Nykter,M., Kramer,R., Wienecke-Baldacchino,A., Sinkkonen,L., Zhou,J.X., Kreisberg,R., Kauffman,S. a, Huang,S. and Shmulevich,I. (2013) Gene-pair expression signatures reveal lineage control. *Nat. Methods*, **10**, 577–583.
24. Bhatia,P., Iovleff,S. and Govaert,G. (2014) blockcluster: An R Package for Model Based Co-Clustering. https://hal.inria.fr/hal-01093554/file/BlockCluster.pdf.
25. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple

Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.

26. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
27. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucl. Acids Res.*, **28**, 27–30.
28. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.
29. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
30. Joy,M.P., Brock,A., Ingber,D.E. and Huang,S. (2005) High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, **2005**, 96–103.
31. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
32. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
33. Rada-Iglesias,A., Bajpai,R., Swigut,T., Brugmann,S.A., Flynn,R.A. and Wysocka,J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283.
34. Hodgin,J.B., Berthier,C.C., John,R., Grone,E., Porubsky,S., Gröne,H.-J., Herzenberg,A.M., Scholey,J.W., Hladunewich,M., Cattran,D.C. *et al.* (2014) The molecular phenotype of endocapillary proliferation: novel therapeutic targets for IgA nephropathy. *PLoS One*, **9**, e103413.
35. Yoneno,K., Hisamatsu,T., Shimamura,K., Kamada,N., Ichikawa,R., Kitazume,M.T., Mori,M., Uo,M., Namikawa,Y., Matsuoka,K. *et al.* (2013) TGR5 signalling inhibits the production of pro-inflammatory cytokines by in vitro differentiated inflammatory and intestinal macrophages in Crohn's disease. *Immunology*, **139**, 19–29.
36. Quadri,M., Fang,M., Picillo,M., Olgiati,S., Breedveld,G.J., Graafland,J., Wu,B., Xu,F., Erro,R., Amboni,M. *et al.* (2013) Mutation in the SYNJ1 gene associated with autosomal recessive, early-onset parkinsonism. *Hum. Mutat.*, **34**, 1208–1215.
37. Drouet,V. and Lesage,S. (2014) Synaptojanin 1 Mutation in Parkinson's Disease Brings Further Insight into the Neuropathological Mechanisms. *Biomed Res. Int.*, **2014**, 289728.
38. Satoh,J.-I., Kino,Y., Yamamoto,Y., Kawana,N., Ishida,T., Saito,Y. and Arima,K. (2014) PLD3 is accumulated on neuritic plaques in Alzheimer's disease brains. *Alzheimers. Res. Ther.*, **6**, 70.
39. Cruchaga,C., Karch,C.M., Jin,S.C., Benitez,B. a, Cai,Y., Guerreiro,R., Harari,O., Norton,J., Budde,J., Bertelsen,S. *et al.* (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, **505**, 550–554.
40. Mabbott,N.A., Baillie,J.K., Brown,H., Freeman,T.C. and Hume,D.A. (2013) An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, **14**, 632.
41. Otani,Y., Yamaguchi,Y., Sato,Y., Furuichi,T., Ikenaka,K., Kitani,H. and Baba,H. (2011) PLD4 is involved in phagocytosis of microglia: Expression and localization changes of PLD4 are correlated with activation state of microglia. *PLoS One*, **6**, e27544.
42. Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, **29**, 150–159.
43. Matoulkova,E., Michalova,E., Vojtesek,B. and Hrstka,R. (2012) The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.*, **9**, 563–576.
44. Cheng,C., Bhardwaj,N. and Gerstein,M. (2009) The relationship between the evolution of microRNA targets and the length of their UTRs. *BMC Genomics*, **10**, 431.
45. Cui,Q., Yu,Z., Pan,Y., Purisima,E.O. and Wang,E. (2007) MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochem. Biophys. Res. Commun.*, **352**, 733–738.
46. Mukherji,S., Ebert,M.S., Zheng,G.X.Y., Tsang,J.S., Sharp,P.A. and van Oudenaarden,A. (2011) MicroRNAs can generate thresholds in target gene expression. *Nat. Genet.*, **43**, 854–859.
47. Stark,A., Brennecke,J., Bushati,N., Russell,R.B. and Cohen,S.M. (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3UTR evolution. *Cell*, **123**, 1133–1146.
48. Georgi,B., Voight,B.F. and Bućan,M. (2013) From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.*, **9**, e1003484.
49. Gjoneska,E., Pfenning,A.R., Mathys,H., Quon,G., Kundaje,A., Tsai,L.-H. and Kellis,M. (2015) Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer/'s disease. *Nature*, **518**, 365–369.
50. Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
51. Joshi,A. (2014) Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs. *BMC Bioinformatics*, **15**, 6591.
52. Hnisz,D., Schuijers,J., Lin,C.Y., Weintraub,A.S., Abraham,B.J., Lee,T.I., Bradner,J.E. and Young,R.A. (2015) Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Mol. Cell*, **58**, 325–370.

Cell type-selective disease-association of genes under high regulatory load

Mafalda Galhardo, Philipp Berninger, Thanh-Phuong Nguyen, Thomas Sauter and Lasse Sinkkonen

Supplementary Information

Supplementary Figures

Supplementary Figure S1: TF load enriches similar bins of enhancer load. Heatmaps of the hypergeometric distribution enrichment significance (adjusted $-\log_{10}$p-value) of genes binned by TF load (y-axis) across different bins based on enhancer load (x-axis) for 9 ENCODE cell lines (from top to bottom, left to right: HUVEC, HCT116, MCF7, A549, HeLaS3, H1hESC, HepG2, K562 and GM12878). Genes were sorted by regulatory load and grouped in bins (y- and x-axis). Bottom or left side bins contain genes with lower regulatory load than top or right side bins. Bins denoted with "1" contain genes with no associated TF or enhancer. The enrichment significance (adjusted $-\log_{10}$p-value) is depicted by the color gradient, increasing from yellow to red (values $\geq$ 50 appear in red). Dark grey represents (adjusted $-\log_{10}$p-value) < 1.3, not considered significant. The significance is evident along the diagonal for all 9 cell lines, denoting the concerted increase between TF and enhancer load, with genes in bins of low TF load enriching the highest for genes in bins of low enhancer load and vice-versa, genes of high TF load enriching the most for high enhancer load genes. Supplementary File 5 contains tables with the enrichment significance (adjusted $-\log_{10}$p-value) obtained and the exact TF load per bin (vertically) or enhancer load per bin (horizontally) for each of 9 cell lines.

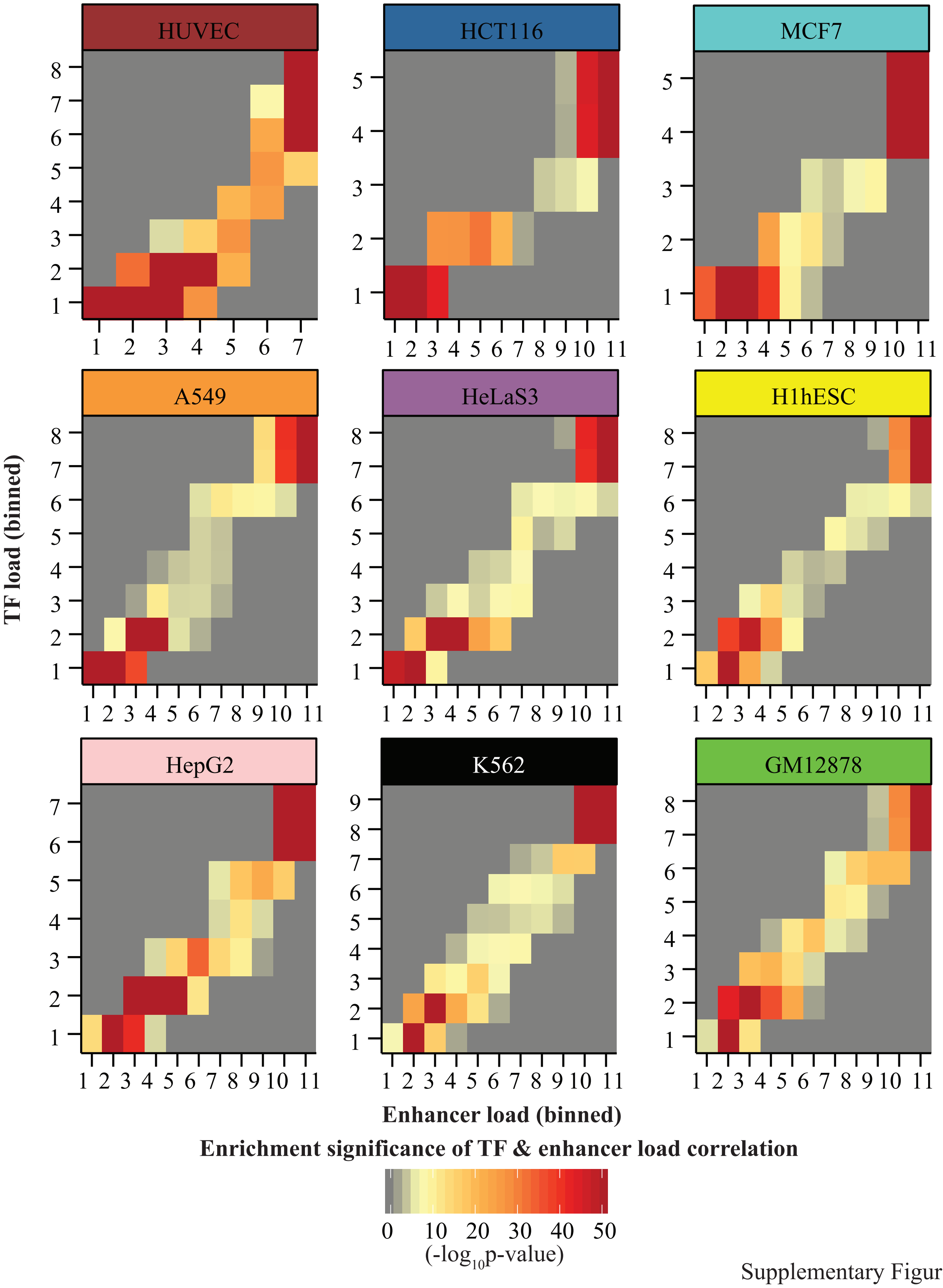
Supplementary Figure S2: High enhancer load genes enrich for disease association also with more stringent disease gene groups. Heatmaps of the hypergeometric distribution enrichment significance (adjusted $-\log_{10}$p-value) of genes binned by enhancer load across 139 samples. Left side bins contain genes with lower enhancer load than bins on the right side. The enrichment significance (adjusted -$\log_{10}$p-value) is depicted by the color gradient, increasing from yellow to red. Grey represents (adjusted -$\log_{10}$p-value) < 1.3 (equivalent to p-value > 0.05), not considered significant. The significance is evident on the bins of highest enhancer load on the right side, with orange and red colours. (A) Results using the set of curated disease genes from DisGeNET version 2, minimum association score of 0.2 (7110 genes of which 5853 were in the background set of 19238 protein coding genes). (B) Results using the set of disease genes from the OMIM database, as of June 2015 (4557 genes of which 3483 were in the background set of 19238 protein coding genes).

Supplementary Figure S3: Genes with highest enhancer load vary across 139 samples. Heatmap of the Jaccard similarity index for the pair-wise comparison of genes in the top enhancer load bin across 139 samples. The heatmap is mirrored along the diagonal. Blue denotes few common while red denotes many common genes on the two sets of highest enhancer load genes from any two samples. The predominance of the blue colour reflects an overall low similarity between the genes with highest enhancer load across samples (average similarity lower than 30%). The bottom and right-side color bars denote groups of samples with the same tissue of origin, color-coded on the bottom.

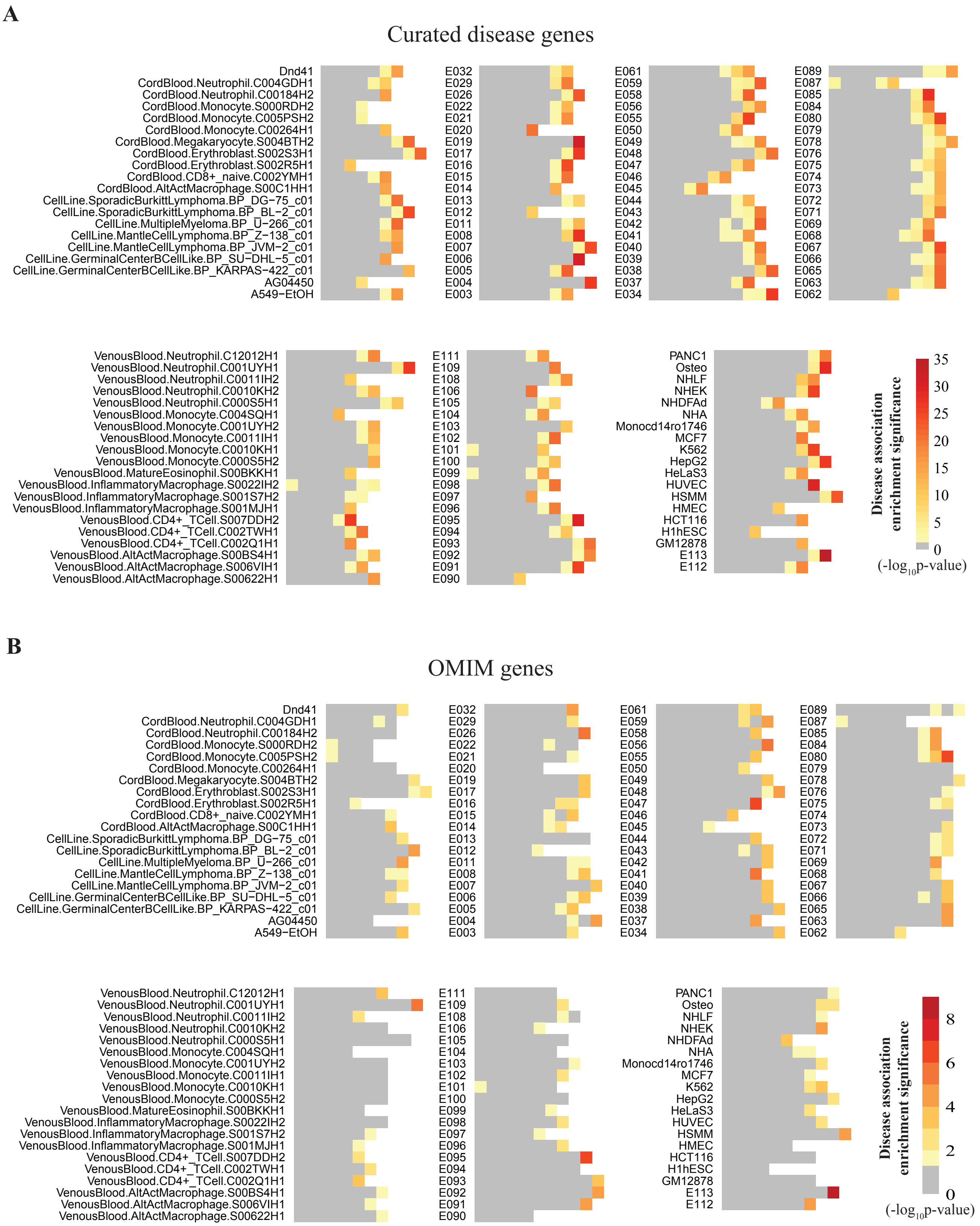
Supplementary Figure S4: Cell type-selective disease-association of genes under high regulatory load. Heatmap from Figure 4 showing the statistical significance (adjusted -log10 p-value) of the disease association enrichment of the high enhancer load genes across 139 samples and 174 diseases with names of the diseases and samples written out for each case. For more details, see Supplementary File 4.

Supplementary Figure S5: Liver disease gene network. Illustration of the reconstructed liver disease gene network containing 3,775 genes and 8,278 interactions. Red nodes represent the high regulatory load genes from the two liver samples (primary liver (E066) and HepG2) and grey nodes the other liver disease genes and their first neighbours in the network. A higher intensity of red color is observed on the central area of the network, reflecting the higher betweenness centrality of the high regulatory load genes as described in Figure 6.

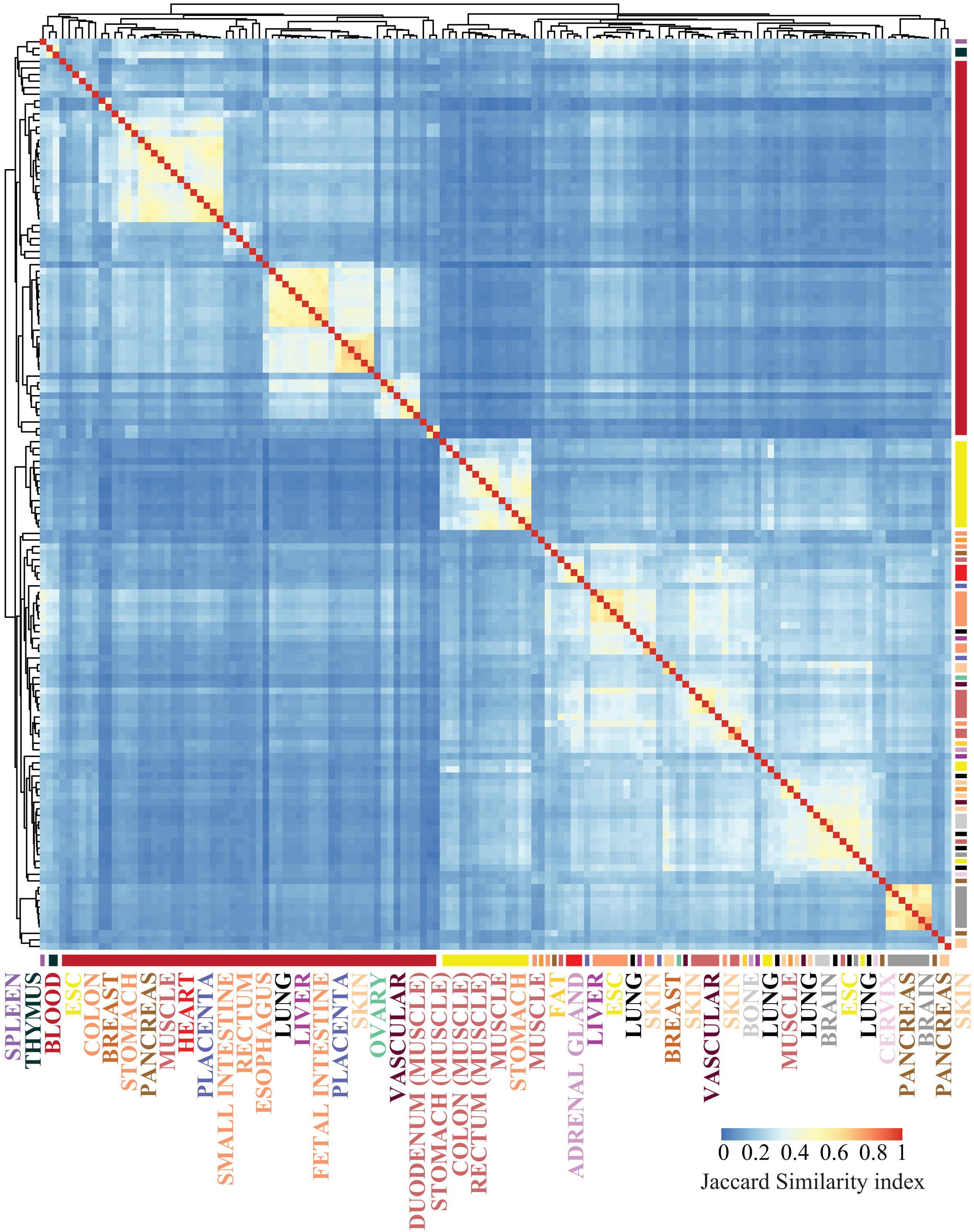
Supplementary Figure S6: HRL genes have longer CDS and transcripts on average. (A) Distributions of CDS lengths in 139 sets of high enhancer load genes from different samples (each depicted by a green line) and in a background set of 16307 CDSs (depicted by the black line). The average CDS length of all mean lengths of the high enhancer load genes was 1816 nt, 24% longer than the average of length of 1455 nt for the background set genes. (B) Distributions of unspliced transcript lengths in 139 sets of high enhancer load genes from different samples (each depicted by a green line) and in a background set of 16307 unspliced transcripts (depicted by the black line). The average unspliced transcript length of all mean lengths of the high enhancer load genes was 105452 nt, 94% longer than the average length of 54451 nt for the background set genes. (C) The CDS lengths (y-axis) the 3'UTR lengths (x-axis) of the transcripts do not show correlation.



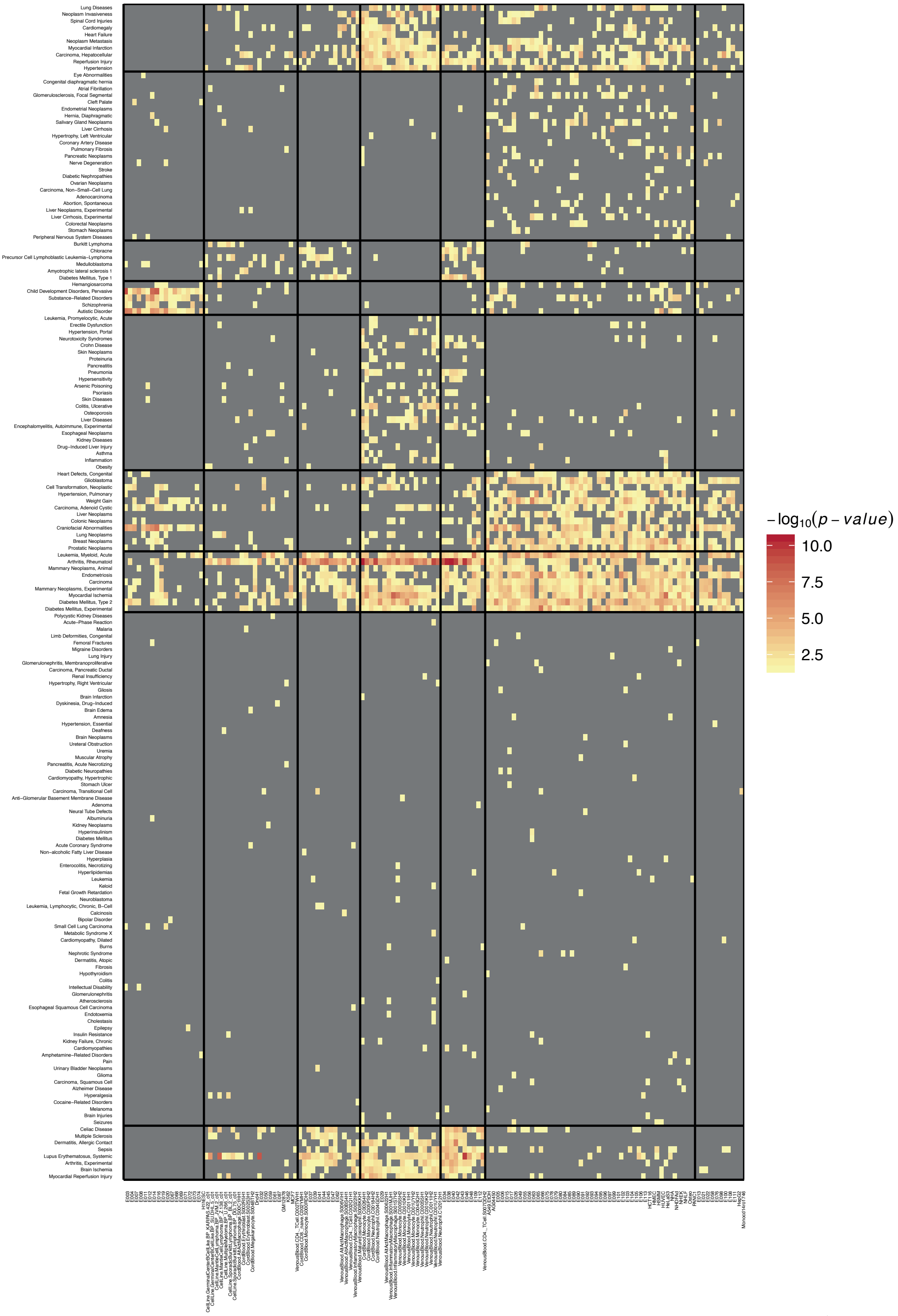
Supplementary Figure S1



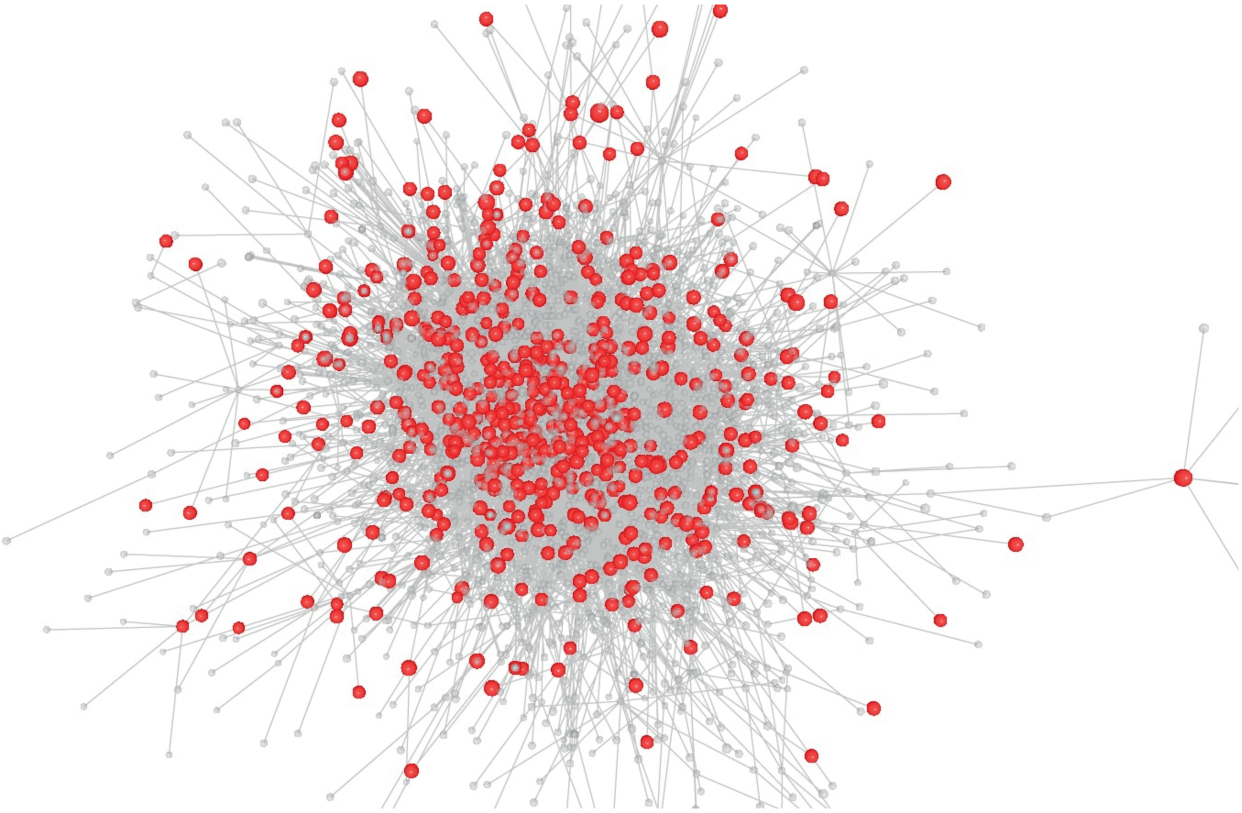
Supplementary Figure S2



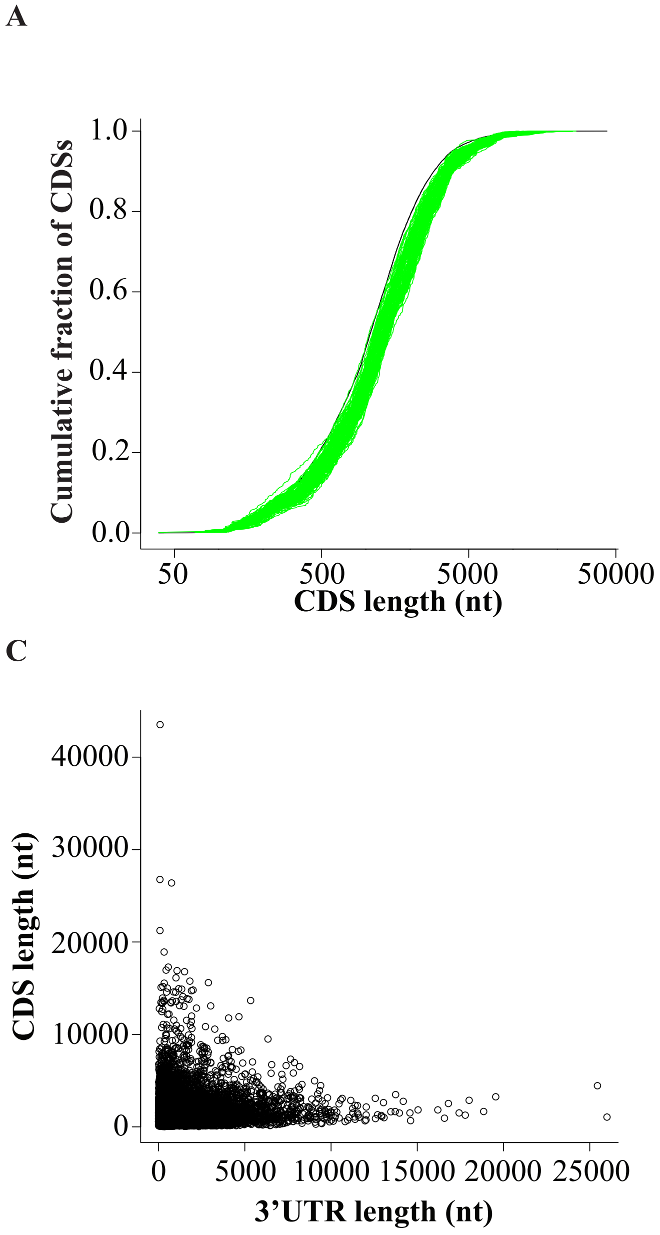
Supplementary Figure S3



Supplementary Figure S4



Supplementary Figure S5



Supplementary Figure S6

4.4 Manuscript III - "IDARE2 - Simultaneous visualization of multi-omics data in Cytoscape"

To facilitate data integration with network visualization, IDARE was upgraded into **IDARE2**, a versatile tool to i) automatically generate image metanodes from diverse user provided inputs and ii) mapping the generated metanodes onto Cytoscape networks through a Cytoscape app with the capability of disentangling large networks into connected sub-networks that are easier to inspect and interpret. Moving between sub-networks can be done by clicking on linker nodes.

This work was mainly undertaken by Thomas Pfau (thomas.pfau@uni.lu) and Jake Lin (jake.lin@uta.fi). My contribution to the work includes the sharing of the initial ideas, discussions and refinements during the generation process and extensive testing of the metanode generation tool.

Manuscript III is integrally presented starting from page 148.

IDARE2 - Simultaneous visualization of multi-omics data in Cytoscape

Thomas Pfau [1,2], Jake Lin [3], Mafalda Galhardo [1] and Thomas Sauter [1,*]

[1]Life Science Research Unit, University of Luxembourg, 162a, Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg

[2]Institute of Complex Systems and Mathematical Biology, University of Aberdeen, Meston Building, AB24 3UE, Aberdeen, United Kingdom

[3]BioMediTech, University of Tampere, Biokatu 8, 33520 Tampere, Finland

Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Summary: Visual integration of experimental data in metabolic networks is an important step to understand their meaning. With genome scale metabolic networks containing thousands of reactions this task becomes increasingly difficult. While databases like KEGG and BioCyc provide curated pathways which allow a navigation of the metabolic landscape of an organism, it is rather laborious to map data directly onto those pathways. There are programs available using these kind of databases as a source for visualisation, however these programs are then restricted to the pathways available in the database. Here, we present a way to overcome these limitations allowing the researcher to visualise multiple data types and time scales in a non-hairball network with linker nodes between subnetworks. The tool can be applied for visualisation of data on any type of network and is not restricted to biological networks.
Availability: http://idare-server.uni.lu
Contact: thomas.sauter@uni.lu

1 INTRODUCTION

With the ever increasing amount of 'omics' data it becomes increasingly important to handle and combine multiple sources of data and interpret the experimental findings. Generally, bioinformatic and statistical processing of omics data yields a set of individual targets, often leaving an open gap to interpretation. Placing those targets into context and visualizing the measurements is key to obtain ideas about the effects of a given treatment. With IDARE (Galhardo et al.(2014)) we introduced an approach to combine numerous sources of information and visualize them in the context of metabolic networks. The concept allows the simultaneous interpretation of multiple experiments and simulations in the context of biological networks, and thus provides an integrative way of visual inspection. Here we extend this concept and provide a convenient possibility for automatically generating images for various types of data and provide an app to incorporate the images into Cytoscape (Shannon et al. (2003); Saito et al. (2012)) networks. With the increasing complexity and completeness of network definitions these networks tend to become very dense, making the visual inspection difficult. Therefore, we provide in addition a method to create connected subnetworks within cytoscape in a way similar to the connections available in KEGG (Kanehisa et al. (2014)) or BioCyc (Caspi et al. (2014)). This allows for a better inspection of data in the view of large networks, while preserving the network structure.

2 RESULTS

The present tool is divided into two distinct parts. The generation of multiomics node images is implemented in MATLAB and available via a webserver. The node generation is run on the high performance computing facility (Varrette et al. (2014)) of the University of Luxembourg, and can thus handle a large amount of simultaneous requests. The second part is the cytoscape application which provides integration of the generated nodes into a cytoscape network, along with several convenience functions.

1. Generation of multiomics visualisation. The IDARE2 webserver (found at http://idare-server.uni.lu) provides a clear interface that allows users to upload processed sets of data along with their description. From the datafiles the webserver then generates image nodes combining the different datasets automatically. The user can select from several ways of data representation within the nodes, including heatmaps, time series, simple itemized representation and graphs. The implementation provides interfaces that allow an easy addition of further data types which the authors are happy to create on request. The input format for all types of representation, except graph representation, is the same, thus allowing the user to quickly create multiple different layouts for data. Input data has to be provided as excel sheets or as tab separated value files. To address unintuitive identifiers (e.g. entrez gene ids for gene nodes) which are often seen in metabolic networks, the user can provide two IDs for each entry, with one used for matching the nodes and the other being presented in the image node (label). As a proof of versatility, the IDARE2 tool successfully reproduces the nodes manually generated in the original IDARE publication (Figure 1) and can be applied to other data without effort. In addition to the image nodes,

[*]to whom correspondence should be addressed

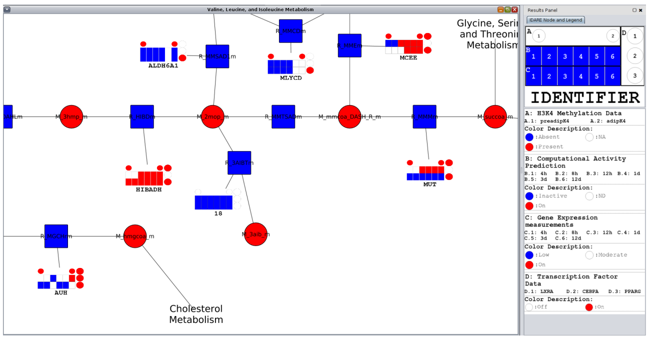


Figure 1. The data used in the original IDARE paper (Galhardo et al. (2014)) visualized within cytoscape using automatically generated nodes. The application automatically generates legends describing the nodes based on the information provided in the data files and during upload.

automatically generated Legends explain the layout of any generated node. This allows the use of inhomogenous datasets, in which multiple node types will be generated and legends for each node type, which can be especially useful, if experimental data is not available for all genes of interest. After submission of the data, the user will be informed upon successful generation of the image nodes and the nodes can then be downloaded. Further information on file formats and additional details can be found in the user guide (http://idare-server.uni.lu/UserGuide.pdf).

2. Cytoscape Application. The cytoscape application consists of two main functionalities:

1. Mapping of the generated image nodes
2. Generation of connected subnetworks from a large network

The first functionality will use the identifiers provided during node generation to map the created images to nodes in the cytoscape network. To allow a greater flexibility, the user is asked which column of the network to use to map the images. This mapping is associated with a specific visualization style that includes the mapping functions. The second functionality of the application allows researchers to disentangle the hairball structure often prevalent in complex networks. Subnetworks can be generated based on common properties of nodes. Both methods are not restricted to metabolic networks, and while the image mapping is generally applicable, the subnetwork generation requires the network to represent a bipartite graph. This allows the application to diverse types of networks ranging from metabolic networks (metabolites and reactions), to social networks (individuals and groups) and others. The user will be asked to select a component type and an interaction type, which are used to determine subnetwork boundaries. It is assumed that interactions can directly belong to a subnetwork (e.g. metabolic pathway), while compounds can be shared. Since the hairball structure is often associated with highly connected compounds, the application allows the selection of compounds which should be excluded from the subnetworks (thus disentangling them). To retain the overall structure of the network, linker nodes are created between compounds shared by interactions belonging to different pathways. This allows the user to follow the network structure and easily identify connections between different subsystems. The links can be followed by double clicking and opening the target network view (if existing). Finally, the app provides a utility function which allows the user to easily add gene and protein nodes to metabolic networks from SBML files. The app is able to read gene-associations provided in the common COBRA style (Schellenberger et al. (2011)) and can interpret bioql annotations for enzymes and genes (like provided by e.g. HMR (Mardinoglu et al. (2014))).

Availability. The IDARE multiomics node generation webserver is freely available at idare-server.uni.lu. Images are generated on the HPC cluster of the University of Luxembourg (Varrette et al. (2014)). The cytoscape application can be downloaded from the cytoscape app store or directly from: idare-server.uni.lu/IDARE.zip. A User Guide is also available on the website which gives further usage details. The image generation software is implemented in MATLAB and available on request.

ACKNOWLEDGEMENT

The authors would like to thank Sylvain Julien and Susann Mudrich for their comments and the HPC Sysadmins for their help in configuring the distributed node generation.

Funding: TS and TP are funded by the Life Science Research Unit, University of Luxembourg. MG is supported by the National Research Foundation of Luxembourg (FNR) (AFR 1276195) and the Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg. JL is funded by the University of Tampere.

Conflict of Interest: none declared.

REFERENCES

Galhardo M, Sinkkonen L, Berninger P, et al. Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. *Nucleic Acids Res* (2014), 42(3), 1474–1496.

Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integratedmodels of biomolecular interaction networks. *Genome Research* (2003), 13(11), 2498–2504.

Saito R, Smoot ME, Ono K, et al. A travel guide to cytoscape plugins. Nat Methods (2012), 9(11), 1069–1076.

Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in kegg. Nucleic Acids Res (2014), 42(Database issue), D199–D205.

Caspi R, Altman T, Billington R, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res (2014), 42(Database issue), D459–D471.

Varrette S, Bouvry P, Cartiaux H, et al. (2014). Management of an academic hpc cluster: The ul experience. In Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014), Bologna, Italy. IEEE.

Schellenberger J, Que R, Fleming RMT, et al. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. Nat Protoc (2011), 6(9), 1290–1307.

Mardinoglu A, Agren R, Kampf C, et al. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat Commun (2014), 5, 3083.

5 Discussion and perspectives

5.1 Overview

The work described in this thesis is centred on the links between genome regulation, metabolism and disease, investigated with integrative approaches. The interplay between the genome, signalling and metabolism orchestrates such diversity of cellular responses and is altered upon disease. Studying this interplay and its (dys)regulation is highly relevant for understanding the pathophysiology of an organism. First, we focussed on the interplay between the gene regulatory and metabolic networks during adipocyte differentiation, a phenotypic change with considerable impact in metabolism and homeostasis, and related with several diseases including metabolic syndrome and obesity. In a second case, we explored the relation between regulatory load and association to disease in order to assess whether the transcriptional control of a gene could predict disease association, thereby contributing to improve current disease gene prioritization routines. Additionally, we looked for properties of the high regulatory load genes differentiating them from other genes, which could relate to the observed link to disease. In both cases, integration facilitated extracting insights from the data relationships, and our systems biology approaches represent one step closer into truly integrative and comprehensive analyses, which we believe are necessary to improve our understanding of human functioning in health and disease, and will become routine in the coming years. Thereby, our integrative analysis allowed the simultaneous visualization of multiple *omics* data types, the extraction of links between multiple objects and biological processes and the understanding of their relationships. In a third case, IDARE2 was presented, a tool envisioned precisely to aid on integrating multi-*omics* data into metanode images and rendering them on biological networks for quick visualization and inspection of properties and relationships. A tool providing such capability is highly useful due to the vastness and complexity of biological processes and networks, with many components and interactions which make it unfeasible for human perception to understand without visual support.

In **"Manuscript I"**, we sought an integrated study of adipocyte differentiation with focus on the interplay between the gene regulatory and metabolic networks, highlighting several dyslipidemia genes to be under combinatorial regulation by the TFs PPARγ, CCAAT/enhancer-binding protein (C/EBP)α or liver X receptor (LXR)α and the miRNAs miR-27a, miR-29a or miR-222. Already here, we interrogated whether the regulatory load was related with disease association, observing an enrichment for vascular-disease-associated genes among metabolic genes with most TFs in a dataset of 10 TFs in HUVEC cells. We believe such integrative approach is more adequate to studying biological processes than very focussed or single factor approaches, which fail to expose how the different components interact within a system and should be decreasingly employed for the study of complex biological phenomena.

In light with work prior and contemporary to **"Manuscript I"** [536–547, 557–560], first published online in November 2013, our analysis is comprehensive and innovative, with both expanded content diversity and integration level.

In regards to content, we experimentally collected data covering a time course of gene expression during adipocyte differentiation, observed to be highly dynamic; the genome-wide binding profiles of three key adipogenic TFs, PPARγ, C/EBPα and LXRα , in adipocytes, observing between $\geqslant 2000$ and $\geqslant 10000$ putative target genes, a number attesting for their prominent regulatory role in adipogenesis; the genome-wide profile of the H3K4me3 histone modification mark in pre-adipocytes and adipocytes, with little observed changes, likely due to the fact that SGBS pre-adipocytes are already committed into differentiating only to adipocytes; and the target genes of the miRNAs miRNA-27a, miRNA-29a and miRNA-222, with a few dozens of key metabolic genes involved in lipid metabolism presenting combinatorial regulation by the miRNAs and TFs.

Overall, our analysis spanned multiple layers from the regulation of transcript levels in adipogenesis.

Combining the experimental data with metabolic modelling (CBM), we used the time course gene expression and the method of Shlomi *et al.* [516] to predict the metabolic activity that parallels the transcriptional cascade largely orchestrated by PPARγ and C/EBPα and ultimately leading to lipid-loaded mature adipocytes. Such prediction of metabolic activity associated with a cellular response (here differentiation) remains still relatively little employed by the scientific community, while technically relatively easy and feasible in most cases. As current proteome activity and metabolic flux measurements do not cover a large portion of the components, metabolic modelling provides a valuable resource for assessing metabolic activity based on already known relationships and newly acquired data, possible to test and improve with additional experiments, namely metabolomics, and to predict perturbation outcomes and generating hypothesis about observed phenomena.

In regards to data integration, we then condensed all data into metabolic pathways, providing an integrated view of adipogenesis through the representation of custom gene metanodes with the different gene-related data and metabolic predictions for the reaction activities from pre-adipocytes to adipocytes embedded on the pathways. This representation allows to percept the flow of metabolic changes upon differentiation and to inspect the network distribution and convergence of the above mentioned regulators on metabolic genes (`http://systemsbiology.uni.lu/idare.html`), being considerably more effective than long lists, tables and individual plots in exposing events, players and relationships.

Such data integration and visualization lead to the perception of a shared and often combinatorial regulatory load on dyslipidemia genes in adipocytes, prompting us to assess the relation between the gene regulatory load and the disease association in a larger dataset. For this purpose, we took public ChIP-seq data from the genome-wide binding of 10 TFs on HUVEC cells, observing an enrichment for association with vascular diseases among metabolic genes with between 6 and 9 TFs ("Manuscript I", page 62), suggesting that genes relevant for disease are under high regulatory control.

Based on this evidence, in order to show that the link between high regulatory load and disease association is general, we then took public TF and active enhancer data from 9 cell lines, the latter including as well data from additional 139 samples spanning 96 tissues and cell types, and tested the enrichment for association with multiple diseases based in DisGeNET, as a function of the regulatory load.

Therefore, in **"Manuscript II"**, we sought the link between the regulatory load of a gene and the likelihood for being disease-associated, observing that genes under higher regulatory load enrich for diseases in a cell type-selective manner across samples, revealing a general principle of intrinsic higher control of key genes, both with TF and active enhancer data, which were positively correlated.

Additionally, we highlighted several properties segregating high regulatory load genes from other genes, such as an average higher participation on KEGG pathways and increased betweenness centrality on a liver disease network for the liver high regulatory load genes, both suggesting HRL genes as signal integrators within biological networks. Of note, HRL genes also presented longer 3'UTRs harboring more binding sites for diverse miRNA families, suggesting a concomitant higher post-transcriptional control. We therefore propose the epigenomic mapping of active enhancers, such as the genome-wide profiling of the H3K27ac mark, as a valuable resource to consider in addition to traditional methods for the prioritization of disease gene candidates, including genome-wide association studies (GWAS) which often provide results with little actionable outcomes.

In light with work prior and contemporary to **"Manuscript II"** [188, 306, 307, 561–563], we present the analysis of a generally larger dataset, covering more cell types than usual, and broader in scope, for instance by considering many diseases and assessing the association to disease beyond genetic variation, as well as exposing other properties of the high regulatory load genes differentiating them from other genes, such as higher average participation on KEGG pathways, betweenness centrality and post-transcriptional regulatory potential. Thereby, we could show a general principle of higher regulation on disease-associated genes, for which further studies to dissect the mechanistic features and evolutionary properties are necessary.

The work presented in **"Manuscript II"** illustrates the importance and usefulness of open access projects, as all the data therein analyzed were obtained from public resources, namely the ENCODE project [12], the BLUEPRINT Epigenome project [316] and the NIH Epigenomic Roadmap project [317]. Indeed, in their integrative analysis of 111 reference human epigenomes [317], the authors focussed on phenotype-associated variants from GWAS studies of diverse traits and disorders and show that "enhancer-associated marks have the greatest ability to distinguish tissue-specific enrichments for regulatory regions, but promoter-, open-chromatin- and transcription-associated marks also have numerous significant enrichments, suggesting that disease variants affect a wide range of processes". This observation brings value to the use of active enhancer data to assess cell type-selective enrichment for disease association as described in **"Manuscript II"**. Furthermore, it suggests that considering additional marks for transcribed (H3K36me3), promoter (H3K4me3 and H3K9ac) and open chromatin (DNase peaks) regions could be useful to capture associations outside annotated enhancer regions. In this way, it would be interesting to perform a similar disease-enrichment analysis ranking genes based on the overall status from enhancer, transcribed, promoter and open chromatin data.

"Manuscript I" and **"Manuscript II"** exemplify how integrative approaches offer the possibility to handle more and varied data while being capable to provide insights about data relationships in better ways, making it easier to understand processes or situations. Currently, visualization tools for multi-*omics* data integration remain few and limited, often requiring a considerable manual effort and the use of multiple tools in order to obtain satisfactory outputs. Further extending this attempt to creating routines for more integrated work, we developed **IDARE2**[(1)], described in **"Manuscript III"**, a tool for automatically generating metanodes depicting user-provided multi-*omics* data including a Cytoscape app for mapping those metanodes into networks that can be collapsed into sub-networks, such as those defined by metabolic pathways, easier to navigate and analyze and inter-exchangeable through clicking on connector nodes. **IDARE2** image metanode generation is extremely versatile, allowing for multiple different input types, which will be automatically arranged into a grid based on user-defined data types and space filling. Upon input data upload, the image generation is done over the HPC facility of the University of Luxembourg [564] without any required human action. After receiving an e-mail notification, the generated metanodes can be downloaded and mapped onto the

[(1)] `http://idare-server.uni.lu/`

respective network using **IDARE2** Cytoscape app, which then allows for data inspection and network analysis, including with other Cytoscape functionalities and apps. Although several freely available tools for mapping data (most often expression) onto pathways exist [527, 565–568], hardly any of them provides the flexibility for data inputs as IDARE2 with custom metanode generation and full integration with network visualization. **IDARE2** is thereby an easy-to-use intuitive tool that will aid researchers visualizing and understanding their data, increasingly complex. We believe that routines such as **IDARE2** custom metanode generation and rapid network visualization will become standard steps in biological data analysis in the near future.

5.2 Adipocyte-related work

Here we studied the differentiation process of adipocytes, known to become impaired with obesity and metabolic syndrome [547]. While the study of adipocyte functioning has provided valuable insights including therapeutic targets, in regards to its association to complex diseases such as obesity, T2DM and metabolic syndrome, in the majority of cases, dietary habits and lifestyle are the major causes for developing such diseases, in particular those habits already during childhood. Therefore, the most effective methods for decreasing the incidence of such diseases include creating awareness among the population and educating for diverse and equilibrated nutrition with abundant exercise.

Currently, more importance goes beyond adipocyte differentiation itself, towards comparing the regulation and metabolism of adipose tissue in lean *versus* obese subjects, with and without T2DM, metabolic syndrome and other related disorders. With clinical data already available at a large-scale, we believe this comparison should be relatively smooth and could provide valuable insights into the regulatory and metabolic alterations occurring during the progression from an healthy lean individual to an overweight diabetic patient, including through simulation of different lifestyles, diet and therapeutics. Our integrated analysis could serve as example or basis for such approaches.

On the interface between cell identity and dynamic adaptation to stimuli, recent work by Fisher *et al.* [569] elegantly exposes the dynamic genomic response of SGBS adipocytes to TNF stimulation mediated by NFκB (RELA, p65 subunit). The authors show an induction of inflammatory genes associated with newly established super-enhancers at the expense of specific genes with cofactor loss from adipocyte super-enhancers. This phenomenon offers a bridge between the inflammatory state and insulin resistance, shedding light into the mechanisms likely taking place in low grade chronic inflammation states such as obesity, in which pro-inflammatory signals lead to adipocyte dysfunction including impaired insulin metabolism [547], possibly through similar mechanisms of adipocyte-super-enhancer silencing upon cofactor redistribution. Furthermore, the authors provide evidence for a cell type-specific repression mechanism in which NFκB selectively redistributes cofactors from high occupancy enhancers, suppressing super-enhancer-associated cell identity genes, which they show also with public data for another four cell lines (A549, IMR90, HeLa and HUVEC), proposing the selective cofactor redistribution from high occupancy enhancers as a general mechanism involved in transcriptional repression associated with activation of signal-dependent transcription factors, namely nuclear receptors. This observation further expands our notion of a dynamic genome regulation, largely revoking a static and rigid idea of chromatin.

5.2.1 Modelling adipocyte metabolism and adipocyte models

During the course of this thesis, two curated adipocyte models were published: one in the context of a multi-tissue model from adipocytes, hepatocytes and myocytes, by Bordbar *et al* [570] in 2011, pioneeringly showing the awareness of the authors for the need of integrative approaches modelling multiple organs; and the second in the context of a comprehensive proteomic dataset used to build an adipocyte-specific metabolic model, by Mardinoglu *et al.* [571] in 2013. Both Bordbar *et al.* and Mardinoglu *et al.* used their models to study clinical cases comparing non-obese *versus* obese diabetic patients and lean *versus* obese patients, respectively.

Therefore, in Bordbar *et al.*, transcriptomic data from the muscle, liver and adipose tissue of diabetic and non-diabetic gastric bypass surgery patients, in fasting state, was obtained and used to build context specific networks for those patients, using GIMME [518], an algorithm mapping transcription data onto a reconstruction and removing reactions associated with absent transcripts, and FVA to further remove reactions that cannot carry flux and determine the flux range of the remaining. The specific multi-tissue models contain between 587 and 705 intracellular reactions and show differences at several pathways and important individual reactions for the obese non-diabetic *versus* obese-diabetic patients, including reactions of metabolites elevated in the blood of diabetic obese, such as fatty acids and lactate, with many active reactions from fatty acid oxidation and carnitine shuttle and inactive lactate dehydrogenase in the hepatocyte and myocyte models of the diabetic obese patients (liver and muscle are unable to utilize lactate as a carbohydrate source).

In Mardinoglu *et al.* proteome data was used to derive a set of proteins and respective enzymes present in adipocytes. Subsequently, transcriptomic data from obese- and non-obese patients (304 patients in total) were used to incorporate differences in lean and obese metabolism in the model (209 female, 95 male) with manual addition of reactions to generate a connected network (in which each reaction can carry flux) which they named iAdipocytes1809, containing 1809 genes, 6160 reactions and 4550 metabolites, a number much larger than those observed in the multi-model of Bordbar *et al.*, including adipocyte-specific metabolic data and a comprehensive review of lipid metabolism with 59 fatty acids. The model was used to predict lipid droplet formation showing impaired metabolism of NEFA and large differences in lipid droplet formation and acetyl-coA metabolism in lean subjects compared to obese subjects.

While both models can be freely obtained, manual curation and ID conversions would be necessary to compare them. A manual reconciliation step to merge these two models also with those we derived for SGBS cells and inspecting to which level the SGBS models capture adipocyte metabolism would be useful to derive a more complete model and to assess to which extent our approach is valuable. Such reconsolidated model could prove useful for developing comprehensive and predictive models for testing metabolic syndrome related clinical cases. Or at least considering the reactions in these models for future studies. Indeed, our SGBS models would need literature curation and additional experimental data in order to validate the presence of the reactions based on gene expression, and also to guarantee no absent reactions occur during adipogenesis. In this context, metabolomics measurements including the uptake rates of exchange metabolites would allow to further constrain the model, while the uptake of metabolites not present in the cell culture medium could be set to a null flux.

5.2.2 BCAAs and adipocytes

The association between obesity, T2DM and insulin resistance is known from long [**Stern1986** , 536, 537, 540], without however having been possible to fully dissect cause and effect order of events, due to the complex setting of these morbidities.

Elevated plasma levels of BCAAs in obese and diabetic patients have been reported for many years [572–574], the causes and mechanisms remaining largely unknown. More recently, the observation that the decrease in the plasma levels of BCAAs following gastric bypass surgery was one of the strongest factors relating with improved insulin resistance further lead researchers to pursue the understanding of the role of BCAAs and their catabolism in the setting of obesity and metabolic syndrome [575, 576].

In 2012, Newgard proposed a model in which BCAAs and fatty acids (FAs) synergize for the setting of metabolic diseases in a high fat diet [577], with a decreased BCAA catabolic capacity by the adipose tissue in obesity (e.g. decreased BCAT2 and BCKDH expression and enzymatic activity, respectively), while increased in muscle, leading to increased plasma levels of the BCAAs and incomplete degradation intermediates, namely C3 and C5 acyl-carnitines with reduced efficiency of fatty acid and glucose oxidation.

As the carbohydrate and fatty acid burden increase with excessive food intake, the progression of a lower BCAA catabolism by adipose tissue increases, further worsening insulin resistance. Newgard also suggests the insulin-sensitizing effects of PPARγ agonists might also relate with its effects in restoring the expression of BCAA catabolic genes.

In our integrated analysis of SGBS adipocyte differentiation, metabolic modelling based on the gene expression between pre-adipocytes and adipocytes predicted an activation of the BCAA pathway upon differentiation. In our *in vitro* setting where the SGBS cells are limited to glucose, amino acids and known supplements from the media, throughout the twelve day course of the differentiation, it is likely that cells take up BCAAs, essential amino acids, for cellular functions, namely protein turnover but also possibly via their catabolism for the generation of acetyl-coA (from isoleucine and leucine) and propionyl-coA (isoleucine and valine), which could be precursors for fatty acid synthesis, and acetoacetyl-coA (leucine), which could be used for cholesterol synthesis, in agreement with a differentiation-induced predicted activation of the cholesterol synthesis pathway based in our metabolic modelling, or an intermediate in ketone body formation.

Interestingly, our integrated analysis of the regulatory and metabolic networks of adipocytes allowed us to realize the combinatorial control by PPARγ and C/EBPα as well as miRNA-29a and miRNA-222 in DBT, member of the BCKDH complex from the BCAA degradation pathway, while other two subunits of the complex, BCKDHB and DLD appear targetted by the three TFs in study, PPARγ, C/EBPα and LXRα, supporting a strong regulation on the BCAA degradation rate limiting step which might well underlie the observed decreased catabolism of BCAAs in adipose tissue of obese individuals.

Interestingly, miRNA-29a has been shown to be up-regulated in diabetic rats [485] with over-expression leading to insulin resistance in 3T3-L1 adipocytes.

In similar lines, more recently, Bagge *et al.* [578] showed on human beta-cells that glucose up-regulates miRNA-29a which in turn decreases glucose-stimula-ted insulin secretion, suggesting the implication of miRNA-29a in the progression from impaired glucose tolerance to type 2 diabetes.

Additionally, miR-222 has recently been reported as a negative regulator of adipogenesis in primary hMSCs [579], in agreement with our observations in SGBS cells.

Therefore, these findings together with our finding of miR-29a targeting DBT (with miR-222) request for an assessment of whether obesity up-regulates miR-29a (and miR-222) in humans, which in turn could underlie the reduced catabolism of BCAAs in the adipose tissue of obese subjects.

5.2.3 Adipocyte-browning could improve energy expenditure

Brown adipose tissue, recently shown to be present and active in small depots of adult humans [412], has acquired much interest as a contributor for body energy expenditure, with raised hopes against obesity and metabolic syndrome.

A curated metabolic model for brown adipose tissue is currently unavailable, and a rigorous model building and curation process to derive a precise model could be very useful to test and simulate the ranges of energy dissipation achieved by activated brown-adipose tissue and the conditions in which it would benefit an individual, a topic which is still largely discussed in the community [580, 581].

Such task was envisaged during the course of the thesis, to compare white and brown adipocyte metabolism and regulation and derive adipocyte models for the two. Due to the scarcity of human brown-adipocyte data available at the time planned for that task and based on interesting results from other work, focus was given to studying the link between high regulatory load and disease association.

Comparing white and brown adipocyte differentiation, their regulation and meta-bolism and link to disease is therefore a very interesting and promising endeavour in regards to finding the potential of brown-adipocytes, or perhaps more indicated, of browning, to increase energy expenditure and reduce the burden from obesity and co-morbidities [424, 582, 583]. The comprehensive adipocyte model by Mardinoglu *et al.* could be a starting point.

5.2.4 Perspectives regarding adipose tissue

Adipose tissue is much more than a storage organ, considerably influencing body homeostasis. Despite many years of research focussing in adipocytes, a resource where to integratively explore and simulate the physiology and metabolism of the adipose tissue in concert with the organism is still missing. As said above, indeed a shift towards multi-tissue or whole organism coupled models is likely to occur in the coming years, allowing to understand how the different tissues influence each other and integrate their functions within the body.

5.3 Work related with high regulatory load genes and link to disease

Here we investigated the link between the transcriptional control of a gene and its likelihood to be disease-associated, taking advantage of public ChIP-seq data from the genome-wide binding of TFs or from the location of active enhancers across multiple cell types. Indeed, we observed a general principle that **genes under higher regulatory control enrich for disease association across cell types**, here shown both via the number of associated TFs and enhancers.

In fact, disease-associated variation converges in regulatory DNA, systematically perturbing TF recognition sequences [25, 584, 585], with 90% locating outside of coding regions and 60% of human autoimmune variants locating within active enhancers of immune cells [561]. In this sense, genes being controlled by more enhancers would have a higher chance of the enhancers containing a SNP and it could be that an accumulation of individual low impact SNPs across regulatory regions, including enhancers, of genes under high regulatory load could synergistically contribute to a disease phenotype, relating with their higher association to disease. Enhancers evolve faster than coding regions and the evidence is for evolution being mostly driven by changes in gene regulation [586–588]. Indeed, the enrichment for disease-associated polymorphisms has been shown for highly interconnected genes in disease specific networks [589], derived from a protein-protein interaction network and differentially expressed genes in 13 complex diseases.

We also observed enrichment for disease association based on genes for which the evidence was not genetic variation, such as “altered expression” and “biomarker”. And when looking at the expression of genes based on the disease association type, they all indeed had an average higher expression compared to the mean of all disease genes, itself already 1.65-fold higher than that of all genes. Therefore, besides (and coupled) with a higher regulatory load, a higher expression also characterizes disease genes.

However, as shown in Figure 5 of **Manuscript II** (page 130), the overlap between top high regulatory load and top highly expressed genes is modest (16.6%). Therefore, a clear next step is to test if highly expressed genes alone enrich for disease association, to test the enrichment with genes that are both HRL and highly expressed and compare them in terms of the individual diseases enriched.

In addition, we also showed a higher average participation in KEGG pathways and a $\geqslant$ 2-fold betweenness centrality (the latter of liver high regulatory load genes in a liver disease network, Figures 6 and S5 of **Manuscript II**, pages 131 and 144, respectively), showing the central role of HRL genes in biological networks possibly as signal integrators [306].

We then looked at the 3'UTR length of HRL genes in comparison to other genes (Figures 7 and S6 of **Manuscript II**, pages 132 and 145, respectively), observing a 39% increase on the average length (across the 139 samples) with more binding sites for diverse miRNA families, suggesting a concomitant higher post-transcriptional regulation. However, this higher 3'UTR length of the HRL genes does not seem to derive from those genes that are also disease-associated, based on an average 3'UTR length of 1695 nt for the HRL genes in all samples *versus* an average of 1650 nt for those HRL genes in all samples that are disease-associated (basically not different).

One could then think of methods to prioritize genes for disease association, and here we highlight the high regulatory load as one promising approach. Gathering and assessing a large panel of other properties, such as high expression, high betweenness centrality, degree or yet other network measures, number of miRNA binding sites and others, could be useful to build a combination of features that would overall better predict disease-associated genes, and improve current outcomes.

When testing the set of high regulatory load genes across samples with a set of human gene orthologs from mouse lethal genes, an enrichment for these lethal orthologs could be observed among the high regulatory load genes, thereby enriched for disease-associated and essential genes.

In 2014, Benayoun *et al.* [590] have shown that the H3K4me3 breadth marks cell identity genes and is associated with higher transcriptional consistency. The concept of H3K4me3 "breadth" is very similar to that used for defining super-enhancers, large clusters of many enhancers characterized by a continuous signal of marks such as Mediator or the H3K27ac. Indeed, as for super-enhancers and cell identity genes, the authors also report broad H3K4me3 deposition on genes essential for the identity and function of a cell type. It would therefore be interesting to know the extent of their overlap and what characterizes their differences in general terms.

In our analysis, we do take into consideration the H3K4me3 mark presence nearby the TSS of a gene in order to include it in further analysis, but unfortunately didn't look at the H3K4me3 breadth as it had not been reported to have a functional role at the time of data processing. Therefore, performing peak calling on the H3K4me3 data using super-enhancer calling settings would allow us to derive the broad peaks of the H3K4me3, which could be compared to those from H3K27ac. Additionally, one could also rank genes based on the H3K4me3 number of peaks, as done for the H3K27ac for the high regulatory load peaks, and perform enrichment for disease association as well. And even taking the set of genes that associated to both broad H3K4me3 and H2K27ac and comparing to previous. These are all very interesting analysis which unfortunately couldn't be concluded in time.

The tool GREAT [532] was used to assign TF or enhancer peaks to genes, based on their “Basal + extension” rule, which defines a basal regulatory domain per gene containing a basal domain from 5 kb upstream and 1 kb downstream from the genes TSS which is extended in both directions to the nearest gene’s basal domain, no longer that 1000 kb. While this setting is likely to capture more realistic scenarios that a one-to-one peak-to-nearest-gene selection, the extent to which it captures real interactions remains unknown. In our setting, by using the H3K4me3 mark to filter out genes in closed chromatin, we take one step into the direction of decreasing false positive associations. In order to obtain experimental evidence for short and long-range regulatory interactions with gene promoters, which could regulate their expression, chromosome conformation capture carbon copy (5C) could be applied, a high-throughput technique to detect looping interactions from chromosome conformation capture (3C) ligation products [591]. Indeed, 5C maps were generated for GM12878, K562 and HeLaS3 cell lines from the ENCODE project, revealing only $\approx$ 7% of looping interactions to be with the nearest gene [321], which clearly points out that associations only to the nearest gene are oversimplified and could fail to reveal regulatory mechanisms.

Comparison to SE-associated genes

Super-enhancers, broad clusters of enhancers, have recently been established to associate with cell identity and key function genes, harbouring transcriptional hotspots with extensive TF co-occupancy and transcriptional activity, and also associated with genetic variants involved in disease [181, 187, 188, 299, 306, 592]. Therefore, a pertinent question would be whether the disease association enrichments are mainly due to SE-associated genes. As shown in Figure 5 A of **Manuscript II**, page 130, this is not the case, with HRL-non-super-enhancer-associated genes still enriching for disease association, with similar significance levels.

On average, 67.9% of super-enhancer-associated genes were included in the top bin of enhancer load from the respective sample, while a lower average of 24.5% of HRL genes are super-enhancer-associated in the respective sample (ranging from 13.5 to 37.9%), a value owing to the lower number of super-enhancer-associated genes (ranging from 300-900) compared to our defined top 10% genes with highest enhancer load (which we considered to define the set of HRL genes per sample, ranging from 1200-1800 genes per sample).

An identical analysis ranking genes based on the length of associated peaks or even combining both the number and length of associated peaks as a surrogate for the enhancer load would be an interesting check for our current results and could prove an asset in the field, being more permissive than the super-enhancer definition, which might be too restrictive based in the enrichments obtained for HRL-non-SE-associated genes, while possibly less affected by technical artifacts or pitfalls.

In fact, the individual disease association enrichments using as input genes associated to super-enhancers reveals a lower number of diseases with a statistically significant enrichment across samples, likely due to the lower set size. In some cases, SE enriched for tissue-related diseases for which statistically significant enrichment with all HRL genes was not achieved, such as for late onset Parkinsons's disease based on super-enhancer data from the substantia nigra.

We observed a low overlap of the genes with highest enhancer load across the 139 samples considered, with an average Jaccard index similarity $<$ 30% for the genes in the top enhancer load bin for any sample pairs, reflecting the cell type relatedness and exclusivity degree of the set of genes with more enhancers in each cell type or tissue, which could then relate with cell type or tissue specific functions. These results are in agreement with those found by the authors of the Roadmap Epigenomics Consortium article describing 111 reference epigenomes (refer to article's Figure 7, page 325), showing a rather cell type-specific subset of enhancers across multiple cell types [317].

Overall, in what compares to SE, HRL genes define a larger set, which could contribute to the observed higher number of diseases enriching per sample based on HRL genes. As testified by a lower overlap of genes in the top enhancer load bin per sample (Jaccard index similarity), the set of HRL genes per sample does conceal cell type uniqueness. We therefore consider the setting of HRL as we used here, or based on peak length as for SE but ranking each gene, or even a combination of peak length and size to rank genes based on the regulatory load, useful as a readout of the likelihood of a gene to be disease-related. Thereby, the HRL could be used for prioritizing novel candidate genes for disease association.

5.4 Perspectives and future work

Developments on metabolic network contextualization methods

Besides the method by Shlomi *et al.*, used within this thesis to predict the metabolic activity of SGBS adipocytes throughout differentiation based on their gene expression, several other metabolic network contextualization methods have been developed (Supplementary file III, non exhaustive listing and description). One of them, published the same year as Shlomi *et al.* is GIMME [518], which tackles the same problem in simpler setting using linear programming instead of MILP, at the cost of requiring an objective function to maximize for, in order to derive a consistent network with highest similarity to the expression data and fulfilling the specified objective function. GIMME was the method used by Bordbar *et al.* to derive their myocyte-hepatocyte-adipocyte model based on expression data and objective functions used as maintenance lower bounds for all cell-specific and multi-tissue type simulations, defined based on published data. As an objective function in context of humans is not easily defined based on the high cellular specialization and diversity of processes, we used Shlomi's method instead of GIMME for deriving SGBS-specific metabolic models based on their gene expression, not requiring an objective function. Nevertheless, we could have adapted human objective functions from published specific models, for instance from HepatoNet [502], in context of the SGBS adipocyte differentiation, or maximizing for triacylglyceride synthesis and lipid droplet formation. As we observed an activation of genes from the cholesterol synthesis pathway, predicted to become active, cholesterol synthesis could also possibly be included into the objective function, including additional experimental measurements to more precisely specify cellular maintenance tasks.

One of the limitations of Shlomi's method is an elevated computation time, increasing with the number of reactions highly and lowly expressed (in regards to those moderately expressed), reaching several hours for networks with few thousand reactions such as Recon1.

Newer methodologies further explore the evidence from *omics* data to derive context-specific metabolic models, based on variable mathematical formulations. Methods relying on simpler linear programming problems are quicker deriving context-specific models from diverse inputs. In particular, FASTCORE [593], developed within our group, remains one of the fastest methods for deriving minimal consistent metabolic models from a generic reconstruction based on a defined set of core reactions.

The accuracy of the resulting models in representing the metabolic activity of the respective cellular system directly depends on the quality and extension of the selected core set. A recent comparison between methods for integration of transcriptomic data into constraint-based models of metabolism highlights that none could robustly outperforming others [594], although a consistent comparison of methods used to predict metabolic activity in human cells is still lacking.

Currently developing methods tackle issues such as the strength of statistical evidence for the expression values, integration with recent technologies such as RNA-seq, siRNA screens and SNP data. Indeed, no unique standard metabolic modelling approach as yet emerged, the choice of methods still largely dependent on the biological system in cause, data type and quality, and of course, biological question.

Nevertheless, FASTCORE-like methods allowing for deriving context-specific sub-networks at the second scale, hold promise to evolve and establish as method of choice within the community, with varied modules tailored for each different data type that can be related to genes, enzymes, reactions or metabolites. In this context, coupling fast prediction of metabolic activity within personalized medicine, monitoring and accompanying the individual throughout the years, could reveal very promising for enhancing treatments, made feasible through FASTCORE-like mathematical formulations.

With the current availability of public expression and multiple other data from diverse clinical settings, including related with metabolic syndrome but also all major types of human diseases, deriving a multi-tissue organ modelling framework, extending Bordbar *et al.* and Mardinoglu *et al.* models, and also including processes beyond intermediate metabolism, such as signalling and transcriptional regulation as well as coupling to clinical data, ultimately containing connected individual models of all body organs and modules for particular processes, should become more and more feasible, and would then allow us to integratively model and simulate human physiology in detail and at an unprecedented scale. Such framework of a "virtual human" model, compiling the well-known facts of human pathophysiology and providing truly systemic simulation of test cases might become standard in the future, possibly even with the generation of one "virtual human" model per person, from birth and continuously updating and integrating all clinical episodes. Such a model could then be tested and simulated in parallel to the person's needs in order to predict affected functions or personalized drug response.

Efforts towards more integrated and global modelling are already ongoing [595–600], including pharmacokinetic and pharmacodynamic integrative frameworks.

In regards to gene prioritization for association to disease, here we present the regulatory load on a gene as a promising readout of the likelihood for a gene to be disease-associated. As we exposed above, finding a combination of features best indicative of this likelihood would possibly greatly help us finding the network of genes mostly contributing for a disease. In regards to the regulatory load, both based on TFs and enhancers, whose binding or location are person and condition dependent, this would imply a case-by-case assessment, in order to obtain a confident readout.

In summary, we present an integrated analysis of human adipogenesis and show that genes under high regulatory load enrich for disease association. In the first case, focus was given to the gene regulatory and metabolic networks, exposing the combinatorial regulation of TFs and miRNAs on lipid metabolism disease genes. In the second case, we show the cell type-selective enrichment for disease of the high regulatory load genes, which appear in more KEGG pathways, have higher betweenness centrality and longer 3'UTR regions with more binding sites for diverse miRNA families than other genes.

Our systems biology approaches allowed us to provide insights about complex processes and could serve as example and be further improved in future studies.

Bibliography

Books

[1] R. Hooke. Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, with Observations and Inquiries Thereupon. Ed. by D. editions Phoenix. 2003rd ed. Courier Corporation, 1665, p. 273. isbn: 0486495647 (cit. on p. 1).

[2] B. Alberts et al. Molecular Biology of the Cell. Ed. by Garland Science - Taylor & Francis Group. 6 th. 2014, pp. 1,464. isbn: 9780815344322 (cit. on pp. 1, 2).

[11] M. Berg, Jeremy et al. Biochemistry. Ed. by Palgrave Macmillan. 8th. WH Freeman, 2015, p. 1120. isbn: 9781319051853 (cit. on p. 2).

[18] B. Palsson. Systems Biology: Properties of Reconstructed Networks. Cambridge Univ Pr, 2006. isbn: 0521859034. doi: `10.2277/0521859034` (cit. on p. 3).

[143] P. W. Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts. Essential Cell Biology. 4th Editio. Taylor & Francis Group, 2013, p. 864. isbn: 9780815344544 (cit. on p. 11).

[405] B. F. Jean-Philippe Bastard. Physiology and Physiopathology of Adipose Tissue. Vol. 28. Springer Science & Business Media, 2012, p. 440. isbn: 2817803434 (cit. on p. 31).

[410] C. Gessner. Historia animalium. Apud Christ. Froschouerum, 1551, pp. 1–1164. doi: `10.5962/bhl.title.68598` (cit. on p. 31).

[533] X.-L. Li and S.-K. Ng. Biological Data Mining in Protein Interaction Networks. Ed. by I. Global. Hershey, PA : Medical Information Science Reference, 2009, p. 450. isbn: 1605663999 (cit. on p. 52).

[534] N. K. Kasabov. Springer Handbook of Bio-/Neuro-Informatics. Ed. by N. K. Kasabov. Vol. 30. Springer Science & Business Media, 2013, p. 1229. isbn: 3642305741 (cit. on p. 52).

[556] R. B. Shiriki Kumanyika. Handbook of Obesity Prevention: A Resource for Health Professionals. Ed. by R. C. B. P. Shiriki Kumanyika PhD, RD, MPH. Springer Science & Business Media, 2007, p. 538. isbn: 0387478604 (cit. on p. 59).

Articles

[3] A Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. Bulletin of the World Health Organization, 79 (8) (2001): 780–90. issn: 0042-9686 (cit. on p. 1).

[4] I. Ezkurdia et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Human molecular genetics, 23 (22) (2014): 5866–78. issn: 1460-2083. doi: 10.1093/hmg/ddu309 (cit. on p. 1).

[5] D. S. Wishart et al. HMDB 3.0–The Human Metabolome Database in 2013. Nucleic acids research, 41 (Database issue) (2013): D801–7. issn: 1362-4962. doi: 10.1093/nar/gks1065 (cit. on pp. 1, 20, 23).

[6] M.-S. Kim et al. A draft map of the human proteome. Nature, 509 (7502) (2014): 575–81. issn: 1476-4687. doi: 10.1038/nature13302 (cit. on p. 1).

[7] M. Wilhelm et al. Mass-spectrometry-based draft of the human proteome. Nature, 509 (7502) (2014): 582–7. issn: 1476-4687. doi: 10.1038/nature13319 (cit. on pp. 1, 2).

[8] F. D. Mast et al. Systems cell biology. The Journal of cell biology, 206 (6) (2014): 695–706. issn: 1540-8140. doi: 10.1083/jcb.201405027 (cit. on pp. 1, 36, 37).

[9] S. T. Keating and A. El-Osta. Epigenetics and Metabolism. Circulation Research, 116 (4) (2015): 715–736. issn: 0009-7330. doi: 10.1161/CIRCRESAHA.116.303936 (cit. on pp. 1, 23).

[10] Y. Saletore et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome biology, 13 (10) (2012): 175. issn: 1474-760X. doi: 10.1186/gb-2012-13-10-175 (cit. on p. 2).

[12] I. Dunham et al. An integrated encyclopedia of DNA elements in the human genome. Nature, 489 (7414) (2012): 57–74. issn: 1476-4687. doi: 10.1038/nature11247 (cit. on pp. 3, 12, 19, 51, 52, 154).

[13] C. M. Rands et al. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS genetics, 10 (7) (2014): e1004525. issn: 1553-7404. doi: 10.1371/journal.pgen.1004525 (cit. on p. 3).

[14] M. L. Mo et al. A genome-scale, constraint-based approach to systems biology of human metabolism. Molecular bioSystems, 3 (9) (2007): 598–603. issn: 1742-206X. doi: 10.1039/b705597h (cit. on p. 3).

[15] B. Palsson. Metabolic systems biology. FEBS letters, 583 (24) (2009): 3900–3904. issn: 1873-3468. doi: 10.1016/j.febslet.2009.09.031 (cit. on p. 3).

[16] J. C. Venter et al. The sequence of the human genome. Science, 291 (5507) (2001): 1304–1351. issn: 0036-8075. doi: 10.1126/science.1058040 (cit. on p. 3).

[17] E. S. Lander et al. Initial sequencing and analysis of the human genome. Nature, 409 (6822) (2001): 860–921. issn: 0028-0836. doi: 10.1038/35057062 (cit. on p. 3).

[19] M. Kellis et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America, 111 (17) (2014): 6131–8. issn: 1091-6490. doi: 10.1073/pnas.1318948111 (cit. on p. 4).

[20] V. Sasisekharan and N. Pattabiraman. Structure of DNA predicted from stereochemistry of nucleoside derivatives. Nature, 275 (5676) (1978): 159–162. issn: 0028-0836. doi: 10.1038/275159a0 (cit. on p. 4).

[21] P. Yakovchuk et al. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic acids research, 34 (2) (2006): 564–574. issn: 1362-4962. doi: 10.1093/nar/gkj454 (cit. on p. 4).

[22] E. Bianconi et al. An estimation of the number of cells in the human body. Annals of human biology, 40 (6) (2013): 463–71. issn: 1464-5033. doi: 10.3109/03014460.2013.807878 (cit. on p. 4).

[23] M. K. Vickaryous and B. K. Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biological Reviews, 81 (03) (2006): 425. issn: 1464-7931. doi: 10.1017/S1464793106007068 (cit. on p. 4).

[24] A. S. Dimas et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science, 325 (5945) (2009): 1246–1250. issn: 1095-9203. doi: 10.1126/science.1174148 (cit. on p. 5).

[25] M. T. Maurano et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science, 337 (6099) (2012): 1190–1195. issn: 1095-9203. doi: 10.1126/science.1222794 (cit. on pp. 5, 160).

[26] K. W. Lee et al. Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? Neuroscience and biobehavioral reviews, 36 (1) (2012): 556–571. issn: 1873-7528. doi: 10.1016/j.neubiorev.2011.09.001 (cit. on p. 5).

[27] A. Portela and M. Esteller. Epigenetic modifications and human disease. Nature biotechnology, 28 (10) (2010): 1057–68. issn: 1546-1696. doi: 10.1038/nbt.1685 (cit. on p. 5).

[28] M. J. Boland et al. Epigenetic regulation of pluripotency and differentiation. Circulation research, 115 (2) (2014): 311–24. issn: 1524-4571. doi: 10.1161/CIRCRESAHA.115.301517 (cit. on pp. 5, 6, 8, 9).

[29] T Cremer and C Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews. Genetics, 2 (4) (2001): 292–301. issn: 1471-0056. doi: 10.1038/35066075 (cit. on p. 5).

[30] K. S. Wendt and F. G. Grosveld. Transcription in the context of the 3D nucleus. Current opinion in genetics & development, 25 (2014): 62–7. issn: 1879-0380. doi: 10.1016/j.gde.2013.11.020 (cit. on p. 5).

[31] R. M. Schultz. The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. Human Reproduction Update, 8 (4) (2002): 323–331. issn: 1355-4786. doi: 10.1093/humupd/8.4.323 (cit. on pp. 5, 13).

[32] K. K. Niakan and K. Eggan. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. Developmental biology, 375 (1) (2013): 54–64. issn: 1095-564X. doi: 10.1016/j.ydbio.2012.12.008 (cit. on pp. 5, 13).

[33] S. Rosa and P. Shaw. Insights into chromatin structure and dynamics in plants. en. Biology, 2 (4) (2013): 1378–410. issn: 2079-7737. doi: 10.3390/biology2041378 (cit. on p. 6).

[34] D. Shlyueva et al. Transcriptional enhancers: from properties to genome-wide predictions. en. Nature reviews. Genetics, 15 (4) (2014): 272–86. issn: 1471-0064. doi: 10.1038/nrg3682 (cit. on p. 6).

[35] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology, 28 (8) (2010): 817–825. issn: 1546-1696. doi: 10.1038/nbt.1662 (cit. on p. 5).

[36] J. Zhu et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell, 152 (3) (2013): 642–54. issn: 1097-4172. doi: 10.1016/j.cell.2012.12.033 (cit. on p. 5).

[37] C. A. Gifford et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell, 153 (5) (2013): 1149–63. issn: 1097-4172. doi: 10.1016/j.cell.2013.04.037 (cit. on p. 5).

[38] D. A. Jackson. Chromatin domains and nuclear compartments: establishing sites of gene expression in eukaryotic nuclei. Molecular biology reports, 24 (3) (1997): 209–20. issn: 0301-4851 (cit. on p. 5).

[39] D Robyr and P Wolffe. Hormone action and chromatin remodelling. Cellular and molecular life sciences : CMLS, 54 (2) (1998): 113–24. issn: 1420-682X (cit. on p. 5).

[40] A. I. Lamond and W. C. Earnshaw. Structure and function in the nucleus. Science, 280 (5363) (1998): 547–53. issn: 0036-8075 (cit. on p. 5).

[41] A. P. Wolffe and D Guschin. Review: chromatin structural features and targets that regulate transcription. Journal of structural biology, 129 (2-3) (2000): 102–22. issn: 1047-8477. doi: 10.1006/jsbi.2000.4217 (cit. on pp. 5, 7).

[42] E. Li. Chromatin modification and epigenetic reprogramming in mammalian development. Nature reviews. Genetics, 3 (9) (2002): 662–73. issn: 1471-0056. doi: 10.1038/nrg887 (cit. on p. 5).

[43] G. J. Narlikar et al. Cooperation between Complexes that Regulate Chromatin Structure and Transcription. Cell, 108 (4) (2002): 475–487. issn: 00928674. doi: 10.1016/S0092-8674(02)00654-2 (cit. on p. 5).

[44] orphanides G. A Unified Theory of Gene Expression. Cell, 108 (4) (2002): 439–451. issn: 00928674. doi: 10.1016/S0092-8674(02)00655-4 (cit. on p. 5).

[45] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 33 (2003): 245–54. doi: 10.1038/ng1089 (cit. on p. 5).

[46] S. I. S. Grewal and D. Moazed. Heterochromatin and epigenetic control of gene expression. Science, 301 (5634) (2003): 798–802. issn: 1095-9203. doi: 10.1126/science.1086887 (cit. on p. 5).

[47] D. Sproul et al. The role of chromatin structure in regulating the expression of clustered genes. Nature reviews. Genetics, 6 (10) (2005): 775–81. issn: 1471-0056. doi: 10.1038/nrg1688 (cit. on p. 5).

[48] S. Venkatesh and J. L. Workman. Histone exchange, chromatin structure and the regulation of transcription. Nature reviews. Molecular cell biology, 16 (3) (2015): 178–89. issn: 1471-0080. doi: 10.1038/nrm3941 (cit. on pp. 5, 7).

[49] T. Phillips and K. Shaw. Chromatin Remodeling in Eukaryotes. Nature Education, 1 (1) (2008): 209 (cit. on p. 6).

[50] M. Magistri et al. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends in genetics : TIG, 28 (8) (2012): 389–396. issn: 0168-9525. doi: 10.1016/j.tig.2012.03.013 (cit. on p. 6).

[51] M. Ha. Understanding the chromatin remodeling code. Plant science : an international journal of experimental plant biology, 211 (2013): 137–45. issn: 1873-2259. doi: 10.1016/j.plantsci.2013.07.006 (cit. on pp. 6, 7).

[52] J. H. Bergmann and D. L. Spector. Long non-coding RNAs: modulators of nuclear structure and function. Current opinion in cell biology, 26 (2014): 10–18. issn: 1879-0410. doi: 10.1016/j.ceb.2013.08.005 (cit. on p. 6).

[53] T. R. Cech and J. A. Steitz. The noncoding RNA revolution-trashing old rules to forge new ones. Cell, 157 (1) (2014): 77–94. issn: 1097-4172. doi: 10.1016/j.cell.2014.03.008 (cit. on p. 6).

[54] R. I. Joh et al. Regulation of histone methylation by noncoding RNAs. Biochimica et biophysica acta, 1839 (12) (2014): 1385–94. issn: 0006-3002. doi: 10.1016/j.bbagrm.2014.06.006 (cit. on p. 6).

[55] K Luger et al. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature, 389 (6648) (1997): 251–60. issn: 0028-0836. doi: 10.1038/38444 (cit. on p. 6).

[56] J. Mellor. Dynamic nucleosomes and gene transcription. Trends in genetics : TIG, 22 (6) (2006): 320–329. issn: 0168-9525. doi: 10.1016/j.tig.2006.03.008 (cit. on p. 7).

[57] J. L. Workman. Nucleosome displacement in transcription. Genes & Development, 20 (15) (2006): 2009–2017. issn: 0890-9369. doi: 10.1101/gad.1435706 (cit. on p. 7).

[58] R. Bargaje et al. Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain. Nucleic acids research, 40 (18) (2012): 8965–8978. issn: 1362-4962. doi: 10.1093/nar/gks665 (cit. on pp. 7, 18).

[59] S. S. Teves et al. Transcribing through the nucleosome. Trends in biochemical sciences, 39 (12) (2014): 577–86. issn: 0968-0004. doi: 10.1016/j.tibs.2014.10.004 (cit. on p. 7).

[60] A. Stein. Nucleosome positioning, gene regulation and disease. Epigenomics, 2 (3) (2010): 351–4. issn: 1750-192X. doi: 10.2217/epi.10.23 (cit. on p. 7).

[61] S. Diermeier et al. TNFα signalling primes chromatin for NF-κB binding and induces rapid and widespread nucleosome repositioning. Genome biology, 15 (12) (2014): 536. issn: 1465-6914. doi: `10.1186/s13059-014-0536-6` (cit. on pp. 7, 18).

[62] Y. K. Chutake et al. Altered nucleosome positioning at the transcription start site and deficient transcriptional initiation in Friedreich ataxia. The Journal of biological chemistry, 289 (22) (2014): 15194–202. issn: 1083-351X. doi: `10.1074/jbc.M114.566414` (cit. on pp. 7, 18).

[63] T. Kouzarides. Chromatin modifications and their function. Cell, 128 (4) (2007): 693–705. issn: 0092-8674. doi: `10.1016/j.cell.2007.02.005` (cit. on p. 7).

[64] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. Cell research, 21 (3) (2011): 381–395. issn: 1748-7838. doi: `10.1038/cr.2011.22` (cit. on p. 7).

[65] P. Tessarz and T. Kouzarides. Histone core modifications regulating nucleosome structure and dynamics. Nature reviews. Molecular cell biology, 15 (11) (2014): 703–8. issn: 1471-0080. doi: `10.1038/nrm3890` (cit. on p. 7).

[66] S. B. Rothbart and B. D. Strahl. Interpreting the language of histone and DNA modifications. Biochimica et biophysica acta, 1839 (8) (2014): 627–43. issn: 0006-3002. doi: `10.1016/j.bbagrm.2014.03.001` (cit. on p. 7).

[67] E. Petty and L. Pillus. Balancing chromatin remodeling and histone modifications in transcription. Trends in genetics : TIG, 29 (11) (2013): 621–629. issn: 0168-9525. doi: `10.1016/j.tig.2013.06.006` (cit. on p. 7).

[68] C. Li and J. Wang. Quantifying Waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. Journal of the Royal Society, Interface / the Royal Society, 10 (89) (2013): 20130787. issn: 1742-5662. doi: `10.1098/rsif.2013.0787` (cit. on pp. 7, 18).

[69] D. J. Huebert and B. E. Bernstein. Genomic views of chromatin. Current opinion in genetics & development, 15 (5) (2005): 476–81. issn: 0959-437X. doi: `10.1016/j.gde.2005.08.001` (cit. on p. 7).

[70] T Jenuwein and C. D. Allis. Translating the histone code. Science, 293 (5532) (2001): 1074–80. issn: 0036-8075. doi: `10.1126/science.1063127` (cit. on p. 7).

[71] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. Nature, 403 (6765) (2000): 41–5. issn: 0028-0836. doi: `10.1038/47412` (cit. on p. 7).

[72] B. M. Turner. Cellular memory and the histone code. Cell, 111 (3) (2002): 285–91. issn: 0092-8674 (cit. on p. 7).

[73] H. T. Spotswood and B. M. Turner. An increasingly complex code. The Journal of clinical investigation, 110 (5) (2002): 577–582. issn: 0021-9738. doi: 10.1172/JCI16547 (cit. on p. 7).

[74] Y. Wang et al. Beyond the double helix: writing and reading the histone code. Novartis Foundation symposium, 259 (2004): 3–17; discussion 17–21, 163–9. issn: 1528-2511 (cit. on p. 7).

[75] E. I. Campos and D. Reinberg. Histones: annotating chromatin. Annual review of genetics, 43 (2009): 559–99. issn: 1545-2948. doi: 10.1146/annurev.genet.032608.103928 (cit. on p. 7).

[76] M. Tan et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell, 146 (6) (2011): 1016–1028. issn: 1097-4172. doi: 10.1016/j.cell.2011.08.008 (cit. on p. 7).

[77] A. M. Arnaudo and B. A. Garcia. Proteomic characterization of novel histone post-translational modifications. Epigenetics & chromatin, 6 (1) (2013): 24. issn: 1756-8935. doi: 10.1186/1756-8935-6-24 (cit. on p. 7).

[78] M. Rye et al. Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. BMC genomics, 15 (2014): 120. issn: 1471-2164. doi: 10.1186/1471-2164-15-120 (cit. on p. 7).

[79] H. Huang et al. SnapShot: Histone Modifications. Cell, 159 (2) (2014): 458–458.e1. issn: 00928674. doi: 10.1016/j.cell.2014.09.037 (cit. on p. 7).

[80] B. E. Bernstein et al. Methylation of histone H3 Lys 4 in coding regions of active genes. Proceedings of the National Academy of Sciences of the United States of America, 99 (13) (2002): 8695–8700. issn: 0027-8424. doi: 10.1073/pnas.082249499 (cit. on p. 8).

[81] T. S. Mikkelsen et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 448 (7153) (2007): 553–60. issn: 1476-4687. doi: 10.1038/nature06008 (cit. on p. 8).

[82] N. D. Heintzman et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature, 459 (7243) (2009): 108–112. issn: 1476-4687. doi: 10.1038/nature07829 (cit. on p. 8).

[83] P. J. Farnham. Insights from genomic profiling of transcription factors. Nature reviews. Genetics, 10 (9) (2009): 605–616. issn: 1471-0064. doi: 10.1038/nrg2636 (cit. on p. 8).

[84] M. P. Creyghton et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America, 107 (50) (2010): 21931–21936. issn: 1091-6490. doi: 10.1073/pnas.1016071107 (cit. on p. 8).

[85] J. Ernst et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473 (7345) (2011): 43–9. issn: 1476-4687. doi: 10.1038/nature09906 (cit. on p. 8).

[86] H. Kimura. Histone modifications for human epigenome analysis. Journal of human genetics, 58 (7) (2013): 439–445. issn: 1435-232X. doi: 10.1038/jhg.2013.66 (cit. on p. 8).

[87] A. Lesne et al. Chromatin fiber allostery and the epigenetic code. Journal of physics. Condensed matter : an Institute of Physics journal, 27 (6) (2015): 064114. issn: 1361-648X. doi: 10.1088/0953-8984/27/6/064114 (cit. on p. 8).

[88] G.-d. Sun et al. Histone lysine methylation in diabetic nephropathy. Journal of diabetes research, 2014 (2014): 654148. issn: 2314-6753. doi: 10.1155/2014/654148 (cit. on p. 8).

[89] Y. Obata et al. Epigenetic modifications of the immune system in health and disease. Immunology and cell biology, 93 (3) (2015): 226–232. issn: 1440-1711. doi: 10.1038/icb.2014.114 (cit. on p. 8).

[90] S. Bekkering et al. The Epigenetic Memory of Monocytes and Macrophages as a Novel Drug Target in Atherosclerosis. Clinical therapeutics (2015). issn: 1879-114X. doi: 10.1016/j.clinthera.2015.01.008 (cit. on p. 8).

[91] W. Tian and Y. Xu. Decoding liver injury: A regulatory role for histone modifications. The international journal of biochemistry & cell biology (2015). issn: 1878-5875. doi: 10.1016/j.biocel.2015.03.009 (cit. on p. 8).

[92] T. G. Gillette and J. A. Hill. Readers, Writers, and Erasers: Chromatin as the Whiteboard of Heart Disease. Circulation Research, 116 (7) (2015): 1245–1253. issn: 0009-7330. doi: 10.1161/CIRCRESAHA.116.303630 (cit. on p. 8).

[93] F Vahid et al. The role dietary of bioactive compounds on the regulation of histone acetylases and deacetylases: A review. Gene, 562 (1) (2015): 8–15. issn: 1879-0038. doi: 10.1016/j.gene.2015.02.045 (cit. on p. 8).

[94] B. E. Bernstein et al. The mammalian epigenome. Cell, 128 (4) (2007): 669–81. issn: 0092-8674. doi: 10.1016/j.cell.2007.01.033 (cit. on p. 8).

[95] S. K. T. Ooi et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature, 448 (7154) (2007): 714–717. issn: 1476-4687. doi: 10.1038/nature05987 (cit. on p. 8).

[96] T. B. Dormann HL and F. W. S, Allis CD, Funabiki H. Dynamic regulation of effector protein binding to histone modifications: the biology of HP1 switching. Cell Cycle, 5 (24) (2006): 2842–51 (cit. on p. 8).

[97] Z. Lu et al. Importance of charge independent effects in readout of the trimethyllysine mark by HP1 chromodomain. Journal of the American Chemical Society, 131 (41) (2009): 14928–31. issn: 1520-5126. doi: 10.1021/ja904951t (cit. on p. 8).

[98] K. P. Koh and A. Rao. DNA methylation and methylcytosine oxidation in cell fate decisions. Current opinion in cell biology, 25 (2) (2013): 152–161. issn: 1879-0410. doi: 10.1016/j.ceb.2013.02.014 (cit. on pp. 8, 9).

[99] D. Jjingo et al. On the presence and role of human gene-body DNA methylation. Oncotarget, 3 (4) (2012): 462–474. issn: 1949-2553 (cit. on p. 8).

[100] M. Ehrlich and M. Lacey. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. Epigenomics, 5 (5) (2013): 553–568. issn: 1750-192X. doi: 10.2217/epi.13.43 (cit. on p. 8).

[101] X. Yang et al. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. Cancer Cell, 26 (4) (2014): 577–90. issn: 15356108. doi: 10.1016/j.ccr.2014.07.028 (cit. on p. 8).

[102] S. Shukla et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature, 479 (7371) (2011): 74–9. issn: 1476-4687. doi: 10.1038/nature10442 (cit. on p. 8).

[103] G. Elliott et al. Intermediate DNA methylation is a conserved signature of genome regulation. Nature communications, 6 (2015): 6363. issn: 2041-1723. doi: 10.1038/ncomms7363 (cit. on p. 8).

[104] B. F. Vanyushin. Enzymatic DNA methylation is an epigenetic control for genetic functions of the cell. Biochemistry (Mosc). 70 (5) (2005): 488–99. issn: 0006-2979 (cit. on p. 8).

[105] H. Meng et al. DNA Methylation, Its Mediators and Genome Integrity. International journal of biological sciences, 11 (5) (2015): 604–617. issn: 1449-2288. doi: 10.7150/ijbs.11218 (cit. on pp. 8, 9).

[106] E. G. Malygin and S. Hattman. DNA methyltransferases: mechanistic models derived from kinetic analysis. Critical reviews in biochemistry and molecular biology, 47 (2) (2012): 97–193. issn: 1549-7798. doi: 10.3109/10409238.2011.620942 (cit. on p. 8).

[107] S. Ito et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science, 333 (6047) (2011): 1300–1303. issn: 1095-9203. doi: 10.1126/science.1210597 (cit. on p. 8).

[108] R. M. Kohli and Y. Zhang. TET enzymes, TDG and the dynamics of DNA demethylation. Nature, 502 (7472) (2013): 472–479. issn: 1476-4687. doi: 10.1038/nature12750 (cit. on p. 8).

[109] E. Li and Y. Zhang. DNA methylation in mammals. Cold Spring Harbor perspectives in biology, 6 (5) (2014): a019133. issn: 1943-0264. doi: 10.1101/cshperspect.a019133 (cit. on pp. 8, 9).

[110] I. Sanli and R. Feil. Chromatin mechanisms in the developmental control of imprinted gene expression. The international journal of biochemistry & cell biology (2015). issn: 1878-5875. doi: 10.1016/j.biocel.2015.04.004 (cit. on pp. 8, 9).

[111] J. D. McGhee and G. D. Ginder. Specific DNA methylation sites in the vicinity of the chicken beta-globin genes. Nature, 280 (5721) (1979): 419–20. issn: 0028-0836 (cit. on p. 9).

[112] P. A. Jones and S. M. Taylor. Cellular differentiation, cytidine analogs and DNA methylation. Cell, 20 (1) (1980): 85–93. issn: 0092-8674 (cit. on p. 9).

[113] J. T. Lee and M. S. Bartolomei. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell, 152 (6) (2013): 1308–23. issn: 1097-4172. doi: 10.1016/j.cell.2013.02.016 (cit. on p. 9).

[114] A. M. Deaton and A. Bird. CpG islands and the regulation of transcription. Genes & development, 25 (10) (2011): 1010–1022. issn: 1549-5477. doi: 10.1101/gad.2037511 (cit. on p. 9).

[115] M. J. Ziller et al. Charting a dynamic DNA methylation landscape of the human genome. Nature, 500 (7463) (2013): 477–481. issn: 1476-4687. doi: 10.1038/nature12433 (cit. on p. 9).

[116] X. Zhang et al. DNA demethylation: where genetics meets epigenetics. Current pharmaceutical design, 20 (11) (2014): 1625–31. issn: 1873-4286 (cit. on p. 9).

[117] J. Zheng et al. DNA methylation: the pivotal interaction between early-life nutrition and glucose metabolism in later life. The British journal of nutrition, 112 (11) (2014): 1850–7. issn: 1475-2662. doi: 10.1017/S0007114514002827 (cit. on p. 9).

[118] B. Illi et al. Chromatin methylation and cardiovascular aging. Journal of molecular and cellular cardiology (2015). issn: 1095-8584. doi: 10.1016/j.yjmcc.2015.02.011 (cit. on p. 9).

[119] S. E. Pinney. Mammalian Non-CpG Methylation: Stem Cells and Beyond. Biology, 3 (4) (2014): 739–751. issn: 2079-7737. doi: 10.3390/biology3040739 (cit. on p. 9).

[120] A Paziewska et al. DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. British journal of cancer, 111 (4) (2014): 781–9. issn: 1532-1827. doi: 10.1038/bjc.2014.337 (cit. on p. 9).

[121] J Wang et al. Hypertensive epigenetics: from DNA methylation to microRNAs. Journal of human hypertension (2015). issn: 1476-5527. doi: 10.1038/jhh.2014.132 (cit. on p. 9).

[122] J. Marín-García and A. T. Akhmedov. Epigenetics of the failing heart. Heart failure reviews (2015). issn: 1573-7322. doi: 10.1007/s10741-015-9483-x (cit. on p. 9).

[123] S. Zaina et al. DNA methylation map of human atherosclerosis. Circulation. Cardiovascular genetics, 7 (5) (2014): 692–700. issn: 1942-3268. doi: 10.1161/CIRCGENETICS.113.000441 (cit. on p. 9).

[124] M. D. P. Valencia-Morales et al. The DNA methylation drift of the atherosclerotic aorta increases with lesion progression. BMC medical genomics, 8 (1) (2015): 7. issn: 1755-8794. doi: 10.1186/s12920-015-0085-1 (cit. on p. 9).

[125] L. Gillberg and C. Ling. The potential use of DNA methylation biomarkers to identify risk and progression of type 2 diabetes. Frontiers in endocrinology, 6 (2015): 43. issn: 1664-2392. doi: 10.3389/fendo.2015.00043 (cit. on p. 9).

[126] B. Gupta and R. D. Hawkins. Epigenomics of autoimmune diseases. Immunology and cell biology, 93 (3) (2015): 271–6. issn: 1440-1711. doi: 10.1038/icb.2015.18 (cit. on p. 9).

[127] P. J. van den Elsen et al. The epigenetics of multiple sclerosis and other related disorders. Multiple sclerosis and related disorders, 3 (2) (2014): 163–75. issn: 2211-0356. doi: 10.1016/j.msard.2013.08.007 (cit. on p. 9).

[128] N. R. Rose and R. J. Klose. Understanding the relationship between DNA methylation and histone lysine methylation. Biochimica et biophysica acta, 1839 (12) (2014): 1362–72. issn: 0006-3002. doi: 10.1016/j.bbagrm.2014.02.007 (cit. on pp. 9, 17).

[129] K. A. Lillycrop et al. DNA methylation, ageing and the influence of early life nutrition. The Proceedings of the Nutrition Society, 73 (3) (2014): 413–21. issn: 1475-2719. doi: 10.1017/S0029665114000081 (cit. on p. 9).

[130] P. R. Mandaviya et al. Homocysteine and DNA methylation: a review of animal and human literature. Molecular genetics and metabolism, 113 (4) (2014): 243–52. issn: 1096-7206. doi: 10.1016/j.ymgme.2014.10.006 (cit. on p. 9).

[131] L. Chakalova and P. Fraser. Organization of transcription. Cold Spring Harbor perspectives in biology, 2 (9) (2010). issn: 1943-0264. doi: 10.1101/cshperspect.a000729 (cit. on p. 9).

[132] J. D. Lewis and E Izaurralde. The role of the cap structure in RNA processing and nuclear export. European journal of biochemistry / FEBS, 247 (2) (1997): 461–9. issn: 0014-2956. doi: DOI:10.1111/j.1432-1033.1997.00461.x (cit. on p. 10).

[133] E. J. Cho et al. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes & development, 11 (24) (1997): 3319–3326. issn: 0890-9369 (cit. on p. 10).

[134] A. G. Matera et al. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nature reviews. Molecular cell biology, 8 (3) (2007): 209–20. issn: 1471-0072. doi: 10.1038/nrm2124 (cit. on p. 10).

[135] D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry, 72 (2003): 291–336. issn: 0066-4154. doi: 10.1146/annurev.biochem.72.121801.161720 (cit. on p. 10).

[136] A. J. Matlin et al. Understanding alternative splicing: towards a cellular code. Nature reviews. Molecular cell biology, 6 (5) (2005): 386–98. issn: 1471-0072. doi: 10.1038/nrm1645 (cit. on p. 10).

[137] S. Clancy. RNA Splicing: Introns, Exons and Spliceosome. Nature Education, 1 (1) (2008): 31 (cit. on p. 10).

[138] A. G. Matera and Z. Wang. A day in the life of the spliceosome. Nature reviews. Molecular cell biology, 15 (2) (2014): 108–121. issn: 1471-0080. doi: 10.1038/nrm3742 (cit. on p. 10).

[139] N. J. Proudfoot et al. Integrating mRNA processing with transcription. Cell, 108 (4) (2002): 501–12. issn: 0092-8674 (cit. on p. 10).

[140] M. Dávila López and T. Samuelsson. Early evolution of histone mRNA 3' end processing. RNA, 14 (1) (2008): 1–10. issn: 1469-9001. doi: 10.1261/rna.782308 (cit. on p. 10).

[141] A. Curinha et al. Implications of polyadenylation in health and disease. Nucleus, 5 (6) (2014): 508–19. issn: 1949-1042. doi: 10.4161/nucl.36360 (cit. on p. 10).

[142] J. M. Vaquerizas et al. A census of human transcription factors: function, expression and evolution. Nature reviews. Genetics, 10 (4) (2009): 252–63. issn: 1471-0064. doi: 10.1038/nrg2538 (cit. on p. 11).

[144] A. H. Brivanlou and J. E. Darnell. Signal transduction and the control of gene expression. Science, 295 (5556) (2002): 813–8. issn: 1095-9203. doi: 10.1126/science.1066355 (cit. on pp. 11, 18).

[145] A. S. Baldwin. Series introduction: the transcription factor NF-kappaB and human disease. The Journal of clinical investigation, 107 (1) (2001): 3–6. issn: 0021-9738. doi: 10.1172/JCI11891 (cit. on p. 11).

[146] S. Neph et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489 (7414) (2012): 83–90. issn: 1476-4687. doi: 10.1038/nature11212 (cit. on p. 11).

[147] H. S. Marinho et al. Hydrogen peroxide sensing, signaling and regulation of transcription factors. Redox biology, 2 (2014): 535–562. issn: 2213-2317. doi: 10.1016/j.redox.2014.02.006 (cit. on p. 11).

[148] M. H. Kagey et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature, 467 (7314) (2010): 430–435. issn: 1476-4687. doi: 10.1038/nature09380 (cit. on p. 11).

[149] B. L. Allen and D. J. Taatjes. The Mediator complex: a central integrator of transcription. Nature reviews. Molecular cell biology, 16 (3) (2015): 155–66. issn: 1471-0080. doi: 10.1038/nrm3951 (cit. on p. 11).

[150] D. Villar et al. Enhancer Evolution across 20 Mammalian Species. Cell, 160 (3) (2015): 554–566. issn: 00928674. doi: 10.1016/j.cell.2015.01.006 (cit. on p. 12).

[151] V Matys et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research, 34 (Database issue) (2006): D108–10. issn: 1362-4962. doi: 10.1093/nar/gkj143 (cit. on p. 12).

[152] D. L. Fulton et al. TFCat: the curated catalog of mouse and human transcription factors. Genome biology, 10 (3) (2009): R29. issn: 1465-6914. doi: 10.1186/gb-2009-10-3-r29 (cit. on p. 12).

[153] A. Mathelier et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic acids research, 42 (Database issue) (2014): D142–7. issn: 1362-4962. doi: 10.1093/nar/gkt997 (cit. on p. 12).

[154] L. Yang et al. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic acids research, 42 (Database issue) (2014): D148–55. issn: 1362-4962. doi: 10.1093/nar/gkt1087 (cit. on p. 12).

[155] A. Sebastian and B. Contreras-Moreira. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics, 30 (2) (2014): 258–65. issn: 1367-4811. doi: 10.1093/bioinformatics/btt663 (cit. on p. 12).

[156] E. Wingender et al. TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic acids research, 43 (Database issue) (2015): D97–102. issn: 1362-4962. doi: 10.1093/nar/gku1064 (cit. on p. 12).

[157] S. T. Whiteside and S Goodbourn. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. Journal of cell science, 104 (Pt 4 (1993): 949–55. issn: 0021-9533 (cit. on p. 12).

[158] C. A. Meier. Regulation of gene expression by nuclear hormone receptors. Journal of receptor and signal transduction research, 17 (1-3) (1997): 319–35. issn: 1079-9893. doi: 10.3109/10799899709036612 (cit. on p. 13).

[159] H Escriva et al. Evolution and diversification of the nuclear receptor superfamily. Annals of the New York Academy of Sciences, 839 (1998): 143–6. issn: 0077-8923 (cit. on p. 13).

[160] M Robinson-Rechavi et al. How many nuclear hormone receptors are there in the human genome? Trends in genetics : TIG, 17 (10) (2001): 554–6. issn: 0168-9525 (cit. on p. 13).

[161] G. A. Francis et al. Nuclear receptors and the control of metabolism. Annual review of physiology, 65 (2003): 261–311. issn: 0066-4278. doi: 10.1146/annurev.physiol.65.092101.142528 (cit. on p. 13).

[162] Z. Zhang et al. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. Genome research, 14 (4) (2004): 580–590. issn: 1088-9051. doi: 10.1101/gr.2160004 (cit. on p. 13).

[163] P. Germain et al. Overview of nomenclature of nuclear receptors. Pharmacological reviews, 58 (4) (2006): 685–704. issn: 0031-6997. doi: 10.1124/pr.58.4.2 (cit. on p. 13).

[164] N. J. McKenna et al. Minireview: Evolution of NURSA, the Nuclear Receptor Signaling Atlas. Molecular endocrinology (Baltimore, Md.) 23 (6) (2009): 740–6. issn: 1944-9917. doi: 10.1210/me.2009-0135 (cit. on p. 13).

[165] S. D. Conzen. Minireview: nuclear receptors and breast cancer. Molecular endocrinology, 22 (10) (2008): 2215–28. issn: 0888-8809. doi: 10.1210/me.2007-0421 (cit. on p. 13).

[166] C. K. Glass and K. Saijo. Nuclear receptor transrepression pathways that regulate inflammation in macrophages and T cells. Nature reviews. Immunology, 10 (5) (2010): 365–76. issn: 1474-1741. doi: 10.1038/nri2748 (cit. on p. 13).

[167] A. Foryst-Ludwig and U. Kintscher. Metabolic impact of estrogen signalling through ERalpha and ERbeta. The Journal of steroid biochemistry and molecular biology, 122 (1-3) (2010): 74–81. issn: 1879-1220. doi: 10.1016/j.jsbmb.2010.06.012 (cit. on p. 13).

[168] R. P. Patel and S. Barnes. Isoflavones and PPAR Signaling: A Critical Target in Cardiovascular, Metastatic, and Metabolic Disease. PPAR research, 2010 (2010): 153252. issn: 1687-4765. doi: 10.1155/2010/153252 (cit. on p. 13).

[169] U. Baschant and J. Tuckermann. The role of the glucocorticoid receptor in inflammation and immunity. The Journal of steroid biochemistry and molecular biology, 120 (2-3) (2010): 69–75. issn: 1879-1220. doi: 10.1016/j.jsbmb.2010.03.058 (cit. on p. 13).

[170] J. P. Overington et al. How many drug targets are there? Nature reviews. Drug discovery, 5 (12) (2006): 993–6. issn: 1474-1776. doi: 10.1038/nrd2199 (cit. on p. 13).

[171] B Mayr and M Montminy. Transcriptional regulation by the phosphorylation-dependent factor CREB. Nature reviews. Molecular cell biology, 2 (8) (2001): 599–609. issn: 1471-0072. doi: 10.1038/35085068 (cit. on p. 13).

[172] M. Thomson et al. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell, 145 (6) (2011): 875–89. issn: 1097-4172. doi: 10.1016/j.cell.2011.05.017 (cit. on p. 13).

[173] N. Yosef and A. Regev. Impulse control: temporal dynamics in gene transcription. Cell, 144 (6) (2011): 886–896. issn: 1097-4172. doi: 10.1016/j.cell.2011.02.015 (cit. on pp. 13, 20).

[174] J. Wang et al. Quantifying the Waddington landscape and biological paths for development and differentiation. Proceedings of the National Academy of Sciences of the United States of America, 108 (20) (2011): 8257–8262. issn: 1091-6490. doi: 10.1073/pnas.1017017108 (cit. on pp. 13, 18).

[175] N. Iovino and G. Cavalli. Rolling ES cells down the Waddington landscape with Oct4 and Sox2. Cell, 145 (6) (2011): 815–7. issn: 1097-4172. doi: 10.1016/j.cell.2011.05.027 (cit. on pp. 13, 18).

[176] E. H. Bresnick et al. Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. Nucleic acids research, 40 (13) (2012): 5819–5831. issn: 1362-4962. doi: 10.1093/nar/gks281 (cit. on p. 13).

[177] A. Champhekar et al. Regulation of early T-lineage gene expression and developmental progression by the progenitor cell transcription factor PU.1. Genes & development, 29 (8) (2015): 832–48. issn: 1549-5477. doi: 10.1101/gad.259879.115 (cit. on p. 13).

[178] C. Moorman et al. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America, 103 (32) (2006): 12027–12032. issn: 0027-8424. doi: 10.1073/pnas.0605003103 (cit. on p. 13).

[179] H. Brunschwig et al. Fine-scale maps of recombination rates and hotspots in the mouse genome. Genetics, 191 (3) (2012): 757–764. issn: 1943-2631. doi: 10.1534/genetics.112.141036 (cit. on p. 13).

[180] R. Siersbæk et al. Transcriptional networks and chromatin remodeling controlling adipogenesis. Trends in endocrinology and metabolism: TEM, 23 (2) (2012): 56–64. issn: 1879-3061. doi: 10.1016/j.tem.2011.10.001 (cit. on pp. 13, 33).

[181] R. Siersbæk et al. Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. Cell reports, 7 (5) (2014): 1443–55. issn: 2211-1247. doi: 10.1016/j.celrep.2014.04.042 (cit. on pp. 13, 162).

[182] H. Stower. Gene expression: Super enhancers. Nature reviews. Genetics, 14 (6) (2013): 367. issn: 1471-0064. doi: 10.1038/nrg3496 (cit. on p. 14).

[183] A. Joshi. Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs. BMC bioinformatics, 15 (1) (2014): 412. issn: 1471-2105. doi: 10.1186/s12859-014-0412-0 (cit. on p. 14).

[184] Z. Liu et al. Enhancer Activation Requires trans-Recruitment of a Mega Transcription Factor Complex. Cell, 159 (2) (2014): 358–373. issn: 00928674. doi: 10.1016/j.cell.2014.08.027 (cit. on p. 14).

[185] S. Pott and J. D. Lieb. What are super-enhancers? Nature Genetics, 47 (1) (2014): 8–12. issn: 1061-4036. doi: 10.1038/ng.3167 (cit. on p. 14).

[186] J.-w. Yin and G. Wang. The Mediator complex: a master coordinator of transcription and cell lineage development. Development, 141 (5) (2014): 977–87. issn: 1477-9129. doi: 10.1242/dev.098392 (cit. on p. 14).

[187] W. A. Whyte et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell, 153 (2) (2013): 307–319. issn: 1097-4172. doi: 10.1016/j.cell.2013.03.035 (cit. on pp. 14, 162).

[188] D. Hnisz et al. Super-enhancers in the control of cell identity and disease. Cell, 155 (4) (2013): 934–947. issn: 1097-4172. doi: 10.1016/j.cell.2013.09.053 (cit. on pp. 14, 153, 162).

[189] R. C. Adam et al. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. Nature, advance on (2015). issn: 0028-0836. doi: 10.1038/nature14289 (cit. on pp. 14, 17).

[190] N. Hah et al. Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. Proceedings of the National Academy of Sciences of the United States of America, 112 (3) (2015): E297–302. issn: 1091-6490. doi: 10.1073/pnas.1424028112 (cit. on p. 14).

[191] C. J. McManus and B. R. Graveley. RNA structure and the mechanisms of alternative splicing. Current opinion in genetics & development, 21 (4) (2011): 373–379. issn: 1879-0380. doi: 10.1016/j.gde.2011.04.001 (cit. on p. 14).

[192] M. W. Medina and R. M. Krauss. Alternative splicing in the regulation of cholesterol homeostasis. Current opinion in lipidology, 24 (2) (2013): 147–152. issn: 1473-6535. doi: 10.1097/MOL.0b013e32835cf284 (cit. on p. 14).

[193] D. Guo et al. RNAa in action: from the exception to the norm. RNA biology, 11 (10) (2014): 1221–5. issn: 1555-8584. doi: 10.4161/15476286.2014.972853 (cit. on p. 14).

[194] P. Björk and L. Wieslander. Mechanisms of mRNA export. Seminars in cell & developmental biology, 32 (2014): 47–54. issn: 1096-3634. doi: 10.1016/j.semcdb.2014.04.027 (cit. on p. 14).

[195] S. F. Mitchell and R. Parker. Principles and properties of eukaryotic mRNPs. Molecular cell, 54 (4) (2014): 547–58. issn: 1097-4164. doi: 10.1016/j.molcel.2014.04.033 (cit. on p. 14).

[196] J. G. Blackinton and J. D. Keene. Post-transcriptional RNA regulons affecting cell cycle and proliferation. Seminars in cell & developmental biology, 34 (2014): 44–54. issn: 1096-3634. doi: 10.1016/j.semcdb.2014.05.014 (cit. on p. 14).

[197] R. Parker and U. Sheth. P bodies and the control of mRNA translation and degradation. Molecular cell, 25 (5) (2007): 635–46. issn: 1097-2765. doi: 10.1016/j.molcel.2007.02.011 (cit. on p. 14).

[198] A. Fleming et al. The carrying pigeons of the cell: exosomes and their role in infectious diseases caused by human pathogens. Pathogens and disease, 71 (2) (2014): 109–20. issn: 2049-632X. doi: 10.1111/2049-632X.12135 (cit. on p. 14).

[199] P. Mitchell. Exosome substrate targeting: the long and short of it. Biochemical Society transactions, 42 (4) (2014): 1129–34. issn: 1470-8752. doi: 10.1042/BST20140088 (cit. on p. 14).

[200] J. Qin and Q. Xu. Functions and application of exosomes. Acta poloniae pharmaceutica, 71 (4) (2014): 537–43. issn: 0001-6837 (cit. on p. 14).

[201] J. J. Moser and M. J. Fritzler. Cytoplasmic ribonucleoprotein (RNP) bodies and their relationship to GW/P bodies. The international journal of biochemistry & cell biology, 42 (6) (2010): 828–43. issn: 1878-5875. doi: 10.1016/j.biocel.2009.11.018 (cit. on p. 14).

[202] M. Olszewska et al. P-bodies and their functions during mRNA cell cycle: mini-review. Cell biochemistry and function, 30 (3) (2012): 177–82. issn: 1099-0844. doi: 10.1002/cbf.2804 (cit. on p. 14).

[203] J. Lu and A. G. Clark. Impact of microRNA regulation on variation in human gene expression. Genome research, 22 (7) (2012): 1243–1254. issn: 1549-5469. doi: 10.1101/gr.132514.111 (cit. on p. 14).

[204] J. Brennecke et al. Principles of microRNA-target recognition. PLoS biology, 3 (3) (2005): e85. issn: 1545-7885. doi: 10.1371/journal.pbio.0030085 (cit. on p. 14).

[205] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. Cell, 136 (2) (2009): 215–233. issn: 1097-4172. doi: 10.1016/j.cell.2009.01.002 (cit. on p. 14).

[206] H. Guo et al. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 466 (7308) (2010): 835–840. issn: 1476-4687. doi: 10.1038/nature09267 (cit. on p. 14).

[207] A Wilczynska and M Bushell. The complexity of miRNA-mediated repression. Cell death and differentiation, 22 (1) (2014): 22–33. issn: 1476-5403. doi: 10.1038/cdd.2014.112 (cit. on p. 14).

[208] S. W. Eichhorn et al. mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. Molecular cell, 56 (1) (2014): 104–15. issn: 1097-4164. doi: 10.1016/j.molcel.2014.08.028 (cit. on p. 14).

[209] A. E. Pasquinelli et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature, 408 (6808) (2000): 86–9. issn: 0028-0836. doi: 10.1038/35040556 (cit. on p. 14).

[210] M Lagos-Quintana et al. Identification of novel genes coding for small expressed RNAs. Science, 294 (5543) (2001): 853–8. issn: 0036-8075. doi: 10.1126/science.1064921 (cit. on p. 14).

[211] A. Stark et al. Identification of Drosophila MicroRNA targets. PLoS biology, 1 (3) (2003): E60. issn: 1545-7885. doi: 10.1371/journal.pbio.0000060 (cit. on p. 14).

[212] M. N. Poy et al. A pancreatic islet-specific microRNA regulates insulin secretion. Nature, 432 (7014) (2004): 226–30. issn: 1476-4687. doi: 10.1038/nature03076 (cit. on p. 14).

[213] M. R. Friedländer et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome biology, 15 (4) (2014): R57. issn: 1465-6914. doi: 10.1186/gb-2014-15-4-r57 (cit. on p. 14).

[214] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research, 42 (Database issue) (2014): D68–73. issn: 1362-4962. doi: 10.1093/nar/gkt1181 (cit. on p. 14).

[215] R. C. Friedman et al. Most mammalian mRNAs are conserved targets of microRNAs. Genome research, 19 (1) (2009): 92–105. issn: 1088-9051. doi: 10.1101/gr.082701.108 (cit. on pp. 14, 54).

[216] E. Bernstein et al. Dicer is essential for mouse development. Nature genetics, 35 (3) (2003): 215–7. issn: 1061-4036. doi: 10.1038/ng1253 (cit. on p. 15).

[217] C. Kanellopoulou et al. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes & development, 19 (4) (2005): 489–501. issn: 0890-9369. doi: 10.1101/gad.1248505 (cit. on p. 15).

[218] S. A. Muljo et al. Aberrant T cell differentiation in the absence of Dicer. The Journal of experimental medicine, 202 (2) (2005): 261–9. issn: 0022-1007. doi: 10.1084/jem.20050678 (cit. on p. 15).

[219] P. K. Rao et al. Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. Circulation research, 105 (6) (2009): 585–594. issn: 1524-4571. doi: 10.1161/CIRCRESAHA.109.200451 (cit. on p. 15).

[220] S. Albinsson et al. MicroRNAs are necessary for vascular smooth muscle growth, differentiation, and function. Arteriosclerosis, thrombosis, and vascular biology, 30 (6) (2010): 1118–1126. issn: 1524-4636. doi: 10.1161/ATVBAHA.109.200873 (cit. on p. 15).

[221] L. A. Medeiros et al. Mir-290-295 deficiency in mice results in partially penetrant embryonic lethality and germ cell defects. Proceedings of the National Academy of Sciences of the United States of America, 108 (34) (2011): 14163–14168. issn: 1091-6490. doi: 10.1073/pnas.1111241108 (cit. on p. 15).

[222] A. E. Pasquinelli. The primary target of let-7 microRNA. Biochemical Society transactions, 41 (4) (2013): 821–4. issn: 1470-8752. doi: 10.1042/BST20130020 (cit. on p. 15).

[223] D. T. Farmer et al. Partially penetrant postnatal lethality of an epithelial specific MicroRNA in a mouse knockout. PloS one, 8 (10) (2013): e76634. issn: 1932-6203. doi: 10.1371/journal.pone.0076634 (cit. on p. 15).

[224] Y. Kawahara. Human diseases caused by germline and somatic abnormalities in microRNA and microRNA-related genes. Congenital anomalies, 54 (1) (2014): 12–21. issn: 1741-4520. doi: 10.1111/cga.12043 (cit. on p. 15).

[225] J. J. Crowley et al. Disruption of the microRNA 137 primary transcript results in early embryonic lethality in mice. Biological psychiatry, 77 (2) (2015): e5–7. issn: 1873-2402. doi: 10.1016/j.biopsych.2014.05.022 (cit. on p. 15).

[226] L. P. Lim et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 433 (7027) (2005): 769–73. issn: 1476-4687. doi: 10.1038/nature03315 (cit. on p. 15).

[227] A. Krek et al. Combinatorial microRNA target predictions. Nature genetics, 37 (5) (2005): 495–500. issn: 1061-4036. doi: 10.1038/ng1536 (cit. on p. 15).

[228] A. Stark et al. Animal microRNAs confer robustness to gene expression and have a significant impact on 3???UTR evolution. Cell, 123 (6) (2005): 1133–1146. issn: 0092-8674. doi: 10.1016/j.cell.2005.11.023 (cit. on p. 15).

[229] D. Baek et al. The impact of microRNAs on protein output. Nature, 455 (7209) (2008): 64–71. issn: 1476-4687. doi: 10.1038/nature07242 (cit. on p. 15).

[230] M. Selbach et al. Widespread changes in protein synthesis induced by microRNAs. Nature, 455 (7209) (2008): 58–63. issn: 1476-4687. doi: 10.1038/nature07228 (cit. on p. 15).

[231] H. Herranz and S. M. Cohen. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes & development, 24 (13) (2010): 1339–44. issn: 1549-5477. doi: 10.1101/gad.1937010 (cit. on p. 15).

[232] S. Mukherji et al. MicroRNAs can generate thresholds in target gene expression. Nature Genetics, 43 (9) (2011): 854–859. issn: 1061-4036. doi: 10.1038/ng.905 (cit. on p. 15).

[233] S. JM et al. Gene expression. MicroRNA control of protein expression noise. Science. 348 (6230) (2015): 128–32 (cit. on p. 15).

[234] E. E. W. Cohen and M. R. Rosner. MicroRNA-regulated feed forward loop network. Cell cycle (Georgetown, Tex.) 8 (16) (2009): 2477–2478. issn: 1551-4005 (cit. on p. 15).

[235] R. Avraham and Y. Yarden. Regulation of signalling by microRNAs. Biochemical Society transactions, 40 (1) (2012): 26–30. issn: 1470-8752. doi: 10.1042/BST20110623 (cit. on pp. 15, 20).

[236] H. Kang and A. Hata. The role of microRNAs in cell fate determination of mesenchymal stem cells: balancing adipogenesis and osteogenesis. BMB reports (2014). issn: 1976-670X (cit. on p. 15).

[237] M. Lagos-Quintana et al. New microRNAs from mouse and human. RNA (New York, N.Y.) 9 (2) (2003): 175–9. issn: 1355-8382 (cit. on p. 15).

[238] A. A. Aravin et al. The small RNA profile during Drosophila melanogaster development. Developmental cell, 5 (2) (2003): 337–50. issn: 1534-5807 (cit. on p. 15).

[239] P. Landgraf et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell, 129 (7) (2007): 1401–1414. issn: 0092-8674. doi: 10.1016/j.cell.2007.04.040 (cit. on p. 15).

[240] A. E. Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nature reviews. Genetics, 13 (4) (2012): 271–82. issn: 1471-0064. doi: 10.1038/nrg3162 (cit. on p. 15).

[241] G. M. Sundaram and P. Sampath. Regulation of context-specific gene expression by posttranscriptional switches. Transcription, 4 (5) (2013): 213–216. issn: 2154-1272 (cit. on p. 15).

[242] J. Hausser and M. Zavolan. Identification and consequences of miRNA-target interactions - beyond repression of gene expression. Nature reviews. Genetics, 15 (9) (2014): 599–612. issn: 1471-0064. doi: 10.1038/nrg3765 (cit. on p. 15).

[243] A. Marson et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell, 134 (3) (2008): 521–533. issn: 1097-4172. doi: 10.1016/j.cell.2008.07.020 (cit. on p. 15).

[244] F. Ozsolak et al. Chromatin structure analyses identify miRNA promoters. Genes & development, 22 (22) (2008): 3172–83. issn: 0890-9369. doi: 10.1101/gad.1706508 (cit. on p. 15).

[245] Y. Lee et al. MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal, 23 (20) (2004): 4051–4060. issn: 0261-4189. doi: 10.1038/sj.emboj.7600385 (cit. on p. 15).

[246] Z. Paroo et al. Biochemical mechanisms of the RNA-induced silencing complex. Cell research, 17 (3) (2007): 187–94. issn: 1748-7838. doi: 10.1038/sj.cr.7310148 (cit. on p. 15).

[247] A. van den Berg et al. RISC-target interaction: cleavage and translational suppression. Biochimica et biophysica acta, 1779 (11) (2008): 668–677. issn: 0006-3002. doi: 10.1016/j.bbagrm.2008.07.005 (cit. on p. 15).

[248] T. Kawamata and Y. Tomari. Making RISC. Trends in biochemical sciences, 35 (7) (2010): 368–76. issn: 0968-0004. doi: 10.1016/j.tibs.2010.03.009 (cit. on p. 15).

[249] J. Azevedo et al. Taking RISCs with Ago hookers. Current opinion in plant biology, 14 (5) (2011): 594–600. issn: 1879-0356. doi: 10.1016/j.pbi.2011.07.002 (cit. on p. 15).

[250] E. Berezikov. Evolution of microRNA diversity and regulation in animals. Nature reviews. Genetics, 12 (12) (2011): 846–60. issn: 1471-0064. doi: 10.1038/nrg3079 (cit. on p. 15).

[251] M. Ha and V. N. Kim. Regulation of microRNA biogenesis. Nature reviews. Molecular cell biology, 15 (8) (2014): 509–24. issn: 1471-0080. doi: 10.1038/nrm3838 (cit. on p. 15).

[252] M. Salmanidis et al. Direct transcriptional regulation by nuclear microRNAs. The international journal of biochemistry & cell biology, 54 (2014): 304–11. issn: 1878-5875. doi: 10.1016/j.biocel.2014.03.010 (cit. on p. 15).

[253] A. M. Gurtan and P. A. Sharp. The role of miRNAs in regulating gene expression networks. Journal of molecular biology, 425 (19) (2013): 3582–3600. issn: 1089-8638. doi: 10.1016/j.jmb.2013.03.007 (cit. on p. 15).

[254] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116 (2) (2004): 281–97. issn: 0092-8674 (cit. on p. 15).

[255] E. A. Clark et al. Concise review: MicroRNA function in multipotent mesenchymal stromal cells. Stem cells, 32 (5) (2014): 1074–82. issn: 1549-4918 (cit. on p. 15).

[256] T. A. Farazi et al. miRNAs in human cancer. The Journal of pathology, 223 (2) (2011): 102–115. issn: 1096-9896. doi: 10.1002/path.2806 (cit. on p. 15).

[257] A. Zampetaki and M. Mayr. MicroRNAs in vascular and metabolic disease. Circulation research, 110 (3) (2012): 508–22. issn: 1524-4571. doi: 10.1161/CIRCRESAHA.111.247445 (cit. on p. 15).

[258] R. Jackstadt and H. Hermeking. MicroRNAs as regulators and mediators of c-MYC function. Biochimica et biophysica acta, 1849 (5) (2015): 544–553. issn: 0006-3002. doi: 10.1016/j.bbagrm.2014.04.003 (cit. on p. 15).

[259] E. Bronze-da Rocha. MicroRNAs expression profiles in cardiovascular diseases. BioMed research international, 2014 (2014): 985408. issn: 2314-6141. doi: 10.1155/2014/985408 (cit. on p. 15).

[260] S. Jia et al. MicroRNAs regulate immune system via multiple targets. Discovery medicine, 18 (100) (2014): 237–47. issn: 1944-7930 (cit. on p. 15).

[261] K Musilova and M Mraz. MicroRNAs in B-cell lymphomas: how a complex biology gets more complex. Leukemia, 29 (5) (2015): 1004–1017. issn: 1476-5551. doi: 10.1038/leu.2014.351 (cit. on p. 15).

[262] P. Arner and A. Kulyté. MicroRNA regulatory networks in human adipose tissue and obesity. Nature reviews. Endocrinology, 11 (5) (2015): 276–288. issn: 1759-5037. doi: 10.1038/nrendo.2015.25 (cit. on pp. 15, 25).

[263] S. Zhao and M.-F. Liu. Mechanisms of microRNA-mediated gene regulation. Science in China. Series C, Life sciences / Chinese Academy of Sciences, 52 (12) (2009): 1111–6. issn: 1862-2798. doi: 10.1007/s11427-009-0152-y (cit. on p. 16).

[264] A. K. L. Leung and P. A. Sharp. Quantifying Argonaute proteins in and out of GW/P-bodies: implications in microRNA activities. Advances in experimental medicine and biology, 768 (2013): 165–182. issn: 0065-2598. doi: 10.1007/978-1-4614-5107-5{_}10 (cit. on p. 16).

[265] I. Salido-Guadarrama et al. MicroRNAs transported by exosomes in body fluids as mediators of intercellular communication in cancer. OncoTargets and therapy, 7 (2014): 1327–1338. issn: 1178-6930. doi: 10.2147/OTT.S61562 (cit. on p. 16).

[266] Y. Li et al. Transport of microRNAs via exosomes. Nature reviews. Cardiology, 12 (4) (2015): 198. issn: 1759-5010. doi: 10.1038/nrcardio.2014.207-c1 (cit. on p. 16).

[267] C. Guay et al. Horizontal transfer of exosomal microRNAs transduce apoptotic signals between pancreatic beta-cells. Cell communication and signaling : CCS, 13 (1) (2015): 17. issn: 1478-811X. doi: 10.1186/s12964-015-0097-7 (cit. on p. 16).

[268] S. Das and M. K. Halushka. Extracellular vesicle microRNA transfer in cardiovascular disease. Cardiovascular pathology : the official journal of the Society for Cardiovascular Pathology (2015). issn: 1879-1336. doi: 10.1016/j.carpath.2015.04.007 (cit. on p. 16).

[269] W. C. Merrick. Mechanism and regulation of eukaryotic protein synthesis. Microbiological reviews, 56 (2) (1992): 291–315. issn: 0146-0749 (cit. on p. 16).

[270] F. Gebauer and M. W. Hentze. Molecular mechanisms of translational control. Nature reviews. Molecular cell biology, 5 (10) (2004): 827–35. issn: 1471-0072. doi: 10.1038/nrm1488 (cit. on p. 16).

[271] B Bilanges and D Stokoe. Mechanisms of translational deregulation in human tumors and therapeutic intervention strategies. Oncogene, 26 (41) (2007): 5973–90. issn: 0950-9232. doi: 10.1038/sj.onc.1210431 (cit. on p. 16).

[272] P. Babitzke et al. Regulation of translation initiation by RNA binding proteins. Annual review of microbiology, 63 (2009): 27–44. issn: 1545-3251. doi: 10.1146/annurev.micro.091208.073514 (cit. on p. 16).

[273] Y. Zhang et al. Coordinated regulation of protein synthesis and degradation by mTORC1. Nature, 513 (7518) (2014): 440–443. issn: 0028-0836. doi: 10.1038/nature13492 (cit. on p. 16).

[274] M. Showkat et al. mTOR Signaling in Protein Translation Regulation: Implications in Cancer Genesis and Therapeutic Interventions. Molecular biology international, 2014 (2014): 686984. issn: 2090-2182. doi: 10.1155/2014/686984 (cit. on p. 16).

[275] S. Kirchner and Z. Ignatova. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature Reviews Genetics, 16 (2) (2014): 98–112. issn: 1471-0056. doi: 10.1038/nrg3861 (cit. on p. 16).

[276] K. Okunishi et al. Inhibition of protein translation as a novel mechanism for prostaglandin E2 regulation of cell functions. FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 28 (1) (2014): 56–66. issn: 1530-6860. doi: 10.1096/fj.13-231720 (cit. on p. 16).

[277] C. A. Piccirillo et al. Translational control of immune responses: from transcripts to translatomes. Nature immunology, 15 (6) (2014): 503–11. issn: 1529-2916. doi: 10.1038/ni.2891 (cit. on p. 16).

[278] J. Wang et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome research, 22 (9) (2012): 1798–1812. issn: 1549-5469. doi: 10.1101/gr.139105.112 (cit. on p. 17).

[279] K. Chen and N. Rajewsky. The evolution of gene regulation by transcription factors and microRNAs. Nature reviews. Genetics, 8 (2) (2007): 93–103. issn: 1471-0056. doi: 10.1038/nrg1990 (cit. on p. 17).

[280] S Arora et al. miRNA-transcription factor interactions: a combinatorial regulation of gene expression. Molecular genetics and genomics : MGG, 288 (3-4) (2013): 77–87. issn: 1617-4623. doi: 10.1007/s00438-013-0734-z (cit. on p. 17).

[281] D. Weichenhan and C. Plass. The evolving epigenome. Human molecular genetics, 22 (R1) (2013): R1–6. issn: 1460-2083. doi: 10.1093/hmg/ddt348 (cit. on p. 17).

[282] D. Holoch and D. Moazed. RNA-mediated epigenetic regulation of gene expression. Nature reviews. Genetics, 16 (2) (2015): 71–84. issn: 1471-0064. doi: 10.1038/nrg3863 (cit. on p. 17).

[283] F. Pelisch et al. RNA metabolism and ubiquitin/ubiquitin-like modifications collide. Briefings in functional genomics, 12 (1) (2013): 66–71. issn: 2041-2657. doi: 10.1093/bfgp/els053 (cit. on p. 17).

[284] K. Ge et al. Transcription coactivator TRAP220 is required for PPAR gamma 2-stimulated adipogenesis. Nature, 417 (6888) (2002): 563–7. issn: 0028-0836. doi: 10.1038/417563a (cit. on pp. 17, 33).

[285] M. Stumpf et al. The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. Proceedings of the National Academy of Sciences of the United States of America, 103 (49) (2006): 18504–18509. issn: 0027-8424. doi: 10.1073/pnas.0604494103 (cit. on p. 17).

[286] L. Sinkkonen et al. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nature Structural & Molecular Biology, 15 (3) (2008): 259–67. issn: 15459985 (cit. on p. 17).

[287] J. M. W. Slack. Conrad Hal Waddington: the last Renaissance biologist? Nature reviews. Genetics, 3 (11) (2002): 889–95. issn: 1471-0056. doi: 10.1038/nrg933 (cit. on p. 18).

[288] D. Haig. Commentary: The epidemiology of epigenetics. International journal of epidemiology, 41 (1) (2012): 13–6. issn: 1464-3685. doi: 10.1093/ije/dyr183 (cit. on p. 18).

[289] F. Nicol-Benoit et al. Drawing a Waddington landscape to capture dynamic epigenetics. Biology of the cell / under the auspices of the European Cell Biology Organization, 105 (12) (2013): 576–84. issn: 1768-322X. doi: 10.1111/boc.201300029 (cit. on p. 18).

[290] J. Baedke. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. Studies in history and philosophy of biological and biomedical sciences, 44 (4 Pt B) (2013): 756–73. issn: 1879-2499. doi: 10.1016/j.shpsc.2013.06.001 (cit. on p. 18).

[291] J. Ladewig et al. Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. Nature reviews. Molecular cell biology, 14 (4) (2013): 225–36. issn: 1471-0080 (cit. on p. 18).

[292] P. Wang et al. Epigenetic state network approach for describing cell phenotypic transitions. Interface focus, 4 (3) (2014): 20130068. issn: 2042-8898. doi: 10.1098/rsfs.2013.0068 (cit. on p. 18).

[293] S. S. P. Rao et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell, 159 (7) (2014): 1665–1680. issn: 00928674. doi: 10.1016/j.cell.2014.11.021 (cit. on p. 18).

[294] M. Iwafuchi-Doi and K. S. Zaret. Pioneer transcription factors in cell reprogramming. Genes & Development, 28 (24) (2014): 2679–2692. issn: 0890-9369. doi: 10.1101/gad.253443.114 (cit. on p. 18).

[295] Modeling transcriptional regulation. Nature genetics, 47 (1) (2015): 1. issn: 1546-1718. doi: 10.1038/ng.3188 (cit. on p. 18).

[296] A. Soufi et al. Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell, 161 (3) (2015): 555–68. issn: 00928674. doi: 10.1016/j.cell.2015.03.017 (cit. on p. 18).

[297] Y. Tao et al. Nucleosome organizations in induced pluripotent stem cells reprogrammed from somatic cells belonging to three different germ layers. BMC biology, 12 (1) (2014): 109. issn: 1741-7007. doi: 10.1186/s12915-014-0109-x (cit. on p. 18).

[298] R. Di Micco et al. Control of embryonic stem cell identity by BRD4-dependent transcriptional elongation of super-enhancer-associated pluripotency genes. Cell reports, 9 (1) (2014): 234–247. issn: 2211-1247. doi: 10.1016/j.celrep.2014.08.055 (cit. on p. 18).

[299] M. Achour et al. Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice. Human molecular genetics (2015). issn: 1460-2083. doi: 10.1093/hmg/ddv099 (cit. on pp. 18, 162).

[300] J. Brown et al. NF-κB Directs Dynamic Super Enhancer Formation in Inflammation and Atherogenesis. Molecular Cell, 56 (2) (2014): 219–31. issn: 10972765. doi: 10.1016/j.molcel.2014.08.024 (cit. on p. 18).

[301] J. Lovén et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell, 153 (2) (2013): 320–334. issn: 1097-4172. doi: 10.1016/j.cell.2013.03.036 (cit. on p. 18).

[302] J. Qian et al. B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. Cell, 159 (7) (2014): 1524–37. issn: 00928674. doi: 10.1016/j.cell.2014.11.013 (cit. on p. 18).

[303] M. R. Mansour et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science, 346 (6215) (2014): 1373–1377. issn: 0036-8075. doi: 10.1126/science.1259037 (cit. on p. 18).

[304] H. Zhou et al. Epstein-Barr virus oncoprotein super-enhancers control B cell growth. Cell host & microbe, 17 (2) (2015): 205–16. issn: 1934-6069. doi: 10.1016/j.chom.2014.12.013 (cit. on p. 18).

[305] E. M. Tomazou et al. Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. Cell reports, 10 (7) (2015): 1082–95. issn: 2211-1247. doi: 10.1016/j.celrep.2015.01.042 (cit. on p. 18).

[306] D. Hnisz et al. Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Molecular cell (2015). issn: 1097-4164. doi: 10.1016/j.molcel.2015.02.014 (cit. on pp. 18, 153, 160, 162).

[307] S. C. J. Parker et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences of the United States of America, 110 (44) (2013): 17921–17926. issn: 1091-6490. doi: 10.1073/pnas.1317023110 (cit. on pp. 18, 153).

[308] G. Vahedi et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. Nature, 520 (7548) (2015): 558–62. issn: 0028-0836. doi: 10.1038/nature14154 (cit. on p. 18).

[309] N. Perrimon et al. Signaling mechanisms controlling cell fate and embryonic patterning. Cold Spring Harbor perspectives in biology, 4 (8) (2012): a005975. issn: 1943-0264. doi: 10.1101/cshperspect.a005975 (cit. on p. 18).

[310] N. J. Martinez and A. J. M. Walhout. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. BioEssays : news and reviews in molecular, cellular and developmental biology, 31 (4)

(2009): 435–445. issn: 1521-1878. doi: 10.1002/bies.200800212 (cit. on p. 19).

[311] S. L. Schreiber and B. E. Bernstein. Signaling network model of chromatin. Cell, 111 (6) (2002): 771–8. issn: 0092-8674 (cit. on p. 18).

[312] C.-C. Lin et al. Crosstalk between transcription factors and microRNAs in human protein interaction network. BMC systems biology, 6 (2012): 18. issn: 1752-0509. doi: 10.1186/1752-0509-6-18 (cit. on p. 18).

[313] I. Kosti et al. An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors. PLoS computational biology, 8 (7) (2012): e1002603. issn: 1553-7358. doi: 10.1371/journal.pcbi.1002603 (cit. on p. 18).

[314] T. G. Kahn et al. Combinatorial interactions are required for the efficient recruitment of pho repressive complex (PhoRC) to polycomb response elements. PLoS genetics, 10 (7) (2014): e1004495. issn: 1553-7404. doi: 10.1371/journal.pgen.1004495 (cit. on p. 18).

[315] A. Cappuccio et al. Combinatorial code governing cellular responses to complex stimuli. Nature communications, 6 (2015): 6847. issn: 2041-1723. doi: 10.1038/ncomms7847 (cit. on p. 18).

[316] J. H. A. Martens and H. G. Stunnenberg. BLUEPRINT: mapping human blood cell epigenomes. Haematologica, 98 (10) (2013): 1487–1489. issn: 1592-8721. doi: 10.3324/haematol.2013.094243 (cit. on pp. 19, 52, 154).

[317] R. E. Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature, 518 (7539) (2015): 317–330. issn: 0028-0836. doi: 10.1038/nature14248 (cit. on pp. 19, 52, 154, 163).

[318] R. E. Thurman et al. The accessible chromatin landscape of the human genome. Nature, 489 (7414) (2012): 75–82. issn: 1476-4687. doi: 10.1038/nature11232 (cit. on p. 19).

[319] M. B. Gerstein et al. Architecture of the human regulatory network derived from ENCODE data. Nature, 489 (7414) (2012): 91–100. issn: 1476-4687. doi: 10.1038/nature11245 (cit. on p. 19).

[320] S. Djebali et al. Landscape of transcription in human cells. Nature, 489 (7414) (2012): 101–8. issn: 1476-4687. doi: 10.1038/nature11233 (cit. on p. 19).

[321] A. Sanyal et al. The long-range interaction landscape of gene promoters. Nature, 489 (7414) (2012): 109–113. issn: 1476-4687. doi: 10.1038/nature11279 (cit. on pp. 19, 162).

[322] S. Neph et al. Circuitry and dynamics of human transcription factor regulatory networks. Cell, 150 (6) (2012): 1274–1286. issn: 1097-4172. doi: 10.1016/j.cell.2012.04.040 (cit. on p. 19).

[323] J. I. Fuxman Bass et al. Human Gene-Centered Transcription Factor Networks for Enhancers and Disease Variants. Cell, 161 (3) (2015): 661–673. issn: 00928674. doi: 10.1016/j.cell.2015.03.003 (cit. on p. 19).

[324] C. S. Greene et al. Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics (2015). issn: 1061-4036. doi: 10.1038/ng.3259 (cit. on p. 19).

[325] E. Pierson et al. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLOS Computational Biology, 11 (5) (2015). Ed. by I. Rigoutsos: e1004220. issn: 1553-7358. doi: 10.1371/journal.pcbi.1004220 (cit. on p. 19).

[326] J. Tsang et al. *MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. - PubMed - NCBI*. 2007. doi: 10.1016/j.molcel.2007.05.018 (cit. on p. 20).

[327] N. Heidari et al. Genome-wide map of regulatory interactions in the human genome. Genome Research, 24 (12) (2014): 1905–1917. issn: 1088-9051. doi: 10.1101/gr.176586.114 (cit. on p. 20).

[328] P. Wang et al. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. Nucleic Acids Research (2015). issn: 0305-1048. doi: 10.1093/nar/gkv398 (cit. on p. 20).

[329] A. Ay et al. Hierarchical decomposition of dynamically evolving regulatory networks. BMC bioinformatics, 16 (1) (2015): 161. issn: 1471-2105. doi: 10.1186/s12859-015-0529-9 (cit. on p. 20).

[330] N. Le Novère. Quantitative and logic modelling of molecular and gene networks. Nature reviews. Genetics, 16 (3) (2015): 146–58. issn: 1471-0064. doi: 10.1038/nrg3885 (cit. on p. 20).

[331] A. Mardinoglu et al. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications, 5 (2014): 3083. issn: 2041-1723. doi: 10.1038/ncomms4083 (cit. on p. 20).

[332] S. Sahoo et al. Modeling the effects of commonly used drugs on human metabolism. The FEBS journal, 282 (2) (2015): 297–317. issn: 1742-4658. doi: 10.1111/febs.13128 (cit. on p. 20).

[333] B. Desvergne et al. Transcriptional regulation of metabolism. Physiological reviews, 86 (2) (2006): 465–514. issn: 0031-9333. doi: `10.1152/physrev.00025.2005` (cit. on p. 21).

[334] F. Wessely et al. Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. Molecular systems biology, 7 (2011): 515. issn: 1744-4292. doi: `10.1038/msb.2011.46` (cit. on p. 21).

[335] C. M. Metallo and M. G. Vander Heiden. Understanding metabolic regulation and its influence on cell physiology. Molecular cell, 49 (3) (2013): 388–398. issn: 1097-4164. doi: `10.1016/j.molcel.2013.01.018` (cit. on p. 22).

[336] K. Ganeshan and A. Chawla. Metabolic regulation of immune responses. Annual review of immunology, 32 (2014): 609–34. issn: 1545-3278. doi: `10.1146/annurev-immunol-032713-120236` (cit. on p. 22).

[337] E. Reznik and C. Sander. Extensive Decoupling of Metabolic Genes in Cancer. PLoS computational biology, 11 (5) (2015): e1004176. issn: 1553-7358. doi: `10.1371/journal.pcbi.1004176` (cit. on p. 22).

[338] L. C. Duffy et al. Progress and challenges in developing metabolic footprints from diet in human gut microbial cometabolism. The Journal of nutrition, 145 (5) (2015): 1123S–1130S. issn: 1541-6100. doi: `10.3945/jn.114.194936` (cit. on p. 22).

[339] R Kaddurah-Daouk and R Weinshilboum. Metabolomic Signatures for Drug Response Phenotypes: Pharmacometabolomics Enables Precision Medicine. Clinical pharmacology and therapeutics, 98 (1) (2015): 71–5. issn: 1532-6535. doi: `10.1002/cpt.134` (cit. on p. 22).

[340] R. Caspi et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research, 42 (Database issue) (2014): D459–71. issn: 1362-4962. doi: `10.1093/nar/gkt1103` (cit. on p. 23).

[341] M. Kanehisa et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research, 42 (Database issue) (2014): D199–205. issn: 1362-4962. doi: `10.1093/nar/gkt1076` (cit. on pp. 23, 53).

[342] A. Chang et al. BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic acids research, 43 (Database issue) (2015): D439–46. issn: 1362-4962. doi: `10.1093/nar/gku1068` (cit. on p. 23).

[343] D. Croft et al. The Reactome pathway knowledgebase. Nucleic acids research, 42 (Database issue) (2014): D472–7. issn: 1362-4962. doi: `10.1093/nar/gkt1102` (cit. on p. 23).

[344] T. Kelder et al. WikiPathways: building research communities on biological pathways. Nucleic acids research, 40 (Database issue) (2012): D1301–7. issn: 1362-4962. doi: 10.1093/nar/gkr1074 (cit. on p. 23).

[345] A. Kamburov et al. The ConsensusPathDB interaction database: 2013 update. Nucleic acids research, 41 (Database issue) (2013): D793–800. issn: 1362-4962. doi: 10.1093/nar/gks1055 (cit. on p. 23).

[346] T. Jewison et al. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. Nucleic Acids Research, 42 (D1) (2013): D478–D484. issn: 0305-1048. doi: 10.1093/nar/gkt1067 (cit. on p. 23).

[347] S. Hino et al. Metabolism-epigenome crosstalk in physiology and diseases. Journal of human genetics, 58 (7) (2013): 410–415. issn: 1435-232X. doi: 10.1038/jhg.2013.57 (cit. on p. 23).

[348] T. Goto et al. Natural compounds regulate energy metabolism by the modulating the activity of lipid-sensing nuclear receptors. Molecular nutrition & food research, 57 (1) (2013): 20–33. issn: 1613-4133. doi: 10.1002/mnfr.201200522 (cit. on p. 24).

[349] A. V. Contreras et al. PPAR-α as a key nutritional and environmental sensor for metabolic adaptation. Advances in nutrition, 4 (4) (2013): 439–452. issn: 2156-5376. doi: 10.3945/an.113.003798 (cit. on p. 24).

[350] E. A. Mazzio and K. F. A. Soliman. Epigenetics and nutritional environmental signals. Integrative and comparative biology, 54 (1) (2014): 21–30. issn: 1557-7023. doi: 10.1093/icb/icu049 (cit. on p. 24).

[351] D. F. Romagnolo et al. Nuclear receptors and epigenetic regulation: opportunities for nutritional targeting and disease prevention. Advances in nutrition (Bethesda, Md.) 5 (4) (2014): 373–85. issn: 2156-5376. doi: 10.3945/an.114.005868 (cit. on p. 24).

[352] M. Giudici et al. Nuclear Receptor Coregulators in Metabolism and Disease. Handbook of experimental pharmacology (2015). issn: 0171-2004. doi: 10.1007/164{_}2015{_}5 (cit. on p. 24).

[353] L. Mouchiroud et al. Transcriptional coregulators: fine-tuning metabolism. Cell metabolism, 20 (1) (2014): 26–40. issn: 1932-7420. doi: 10.1016/j.cmet.2014.03.027 (cit. on p. 24).

[354] L. Galdieri et al. Protein acetylation and acetyl coenzyme a metabolism in budding yeast. Eukaryotic cell, 13 (12) (2014): 1472–1483. issn: 1535-9786. doi: 10.1128/EC.00189-14 (cit. on p. 24).

[355] J.-W. Kim and C. V. Dang. Multifaceted roles of glycolytic enzymes. Trends in biochemical sciences, 30 (3) (2005): 142–50. issn: 0968-0004. doi: 10.1016/j.tibs.2005.01.005 (cit. on p. 24).

[356] N. Reynolds et al. Transcriptional repressors: multifaceted regulators of gene expression. Development (Cambridge, England), 140 (3) (2013): 505–12. issn: 1477-9129. doi: 10.1242/dev.083105 (cit. on p. 24).

[357] S. J. H. Ricoult and B. D. Manning. The multifaceted role of mTORC1 in the control of lipid metabolism. EMBO reports, 14 (3) (2013): 242–251. issn: 1469-3178. doi: 10.1038/embor.2013.5 (cit. on p. 24).

[358] L. Hebbard and B. Ranscht. Multifaceted roles of adiponectin in cancer. Best practice & research. Clinical endocrinology & metabolism, 28 (1) (2014): 59–69. issn: 1878-1594. doi: 10.1016/j.beem.2013.11.005 (cit. on p. 24).

[359] A. Tchicaya and N. Lorentz. Vivre au luxembourg. CEPS/INSTEAD, 66 (2010) (cit. on p. 24).

[360] D. J. Hunter. Gene-environment interactions in human diseases. Nature reviews. Genetics, 6 (4) (2005): 287–98. issn: 1471-0056. doi: 10.1038/nrg1578 (cit. on p. 25).

[361] C Lavebratt et al. Epigenetic regulation in obesity. International journal of obesity (2005), 36 (6) (2012): 757–65. issn: 1476-5497. doi: 10.1038/ijo.2011.178 (cit. on p. 25).

[362] S. Podder and T. C. Ghosh. Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. Molecular biology and evolution, 27 (4) (2010): 934–41. issn: 1537-1719. doi: 10.1093/molbev/msp297 (cit. on p. 25).

[363] T. Rolland et al. A Proteome-Scale Map of the Human Interactome Network. Cell, 159 (5) (2014): 1212–1226. issn: 00928674. doi: 10.1016/j.cell.2014.10.050 (cit. on p. 25).

[364] M. Vidal et al. Interactome networks and human disease. Cell, 144 (6) (2011): 986–998. issn: 1097-4172. doi: 10.1016/j.cell.2011.02.016 (cit. on p. 25).

[365] J. L. Thorne and M. J. Campbell. Nuclear receptors and the Warburg effect in cancer. International journal of cancer. Journal international du cancer (2014). issn: 1097-0215. doi: 10.1002/ijc.29012 (cit. on p. 25).

[366] K.-I. Goh et al. The human disease network. Proceedings of the National Academy of Sciences of the United States of America, 104 (21) (2007): 8685–8690. issn: 0027-8424. doi: 10.1073/pnas.0701361104 (cit. on pp. 25, 27).

[367] J. Park et al. The impact of cellular networks on disease comorbidity. Molecular systems biology, 5 (2009): 262. issn: 1744-4292. doi: 10.1038/msb.2009.16 (cit. on p. 25).

[368] N. Sahni et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell, 161 (3) (2015): 647–660. issn: 00928674. doi: 10.1016/j.cell.2015.04.013 (cit. on p. 25).

[369] L. I. Furlong. Human diseases through the lens of network biology. Trends in genetics : TIG, 29 (3) (2013): 150–159. issn: 0168-9525. doi: 10.1016/j.tig.2012.11.004 (cit. on p. 25).

[370] M. Gustafsson et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome medicine, 6 (10) (2014): 82. issn: 1756-994X. doi: 10.1186/s13073-014-0082-6 (cit. on p. 25).

[371] S. D. Ghiassian et al. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLOS Computational Biology, 11 (4) (2015). Ed. by A. Rzhetsky: e1004120. issn: 1553-7358. doi: 10.1371/journal.pcbi.1004120 (cit. on p. 25).

[372] A.-L. Barabási et al. Network medicine: a network-based approach to human disease. Nature reviews. Genetics, 12 (1) (2011): 56–68. issn: 1471-0064. doi: 10.1038/nrg2918 (cit. on pp. 26, 27).

[373] A. Hamosh et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 33 ((Database issue)) (2005): D514–7. issn: 03051048. doi: 10.1093/nar/gki033 (cit. on p. 25).

[374] The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Research, 43 (Database issue) (2014): D204–12. issn: 0305-1048. doi: 10.1093/nar/gku989 (cit. on p. 25).

[375] A. P. Davis et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic acids research, 43 (Database issue) (2015): D914–20. issn: 1362-4962. doi: 10.1093/nar/gku935 (cit. on pp. 25, 54).

[376] M. J. Landrum et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, 42 (Database issue) (2014): D980–5. issn: 1362-4962. doi: 10.1093/nar/gkt1113 (cit. on p. 25).

[377] K. G. Becker et al. The genetic association database. Nature genetics, 36 (5) (2004): 431–2. issn: 1061-4036. doi: 10.1038/ng0504-431 (cit. on p. 25).

[378] J. Pinero et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015 (2015): bav028–bav028. issn: 1758-0463. doi: 10.1093/database/bav028 (cit. on pp. 26, 52).

[379] D.-S. Lee et al. The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences of the United States of America, 105 (29) (2008): 9880–5. issn: 1091-6490. doi: 10.1073/pnas.0802208105 (cit. on p. 27).

[380] P. Braun et al. Networking metabolites and diseases. Proceedings of the National Academy of Sciences of the United States of America, 105 (29) (2008): 9849–9850. issn: 1091-6490. doi: 10.1073/pnas.0805644105 (cit. on p. 27).

[381] J. Menche et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science, 347 (6224) (2015): 1257601. issn: 1095-9203. doi: 10.1126/science.1257601 (cit. on p. 27).

[382] X. Zhou et al. Human symptoms-disease network. Nature communications, 5 (2014): 4212. issn: 2041-1723. doi: 10.1038/ncomms5212 (cit. on p. 27).

[383] M. Ehrlich. DNA hypomethylation in cancer cells. Epigenomics, 1 (2) (2009): 239–259. issn: 1750-192X. doi: 10.2217/epi.09.33 (cit. on p. 27).

[384] W.-H.-O.-D. of Noncommunicable-Disease-Surveillance. *Definition, diagnosis and classification of diabetes mellitus and its complications : report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus*. Tech. rep. 1999 (cit. on p. 28).

[385] M. H. Abu Bakar et al. Metabolomics - the complementary field in systems biology: a review on obesity and type 2 diabetes. Molecular bioSystems, 11 (7) (2015): 1742–74. issn: 1742-2051. doi: 10.1039/c5mb00158g (cit. on p. 28).

[386] J. A. Martínez et al. Epigenetics in adipose tissue, obesity, weight loss, and diabetes. Advances in nutrition, 5 (1) (2014): 71–81. issn: 2156-5376. doi: 10.3945/an.113.004705 (cit. on p. 28).

[387] D. E. Schones et al. Chromatin Modifications Associated With Diabetes and Obesity. Arteriosclerosis, thrombosis, and vascular biology, 35 (7) (2015): 1557–61. issn: 1524-4636. doi: 10.1161/ATVBAHA.115.305041 (cit. on pp. 28, 29).

[388] H Beck-Nielsen and L. C. Groop. Metabolic and genetic characterization of prediabetic states. Sequence of events leading to non-insulin-dependent diabetes mellitus. The Journal of clinical investigation, 94 (5) (1994): 1714–1721. issn: 0021-9738. doi: 10.1172/JCI117518 (cit. on p. 28).

[389] H. Ding and C. R. Triggle. Endothelial cell dysfunction and the vascular complications associated with type 2 diabetes: assessing the health of the endothelium. Vascular health and risk management, 1 (1) (2005): 55–71. issn: 1176-6344 (cit. on p. 28).

[390] D. Popov. Endothelial cell dysfunction in hyperglycemia: Phenotypic change, intracellular signaling modification, ultrastructural alteration, and potential clinical outcomes. International Journal of Diabetes Mellitus, 2 (3) (2010): 189–195. issn: 18775934. doi: 10.1016/j.ijdm.2010.09.002 (cit. on p. 28).

[391] A. Avogaro et al. Endothelial dysfunction in diabetes: the role of reparatory mechanisms. Diabetes care, 34 Suppl 2 (2011): S285–90. issn: 1935-5548. doi: 10.2337/dc11-s239 (cit. on p. 28).

[392] M. Pangare and A. Makino. Mitochondrial function in vascular endothelial cell in diabetes. Journal of smooth muscle research, 48 (1) (2012): 1–26. issn: 1884-8796 (cit. on p. 28).

[393] M. Virji and D. J. Hill. In vitro models of infection II–human umbilical vein endothelial cells (HUVECs) system. Methods in molecular medicine, 71 (2003): 297–314. issn: 1543-1894 (cit. on p. 28).

[394] H.-J. Park et al. Human umbilical vein endothelial cells and human dermal microvascular endothelial cells offer new insights into the relationship between lipid metabolism and angiogenesis. Stem cell reviews, 2 (2) (2006): 93–102. issn: 1550-8943. doi: 10.1007/s12015-006-0015-x (cit. on p. 28).

[395] M. R. Richardson et al. Venous and arterial endothelial proteomics: mining for markers and mechanisms of endothelial diversity. Expert review of proteomics, 7 (6) (2010): 823–831. issn: 1744-8387. doi: 10.1586/epr.10.92 (cit. on p. 28).

[396] A. Le Guelte and J. Gavard. Role of endothelial cell-cell junctions in endothelial permeability. Methods in molecular biology, 763 (2011): 265–79. issn: 1940-6029. doi: 10.1007/978-1-61779-191-8{_}18 (cit. on p. 28).

[397] I. Hameed et al. Type 2 diabetes mellitus: From a metabolic disorder to an inflammatory condition. World journal of diabetes, 6 (4) (2015): 598–612. issn: 1948-9358. doi: 10.4239/wjd.v6.i4.598 (cit. on p. 29).

[398] K. J. Chang-Chen et al. Beta-cell failure as a complication of diabetes. Reviews in endocrine & metabolic disorders, 9 (4) (2008): 329–43. issn: 1389-9155. doi: 10.1007/s11154-008-9101-5 (cit. on p. 29).

[399] P. K. Crane et al. Glucose levels and risk of dementia. The New England journal of medicine, 369 (6) (2013): 540–548. issn: 1533-4406. doi: 10.1056/NEJMoa1215740 (cit. on p. 29).

[400] C. Ballard et al. Alzheimer's disease. Lancet, 377 (9770) (2011): 1019–31. issn: 1474-547X. doi: 10.1016/S0140-6736(10)61349-9 (cit. on p. 29).

[401] A. Kleinridders et al. Insulin action in brain regulates systemic metabolism and brain function. Diabetes, 63 (7) (2014): 2232–2243. issn: 1939-327X. doi: 10.2337/db14-0568 (cit. on p. 29).

[402] C. Carvalho et al. Increased susceptibility to amyloid-β toxicity in rat brain microvascular endothelial cells under hyperglycemic conditions. Journal of Alzheimer's disease : JAD, 38 (1) (2014): 75–83. issn: 1875-8908. doi: 10.3233/JAD-130464 (cit. on p. 30).

[403] S. L. Macauley et al. Hyperglycemia modulates extracellular amyloid-β concentrations and neuronal activity in vivo. The Journal of clinical investigation, 125 (6) (2015): 2463–2467. issn: 1558-8238. doi: 10.1172/JCI79742 (cit. on p. 30).

[404] J. Moitra et al. Life without white fat: a transgenic mouse. Genes & development, 12 (20) (1998): 3168–3181. issn: 0890-9369. doi: 10.1101/gad.12.20.3168 (cit. on p. 30).

[406] G. De Pergola and F. Silvestris. Obesity as a major risk factor for cancer. Journal of obesity, 2013 (2013): 291546. issn: 2090-0716. doi: 10.1155/2013/291546 (cit. on p. 31).

[407] M. H. Fonseca-Alaniz et al. Adipose tissue as an endocrine organ: from theory to practice. Jornal de Pediatria, 83 (5 Suppl) (2007): S192–203. issn: 0021-7557. doi: 10.2223/JPED.1709 (cit. on p. 31).

[408] S. E. Wozniak et al. Adipose tissue: the new endocrine organ? A review article. Digestive diseases and sciences, 54 (9) (2009): 1847–56. issn: 1573-2568. doi: 10.1007/s10620-008-0585-3 (cit. on p. 31).

[409] S. P. Poulos et al. The development and endocrine functions of adipose tissue. Molecular and cellular endocrinology, 323 (1) (2010): 20–34. issn: 1872-8057. doi: 10.1016/j.mce.2009.12.011 (cit. on pp. 31, 32, 59).

[411] D Tews and M Wabitsch. Renaissance of brown adipose tissue. Hormone research in pædiatrics, 75 (4) (2011): 231–9. issn: 1663-2826. doi: 10.1159/000324806 (cit. on p. 31).

[412] J. Nedergaard et al. Unexpected evidence for active brown adipose tissue in adult humans. American journal of physiology. Endocrinology and metabolism, 293 (2) (2007): E444–52. issn: 0193-1849. doi: 10.1152/ajpendo.00691.2006 (cit. on pp. 31, 159).

[413] B. Cannon and J. Nedergaard. Developmental biology: Neither fat nor flesh. Nature, 454 (7207) (2008): 947–8. issn: 1476-4687. doi: 10.1038/454947a (cit. on p. 31).

[414] A. M. Cypess et al. Identification and Importance of Brown Adipose Tissue in Adult Humans. The New England journal of medicine, 360 (15) (2009): 1509–1517. issn: 1533-4406. doi: 10.1056/NEJMoa0810780 (cit. on p. 31).

[415] P. Lee et al. Brown adipose tissue in adult humans: a metabolic renaissance. Endocrine reviews, 34 (3) (2013): 413–38. issn: 1945-7189. doi: 10.1210/er.2012-1081 (cit. on p. 31).

[416] P. Seale et al. PRDM16 controls a brown fat/skeletal muscle switch. Nature, 454 (7207) (2008): 961–967. issn: 1476-4687. doi: 10.1038/nature07182 (cit. on pp. 31, 32).

[417] Y.-H. Tseng et al. New role of bone morphogenetic protein 7 in brown adipogenesis and energy expenditure. Nature, 454 (7207) (2008): 1000–1004. issn: 1476-4687. doi: 10.1038/nature07221 (cit. on p. 31).

[418] C Guerra et al. Emergence of brown adipocytes in white fat in mice is under genetic control. Effects on body weight and adiposity. The Journal of clinical investigation, 102 (2) (1998): 412–420. issn: 0021-9738. doi: 10.1172/JCI3155 (cit. on p. 31).

[419] B Cousin et al. Occurrence of brown adipocytes in rat white adipose tissue: molecular and morphological characterization. Journal of cell science, 103 (Pt 4 (1992): 931–42. issn: 0021-9533 (cit. on p. 31).

[420] L. P. Kozak. The genetics of brown adipocyte induction in white fat depots. Frontiers in endocrinology, 2 (2011): 64. issn: 1664-2392. doi: 10.3389/fendo.2011.00064 (cit. on p. 31).

[421] J. Wu et al. Adaptive thermogenesis in adipocytes: is beige the new brown? Genes & development, 27 (3) (2013): 234–250. issn: 1549-5477. doi: 10.1101/gad.211649.112 (cit. on p. 31).

[422] M. Giralt and F. Villarroya. White, brown, beige/brite: different adipose cells for different functions? Endocrinology, 154 (9) (2013): 2992–3000. issn: 1945-7170. doi: 10.1210/en.2013-1403 (cit. on p. 31).

[423] M. Rosell et al. Brown and white adipose tissues: intrinsic differences in gene expression and response to cold exposure in mice. American journal of physiology. Endocrinology and metabolism, 306 (8) (2014): E945–64. issn: 1522-1555. doi: 10.1152/ajpendo.00473.2013 (cit. on p. 31).

[424] R. Cereijo et al. Thermogenic brown and beige/brite adipogenesis in humans. Annals of medicine, 47 (2) (2015): 169–77. issn: 1365-2060. doi: 10.3109/07853890.2014.952328 (cit. on pp. 31, 159).

[425] S. Cinti. Transdifferentiation properties of adipocytes in the adipose organ. American journal of physiology. Endocrinology and metabolism, 297 (5) (2009): E977–86. issn: 1522-1555. doi: 10.1152/ajpendo.00183.2009 (cit. on p. 31).

[426] A. Giordano et al. White, brown and pink adipocytes: the extraordinary plasticity of the adipose organ. European journal of endocrinology / European Federation of Endocrine Societies, 170 (5) (2014): R159–71. issn: 1479-683X. doi: 10.1530/EJE-13-0945 (cit. on p. 31).

[427] B. Gustafson and U. Smith. Regulation of white adipogenesis and its relation to ectopic fat accumulation and cardiovascular risk. Atherosclerosis, 241 (1) (2015): 27–35. issn: 1879-1484. doi: 10.1016/j.atherosclerosis.2015.04.812 (cit. on p. 32).

[428] S. Gesta et al. Developmental origin of fat: tracking obesity to its source. Cell, 131 (2) (2007): 242–256. issn: 0092-8674. doi: 10.1016/j.cell.2007.10.004 (cit. on p. 32).

[429] N. Billon and C. Dani. Developmental origins of the adipocyte lineage: new insights from genetics and genomics studies. Stem cell reviews, 8 (1) (2012): 55–66. issn: 1558-6804. doi: 10.1007/s12015-011-9242-x (cit. on p. 32).

[430] J. Sanchez-Gurmaches et al. PTEN loss in the Myf5 lineage redistributes body fat and reveals subsets of white adipocytes that arise from Myf5 precursors. Cell metabolism, 16 (3) (2012): 348–362. issn: 1932-7420. doi: 10.1016/j.cmet.2012.08.003 (cit. on p. 32).

[431] R. Berry and M. S. Rodeheffer. Characterization of the adipocyte cellular lineage inÂăvivo. Nature Cell Biology, 15 (3) (2013): 302–308. issn: 1465-7392. doi: 10.1038/ncb2696 (cit. on p. 32).

[432] J. Sanchez-Gurmaches and D. A. Guertin. Adipocyte lineages: tracing back the origins of fat. Biochimica et biophysica acta, 1842 (3) (2014): 340–351. issn: 0006-3002. doi: 10.1016/j.bbadis.2013.05.027 (cit. on p. 32).

[433] D. C. Berry et al. The developmental origins of adipose tissue. Development (Cambridge, England), 140 (19) (2013): 3939–49. issn: 1477-9129. doi: 10.1242/dev.080549 (cit. on p. 32).

[434] G. Frühbeck et al. BAT: a new target for human obesity? Trends in pharmacological sciences, 30 (8) (2009): 387–96. issn: 1873-3735. doi: 10.1016/j.tips.2009.05.003 (cit. on p. 33).

[435] M. Laudes. Role of WNT signalling in the determination of human mesenchymal stem cells into preadipocytes. Journal of molecular endocrinology, 46 (2) (2011): R65–72. issn: 1479-6813. doi: 10.1530/JME-10-0169 (cit. on p. 32).

[436] A. G. Cristancho and M. A. Lazar. Forming functional fat: a growing understanding of adipocyte differentiation. Nature reviews. Molecular cell biology, 12 (11) (2011): 722–34. issn: 1471-0080. doi: 10.1038/nrm3198 (cit. on pp. 32, 33).

[437] A. W. James. Review of Signaling Pathways Governing MSC Osteogenic and Adipogenic Differentiation. Scientifica, 2013 (2013): 684736. issn: 2090-908X. doi: 10.1155/2013/684736 (cit. on p. 32).

[438] M. M. Musri and M. Párrizas. Epigenetic regulation of adipogenesis. Current opinion in clinical nutrition and metabolic care, 15 (4) (2012): 342–9. issn: 1473-6519. doi: 10.1097/MCO.0b013e3283546fba (cit. on p. 33).

[439] R. R. Bowers et al. Stable stem cell commitment to the adipocyte lineage by inhibition of DNA methylation: role of the BMP-4 gene. Proceedings of the National Academy of Sciences of the United States of America, 103 (35) (2006): 13022–13027. issn: 0027-8424. doi: 10.1073/pnas.0605789103 (cit. on p. 33).

[440] D. J. Steger et al. Propagation of adipogenic signals through an epigenomic transition state. Genes & development, 24 (10) (2010): 1035–1044. issn: 1549-5477. doi: 10.1101/gad.1907110 (cit. on p. 33).

[441] R. Siersbæk et al. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. The EMBO journal, 30 (8) (2011): 1459–1472. issn: 1460-2075. doi: 10.1038/emboj.2011.65 (cit. on p. 33).

[442] J.-E. Lee and K. Ge. Transcriptional and epigenetic regulation of PPARγ expression during adipogenesis. Cell & bioscience, 4 (2014): 29. issn: 2045-3701. doi: 10.1186/2045-3701-4-29 (cit. on p. 33).

[443] J. Ho Lee et al. TonEBP suppresses adipogenesis and insulin sensitivity by blocking epigenetic transition of PPARγ2. Scientific reports, 5 (2015): 10937. issn: 2045-2322. doi: 10.1038/srep10937 (cit. on pp. 33, 34).

[444] E. D. Rosen et al. Transcriptional regulation of adipogenesis. Genes & development, 14 (11) (2000): 1293–307. issn: 0890-9369 (cit. on p. 34).

[445] E. D. Rosen and O. a. MacDougald. Adipocyte differentiation from the inside out. Nature reviews. Molecular cell biology, 7 (12) (2006): 885–96. issn: 1471-0072. doi: 10.1038/nrm2066 (cit. on p. 34).

[446] P Tontonoz et al. Stimulation of adipogenesis in fibroblasts by PPAR gamma 2, a lipid-activated transcription factor. Cell, 79 (7) (1994): 1147–56. issn: 0092-8674 (cit. on p. 34).

[447] N Marx et al. PPARgamma activation in human endothelial cells increases plasminogen activator inhibitor type-1 expression: PPARgamma as a potential mediator in vascular disease. Arteriosclerosis, thrombosis, and vascular biology, 19 (3) (1999): 546–51. issn: 1079-5642 (cit. on p. 34).

[448] C. J. Nicol et al. PPARgamma in endothelial cells influences high fat diet-induced hypertension. American journal of hypertension, 18 (4 Pt 1) (2005): 549–56. issn: 0895-7061. doi: 10.1016/j.amjhyper.2004.10.032 (cit. on p. 34).

[449] L. Széles et al. PPARgamma in immunity and inflammation: cell types and diseases. Biochimica et biophysica acta, 1771 (8) (2007): 1014–30. issn: 0006-3002. doi: 10.1016/j.bbalip.2007.02.005 (cit. on p. 34).

[450] P. Tontonoz and B. M. Spiegelman. Fat and beyond: the diverse biology of PPARgamma. Annual review of biochemistry, 77 (2008): 289–312. issn: 0066-4154. doi: 10.1146/annurev.biochem.77.061307.091829 (cit. on p. 34).

[451] J.-H. Kim et al. The multifaceted factor peroxisome proliferator-activated receptor γ (PPARγ) in metabolism, immunity, and cancer. Archives of pharmacal research, 38 (3) (2015): 302–12. issn: 0253-6269. doi: 10.1007/s12272-015-0559-x (cit. on p. 34).

[452] B. A. Neuschwander-Tetri et al. Troglitazone-induced hepatic failure leading to liver transplantation. A case report. Annals of internal medicine, 129 (1) (1998): 38–41. issn: 0003-4819 (cit. on p. 34).

[453] C. V. Rizos et al. How safe is the use of thiazolidinediones in clinical practice? Expert opinion on drug safety, 8 (1) (2009): 15–32. issn: 1744-764X. doi: 10.1517/14740330802597821 (cit. on p. 34).

[454] D. Bilik et al. Thiazolidinediones and fractures: evidence from translating research into action for diabetes. The Journal of clinical endocrinology and metabolism, 95 (10) (2010): 4560–5. issn: 1945-7197. doi: 10.1210/jc.2009-2638 (cit. on p. 34).

[455] M. Lu et al. Brain PPAR-γ promotes obesity and is required for the insulin-sensitizing effect of thiazolidinediones. Nature medicine, 17 (5) (2011): 618–622. issn: 1546-170X. doi: 10.1038/nm.2332 (cit. on p. 34).

[456] S. N. Friedland et al. The cardiovascular effects of peroxisome proliferator-activated receptor agonists. The American journal of medicine, 125 (2) (2012): 126–33. issn: 1555-7162. doi: 10.1016/j.amjmed.2011.08.025 (cit. on p. 34).

[457] C. Bosetti et al. Cancer risk for patients using thiazolidinediones for type 2 diabetes: a meta-analysis. The oncologist, 18 (2) (2013): 148–156. issn: 1549-490X. doi: 10.1634/theoncologist.2012-0302 (cit. on p. 34).

[458] J. Nedergaard et al. PPARgamma in the control of brown adipocyte differentiation. Biochimica et biophysica acta, 1740 (2) (2005): 293–304. issn: 0006-3002. doi: 10.1016/j.bbadis.2005.02.003 (cit. on p. 34).

[459] A. Koppen and E. Kalkhoven. Brown vs white adipocytes: the PPARgamma coregulator story. FEBS letters, 584 (15) (2010): 3250–9. issn: 1873-3468. doi: 10.1016/j.febslet.2010.06.035 (cit. on p. 34).

[460] A. G. Atanasov et al. Honokiol: a non-adipogenic PPARγ agonist from nature. Biochimica et biophysica acta, 1830 (10) (2013): 4813–9. issn: 0006-3002. doi: 10.1016/j.bbagen.2013.06.021 (cit. on p. 34).

[461] L. Wang et al. Natural product agonists of peroxisome proliferator-activated receptor gamma (PPARγ): a review. Biochemical pharmacology, 92 (1) (2014): 73–89. issn: 1873-2968. doi: 10.1016/j.bcp.2014.07.018 (cit. on p. 34).

[462] S. Sugii et al. PPARgamma activation in adipocytes is sufficient for systemic insulin sensitization. Proceedings of the National Academy of Sciences of the United States of America, 106 (52) (2009): 22504–22509. issn: 1091-6490. doi: 10.1073/pnas.0912487106 (cit. on p. 34).

[463] S. Koschmieder et al. Dysregulation of the C/EBPalpha differentiation pathway in human cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 27 (4) (2009): 619–628. issn: 1527-7755. doi: 10.1200/JCO.2008.17.9812 (cit. on p. 34).

[464] H. G. Linhart et al. C/EBPalpha is required for differentiation of white, but not brown, adipose tissue. Proceedings of the National Academy of Sciences of the United States of America, 98 (22) (2001): 12532–12537. issn: 0027-8424. doi: 10.1073/pnas.211416898 (cit. on p. 34).

[465] E. D. Rosen. The transcriptional basis of adipocyte development. Prostaglandins, Leukotrienes and Essential Fatty Acids, 73 (1) (2005): 31–4. issn: 0952-3278. doi: 10.1016/j.plefa.2005.04.004 (cit. on p. 34).

[466] L. Wang et al. Liver X receptors in the central nervous system: from lipid homeostasis to neuronal degeneration. Proceedings of the National Academy of Sciences of the United States of America, 99 (21) (2002): 13878–13883. issn: 0027-8424. doi: 10.1073/pnas.172510899 (cit. on p. 35).

[467] R. P. Koldamova et al. The liver X receptor ligand T0901317 decreases amyloid beta production in vitro and in a mouse model of Alzheimer's disease. The Journal of biological chemistry, 280 (6) (2005): 4079–88. issn: 0021-9258. doi: 10.1074/jbc.M411420200 (cit. on p. 35).

[468] C. Gabbi et al. Action mechanisms of Liver X Receptors. Biochemical and biophysical research communications, 446 (3) (2014): 647–50. issn: 1090-2104. doi: 10.1016/j.bbrc.2013.11.077 (cit. on p. 35).

[469] S. D. Lee and P. Tontonoz. Liver X receptors at the intersection of lipid metabolism and atherogenesis. Atherosclerosis, 242 (1) (2015): 29–36. issn: 1879-1484. doi: 10.1016/j.atherosclerosis.2015.06.042 (cit. on p. 35).

[470] J. J. Repa and D. J. Mangelsdorf. The liver X receptor gene team: potential new players in atherosclerosis. Nature medicine, 8 (11) (2002): 1243–8. issn: 1078-8956. doi: 10.1038/nm1102-1243 (cit. on p. 35).

[471] J Laurencikiene and M Rydén. Liver X receptors and fat cell metabolism. International journal of obesity (2005), 36 (12) (2012): 1494–1502. issn: 1476-5497. doi: 10.1038/ijo.2012.21 (cit. on p. 35).

[472] C Hilton et al. MicroRNAs in adipose tissue: their role in adipogenesis and obesity. International journal of obesity (2005), 37 (3) (2013): 325–32. issn: 1476-5497. doi: 10.1038/ijo.2012.59 (cit. on p. 35).

[473] J. Chen et al. The role of miRNAs in the differentiation of adipose-derived stem cells. Current stem cell research & therapy, 9 (3) (2014): 268–79. issn: 1574-888X (cit. on p. 35).

[474] D. Hamam et al. microRNAs as regulators of adipogenic differentiation of mesenchymal stem cells. Stem cells and development, 24 (4) (2015): 417–425. issn: 1557-8534. doi: 10.1089/scd.2014.0331 (cit. on p. 35).

[475] E. Smolle and J. Haybaeck. Non-coding RNAs and lipid metabolism. International journal of molecular sciences, 15 (8) (2014): 13494–13513. issn: 1422-0067. doi: 10.3390/ijms150813494 (cit. on p. 35).

[476] C. Esau et al. MicroRNA-143 regulates adipocyte differentiation. The Journal of biological chemistry, 279 (50) (2004): 52361–5. issn: 0021-9258. doi: 10.1074/jbc.C400438200 (cit. on p. 35).

[477] L. Chen et al. MicroRNA-143 regulates adipogenesis by modulating the MAP2K5-ERK5 signaling. Scientific reports, 4 (2014): 3819. issn: 2045-2322. doi: 10.1038/srep03819 (cit. on p. 35).

[478] T. Sun et al. MicroRNA let-7 Regulates 3T3-L1 Adipogenesis. Molecular endocrinology Baltimore Md, 23 (6) (2009): 925–931 (cit. on p. 35).

[479] J. Wei et al. let-7 enhances osteogenesis and bone formation while repressing adipogenesis of human stromal/mesenchymal stem cells by regulating HMGA2. Stem cells and development, 23 (13) (2014): 1452–1463. issn: 1557-8534. doi: 10.1089/scd.2013.0600 (cit. on p. 35).

[480] Q. Lin et al. A role of miR-27 in the regulation of adipogenesis. The FEBS journal, 276 (8) (2009): 2348–58. issn: 1742-4658 (cit. on p. 35).

[481] M. Karbiener et al. microRNA miR-27b impairs human adipocyte differentiation and targets PPARgamma. Biochemical and biophysical research communications, 390 (2) (2009): 247–51. issn: 1090-2104. doi: 10.1016/j.bbrc.2009.09.098 (cit. on p. 35).

[482] Y. H. Son et al. Regulation of Adipocyte Differentiation via MicroRNAs. Endocrinology and metabolism (Seoul, Korea), 29 (2) (2014): 122–135. issn: 2093-596X. doi: 10.3803/EnM.2014.29.2.122 (cit. on p. 35).

[483] M Garofalo et al. miR221/222 in cancer: their role in tumor progression and response to therapy. Current molecular medicine, 12 (1) (2012): 27–33. issn: 1875-5666 (cit. on p. 35).

[484] W.-J. Chen et al. The magic and mystery of microRNA-27 in atherosclerosis. Atherosclerosis, 222 (2) (2012): 314–23. issn: 1879-1484. doi: 10.1016/j.atherosclerosis.2012.01.020 (cit. on p. 35).

[485] A. He et al. Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes. Molecular endocrinology, 21 (11) (2007): 2785–94. issn: 0888-8809. doi: 10.1210/me.2007-0167 (cit. on pp. 35, 158).

[486] S. P. Poulos et al. Cell line models for differentiation: preadipocytes and adipocytes. Experimental biology and medicine (2010): 1185–1193. issn: 1535-3699. doi: 10.1258/ebm.2010.010063 (cit. on p. 35).

[487] H Green and O Kehinde. An established preadipose cell line and its differentiation in culture. II. Factors affecting the adipose conversion. Cell, 5 (1) (1975): 19–27. issn: 0092-8674 (cit. on p. 35).

[488] M Wabitsch et al. Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. International Journal of Obesity, 25 (1) (2001): 8–15. issn: 0307-0565 (cit. on p. 35).

[489] P. Fischer-Posovszky et al. Human SGBS cells - a unique tool for studies of human fat cell biology. Obesity facts, 1 (4) (2008): 184–189. issn: 1662-4025. doi: 10.1159/000145784 (cit. on p. 35).

[490] E. H. Allott et al. The SGBS cell strain as a model for the in vitro study of obesity and cancer. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico, 14 (10) (2012): 774–82. issn: 1699-3055. doi: 10.1007/s12094-012-0863-6 (cit. on p. 35).

[491] M. Altaf-Ul-Amin et al. Systems biology in the context of big data and networks. BioMed research international, 2014 (2014): 428570. issn: 2314-6141. doi: 10.1155/2014/428570 (cit. on p. 36).

[492] V. Memisevic et al. An integrative approach to modeling biological networks. Journal of integrative bioinformatics, 7 (3) (2010). issn: 1613-4516. doi: 10.2390/biecoll-jib-2010-120 (cit. on p. 37).

[493] B. Papp et al. Systems-biology approaches for predicting genomic evolution. Nature reviews. Genetics, 12 (9) (2011): 591–602. issn: 1471-0064. doi: 10.1038/nrg3033 (cit. on p. 37).

[494] D. Machado et al. Modeling formalisms in Systems Biology. AMB Express, 1 (2011): 45. issn: 2191-0855. doi: 10.1186/2191-0855-1-45 (cit. on p. 37).

[495] V. Chelliah et al. BioModels: ten-year anniversary. Nucleic Acids Research, 43 (D1) (2014): D542–D548. issn: 0305-1048. doi: 10.1093/nar/gku1181 (cit. on p. 38).

[496] S. M. Keating and N. Le Novère. Supporting SBML as a model exchange format in software applications. Methods in molecular biology, 1021 (2013): 201–25. issn: 1940-6029. doi: 10.1007/978-1-62703-450-0{_}11 (cit. on p. 38).

[497] N. C. Duarte et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences of the United States of America, 104 (6) (2007): 1777–82. issn: 0027-8424 (cit. on p. 38).

[498] H. Ma et al. The Edinburgh human metabolic network reconstruction and its functional analysis. Molecular systems biology, 3 (135) (2007): 135. issn: 1744-4292. doi: 10.1038/msb4100177 (cit. on p. 38).

[499] T. Hao et al. Compartmentalization of the Edinburgh Human Metabolic Network. BMC bioinformatics, 11 (2010): 393. issn: 1471-2105. doi: 10.1186/1471-2105-11-393 (cit. on p. 39).

[500] I. Thiele et al. A community-driven global reconstruction of human metabolism. Nature biotechnology, 31 (5) (2013): 419–425. issn: 1546-1696. doi: 10.1038/nbt.2488 (cit. on p. 39).

[501] L. Jerby et al. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Molecular Systems Biology, 6 (401) (2010): 1–9. issn: 1744-4292. doi: 10.1038/msb.2010.56 (cit. on p. 39).

[502] C. Gille et al. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Molecular systems biology, 6 (1) (2010): 411. issn: 1744-4292. doi: 10.1038/msb.2010.62 (cit. on pp. 39, 164).

[503] N. E. Lewis et al. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nature Biotechnology, 28 (12) (2010): 1279–1285. issn: 1087-0156. doi: 10.1038/nbt.1711 (cit. on p. 39).

[504] R. Agren et al. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS computational biology, 8 (5) (2012): e1002518. issn: 1553-7358. doi: 10.1371/journal.pcbi.1002518 (cit. on p. 39).

[505] B. J. Schmidt et al. GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics, 29 (22) (2013): 2900–2908. issn: 1367-4811. doi: 10.1093/bioinformatics/btt493 (cit. on p. 39).

[506] M. Hadi and S.-A. Marashi. Reconstruction of a generic metabolic network model of cancer cells. Molecular bioSystems, 10 (11) (2014): 3014–21. issn: 1742-2051. doi: 10.1039/c4mb00300d (cit. on p. 39).

[507] I. Thiele and B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction - Supplemental tables. Nature protocols, 5 (1) (2010): 1–22. issn: 1750-2799. doi: 10.1038/nprot.2009.203 (cit. on p. 39).

[508] N. D. Price et al. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature reviews. Microbiology, 2 (11) (2004): 886–897. issn: 1740-1526. doi: 10.1038/nrmicro1023 (cit. on pp. 40, 41).

[509] J. D. Trawick and C. H. Schilling. Use of constraint-based modeling for the prediction and validation of antimicrobial targets. Biochemical pharmacology, 71 (7) (2006): 1026–35. issn: 0006-2952. doi: 10.1016/j.bcp.2005.10.049 (cit. on p. 40).

[510] M. Terzer et al. Genome-scale metabolic networks. Wiley interdisciplinary reviews. Systems biology and medicine, (3) (2009): 285–297. issn: 1939-005X. doi: 10.1002/wsbm.37 (cit. on p. 40).

[511] E. Grafahrend-Belau et al. FBA-SimVis: interactive visualization of constraint-based metabolic models. Bioinformatics, 25 (20) (2009): 2755–2757. issn: 1367-4811. doi: 10.1093/bioinformatics/btp408 (cit. on p. 40).

[512] J. Thakar et al. Constraint-based network model of pathogen-immune system interactions. Journal of the Royal Society, Interface / the Royal Society, 6 (36) (2009): 599–612. issn: 1742-5662. doi: 10.1098/rsif.2008.0363 (cit. on p. 40).

[513] S. M. Kelk et al. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. Scientific reports, 2 (2012): 580. issn: 2045-2322. doi: 10.1038/srep00580 (cit. on p. 40).

[514] J. M. Lee et al. Flux balance analysis in the era of metabolomics. Briefings in bioinformatics, 7 (2) (2006): 140–50. issn: 1467-5463. doi: 10.1093/bib/bbl007 (cit. on p. 41).

[515] J. Schellenberger et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature protocols, 6 (9) (2011): 1290–1307. issn: 1750-2799. doi: 10.1038/nprot.2011.308 (cit. on p. 41).

[516] T. Shlomi et al. Network-based prediction of human tissue-specific metabolism. Nature biotechnology, 26 (9) (2008): 1003–1010. issn: 1546-1696. doi: 10.1038/nbt.1487 (cit. on pp. 41, 49, 60, 152).

[517] H. Zur et al. iMAT: An Integrative Metabolic Analysis Tool. Bioinformatics (2010): 3–5 (cit. on pp. 41, 60).

[518] S. A. Becker and B. O. Palsson. Context-specific metabolic networks are consistent with experiments. PLoS computational biology, 4 (5) (2008): e1000082. issn: 1553-7358. doi: 10.1371/journal.pcbi.1000082 (cit. on pp. 42, 156, 163).

[519] A. Alyass et al. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Medical Genomics, 8 (1) (2015): 33. issn: 1755-8794. doi: 10.1186/s12920-015-0108-y (cit. on p. 42).

[520] J. M. Fonville et al. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. Journal of Chemometrics, 24 (11-12) (2010): n/a–n/a. issn: 08869383. doi: 10.1002/cem.1359 (cit. on p. 42).

[521] M. A. Ovacik and I. P. Androulakis. On the Potential for Integrating Gene Expression and Metabolic Flux Data. Current Bioinformatics, 3 (2008): 1–7 (cit. on p. 42).

[522] J. S. Hamid et al. Data Integration in Genetics and Genomics: Methods and Challenges. Human Genomics and Proteomics, 2009 (2009): 1–13. issn: 1757-4242. doi: 10.4061/2009/869093 (cit. on p. 42).

[523] J. Tang et al. Integrating post-genomic approaches as a strategy to advance our understanding of health and disease. Genome medicine, 1 (3) (2009): 35. issn: 1756-994X. doi: 10.1186/gm35 (cit. on p. 42).

[524] P. Shannon et al. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, (Karp 2001) (2003): 2498–2504. doi: 10.1101/gr.1239303.metabolite (cit. on p. 42).

[525] P. D. Karp et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in bioinformatics, 11 (1) (2010): 40–79. issn: 1477-4054. doi: 10.1093/bib/bbp043 (cit. on p. 42).

[526] R. E. Curtis et al. TVNViewer: An interactive visualization tool for exploring net- works that change over time or space (2011): 1–2 (cit. on p. 42).

[527] G. Sales et al. Graphite Web: Web tool for gene set analysis exploiting pathway topology. Nucleic acids research, 41 (Web Server issue) (2013): W89–97. issn: 1362-4962. doi: `10.1093/nar/gkt386` (cit. on pp. 42, 155).

[528] N. Töpfer et al. Integration of metabolomics data into metabolic networks. Frontiers in Plant Science, 6 (2015): 49. issn: 1664-462X. doi: `10.3389/fpls.2015.00049` (cit. on p. 42).

[529] M. Fondi and P. Liò. Multi -omics and metabolic modelling pipelines: challenges and tools for systems microbiology. Microbiological Research, 171 (2015): 52–64. issn: 09445013. doi: `10.1016/j.micres.2015.01.003` (cit. on p. 42).

[530] S. Imam et al. Data-driven integration of genome-scale regulatory and metabolic network models. Frontiers in microbiology, 6 (2015): 409. issn: 1664-302X. doi: `10.3389/fmicb.2015.00409` (cit. on p. 42).

[531] K. Michelsen et al. *Promoting better integration of health information systems: best practices and challenges.* Tech. rep. 2015, p. 40 (cit. on p. 45).

[532] C. Y. McLean et al. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology, 28 (5) (2010): 495–501. issn: 1546-1696. doi: `10.1038/nbt.1630` (cit. on pp. 51–53, 162).

[535] T. S. Keshava Prasad et al. Human Protein Reference Database–2009 update. Nucleic Acids Research, 37 (Database) (2009): D767–D772. issn: 0305-1048. doi: `10.1093/nar/gkn892` (cit. on p. 54).

[536] B Jeanrenaud. Hyperinsulinemia in obesity syndromes: its metabolic consequences and possible etiology. Metabolism: clinical and experimental, 27 (12 Suppl 2) (1978): 1881–92. issn: 0026-0495 (cit. on pp. 59, 152, 157).

[537] P Lönnroth et al. Insulin binding and responsiveness in fat cells from patients with reduced glucose tolerance and type II diabetes. Diabetes, 32 (8) (1983): 748–54. issn: 0012-1797 (cit. on pp. 59, 152, 157).

[538] P Engfeldt et al. Effect of fasting on insulin receptor binding and insulin action in different human subcutaneous fat depots. The Journal of clinical endocrinology and metabolism, 60 (5) (1985): 868–73. issn: 0021-972X. doi: `10.1210/jcem-60-5-868` (cit. on pp. 59, 152).

[539] M. P. Stern and S. M. Haffner. Body fat distribution and hyperinsulinemia as risk factors for diabetes and cardiovascular disease. Arteriosclerosis (Dallas, Tex.) 6 (2) (1986): 123–30. issn: 0276-5047 (cit. on pp. 59, 152).

[540] P Björntorp. Adipose tissue distribution, plasma insulin, and cardiovascular disease. Diabète & métabolisme, 13 (3 Pt 2) (1987): 381–5. issn: 0338-1684 (cit. on pp. 59, 152, 157).

[541] K Landin et al. Increased insulin resistance and fat cell lipolysis in obese but not lean women with a high waist/hip ratio. European journal of clinical investigation, 20 (5) (1990): 530–5. issn: 0014-2972 (cit. on pp. 59, 152).

[542] S. B. Pedersen et al. Abdominal obesity is associated with insulin resistance and reduced glycogen synthetase activity in skeletal muscle. Metabolism: clinical and experimental, 42 (8) (1993): 998–1005. issn: 0026-0495 (cit. on pp. 59, 152).

[543] S Reynisdottir et al. Multiple lipolysis defects in the insulin resistance (metabolic) syndrome. The Journal of clinical investigation, 93 (6) (1994): 2590–2599. issn: 0021-9738. doi: 10.1172/JCI117271 (cit. on pp. 59, 152).

[544] J. P. Després. Dyslipidaemia and obesity. Baillière's clinical endocrinology and metabolism, 8 (3) (1994): 629–60. issn: 0950-351X (cit. on pp. 59, 152).

[545] M Halle et al. Importance of TNF-alpha and leptin in obesity and insulin resistance: a hypothesis on the impact of physical exercise. Exercise immunology review, 4 (1998): 77–94. issn: 1077-5552 (cit. on pp. 59, 152).

[546] Y. Matsuzawa et al. Importance of adipocytokines in obesity-related diseases. Hormone research, 60 Suppl 3 (2003): 56–9. issn: 0301-0163. doi: 74502 (cit. on pp. 59, 152).

[547] A. Guilherme et al. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. Nature reviews. Molecular cell biology, 9 (5) (2008): 367–377. issn: 1471-0080. doi: 10.1038/nrm2391 (cit. on pp. 59, 152, 155, 156).

[548] G STEINER and G. F. CAHILL. ADIPOSE TISSUE PHYSIOLOGY. Annals of the New York Academy of Sciences, 110 (1963): 749–53. issn: 0077-8923 (cit. on p. 59).

[549] F. M. Gregoire et al. Understanding adipocyte differentiation. Physiological reviews, 78 (3) (1998): 783–809. issn: 0031-9333 (cit. on p. 59).

[550] J. M. Ntambi and Y.-c. Kim. Adipocyte Function , Differentiation and Metabolism. The Journal of nutrition (2000): 3122–3126 (cit. on p. 59).

[551] H. S. Camp et al. Adipogenesis and fat-cell function in obesity and diabetes. Trends in molecular medicine, 8 (9) (2002): 442–7. issn: 1471-4914 (cit. on p. 59).

[552] K. N. Frayn et al. Integrative physiology of human adipose tissue. Int J Obes Relat Metab Disord. 27 (8) (2003): 875–88. doi: 10.1038/sj.ijo.0802326 (cit. on p. 59).

[553] S. R. Farmer. Transcriptional control of adipocyte formation. Cell metabolism, 4 (4) (2006): 263–273. issn: 1550-4131. doi: 10.1016/j.cmet.2006.07.001 (cit. on p. 59).

[554] K. M. McTigue et al. Obesity in older adults: a systematic review of the evidence for diagnosis and treatment. Obesity (Silver Spring, Md.) 14 (9) (2006): 1485–97. issn: 1930-7381. doi: 10.1038/oby.2006.171 (cit. on p. 59).

[555] A. Fernández-Quintela et al. The role of dietary fat in adipose tissue metabolism. Public health nutrition, 10 (10A) (2007): 1126–31. issn: 1368-9800. doi: 10.1017/S1368980007000602 (cit. on p. 59).

[557] R. J. Perera et al. Identification of novel PPARgamma target genes in primary human adipocytes. Gene, 369 (2006): 90–9. issn: 0378-1119. doi: 10.1016/j.gene.2005.10.021 (cit. on p. 152).

[558] R. Nielsen et al. Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. Genes & development, 22 (21) (2008): 2953–2967. issn: 0890-9369. doi: 10.1101/gad.501108 (cit. on p. 152).

[559] M. I. Lefterova et al. Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages. Molecular and cellular biology, 30 (9) (2010): 2078–2089. issn: 1098-5549. doi: 10.1128/MCB.01651-09 (cit. on p. 152).

[560] T. S. Mikkelsen et al. Comparative epigenomic analysis of murine and human adipogenesis. Cell, 143 (1) (2010): 156–169. issn: 1097-4172. doi: 10.1016/j.cell.2010.09.006 (cit. on p. 152).

[561] K. K.-H. Farh et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518 (7539) (2014): 337–43. issn: 0028-0836. doi: 10.1038/nature13835 (cit. on pp. 153, 160).

[562] L. Pasquali et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nature genetics, 46 (2) (2014): 136–143. issn: 1546-1718. doi: 10.1038/ng.2870 (cit. on p. 153).

[563] S. Johnatty et al. Genome-wide analysis identifies novel loci associated with ovarian cancer outcomes: findings from the Ovarian Cancer Association Consortium. Clinical cancer research : an official journal of the American Association for Cancer Research (2015). issn: 1078-0432. doi: 10.1158/1078-0432.CCR-15-0632 (cit. on p. 153).

[564] S. Varrette et al. Management of an Academic HPC Cluster: The UL Experience (2014) (cit. on p. 154).

[565] F. García-Alcalde et al. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics, 27 (1) (2010): 137–139. issn: 1367-4811. doi: 10.1093/bioinformatics/btq594 (cit. on p. 155).

[566] M. Kanehisa et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, 40 (Database issue) (2012): D109–14. issn: 1362-4962. doi: 10.1093/nar/gkr988 (cit. on p. 155).

[567] G. Manyam et al. KPP: KEGG Pathway Painter. BMC systems biology, 9 Suppl 2 (2015): S3. issn: 1752-0509. doi: 10.1186/1752-0509-9-S2-S3 (cit. on p. 155).

[568] J. R. Karr et al. NetworkPainter: dynamic intracellular pathway animation in Cytobank. BMC bioinformatics, 16 (1) (2015): 172. issn: 1471-2105. doi: 10.1186/s12859-015-0602-4 (cit. on p. 155).

[569] S. F. Schmidt et al. Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. Genome Research (2015): gr.188300.114. issn: 1088-9051. doi: 10.1101/gr.188300.114 (cit. on p. 155).

[570] A. Bordbar et al. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC systems biology, 5 (1) (2011): 180. issn: 1752-0509. doi: 10.1186/1752-0509-5-180 (cit. on p. 156).

[571] A. Mardinoglu et al. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. Molecular Systems Biology, 9 (649) (2013): 649. issn: 1744-4292. doi: 10.1038/msb.2013.5 (cit. on p. 156).

[572] P Felig et al. Plasma amino acid levels and insulin secretion in obesity. The New England journal of medicine, 281 (15) (1969): 811–6. issn: 0028-4793. doi: 10.1056/NEJM196910092811503 (cit. on p. 157).

[573] G Forlani et al. Insulin-dependent metabolism of branched-chain amino acids in obesity. Metabolism: clinical and experimental, 33 (2) (1984): 147–50. issn: 0026-0495 (cit. on p. 157).

[574] W. C. Abbott et al. The effect of dextrose and amino acids on respiratory function and energy expenditure in morbidly obese patients following gastric bypass surgery. The Journal of surgical research, 41 (3) (1986): 225–35. issn: 0022-4804 (cit. on p. 157).

[575] B. Laferrère et al. Differential metabolic impact of gastric bypass surgery versus dietary intervention in obese diabetic subjects despite identical weight loss. Science translational medicine, 3 (80) (2011): 80re2. issn: 1946-6242. doi: 10.1126/scitranslmed.3002043 (cit. on p. 157).

[576] M. A. Lips et al. Roux-en-Y gastric bypass surgery, but not calorie restriction, reduces plasma branched-chain amino acids in obese women independent of weight loss or the presence of type 2 diabetes. Diabetes care, 37 (12) (2014): 3150–6. issn: 1935-5548. doi: 10.2337/dc14-0195 (cit. on p. 157).

[577] C. B. Newgard. Interplay between lipids and branched-chain amino acids in development of insulin resistance. Cell metabolism, 15 (5) (2012): 606–614. issn: 1932-7420. doi: 10.1016/j.cmet.2012.01.024 (cit. on p. 158).

[578] A. Bagge et al. MicroRNA-29a is up-regulated in beta-cells by glucose and decreases glucose-stimulated insulin secretion. Biochemical and biophysical research communications, 426 (2) (2012): 266–272. issn: 1090-2104. doi: 10.1016/j.bbrc.2012.08.082 (cit. on p. 158).

[579] M. Skårn et al. Adipocyte differentiation of human bone marrow-derived stromal cells is modulated by microRNA-155, microRNA-221, and microRNA-222. Stem cells and development, 21 (6) (2012): 873–83. issn: 1557-8534. doi: 10.1089/scd.2010.0503 (cit. on p. 158).

[580] A. Warner and J. Mittag. Brown fat and vascular heat dissipation: The new cautionary tail. Adipocyte, 3 (3) (2014): 221–223. issn: 2162-3945. doi: 10.4161/adip.28815 (cit. on p. 159).

[581] T. J. Schulz and Y.-H. Tseng. Systemic control of brown fat thermogenesis: integration of peripheral and central signals. Annals of the New York Academy of Sciences, 1302 (2013): 35–41. issn: 1749-6632. doi: 10.1111/nyas.12277 (cit. on p. 159).

[582] S. Rajan et al. Adipocyte transdifferentiation and its molecular targets. Differentiation; research in biological diversity, 87 (5) (2014): 183–92. issn: 1432-0436. doi: 10.1016/j.diff.2014.07.002 (cit. on p. 159).

[583] J. Nedergaard and B. Cannon. The Browning of White Adipose Tissue: Some Burning Issues. Cell Metabolism, 20 (3) (2014): 396–407. issn: 15504131. doi: 10.1016/j.cmet.2014.07.005 (cit. on p. 159).

[584] M. A. Schaub et al. Linking disease associations with regulatory information in the human genome. Genome research, 22 (9) (2012): 1748–1759. issn: 1549-5469. doi: 10.1101/gr.136127.111 (cit. on p. 160).

[585] B. Vernot et al. Personal and population genomics of human regulatory variation. Genome Research, 22 (2012): 1689–1697. issn: 10889051. doi: 10.1101/gr.134890.111 (cit. on p. 160).

[586] D. A. Kleinjan and V. van Heyningen. Long-range control of gene expression: emerging mechanisms and disruption in disease. American journal of human genetics, 76 (1) (2005): 8–32. issn: 0002-9297. doi: 10.1086/426833 (cit. on p. 160).

[587] M. Rubinstein and F. S. J. de Souza. Evolution of transcriptional enhancers and animal diversity. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 368 (1632) (2013): 20130017. issn: 1471-2970. doi: 10.1098/rstb.2013.0017 (cit. on p. 160).

[588] A. M. Cheatle Jarvela and V. F. Hinman. Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. EvoDevo, 6 (1) (2015): 3. issn: 2041-9139. doi: 10.1186/2041-9139-6-3 (cit. on p. 160).

[589] F. Barrenäs et al. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. - PubMed - NCBI. Genome biology, 13 (6) (2012): R46. issn: 1465-6914. doi: 10.1186/gb-2012-13-6-r46 (cit. on p. 160).

[590] B. A. Benayoun et al. H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. Cell, 158 (3) (2014): 673–688. issn: 00928674. doi: 10.1016/j.cell.2014.06.027 (cit. on p. 161).

[591] J. Dostie et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research, 16 (10) (2006): 1299–1309. issn: 1088-9051. doi: 10.1101/gr.5571506 (cit. on p. 162).

[592] S. C. J. Parker et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences of the United States of America, 110 (44) (2013): 17921–6. issn: 1091-6490 (cit. on p. 162).

[593] N. Vlassis et al. Fast reconstruction of compact context-specific metabolic network models. PLoS computational biology, 10 (1) (2014): e1003424. issn: 1553-7358. doi: 10.1371/journal.pcbi.1003424 (cit. on p. 164).

[594] D. Machado and M. Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS computational biology, 10 (4) (2014): e1003580. issn: 1553-7358. doi: 10.1371/journal.pcbi.1003580 (cit. on p. 164).

[595] M. Krauss et al. Integrating cellular metabolism into a multiscale whole-body model. PLoS computational biology, 8 (10) (2012): e1002750. issn: 1553-7358. doi: 10.1371/journal.pcbi.1002750 (cit. on p. 165).

[596] C. P. Fisher et al. QSSPN: dynamic simulation of molecular interaction networks describing gene regulation, signalling and whole-cell metabolism in human cells. Bioinformatics, 29 (24) (2013): 3181–3190. issn: 1367-4811. doi: 10.1093/bioinformatics/btt552 (cit. on p. 165).

[597] A. Henney and H. Coaker. The Virtual Liver Network: systems understanding from bench to bedside. Future medicinal chemistry, 6 (16) (2014): 1735–40. issn: 1756-8927 (cit. on p. 165).

[598] A. Naik et al. SteatoNet: the first integrated human metabolic model with multi-layered regulation to investigate liver-associated pathologies. PLoS computational biology, 10 (12) (2014): e1003993. issn: 1553-7358. doi: 10.1371/journal.pcbi.1003993 (cit. on p. 165).

[599] A. Kumar et al. Multi-tissue computational modeling analyzes pathophysiology of type 2 diabetes in MKR mice. PloS one, 9 (7) (2014): e102319. issn: 1932-6203. doi: 10.1371/journal.pone.0102319 (cit. on p. 165).

[600] P. Samdani et al. A Comprehensive Inter-Tissue Crosstalk Analysis Underlying Progression and Control of Obesity and Diabetes. Scientific reports, 5 (2015): 12340. issn: 2045-2322. doi: 10.1038/srep12340 (cit. on p. 165).

Appendix

Genomics Data 2 (2014) 246–248

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Transcriptomics profiling of human SGBS adipogenesis

Mafalda Galhardo [a], Lasse Sinkkonen [a], Philipp Berninger [b], Jake Lin [c], Thomas Sauter [a,*], Merja Heinäniemi [d,*]

[a] Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg
[b] Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland
[c] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, House of Biomedicine, 7 Avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg
[d] Institute of Biomedicine, School of Medicine, University of Eastern Finland, FI-70120 Kuopio, Finland

ARTICLE INFO

Article history:
Received 29 June 2014
Accepted 13 July 2014
Available online 7 August 2014

Keywords:
Adipocyte
Differentiation
Metabolism

ABSTRACT

Obesity is an ever-growing epidemic where tissue homeostasis is influenced by the differentiation of adipocytes that function in lipid metabolism, endocrine and inflammatory processes. While this differentiation process has been well-characterized in mice, limited data is available from human cells. Applying microarray expression profiling in the human SGBS pre-adipocyte cell line, we identified genes with differential expression during differentiation in combination with constraint-based modeling of metabolic pathway activity. Here we describe the experimental design and quality controls in detail for the gene expression and related results published by Galhardo et al. in Nucleic Acids Research 2014 associated with the data uploaded to NCBI Gene Expression Omnibus (GSE41352).

Specifications [standardized info for the reader]; where applicable, please follow the Ontology for Biomedical Investigations: http://obi-ontology.org/page/Main_Page

Organism/cell line/tissue	Human/SGBS pre-adipocyte/adipose tissue
Sex	Male
Sequencer or array type	Illumina HumanHT-12V3.0 expression beadchip
Data format	Raw and analyzed data
Experimental factors	Time point of differentiation to adipocytes. Cells were cultured 2 days in serum-free OF medium prior to differentiation
Experimental features	Time series of differentiation (20 samples, 7 time points in duplicate or triplicate). SGBS pre-adipocyte cells originate from patient with SGB syndrome. See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 for more details on differentiation protocol and origin of cells.
Consent	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]
Sample source location	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41352.

Experimental design, materials and methods

Cell differentiation and experimental design

Gene expression levels during adipocyte differentiation were obtained by stimulating the SGBS pre-adipocyte cell line with a mix of differentiation inducing compounds and collecting RNA samples at 0, 4, 8 and 12 h and on days 1, 3 and 12 of adipogenesis for hybridization on Illumina HT-12 microarrays. Triplicate samples were prepared following the differentiation protocol modified from [2] (exception is 12 h time point that has only duplicate samples) as shown in Table 1. SGBS cells differentiate within 10–12 days as determined by microscopic analysis (Oil red O staining). At this time point the cells are filled with small sized lipid droplets and are most responsive, whereas at later time points (20 days) the lipid droplets fuse and cells are less active (personal communication, Dr. Martin Wabitsch).

Specifically, SGBS cells were cultured in Dulbecco's modified Eagle's medium (DMEM)/Nutrient Mix F12 (Gibco) containing 8 mg/L biotin, 4 mg/L pantothenate, 0.1 mg/mg streptomycin and 100 U/mL penicillin (OF medium) supplemented with 10% FBS in a humidified 95% air/5% CO_2 incubator. The cells were seeded into 10 cm plates, which were coated with a solution of 10 μL/mL fibronectin and 0.05% gelatine in phosphate-buffered saline. Confluent cells were cultured in serum-free OF medium for 2 days followed by stimulation to differentiate with OF media supplemented with 0.01 mg/mL human transferrin, 200 nM T3, 100 nM cortisol, 20 nM insulin, 500 μM IBMX and 100 nM rosiglitazone (Cayman Chemicals). After day 4, the differentiating cells

* Corresponding authors.
E-mail addresses: thomas.sauter@uni.lu (T. Sauter), merja.heinaniemi@uef.fi (M. Heinäniemi).

http://dx.doi.org/10.1016/j.gdata.2014.07.004

Table 1
Microarray sample description from the SGBS pre-adipocyte differentiation experiment (GSE41578). GEO sample identifiers are presented for the 20 samples prepared, as well as their differentiation time point and replicate number.

Sample name	GSM identifier	Title	Time	Replicate
Sample 1	GSM1015366	SGBS_day0_1	0 h	1
Sample 2	GSM1015367	SGBS_day0_2	0 h	2
Sample 3	GSM1015368	SGBS_day0_3	0 h	3
Sample 4	GSM1015369	SGBS_4h_1	4 h	1
Sample 5	GSM1015370	SGBS_4h_2	4 h	2
Sample 6	GSM1015371	SGBS_4h_3	4 h	3
Sample 7	GSM1015372	SGBS_8h_1	8 h	1
Sample 8	GSM1015373	SGBS_8h_2	8 h	2
Sample 9	GSM1015374	SGBS_8h_3	8 h	3
Sample 10	GSM1015375	SGBS_12h_1	12 h	1
Sample 11	GSM1015376	SGBS_12h_2	12 h	2
Sample 12	GSM1015377	SGBS_day1_1	Day 1	1
Sample 13	GSM1015378	SGBS_day1_2	Day 1	2
Sample 14	GSM1015379	SGBS_day1_3	Day 1	3
Sample 15	GSM1015380	SGBS_day3_1	Day 3	1
Sample 16	GSM1015381	SGBS_day3_2	Day 3	2
Sample 17	GSM1015382	SGBS_day3_3	Day 3	3
Sample 18	GSM1015383	SGBS_day12_1	Day 12	1
Sample 19	GSM1015384	SGBS_day12_2	Day 12	2
Sample 20	GSM1015385	SGBS_day12_3	Day 12	3

were kept in OF media supplemented with 0.01 mg/mL human transferrin, 100 nM cortisol and 20 nM insulin.

Gene expression analysis

Total RNA was extracted using TriSure (Bioline). 1 mL of TriSure was added per a confluent 10 cm dish to lyse the cells. RNA was extracted with 200 μL chloroform and precipitated from the aqueous phase with 400 μL isopropanol by incubating at −20 °C overnight. The longer isopropanol incubation allowed the precipitation of microRNAs and other small RNAs from the same samples. The total RNA samples were processed according to the manufacturer instructions to prepare cDNA that was hybridized on microarrays (Turku Centre for Biotechnology, Microarray and Sequencing Facility, Turku, Finland). Total RNA integrity was confirmed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Data processing and normalization

The raw data files were processed and quality controlled using the R/Bioconductor lumi package. Raw and normalized expression values are available via GEO (GSE41352). Control probe data was included and used to background correct the signal values with the lumiB "bgAdjust" method. We provide this data and sample data in a format that is directly compatible with the lumi analysis package through our web resource at http://systemsbiology.uni.lu/idare.html. The data was then transformed with the "vst" method and normalized with robust spline normalization (rsn) method. The probe intensity value distribution and sample relation are plotted in Figs. 1 and 2, with sample naming described in Table 1. No outliers were detected based on data value range at this step and the samples clustered according to the biological sample group. The code that can be used to download processed data from GEO or to process them from the files that we provide through our website is available (see Additional Data File 1).

Statistical analysis

The negative probe signals were used to filter non-expressed genes. Only genes that had a detection p-value < 0.05 within all samples of at least one time point were selected for statistical analysis, resulting in a total of 12756 detected probes. The statistical analysis was performed using the R/Bioconductor limma package. The F-test was used to assess

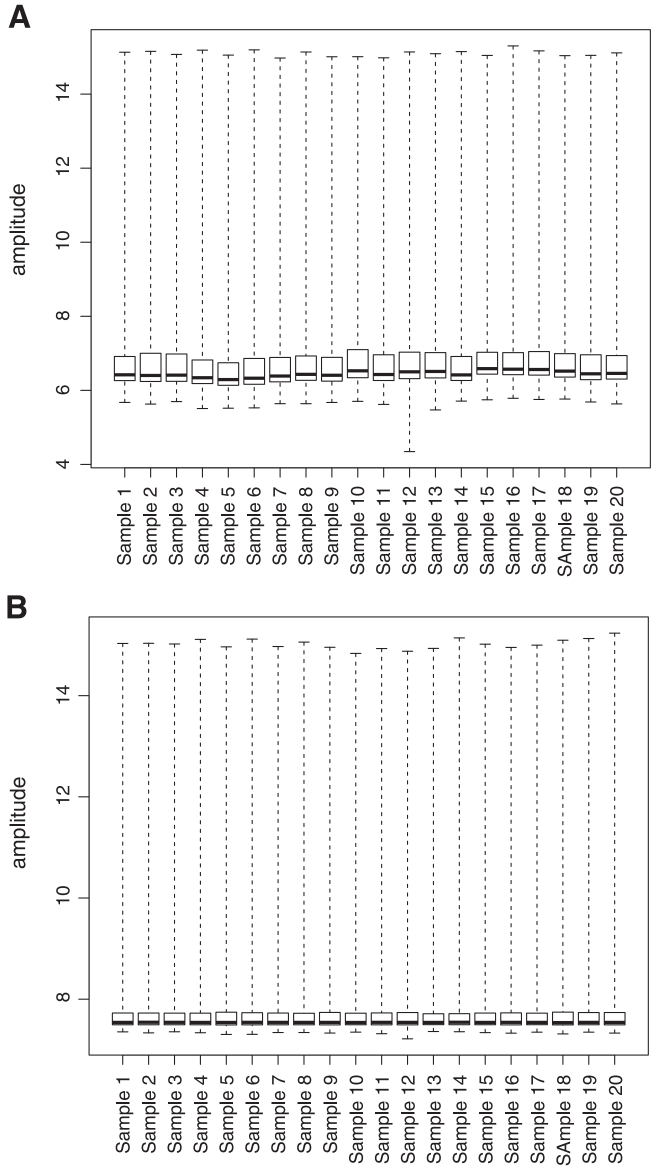


Fig. 1. Probe intensity plots for the 20 SGBS differentiation samples in GSE41578. A) Box plots of raw probe intensities. B) Box plots of normalized probe intensities indicate the absence of outliers and comparable data mean intensities.

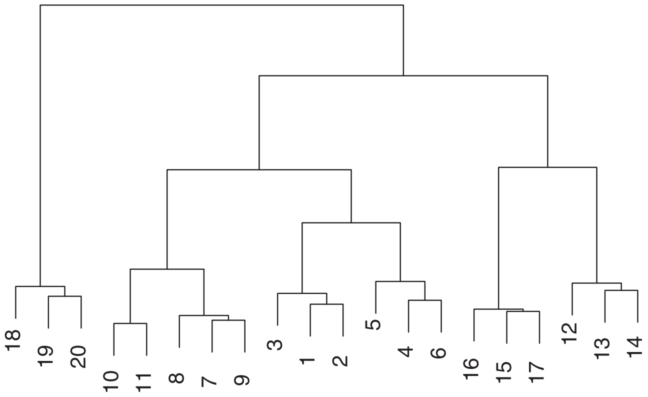


Fig. 2. Hierarchical clustering of the SGBS differentiation microarray samples. The dendrogram shows high similarity between replicates and grouping based on differentiation time progression.

significance of overall dynamic response over the differentiation while a two-tailed t-test was performed to compare specific time points to day 0 undifferentiated cells. In both analyses Benjamini–Hochberg adjusted p-value < 0.01 was considered statistically significant. In total, 1936 Refseq transcripts changed their expression more than 2-fold up or down during the differentiation time series. The code that can be used to filter non-expressed genes and to perform the statistical analysis is available (see Additional Data File 1).

Several of these genes were metabolic genes, represented by 2-fold more differentially expressed genes compared to other gene categories with similar numbers of genes (extracted from the GO Online SQL Environment, as of 12th of August 2013: cell projection, envelope, locomotion and receptor activity).

Analysis of metabolic genes in Recon1

The annotation data from Recon1 was obtained and checked against the current EntrezGene and Refseq annotations (hg19 Refseq; Feb 02 2012). The reaction to gene mappings were updated with current gene IDs (see Table S1). Withdrawn IDs and pseudogenes present a difficulty in the Recon1 annotation. As there were only few such genes (see Table S1), they were left out from visualizations and assigned expression level 0 in modeling. *LPIN1* was missing and due to its central role in adipocytes, it was added to the triacylglycerol pathway reaction catalyzed by *Phosphatidic Acid Phosphatase* (PPAP).

The expression profiles of metabolic genes (from Recon 1 [3]) or TFs (from [4]) were clustered for visualization using self-organizing maps (GEDI software [5]) and AutoSOME [6] as instructed in the tool documentation. The settings to reproduce the results presented in [1] were the following: GEDI grid size was adjusted based on input gene number and settings were tuned in order to minimize data missing grid points (gene density map) (see Table S2). AutoSOME GUI was used following the description in the manual without data filtering. Clustering was done for columns (samples) on “precision” mode, with the “Fuzzy Cluster Network” option and network visualization with Cytoscape [7]. Enriched pathways of the human metabolic reconstruction [3] were determined using a hypergeometric test.

A consistent version of the generic human metabolic model Recon1 [3] was used as modeling platform for prediction of network activity distributions. The Recon1 model was downloaded from the BiGG database [8] (04.11.11) and the consistent version was derived using the function “reduceModel” from the COBRA toolbox 2.0 [9], which resulted in the exclusion of 1273 reactions (34%) of the initial model (Table S3). To include the microarray data as soft-constraints for reaction activity prediction, the probes were mapped to Entrez Gene IDs. First, continuous log2 normalized expression values for the probes were discretized into three categories: lowly expressed (-1), moderately expressed (0) and highly expressed (1) based on the mean expression ± 0.5 $*$ standard deviation cutoffs across all arrays. Then, one unique discretized value per gene was selected taking the rounded discretized mean of all probes for a gene. Each gene was then assigned to the Recon1 reaction based on gene–protein-reaction associations.

Discussion

Here we describe a time series dataset of human SGBS pre-adipocyte differentiation. This dataset is comprised of whole transcriptome gene expression profiling data derived using the Illumina BeadArrays. We demonstrated differential expression that was particularly prevalent among metabolic genes. Moreover, discretization of the metabolic gene expression levels allowed using them as soft-constrains for metabolic activity modeling. Further, this dataset is part of a GEO SuperSeries (GSE41578) and we have used it in combination with next-generation sequencing data and microRNA expression profiles to associate putative regulators to the metabolic genes in [1]. To further analyze the data in an integrative manner, we introduced gene metanodes and the web portal IDARE (Integrated Data Nodes or Regulation) in [1] for interactive data exploration of various data types within the metabolic network context, available at http://systemsbiology.uni.lu/idare.html, including a detailed user guide. The data could be similarly analyzed to interrogate the regulation of other pathways. Results from the data have increased our understanding of human adipogenesis.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.07.004.

Acknowledgments

We would like to acknowledge the following funding sources: Fonds National de la Recherche Luxembourg [AFR, AM2c and C08/BM/01]; funding from the University of Eastern Finland and the Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg.

References

[1] M. Galhardo, L. Sinkkonen, P. Berninger, J. Lin, T. Sauter, M. Heinäniemi, Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. Nucleic Acids Res. 42 (2014) 1474–1496.

[2] M. Wabitsch, R.E. Brenner, I. Melzner, M. Braun, P. Möller, E. Heinze, K.M. Debatin, H. Hauner, Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. Int. J. Obes. Relat. Metab. Disord. 25 (2001) 8–15.

[3] N.C. Duarte, S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, B.Ø. Palsson, Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 1777–1782.

[4] M. Heinäniemi, M. Nykter, R. Kramer, A. Wienecke-Baldacchino, L. Sinkkonen, J.X. Zhou, R. Kreisberg, S.A. Kauffman, S. Huang, I. Shmulevich, Gene-pair expression signatures reveal lineage control. Nat. Methods 10 (2013) 577–583.

[5] G.S. Eichler, S. Huang, D.E. Ingber, Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics 19 (2003) 2321–2322.

[6] A.M. Newman, J.B. Cooper, AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinforma. 11 (2010) 117.

[7] M. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27 (2011) 431–432.

[8] J. Schellenberger, J.O. Park, T.C. Conrad, B.Ø. Palsson, BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11 (2010) 213.

[9] J. Schellenberger, R. Que, R.M.T. Fleming, I. Thiele, J.D. Orth, A.M. Feist, D.C. Zielinski, A. Bordbar, N.E. Lewis, S. Rahmanian, J. Kang, D.R. Hyduke, B.Ø. Palsson, Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. 6 (2011) 1290–1307.

Genomics Data 2 (2014) 230–236

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

ChIP-seq profiling of the active chromatin marker H3K4me3 and PPARγ, CEBPα and LXR target genes in human SGBS adipocytes

Mafalda Galhardo [a], Lasse Sinkkonen [a], Philipp Berninger [b], Jake Lin [c], Thomas Sauter [a,*], Merja Heinäniemi [d,*]

[a] Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg
[b] Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland
[c] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, House of Biomedicine, 7 Avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg
[d] Institute of Biomedicine, School of Medicine, University of Eastern Finland, FI-70120 Kuopio, Finland

ARTICLE INFO

Article history:
Received 29 June 2014
Accepted 10 July 2014
Available online 6 August 2014

Keywords:
Adipocyte
ChIP-seq
Transcription factor

ABSTRACT

Transcription factors (TFs) represent key factors to establish a cellular phenotype. It is known that several TFs could play a role in disease, yet less is known so far how their targets overlap. We focused here on identifying the most highly induced TFs and their putative targets during human adipogenesis. Applying chromatin immunoprecipitation coupled with deep sequencing (ChIP-Seq) in the human SGBS pre-adipocyte cell line, we identified genes with binding sites in their vicinity for the three TFs studied, PPARγ, CEBPα and LXR. Here we describe the experimental design and quality controls in detail for the deep sequencing data and related results published by Galhardo et al. in Nucleic Acids Research 2014 [1] associated with the data uploaded to NCBI Gene Expression Omnibus (GSE41578).

Introduction

Specifications [*standardized info for the reader*] where applicable, please follow the Ontology for Biomedical Investigations: http://obi-ontology.org/page/Main_Page

Organism/cell line/tissue	Human/SGBS preadipocyte/adipose tissue
Sex	Male
Sequencer or array type	Illumina Genome Analyzer II
Data format	Raw and analyzed data
Experimental factors	ChIP-antibody used
Experimental features	Genome-wide binding or chromatin marker level (6 samples, including input control). SGBS preadipocyte cells originate from a patient with SGB syndrome. See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2] for more details on differentiation protocol and origin of cells
Consent	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]
Sample source location	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41578
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41629
http://www.ncbi.nlm.nih.gov/sra?term=SRP016497

Experimental design, materials and methods

Cell differentiation and experimental design

Chromatin was collected at day 0 and day 10 of adipogenesis for ChIP. SGBS cells differentiate within 10–12 days as determined by microscopic analysis (Oil Red O Staining). At this time point the cells are filled with small sized lipid droplets and are most responsive, whereas at later time points (20 days) the lipid droplets fuse and cells are less active (personal communication, Dr. Martin Wabitsch).

Specifically, SGBS cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM)/Nutrient Mix F12 (Gibco) containing 8 mg/l biotin, 4 mg/l pantothenate, 0.1 mg/mg streptomycin and 100 U/ml penicillin (OF medium) supplemented with 10% FBS in a humidified 95% air/5% CO_2 incubator. The cells were seeded into 10 cm plates, which were coated with a solution of 10 μl/ml fibronectin and 0.05% gelatine in phosphate-buffered saline. Confluent cells were cultured in serum-free OF medium for 2 days followed by stimulation to differentiate with OF media supplemented with 0.01 mg/ml human transferrin, 200 nM T3, 100 nM cortisol, 20 nM insulin, 500 μM IBMX and 100 nM rosiglitazone (Cayman Chemicals). After day 4, the differentiating cells

* Corresponding authors.
E-mail addresses: thomas.sauter@uni.lu (T. Sauter), merja.heinaniemi@uef.fi (M. Heinäniemi).

http://dx.doi.org/10.1016/j.gdata.2014.07.002
2213-5960/© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

were kept in OF media supplemented with 0.01 mg/ml human transferrin, 100 nM cortisol and 20 nM insulin.

Chromatin immunoprecipitation

Nuclear proteins were cross-linked to DNA by adding formaldehyde directly to the medium to a final concentration of 1% for 8 min at room temperature. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M and incubating for 5 min at room temperature on a rocking platform. The medium was removed and the cells were washed twice with ice-cold PBS. The cells were then collected in lysis buffer (1% SDS, 10 mM EDTA, protease inhibitors, 50 mM Tris–HCl, pH 8.1) and the lysates were sonicated by a Bioruptor UCD-200 (Diagenode, Liege, Belgium) to result in DNA fragments of 200 to 500 bp in length. Cellular debris was removed by centrifugation and the lysates were diluted 1:10 in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, protease inhibitors, 16.7 mM Tris–HCl, pH 8.1). Chromatin solutions were incubated overnight at 4 °C with rotation with antibodies against H3K4me3 (4 μl per IP of 17–614, Millipore, Billerica, MA, USA), PPARγ (mixture of 0.5 μl per IP of sc-7196x, Santa Cruz Biotechnologies, Santa Cruz, CA, USA and 5 μl per IP of 101700, Cayman, Ann Arbor, MI USA), CEBPα (5 μl per IP of sc-61, Santa Cruz Biotechnologies), and LXRα (5 μl per IP, kind gift from Eckardt Treuter, Karolinska Institute, Stockholm, Sweden). The LXR antibody recognizes also LXRβ that maintains a constant low level of expression during differentiation. The immuno-complexes were collected with 20 μl of MagnaChIP protein A beads (Millipore) for 1 h at 4 °C with rotation. Non-specific background was removed by incubating the MagnaChIP protein A beads overnight at 4 °C with rotation in the presence of BSA (250 μg/ml). The beads were washed sequentially for 3 min by rotation with 1 ml of the following buffers: low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris–HCl, pH 8.1), high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl, 20 mM Tris–HCl, pH 8.1) and LiCl wash buffer (0.25 M LiCl, 1% Nonidet P-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris–HCl, pH 8.1). Finally, the beads were washed twice with 1 ml TE buffer (1 mM EDTA, 10 mM Tris–HCl, pH 8.1). The immuno-complexes were then eluted by adding 500 μl of elution buffer (25 mM Tris–HCl, pH 7.5, 10 mM EDTA, 0.5% SDS) and incubating for 30 min on rotation. The cross-linking was reversed and the remaining proteins were digested by adding 2.5 μl of proteinase K (Fermentas) to a final concentration of 80 μg/ml and incubating overnight at 65 °C. The DNA was recovered by phenol/chloroform/isoamyl alcohol (25:24:1) extractions and precipitated with 0.1 volume of 3 M sodium acetate, pH 5.2, and 2 volumes of ethanol using glycogen as carrier. Immunoprecipitated chromatin DNA was then used as a template for real-time quantitative PCR or for library preparation and sequencing (performed at EMBL core facility).

Data processing and alignment

Sequencing reads were quality controlled using the FASTQC software v.0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The quality scores were consistently high along the read length and the samples had overall good quality based on multiple metrics (see Supplementary data file 1, Figs. 1–5). Possible clonality in the PCR step of library preparation was evaluated by counting reads mapping per

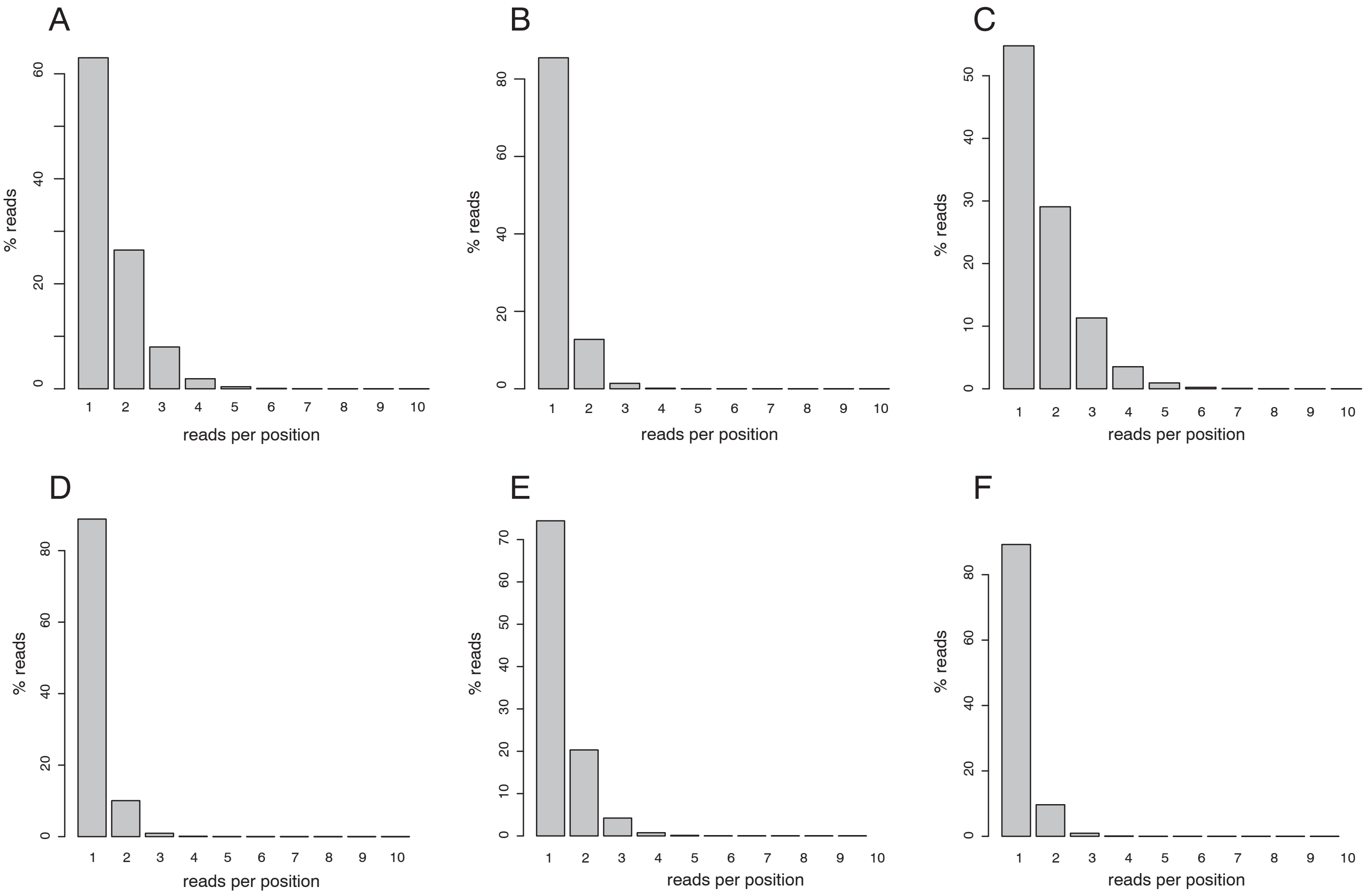


Fig. 1. Analysis of sample clonality. Histograms of clonal read depth are shown for the ChIP-seq samples. The bars indicate the number of reads per unique position. In an ideal ChIP-seq experiment there is a high fraction of single reads per position. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

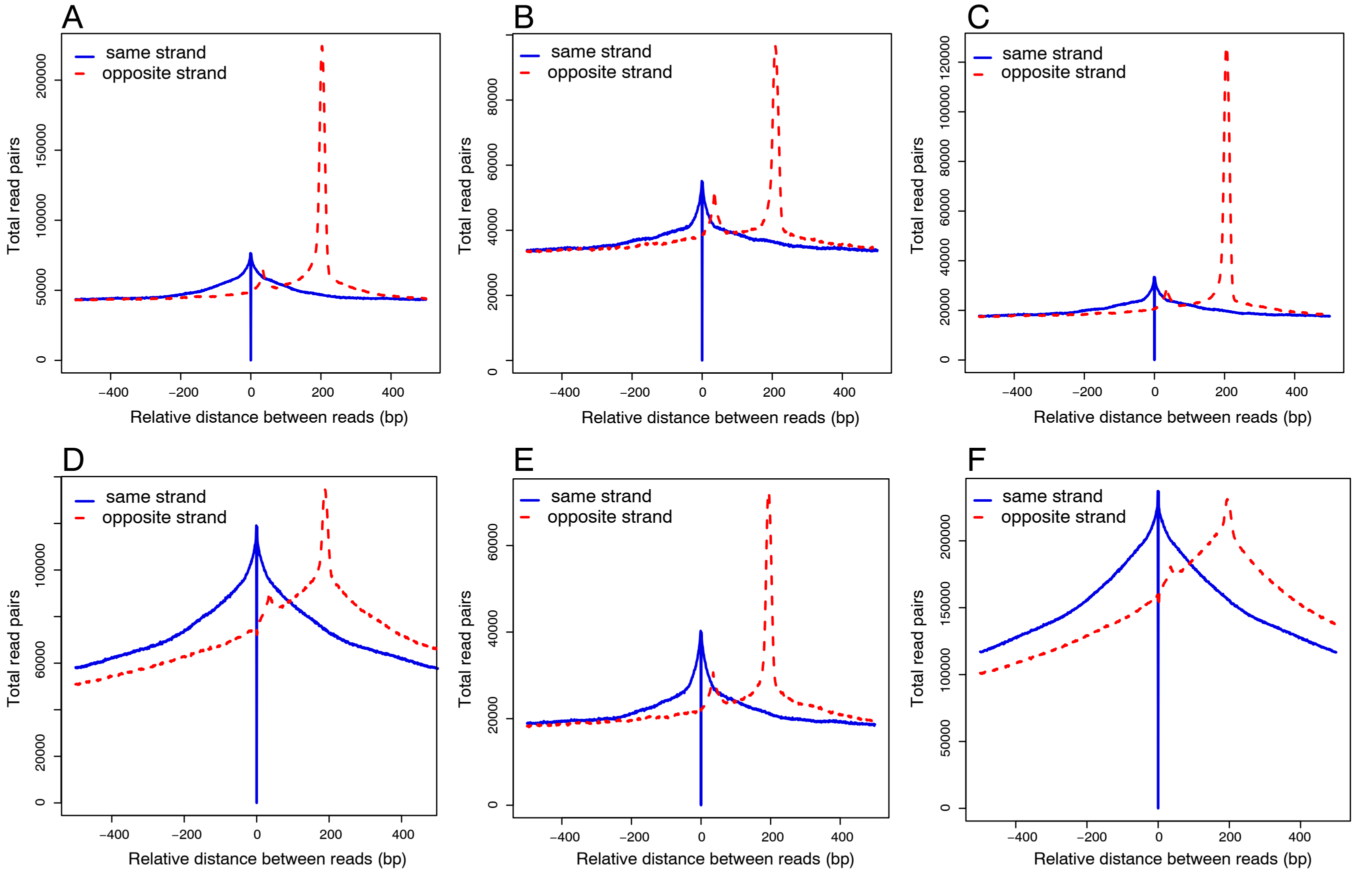


Fig. 2. Analysis of fragment length. The relative distance of reads mapping to ChIP-seq signal maximal from the two strands (positive and negative) is shown for the ChIP-seq samples. In a typical ChIP-seq experiment the peaks from opposite strands are 100–300 bp separated. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

genomic position (Fig. 1). This revealed some degree of clonal amplification in the samples for PPARγ and LXR. Therefore, we chose to include a stack collapsing step to the preprocessing. The fragment length was estimated based on distance between reads mapping to positive and negative strands at peak locations (Fig. 2) and agreed well in each sample with the expected size of approximately 200 bp. When examining the read base pair content (Figs. 3 and 4), a deviation from the expected GC-content was observed in the input sample of SGBS cells and this sample was replaced in the downstream analysis by a new input obtained from similarly differentiated cells. The slightly higher GC-content in H3K4me3 peaks is expected as these reads derive from promoter proximal regions that have typically higher GC-content than the rest of the genome.

Specifically, the FASTX software v.0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to remove read artifacts and those reads that had low quality base pair calling (minimum quality score of phred 10 across the read length was required) and to collapse read stacks. Subsequently, reads were aligned to the hg19 human genome using the Bowtie software v0.1.25 [3] with the following settings: one mismatch was allowed, maximum three locations in the genome were allowed, and the highest quality match was reported. This resulted in 9608582 mapped reads for PPARγ, 19889853 for CEBPα, 15375177 for LXR, 12253403 for adipocyte H3K4me3, 12550706 for preadipocyte H3K4me3 and 18109349 for input. A script that downloads the deposited reads from the NCBI SRA database and produces aligned reads with these settings is provided (see Additional Data File 2).

Analysis of ChIP-seq signal

The H3K4me3 histone mark is often found at active transcription start sites (TSS). We were interested to assign each gene to H3K4me3-positive vs -negative categories. For this purpose, the mixture modeling approach implemented in the EpiChIP software v.0.9.7 [4] was applied. The results have been presented elsewhere [1] and a thorough user-guide is available from the tool website (http://epichip.sourceforge.net/tutorial.html). As instructed in the user guide, differently sized windows around the TSS regions were quantified to choose a proper window size. The region −750 to +1250 centered at Refseq TSS coordinates had the highest amount of signal and was therefore chosen for signal quantification. Two distributions were clearly visible: the higher values corresponding to actual chromatin marker signal distribution separated from the background distribution. The software then assigns a probability for each TSS region that indicates whether it corresponds to the signal distribution. We employed the default settings and used the noise, unclassified and signal results to compare the preadipocyte and the adipocyte TSS activity.

TF peak detection was performed using the Quest software v.2.4 [5]. To allow configuring all settings, we turned on the advanced mode. Parameters were generated using the command QuEST_2.4/generate_QuEST_parameters.pl -bowtie_align_ChIP sample.bowtie -bowtie_align_RX_noIP input.bowtie -gt genome_table_hg19 -ap sampleFolder -ChIP_name sampleChIP –advanced. Default settings were otherwise applied except for the mappable genome fraction

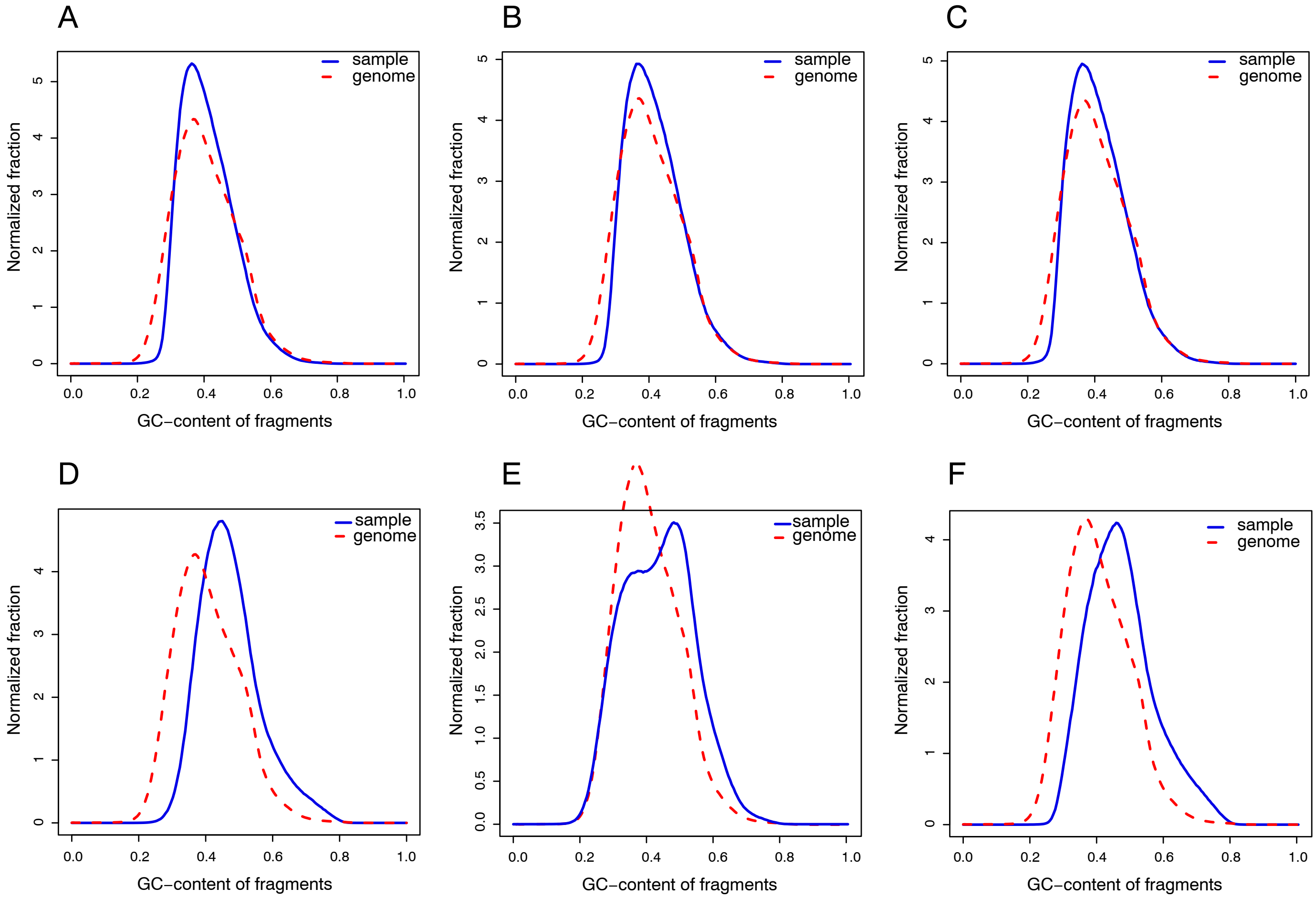


Fig. 3. GC-content of reads. The GC-content of the reads compared to that of the genome (hg19) is shown for the ChIP-seq samples. Typical ChIP-seq experiments with TF antibodies show a distribution that closely resembles that of random fragments from the genome. Gene proximal areas typically have higher GC-content and this is reflected in the H3K4me3 samples that have peaks nearby transcription start sites. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

(set to 0.88) and enrichment (ChIP enrichment set to 15 and ChIP to background enrichment to 2.5). The final peak lists were filtered to remove peaks with q-value (fdr) above 0.001 (−logQvalue > 3). Based on examining the signal wiggle files, cut-offs for low-occupancy (enrichment > 15) and high occupancy (enrichment > 30) binding sites were defined. Finally, using the UCSC Table Browser, we obtained a file (group: Repeats, track: RepeatMaster) corresponding to satellite repeats (#filter: rmsk.repClass = 'satellite') and removed peaks overlapping these regions.

TF motif detection by the MEME-ChIP software [6] was performed using the high occupancy peaks. We used the setting -nmotifs 10 -minw 6 -maxw 30 and matched the motifs found to the JASPAR 2009 CORE database. MEME analysis using 600 randomly chosen trimmed (central 100 bp) input sequences revealed the respective TF motif as top motif present in the sample in each case (Fig. 6). The canonical binding sites matched to the TF analyzed were: MA0065.2 PPARG::RXRA and MA0065.1 PPARG::RXRA for PPARγ peaks; CEBPα: MA0102.2 (CEBPA), MA0102.1 (Cebpa) for CEBPα peaks; while a close match to motif reported by Feldman et al. [7] was found for LXR. This indicates that the antibody collection and downstream analysis were successfully performed. A script to run the motif analysis is provided (see Additional data file 2).

Discussion

Here we describe deep sequencing data obtained from human SGBS preadipocyte differentiation. This dataset is composed of data derived using the Illumina Genome Analyzer II. We demonstrated genome-wide binding pattern of key adipogenic TFs that were shown to co-occupy several loci. Further, this dataset is part of a GEO Superseries (GSE41578) and we have used it in combination with gene expression data to associate putative target genes in [1]. To further analyze the data in an integrative manner, we introduced the web portal IDARE (Integrated Data Nodes or Regulation) in [1] for interactive data exploration of the results within the metabolic network context, available at http://systemsbiology.uni.lu/idare.html, including a detailed user guide. Direct links to our ChIP-seq track hub which can be used to visualize the signal at any genomic loci are available from this website. We also provide results on motif analysis presented in this paper that can be used for further analysis

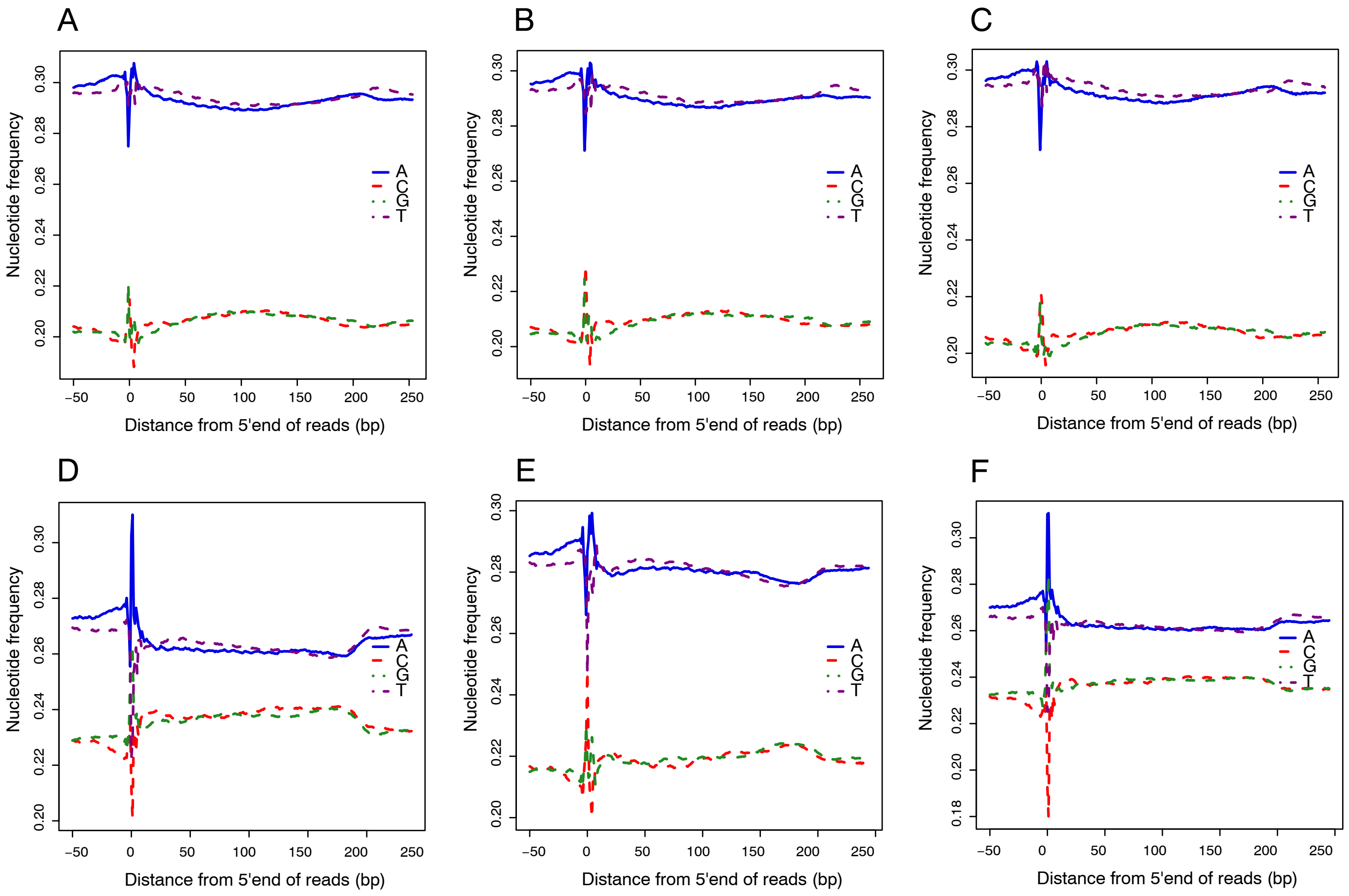


Fig. 4. Nucleotide frequencies along the read length. The frequency of each nucleotide along the read is plotted for each ChIP-seq sample. The frequencies of A and T are typically very similar to those of G and C. Gene proximal areas typically have higher GC-content and this is reflected in the H3K4me3 samples that have peaks nearby transcription start sites. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

of combinatorial TF binding. Results from the data have increased our understanding of the TF-mediated control of human adipogenesis.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.07.002.

Acknowledgments

We would like to acknowledge the following funding sources: Fonds National de la Recherche Luxembourg [AFR, AM2c and C08/BM/01]; University of Eastern Finland and Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg.

References

[1] M. Galhardo, L. Sinkkonen, P. Berninger, J. Lin, T. Sauter, M. Heinäniemi, Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. Nucleic Acids Res. 42 (2014) 1474–1496.

[2] M. Wabitsch, R.E. Brenner, I. Melzner, M. Braun, P. Möller, E. Heinze, K.M. Debatin, H. Hauner, Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. Int. J. Obes. Relat. Metab. Disord. 25 (2001) 8–15.

[3] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25.

[4] D. Hebenstreit, M. Gu, S. Haider, D.J. Turner DJ, P. Liò, S.A. Teichmann, EpiChIP: gene-by-gene quantification of epigenetic modification levels. Nucleic Acids Res. 39 (2011) e27.

[5] A. Valouev, D.S. Johnson DS, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, A. Sidow, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods 5 (2008) 829–834.

[6] P. Machanick, T.L. Bailey, MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27 (2011) 1696–1697.

[7] R. Feldmann, C. Fischer, V. Kodelja, S. Behrens, S. Haas, M. Vingron, B. Timmermann, A. Geikowski, S. Sauer, Genome-wide analysis of LXRα activation reveals new transcriptional networks in human atherosclerotic foam cells. Nucleic Acids Res. 41 (2013) 3518–3531.

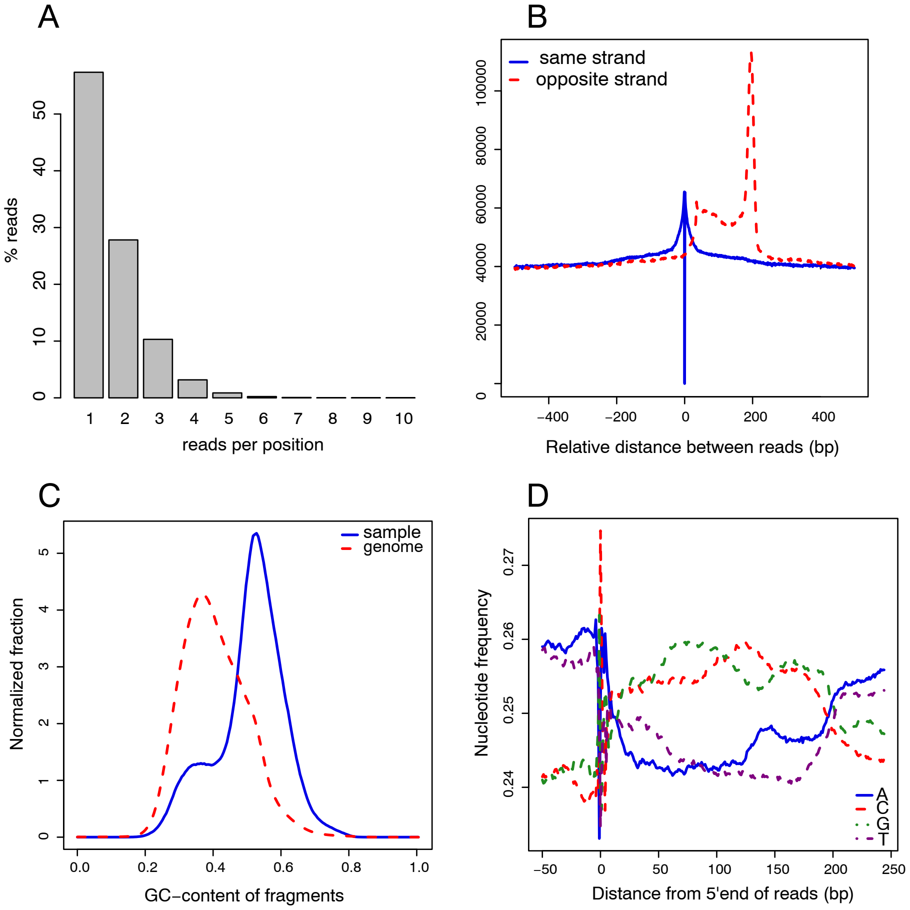


Fig. 5. Quality control data for discarded sample. One input sample did not match the other samples in terms of the quality results. As in Figs. 1–4, analysis of sample clonality is shown in A, analysis of fragment length in B, GC-content in C and nucleotide frequencies along the read length in D.

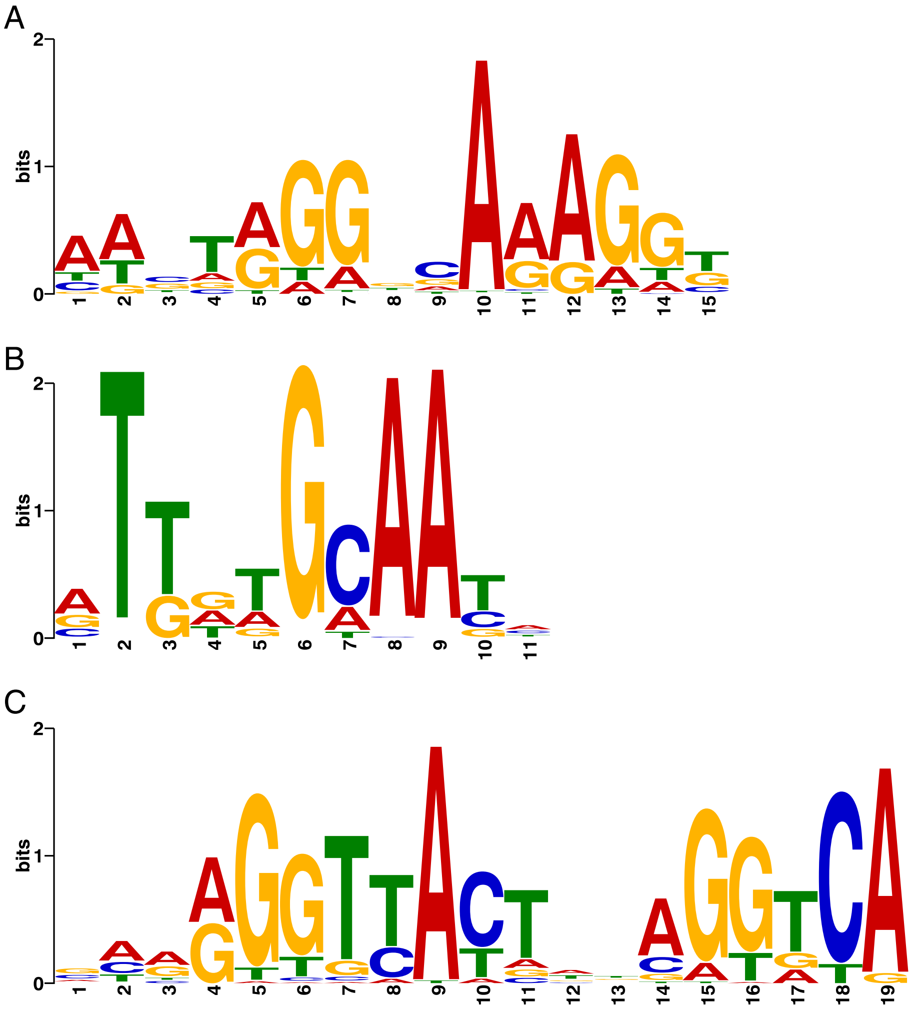


Fig. 6. TF *de novo* motif analysis. The top motifs detected in PPARg (in A), CEBPa (in B) and LXR (in C) peaks are shown. Letter height indicates information content in bits. Those positions with high information content are typically well conserved between binding sites and correspond to protein–DNA contacts.

The IDARE Cytoscape Application and Web Server

Thomas Pfau[1,2], Jake Lin[3], Mafalda Galhardo[1], and Thomas Sauter[1*]

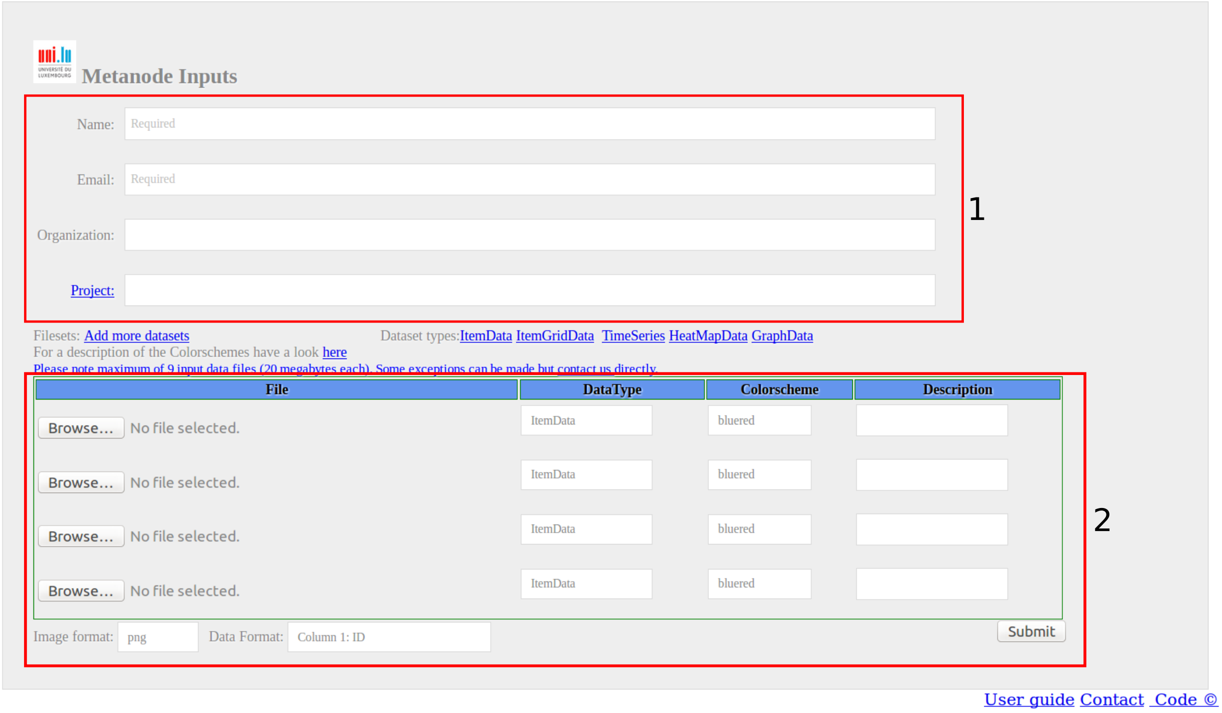
[1]Life Sciences Research Unit, University of Luxembourg
[2]Institute of Complex Systems and Mathematical Biology, University of Aberdeen
[3]Computational Biology, BioMediTech, University of Tampere
[*]Corresponding author - thomas.sauter@uni.lu

March 18, 2015

Chapter 1

The IDARE Webserver for integrated OMICS node generation

The IDARE web server (found at `http://idare-server.uni.lu/`) provides an easy to use interface to allow the automated generation of images comprising multiple sources of information. In this section we highlight the most important features of the user interface and explain what the different fields are for. The layout of the webserver interface can be seen in Figure 1.1



Metanode Inputs

Name: Required

Email: Required

Organization:

Project:

Filesets: Add more datasets Dataset types:ItemData ItemGridData TimeSeries HeatMapData GraphData

For a description of the Colorschemes have a look here

Please note maximum of 9 input data files (20 megabytes each). Some exceptions can be made but contact us directly.

File	DataType	Colorscheme	Description
Browse... No file selected.	ItemData	bluered	
Browse... No file selected.	ItemData	bluered	
Browse... No file selected.	ItemData	bluered	
Browse... No file selected.	ItemData	bluered	

Image format: png Data Format: Column 1: ID

Figure 1.1: This is the Webinterface of the automated image node generation software. The upper area (1) contains the fields to enter the user data as detailed in Secion 1.1. The lower area (2) covers all settings for specific properties of the generated nodes and allows the upload of data files for processing which are explained in Section 1.2.

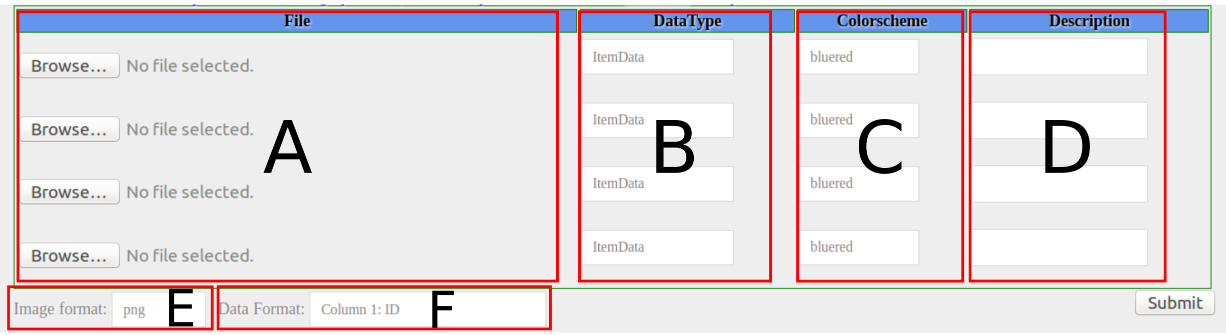


Image format: png

Data Format: Column 1: ID

Figure 1.2: The Datafields for input data. The Files to be uploaded are selected in Field A. The Dropdown Selectors B and C allow the selection of the type of data and the colorscheme used (explained in detail in Table I). Field D allows the supply of a description of the dataset while the dropdowns E and F define what type of images are generated and what format (single or dual-id) is used (see 1.2.1).

Field	Description
File (A)	The files containing the different data sets that will be used for the creation of the graphical nodes.
DataType(B)	The type of data in the File. There are currently 5 different types of data that can be processed which are detailed in Section 1.2.2
ColorScheme (C)	The Colorscheme used for filling the entry representations. If data is used that is around 0 (and not too skewed to one side) the image generator tries to set the middle color to 0.
Description (D)	The description provided here will be used as the header of the legend for this dataset.
Image Format (E)	The format of the images generated. Either PNG or SVG (support vector graphics).
Data Format (F)	Whether a 1 Column or 2 Column Item ID will be assumed. This has to be the same for all datasets (see Section 1.2.1)

Table I: The Options on the User interface of the image generation server.

1.1 User Data Input

User data is divided into a couple of categories and the only strictly necessary data is the name and email address. The latter will be used to inform and contact the user and provide updates regarding the image generation. The generation time for the images is dependent on the number of datasets and the number of entries in the datasets. It commonly takes about 5 minutes for one Dataset of 1000 entries.

1.2 Project Data Files and Properties

There are Multiple options associated with the datasets that can be used when transmiting data to the server. The different options are listed in Table I. Figure 1.2 gives an overview of the options available in the interface.

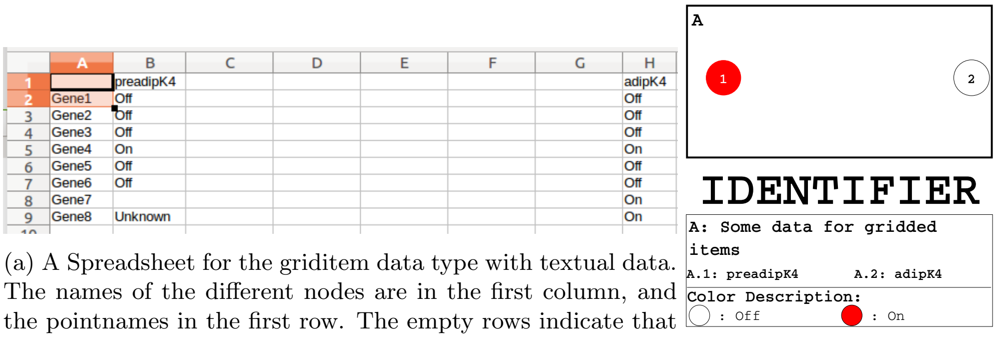
1.2.1 File Format

In general, the image generation server can use two types of input files:

1. Excel Spreadsheets

2. CSV Files with values separated by tabulator.

The system expects the first row to contain information on the different datapoints of the supplied data. These will be used as labels for the legend of the respective dataset. There are two different ways for



	A	B	C	D	E	F	G	H
1		preadipK4						adipK4
2	Gene1	Off						Off
3	Gene2	Off						Off
4	Gene3	Off						Off
5	Gene4	On						On
6	Gene5	Off						Off
7	Gene6	Off						Off
8	Gene7							On
9	Gene8	Unknown						On

(a) A Spreadsheet for the griditem data type with textual data. The names of the different nodes are in the first column, and the pointnames in the first row. The empty rows indicate that there are 5 unused items between the first and the last, which allows alignment of data.)

(b) A node image (Top) and the legend image for Griddata from 1.3a

Figure 1.3: An example of the data generating a graph node and the resulting node legend

labeling the nodes. The default option is, that the node labels will be the entries in the first column (starting from row 2). For use in the cytoscape app these identifiers have to match to one column of the generated network so that they can be matched to the appropriate nodes in the network. The second option for labeling is that one column containing the cytoscape node ids is provided and the second column contains the labels that will be displayed on the node images. This is particularly important since often published networks use more computer readable identifiers that are hard to interpret by a user. It allows the user to use gene symbols as labels while interacting with a network that e.g. only contains ENSEMBLE IDs, making visual inspection easier. The remaining file should consist of either up to five different values (e.g. 'on', 'off', 'unknown') or numeric values. These values will be used to generate the nodes. Empty entries will be interpreted as missing values and the respective positions will not be printed in the generated images. If white is also a color in the selected colorscheme, there will be no difference between a node with a value mapped to white and without a value. If there is a completely empty column, this column will still be included in the resulting nodes (see Figure 1.3).

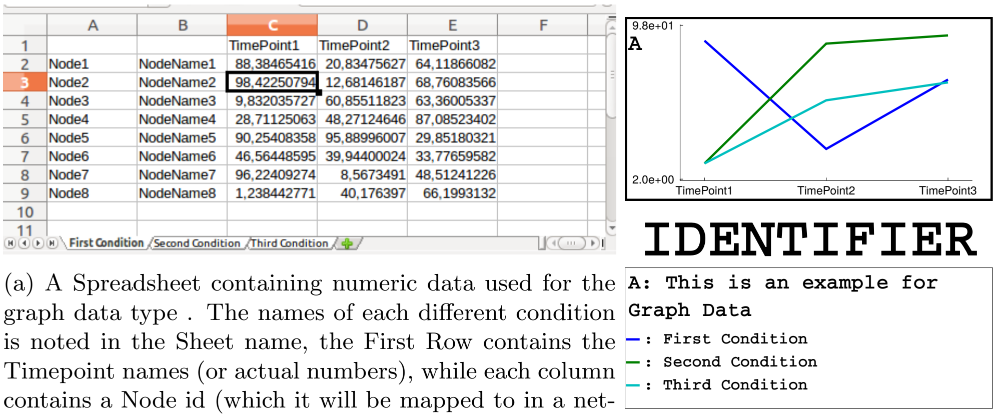
This can be used to match two datasets which have values that only partially overlap. An example would be two time series where one series has timepoint 1h,2h,4h, and 8h, while the other has 1h,4h,6h and 8h. In this instance adding an empty row behind 4h in Dataset 1 and one after 1h in dataset 2 will lead to matched positions. There is one Dataset Type that is different: GraphData. For graphs with more than one line, only excel sheets can be used. This is necessary, as the sheets in an excel file will be interpreted as independent datasets for each line on the graph and no sheets are available for csv. An example for a GraphData file is presented in Figure 1.4.

1.2.2 Data types for images

There are five different types of data implemented at the moment:

1. ItemData
2. ItemGridData
3. TimeSeriesData
4. HeatMapData
5. GraphData

An overview of the properties is provided in Table II. ItemData is the simplest type of data and where each column will be represented by a circle. The System assumes, that the order in which these items



	A	B	C	D	E	F
1			TimePoint1	TimePoint2	TimePoint3	
2	Node1	NodeName1	88,38465416	20,83475627	64,11866082	
3	Node2	NodeName2	98,42250794	12,68146187	68,76083566	
4	Node3	NodeName3	9,832035727	60,85511823	63,36005337	
5	Node4	NodeName4	28,71125063	48,27124646	87,08523402	
6	Node5	NodeName5	90,25408358	95,88996007	29,85180321	
7	Node6	NodeName6	46,56448595	39,94400024	33,77659582	
8	Node7	NodeName7	96,22409274	8,5673491	48,51241226	
9	Node8	NodeName8	1,238442771	40,176397	66,1993132	
10						
11						

First Condition / Second Condition / Third Condition

A: This is an example for Graph Data

— : First Condition

— : Second Condition

— : Third Condition

(a) A Spreadsheet containing numeric data used for the graph data type . The names of each different condition is noted in the Sheet name, the First Row contains the Timepoint names (or actual numbers), while each column contains a Node id (which it will be mapped to in a network) and a node name. The first row and the first two columns have to be the same in all sheets of the spreadsheet (if no label column is supplied, only the first column has to be identical)

(b) A node image (Top) and the legend image for the data from 1.4a

Figure 1.4: An example of the data generating a graph node and the resulting node legend

Data Type	Description	Properties
ItemData	Data that is individual and for which the different entries are commonly unconnected.	Circles, Edge
ItemGridData	Itemized Data that is in a specific Grid	Circles, Center
TimeSeriesData	Data that is connected and displayable in one row	Boxes, Center
HeatMapData	Data that can be rearranged but is somewhat connected and displayed in the style of a heatmap	Boxes, Center or Edge
GraphData	Data that is displayed as a line graph. A similar scale is necessary	Graph, Center

Table II: A short description of the currently available types of data visualisations.

are represented is not important and will place them at any position that is left after placing all other datasets. ItemGridData is similar to item data, generating one circle per column in the dataset file. The difference is, that the System assume that there is an order to this data and will place it into one of the central rows. TimeSeries data is similar to ItemGridData but the representation is done with connected boxes making the connection between the different values more obvious. HeatMapData is also using boxes but can be placed at any position convenient for the system. In this respect it is similar to ItemData. The final data type is GraphData. This type of data will generate a graph with one line plot for each sheet in the provided excel sheet. If the Column IDs are numeric values it will use these values to place the points at the appropriate positions on the graph.

1.2.3 Processing and Results

Once the jobs are submitted the user will be informed of the progress by email and a download link will be provided once the files are generated.

Chapter 2

The IDARE2 Cytoscape Application

The IDARE2 cytoscape app was designed to allow the mapping of the images produced by the image generation webserver onto cytoscape networks. It can be downloaded directly from the idare webserver at `http://idare-server.uni.lu/IDARE.zip`. This allows the use of the vast amount of apps available for network analysis in cytoscape while visualising experimental data in an easy to use, context specific fashion. The app contains two main modules:

1. The subnetwork extractor tool and
2. the image mapping tool and

This chapter will detail the two tools and their possibilities and give guidance to their use. In addition, a small helper function specific to SBML Files in COBRA format is provided.

2.1 The Subnetwork extractor tool

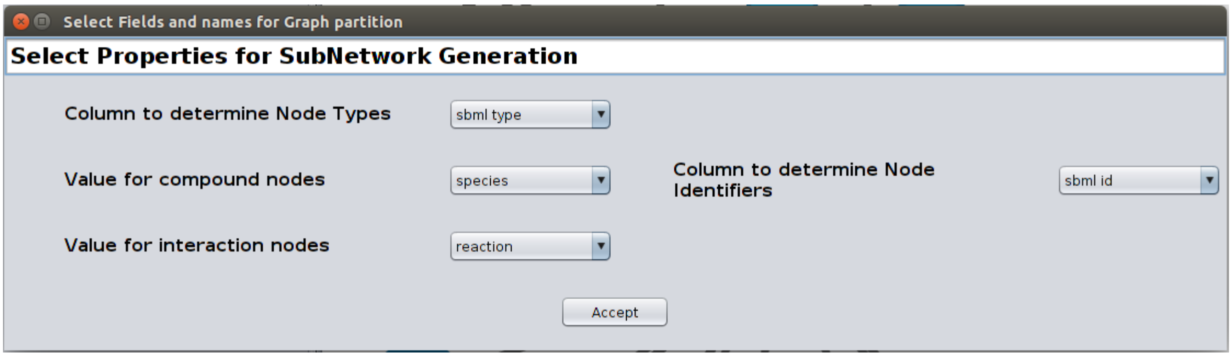
One issue often found in large networks is that visualisation becomes difficult due to the enourmous amounts of interactions, leading to the classical image of the network-hairball. While some properties might still be visible on that level, a detailed view of a subnetwork often allows a better interpretation of the observed fluxes through a network. With biological models, there is commonly known information of pathways which form connected subnetworks. It is easy to restrict a network to the reactions found in such a pathway if only that specific pathway is of interest to the researcher. In doing so, it is however often problematic to keep track of branching pathways, which leads to a loss of the overview.

The subnetwork extractor allows the user to extract subnetworks while keeping connections to the general network. This is achieved by creating links between different subnetworks that point to the position of the linked metabolite in the other network. Thus it is possible to follow the course of flux, or the general network structure by following the links generated. The method is applicable to any bi-partite network, that can be divided into “compound” and “interaction” nodes.

The following figures provide an example of the application of the subnetwork extractor tool. In essence, the user selects one column of the Cytoscape network table which has to contain the information which class each node belongs to (Fig. 2.1).

This information is also detailed in Table I.

Then the classes representing compounds and interactions, respectively, are selected from the entries in that column. If there are further values in that column, those are ignored for definition of the subnetworks. However, nodes will be included in the networks if they are connected to any of the subnetwork nodes and not defined to be in another subnetwork.



Select Properties for SubNetwork Generation

Column to determine Node Types: sbml type

Value for compound nodes: species

Value for interaction nodes: reaction

Column to determine Node Identifiers: sbml id

Figure 2.1: On the left: The property selection screen for the column used for compounds and interactions and the identifiers for the compounds and interactions. Right: The user is asked to select which row is considered for the unique identifiers in the network.

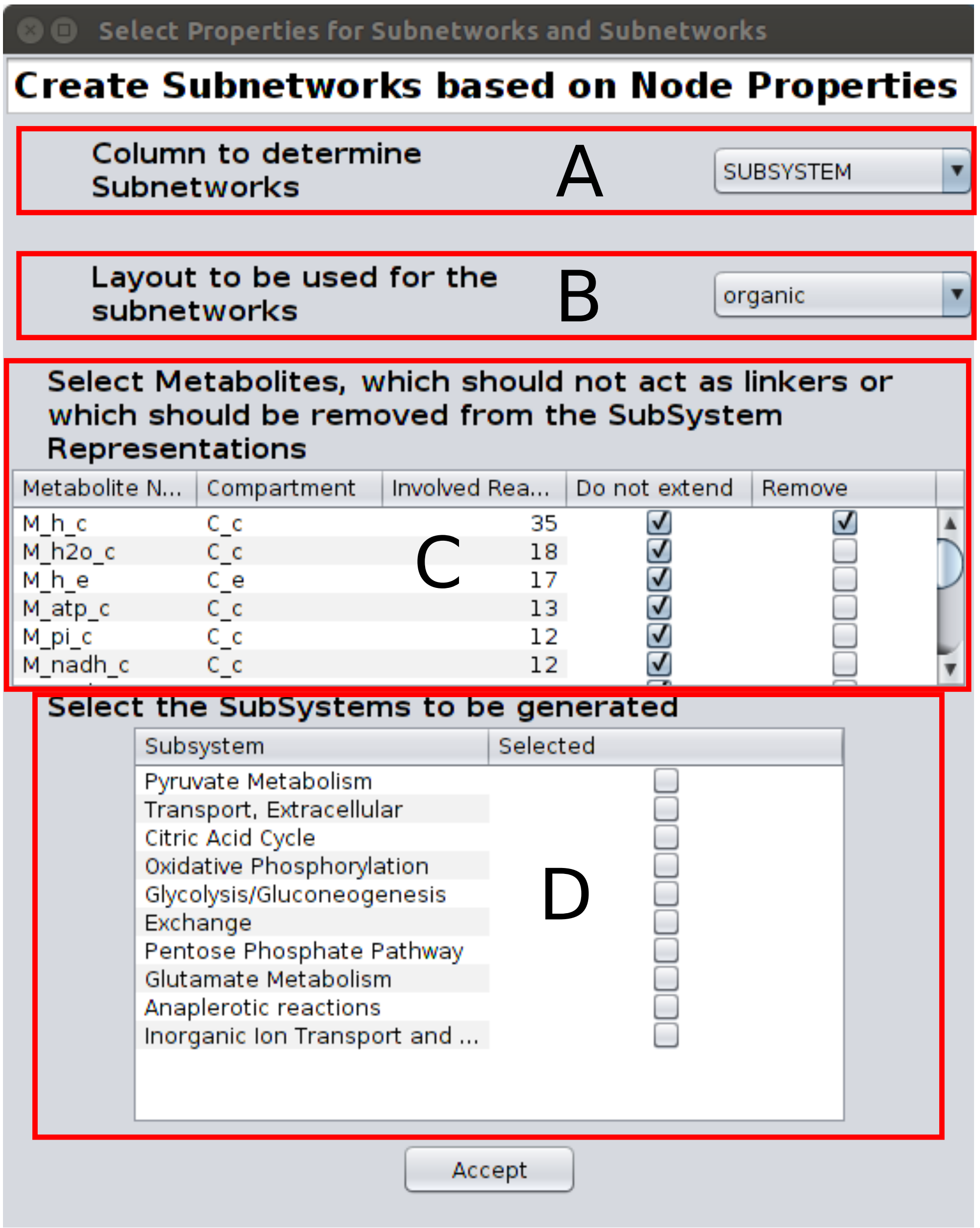
Property	Default	Description
IDARENodeType	sbml type	The type of a node used by IDARE (important for compound/interaction selection)
IDARENodeName	sbml id	The id of a node used by IDARE (images are matached to these IDs)
IDARELinkTargets	-	This column defines which other node a linkernode is linking to. Only read during loading.
IDARELinkTargetID	-	This is the ID of the node for linking purposes. Only read during loading.
IDARETargetSubsystem	-	The name of the target Subsystem of a Linker Node.

Table I: Cytoscape columns used by IDARE for data management

A second column has to be selected that will be used to determine the membership of an interaction in a given subnetwork. This column will also be used to determine the available subnetworks.

Finally, the user is presented with a selection screen (see Figure 2.2) , containing all compounds and all subnetworks. This selection screen allows the definition of compounds, which should not become part of the final subnetworks (like e.g protons or water, which are abundant and make layouting and visual inspection difficult). It also allows the definition of compounds which should not be considered when creating links between subnetworks. This is useful if considering e.g. metabolites like glyceraldehyde-3-phosphate. While removing it from the pathways would lead to gaps in the flow, branching would lead to the inclusion of an enourmous amount of links. While this can be desireable the tool assumes that very common compounds are not supposed to branch and the most common are to be removed. However the final choice of compounds to be removed/declared as non branching can be done by the user in the selection screen.

In addition, the selection screen allows the user to select which subnetworks should be generated. When all selections are made, the subnetwork generator will create one additional network view for each subnetwork created. It will also generate linker nodes for each combination of subnetworks sharing branching compounds. Double clicking on one of these links will automatically open the respective network view, centered on the compound in the opened view corresponding to the compound that was connected to the link in the origin network.



Create Subnetworks based on Node Properties

Column to determine Subnetworks: SUBSYSTEM

Layout to be used for the subnetworks: organic

Select Metabolites, which should not act as linkers or which should be removed from the SubSystem Representations

Metabolite N...	Compartment	Involved Rea...	Do not extend	Remove
M_h_c	C_c	35	☑	☑
M_h2o_c	C_c	18	☑	☐
M_h_e	C_e	17	☑	☐
M_atp_c	C_c	13	☑	☐
M_pi_c	C_c	12	☑	☐
M_nadh_c	C_c	12	☑	☐

Select the SubSystems to be generated

Subsystem	Selected
Pyruvate Metabolism	☐
Transport, Extracellular	☐
Citric Acid Cycle	☐
Oxidative Phosphorylation	☐
Glycolysis/Gluconeogenesis	☐
Exchange	☐
Pentose Phosphate Pathway	☐
Glutamate Metabolism	☐
Anaplerotic reactions	☐
Inorganic Ion Transport and ...	☐

Figure 2.2: Field A: The Selection of the column used to determine the subnetworks (The default selected item is “SUBSYSTEMS” if it is present. According to this selection, Field D will adjust the available networks. Field B is the selector for the column used to select a layouting algorithm for the newly created subnetworks. Field C provides the user with the options which compounds to consider in subnetwork generation (for either linking or in general)

2.2 Image mapping tool

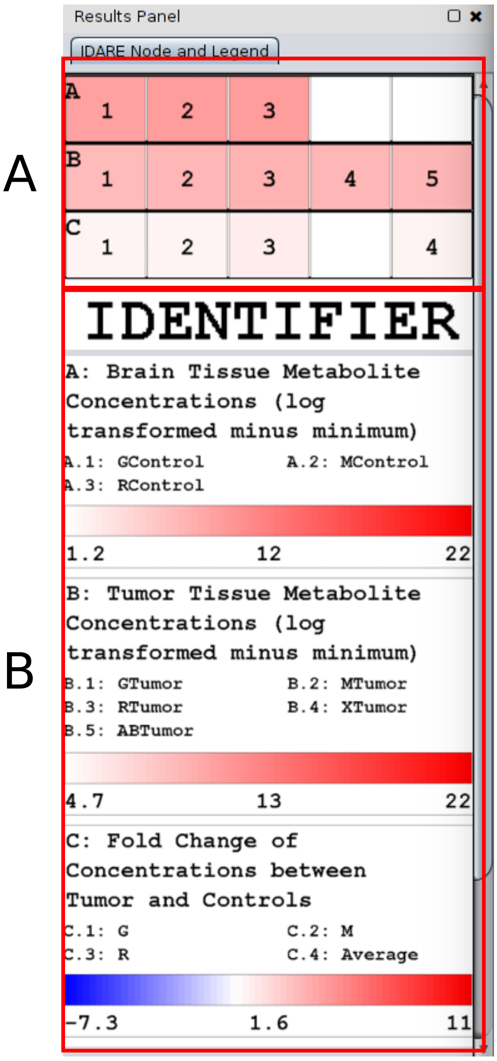
The second tool provided by the app is the image mapping tool. There are two parts of this tool, an image storage and a visual mapping function. The images can be loaded from the archive generated by the webserver using the function Apps → IDARE → Load Metanodes (which can be seen in figures ...). Multiple Node files can be loaded and will be managed by the IDARE app. Upon loading, the user is asked to define which column to use to map the IDs.

If the app encounters images for nodes which are already associated with an image the new image will override the old one. When loading images the system will try to determine whether the current network is already set up for IDARE. I.e. it will check, whether IDARE-specific columns exist in the network table which are used to match images and determine layout properties. If the network is not set up (or not completely set up), the user will be asked to provide information for the setup.

This information encompasses:

1. The column containing the id that the images are mapped against.
2. The column containing the information about the type of the node.
3. The identifiers in the column specified in (2) that stand for compounds and interactions, respectively.

While only the first is strictly necessary for the image mapping, the IDARE Visual style will assume certain columns to be present and it is necessary to initialize these. After loading the images, the app will map them directly on the nodes indicated by the image names. It will also associate each node with a legend that will be displayed in the cytoscape results panel, when the respective node is selected. The legend will contain detailed information about the fields of the node (see Figure 2.3).



IDENTIFIER

A: Brain Tissue Metabolite Concentrations (log transformed minus minimum)

A.1: GControl A.2: MControl
A.3: RControl

B: Tumor Tissue Metabolite Concentrations (log transformed minus minimum)

B.1: GTumor B.2: MTumor
B.3: RTumor B.4: XTumor
B.5: ABTumor

C: Fold Change of Concentrations between Tumor and Controls

C.1: G C.2: M
C.3: R C.4: Average

Figure 2.3: The legend displayed when selecting an image node. A: a node representation with Identifiers for each field used. B: Descriptions of the fields.

2.3 The IDARE Visual Style

The IDARE Visual Style is closely linked to the image mapping tool. It is necessary to apply the mappings between images and nodes. In addition, it applies more appropriate styles to the linking nodes, displaying only the name. Finally there are some layou choices made for nodes: Nodes with the type reaction are visualised as blue squares, species are visualised as red circles, genes are visualised as yellow diamonds and proteins are visualised as green hexagons. In addition, undirected edges will be marked by arrows at both ends and directed edges by an arrow at the target node. This makes the visual style focused on metabolic networks. However, the user is free to adapt the style to her needs.

2.4 COBRA specific SBML reader

The COBRA SBML specification defines several fields in the notes section of SBML files which are specific to metabolic reconstructions. In particular those fields include information about genes (GENE_ASSOCIATION and GENE_LIST). To visualize metabolic models, this information is very useful, as it will give a link between the level of expression data and metabolism. Unfortunately the normal SBML import of Cytoscape completly ignores (and even discards) any information that is not directly associated with species and reactions. Thus the supplied SBML Annotation task allows to retrieve this information from an SBML file and add it to the model. It will also add information from all other COBRA fields like CHARGE, FORMULA, AUTHORS, EC Number or SUBSYSTEM. Especially the latter is very useful when trying to create subnetworks as it often contains the common definitions of metabolic pathways.