Epigenetic control of metabolic identity across cell types

Maria Pires Pacheco[1], Déborah Gerard[1], Riley J. Mangan[2,3,4], Alec R. Chapman[2,3], Dennis Hecker[5], Manolis Kellis[2,3], Marcel H. Schulz[5], Lasse Sinkkonen[1*], Thomas Sauter[1*]

[1] Department of Life Sciences and Medicine, University of Luxembourg, 6, Avenue du Swing, L-4367 Belvaux, Luxembourg

[2] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

[3] Broad Institute of MIT and Harvard, Cambridge, MA, 02139, USA

[4] Genetics Training Program, Harvard Medical School, Boston MA, 02115, USA

[5] Institute for Computational Genomic Medicine and Institute of Cardiovascular Regeneration, Medical Faculty, Goethe University, 60590 Frankfurt am Main, German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany

* Corresponding authors: thomas.sauter@uni.lu; lasse.sinkkonen@uni.lu

ORCIDs:

Maria Pires Pacheco: https://orcid.org/0000-0001-7956-8093

Déborah Gerard: https://orcid.org/0000-0002-6083-6614

Riley J. Mangan: https://orcid.org/0000-0003-3342-3934

Alec R. Chapman: https://orcid.org/0000-0002-6580-4005

Dennis Hecker: https://orcid.org/0000-0003-0272-243X

Manolis Kellis: https://orcid.org/0000-0001-7113-9630

Marcel H. Schulz: https://orcid.org/0000-0002-1252-3656

Lasse Sinkkonen: https://orcid.org/0000-0002-4223-3027

Thomas Sauter: https://orcid.org/0000-0001-8225-2954

Abstract

Constraint-based network modelling is a powerful tool for analysing cellular metabolism at genomic scale. Here, we conducted an integrative analysis of metabolic networks reconstructed from RNA-seq data with paired epigenomic data from the EpiATLAS resource of the International Human Epigenome Consortium (IHEC). Applying a state-of-the-art contextualisation algorithm, we reconstructed metabolic networks across 1,555 samples corresponding to 58 tissues and cell types. Analysis of these networks revealed the distribution of metabolic functionalities across human cell types and provides a compendium of human metabolic activity. This integrative approach allowed us to define, across tissues and cell types, i) reactions that fulfil the basic metabolic processes (core metabolism), and ii) cell type-specific functions (unique metabolism), that shape the metabolic identity of a cell or a tissue. Integration with EpiATLAS-derived cell type-specific gene-level chromatin states and enhancer-gene interactions identified enhancers, transcription factors, and key nodes controlling core and unique metabolism. Transport and first reactions of pathways were enriched for high expression, active chromatin state, and Polycomb-mediated repression in cell types where pathways are inactive,

suggesting that key nodes are targets of repression. This integrative analysis forms the basis for identifying regulation points that control metabolic identity in human cells.

Introduction

Identifying the molecular hallmarks of cell identity is crucial for understanding cellular differentiation in development and disease and can inform strategies for cellular reprogramming. Although the concept of cell identity remains debated, it is often inferred from cellular functions and phenotypes governed by gene expression patterns maintained by identity-specifying transcription factors (TFs) [1].

Metabolism is a key determinant of cell identity. For example, in Type II Diabetes, high blood glucose levels can disrupt beta cell identity, resulting in loss of beta cell-specific gene expression and gain of gene expression programs of other islet cell types [2]. Similarly, oxidative phosphorylation and ROS signalling drives the differentiation of intestinal stem cells into Paneth cells [3]. The centrality of metabolism in cellular identity is further supported by the observation that culture media, rather than tissue of origin, governs pluripotent stem cell identity.

Metabolic identity extends the cellular identity concept to the activities of biochemical pathways in a cell. Metabolic identity includes the core metabolism, a set of biochemical reactions shared across cell types that maintain energy and chemical homeostasis, and unique metabolism, or cell type-specific metabolic processes in differentiated cells, which correspond to specialized tasks including alcohol detoxification in the liver and melanin production in skin melanocytes.

Metabolic network modelling, which integrates cell type-specific epigenomic maps with the biochemical reactome, is a powerful approach to capture metabolic identity across cell types and identity disruptions in disease states[4–10]. Metabolic models provide key insights into how the inactivation of rate-limiting enzymes controls metabolic flux, interactions between metabolic pathways and compartments, and the ultimate capacity of cells to perform metabolic functions [11].

Cell-type-specific metabolism is achieved through precise regulation of pathway enzymes through the epigenetic control of gene expression, post-translational modifications, and compartmental sequestration of metabolites and co-factors. In support of this, we previously demonstrated that cells place high regulatory loads on rate-limiting pathway enzymes [12]. Specifically, loci encoding transporter proteins that control substrate availability show high H3K27ac signal in human macrophages [13].

More generally, cell-type-specific gene expression is maintained by regulatory elements such as promoters, enhancers, and silencers, whose activities have been mapped genome-wide through comprehensive ChIP-seq, ATAC-seq, and DNase-seq profiling of histone modifications and open chromatin generated by the Roadmap Epigenomics [14], ENCODE [15], and other consortia. The recent International Human Epigenome Consortium EpiATLAS resource [16] greatly expands these resources, providing an unprecedented opportunity to dissect the regulatory control of metabolism. Notably, enhancer-gene interactions inferred in the EpiATLAS resource using the generalized Activity-By-Contact (ABC) model [17] provide opportunities to connect regulatory elements, and their upstream transcription factor regulators, to metabolic genes.

Probabilistic graphical model methods, including ChromHMM [18] and Segway [19], infer cell type-specific chromatin states from combinatorial and spatial patterns of epigenomic signals. While ChromHMM and Segway define chromatin states at the level of individual loci, the recent ChromGene [20] method defines gene-level chromatin state by integrating histone modification signals throughout the gene body and flanking regions using a mixture of hidden Markov models [21]. ChromGene defines six active

states (1-6), one poised state (7), and five repressive states (8-12) **(Supplementary Table 1)**, based on combinations of ten histone modifications. This gene-level view provides better insight into which genes encoding for enzymes and transporters in the metabolic networks are active at the chromatin level, and which are transcriptionally repressed.

In this work, we reconstructed metabolic models across 1,555 biosamples from the IHEC EpiATLAS resource spanning 58 human cell types and tissues to identify core and tissue-specific metabolic reactions. We leveraged EpiATLAS enhancer-gene interactions to discover regulatory elements and upstream transcription factors controlling these processes. We observe that transporter genes are enriched for cell type-specific regulatory control and that many core metabolic pathways expressed across tissue types exhibit cell-type-specific regulatory control, providing a mechanistic basis for tissue-specific metabolic reprogramming of core pathways.

Results

Metabolic models capture core and tissue-specific metabolism

We used RNA-seq data of 1,555 samples from the EpiATLAS [16] resource to reconstruct metabolic models that served as scaffolds to map gene-level chromatin states (ChromGene) and enhancer activity data from the corresponding samples to the human metabolic network to define: 1) the core metabolic biochemical reactions across cell types and tissues, 2) the metabolic identity for individual cell types and tissues, and 3) the regulatory mechanisms controlling this identity **(Figure 1)**.

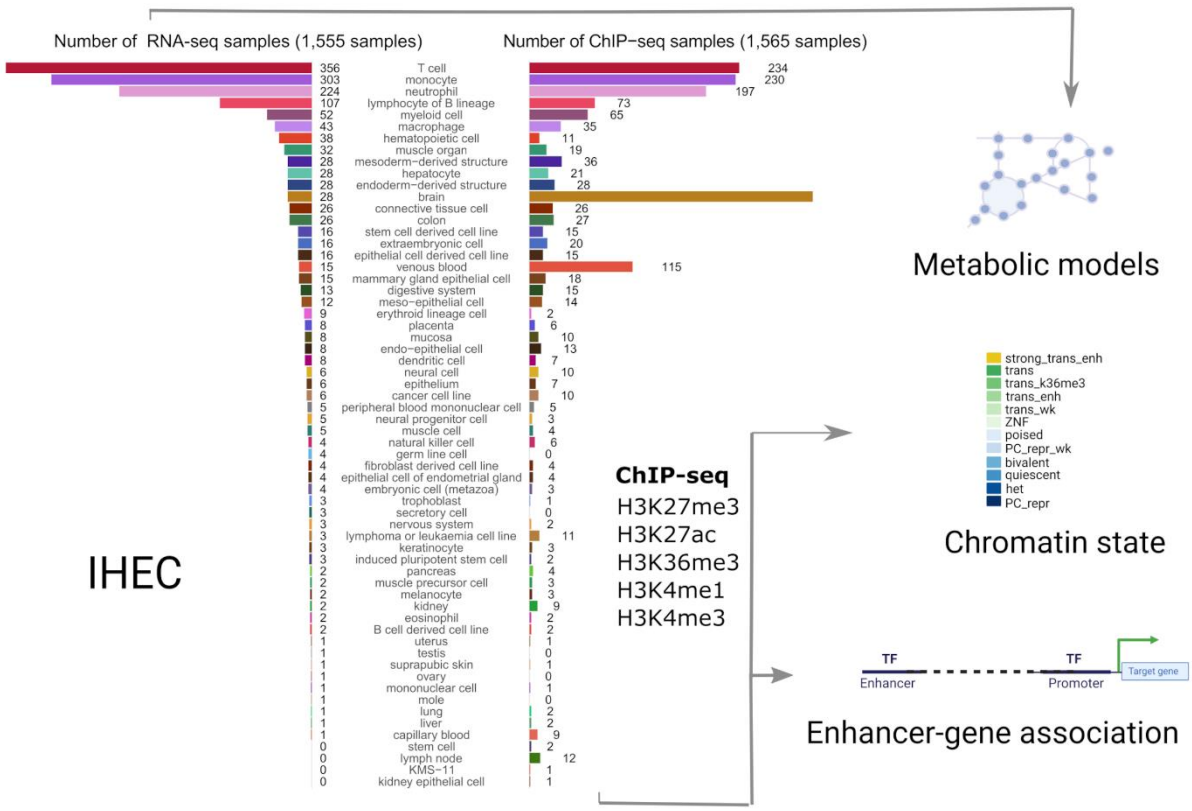


Figure 1: Systematic modelling of epigenetic control of metabolism. The number of integrated RNA-seq and ChIP-seq samples for each tissue and cell type is represented as bar plots. We used RNA-seq profiles (left bar plot) from 58 tissues and cell types to reconstruct metabolic networks with rFASTCORMICS [5]. These models were then used as scaffolds to map enhancer-activity scores (1,565 samples across 56 cell types and tissues, right bar

plot) as well as gene-level ChromGene chromatin state annotations (1,698 samples across 57 cell types and tissues, not shown as samples distribution is similar to the enhancer activity scores) to identify key regulation points. 12 ChromGene states were defined based on the integration of the chromatin marks: strong trans enh (highest expression), trans, and trans wk (transcribed but at different expression levels depending on the presence of H3K27ac H3K4me1, and H3K36m3), ZNF (Zinc Finger-related genes), poised, PC repr wk (weak Polycomb-mediated repression), bivalent (high level of repressive mark H3K27me3, and moderate levels of H3K4me1 and H3K4me3), het (genes associated with heterochromatin), PC repr (Polycomb-mediated repressed genes).

These metabolic reconstructions span 58 tissues and cell types. Blood cells account for the largest fraction of these models, with 356 T cell, 303 monocyte, 224 neutrophil, and 107 lymphocyte B lineage models **(Figure 1, left bar plot).** The median number of model reactions per cell and tissue type ranged from 2,360 (T cell) to 4,252 (liver) **(Figure 2A)**. The twelve model groups with the lowest median number of reactions are models for blood and immune cells, notably T cell, lymphocyte of B lineage, neutrophil, venous blood, and natural killer cell models **(Figure 2A)**. By contrast, hepatocyte, trophoblast, placenta, colon, and kidney models have the highest number of reactions, an indication of their greater metabolic complexity. The substantial size of the liver and hepatocyte models is consistent with the large number of liver-specific metabolic tasks such as drug metabolism, bile acid synthesis, homeostasis of carbohydrates, amino acids, lipids, and fat-soluble vitamins [22,23].

While many basic metabolic processes are shared across tissues and cell types, we hypothesized that individual tissues and cell types (1) fulfil core metabolic processes through tissue-specific reactions and regulatory mechanisms and (2) have tissue-specific metabolic processes for tissue-specific functions (*e.g.*, bile acid synthesis in hepatocytes). In support of these hypotheses, we observed that a t-SNE embedding of samples by metabolic models distinguished cell-type clusters (**Figure 2B**). This result is further supported by greater Jaccard similarity (in reaction composition) within tissue types than between tissue types **(Supplementary Figure S1A**). Notably, hepatocytes and neutrophils define the extremes of the first principal component (**Supplementary Figure S1B**), suggesting a highly distinct metabolism from other cell types. Taken together, these data demonstrate that metabolic models capture the cell type-specific metabolic character of cell types.

We next defined the set of 'core' metabolic reactions as the set of 541 reactions (from 10,600 total reactions) in Recon3D [24] which were found in over 99% of models. These include reactions in pathways required for cellular homeostasis, including reactive oxygen species (ROS) detoxification, fatty acid synthesis, squalene and cholesterol synthesis, purine synthesis, and nucleotide interconversion. Moreover, pathways implicated in the maintenance of the extracellular matrix, such as chondroitin sulphate metabolism, and energy metabolism, were present ubiquitously across all cell types and tissues **(Figure 3A)**.

Next, we computed the number of reactions present in at least 90% of samples in each tissue or cell type. 253 reactions fulfil this criterion include reactions in fatty acid synthesis and oxidation, inositol phosphate metabolism, and glycolysis/gluconeogenesis. Notably, even for pathways commonly shared across all tissues and cell types, some reactions were specifically included only in some models (**Figure 3B and Supplementary Figure S2).** In contrast to these highly common reactions, 4,799 reactions were included in less than 1% of models (3,944 were never included, comprising redundant exchange and transporter reactions, drug metabolism reactions, and reactions not under gene control and thus not supported by the data) **(Supplementary Table S3)**. Some bile acid synthesis, drug metabolism, cytochrome C, steroid metabolism, and fatty acid oxidation pathways were unique to a single tissue or cell type, namely hepatocytes. However, we observed that other reactions in these

pathways were present also in other tissues, suggesting that pathway activation or repression may occur at specific key reactions rather than at the level of the whole pathway.

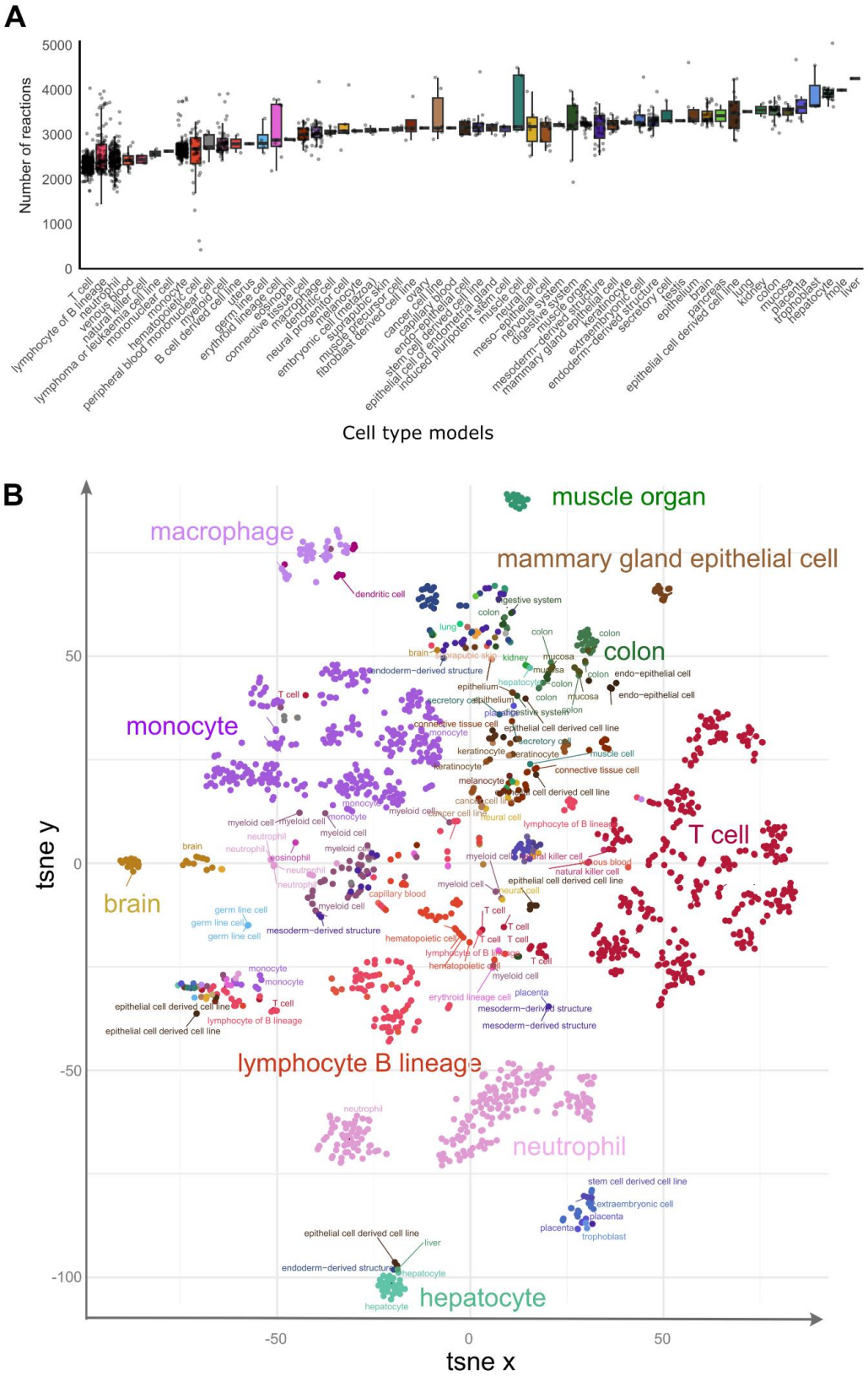


Figure 2: Sample models capture metabolic complexity across tissues and cell types. A) The distribution of the number of model reactions per cell type and tissue is represented as a box plot. The number of reactions in each

model indicates the complexity of the metabolism of cell types and tissues. Error bars represent the upper and lower quartiles. Outliers are data points outside the range defined by Q1 - 1.5 interquartile range (IQR) and Q3 + 1.5 IQR. B) t-SNE embedding of sample models separates tissues and cell types. Sub-clusters are visible for models of some cell types such as T cells, neutrophils, lymphocytes of B lineage, and monocytes.

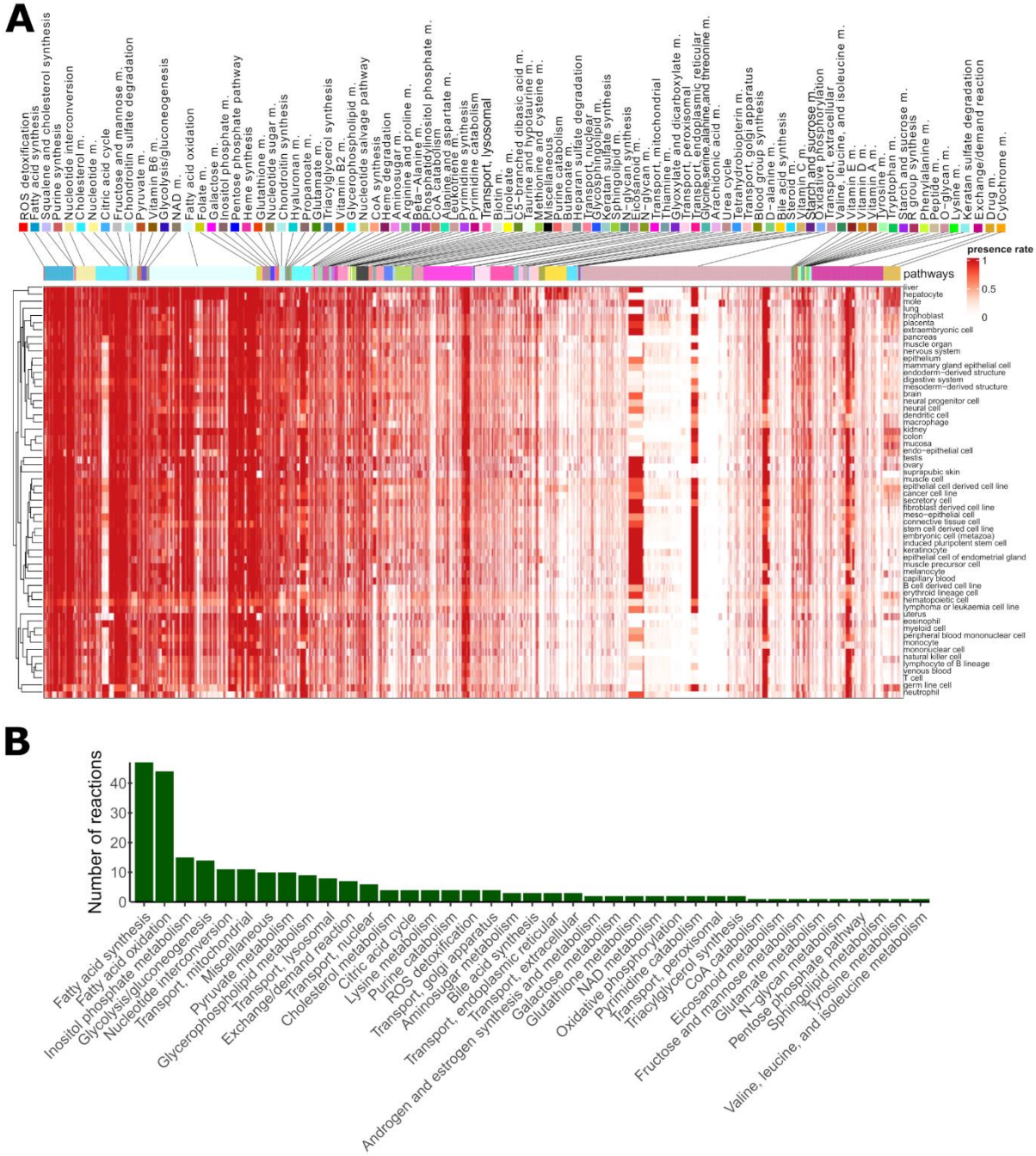


Figure 3: Metabolic models capture core and tissue-specific metabolic reaction pathways. A) Presence rates computed for each reaction and tissue type. Pathways were ranked according to their mean presence rate across models. Columns represent the reactions included in at least one of the models and rows represent the cell and tissue type. Some pathways (*e.g.*, ROS, fatty acid synthesis and squalene pathways) were included in most metabolic models, while other pathways (*e.g.*, drug metabolism, bile acid synthesis) were unique to hepatocytes. Even among the pathways with the highest presence rate, some reactions have a cell type-specific or sample-specific expression pattern. We observe a similar pattern for unique pathways (*e.g.*, bile, drug metabolism)

which contain besides unique reactions a small fraction of ubiquitous reactions. B) The number of reactions per pathway present in at least 90% of the tissue and cell type models. Reactions of central carbon metabolism are among the reactions with the highest presence rate.

Lymphoma or leukaemia cell lines, as well as blood and immune cells, have a more compact metabolism (*i.e.*, smaller number of reactions) **(Figure 3A)**, but also tend to have a lower reaction presence rate in most pathways, as evidenced by the brighter horizontal lines in **Figure 3A.** This finding suggests that, for some samples, these reactions with a lower presence rate can be shut down, possibly due to environmental signals or, in case of cancer cells, due to mutations. Some branches of fatty acid oxidation (FAO) were absent or had low presence rates in most cell types except for liver, placenta, trophoblast, extraembryonic cell, and mole. Finally, the reaction presence rate of metabolic models correlates with the median expression levels associated with the metabolic genes mapped to Recon3D, demonstrating the sensitivity of model building pipeline **(Supplementary Figure S3)**. Taken together, the reconstructed metabolic models reveal core and tissue-specific metabolic characteristics.

Core and unique metabolic genes are associated with higher enhancer activity and expression levels than non-housekeeping genes shared by multiple tissues

While genes controlling core metabolism, by definition, must be expressed ubiquitously, we sought to determine whether these commonly expressed genes are differentially regulated across cell types. We defined a set of core genes (n=3,530) as genes with active ChromGene [20] annotation across all EpiATLAS cell types and tissues (Methods). We also defined a set of core metabolic genes (n=306) as the subset of core genes present in Recon3D. Similarly, we defined unique (n=1,013) and unique metabolic genes (n=121) as the set of genes active in only a single tissue or cell type. The core gene set was enriched for genes implicated in the metabolism of RNA and processing of introns, while core metabolic genes were enriched in mRNA splicing, translation, energy production, biosynthesis pathways, lipid metabolism, and ROS metabolism **(Supplementary Table S3)**. Both core gene sets were strongly enriched for common essential genes obtained within the DepMap project [25] **(Figure 4A)**. 934 out of 1,549 common essential genes and 48 out of 79 common essential metabolic genes were core and core metabolic genes, respectively. Core genes and core metabolic genes together captured 60% of the common essential genes. The missing 31 essential metabolic genes were associated with active ChromGene states (1-6) in most tissues and cell types, with one exception, gamma-glutamyl transferase (*GGTLC2*), which carried a repressive chromatin state across most samples (**Supplementary Figure S4)**. We observed similar patterns for non-metabolic core essential genes: except for a few genes that are associated with a repressive state in most cell types, core essential genes carry activating states (ChromGene state 1-6) in most cell types (**Supplementary Figure S4)**. The enrichment for overlaps between common essential and core genes implicates these genes in key metabolic pathways that are highly conserved across cell types and essential for cell survival.

We next used the generalized Activity-By-Contact (ABC) method [17,26] to examine the enhancers linked to core and core metabolic genes. For each gene, we computed an enhancer activity score by summing the activity of all enhancers predicted to interact with a gene per cell type or tissue (Methods). Consistently, higher median enhancer activity scores were observed for the core and core metabolic genes **(Figure 4B)**. Also, the increase of the median gene expression was more pronounced for the core and core metabolic genes compared to all genes and all metabolic genes **(Figure 4C).** This is consistent with the previous observation that core fitness genes have substantially higher expression than non-fitness genes [27,28]. Also, core metabolic genes are associated, by definition, with activating ChromGene states 1-6 and a higher enhancer activity **(Figure 4D)**. There was a substantial number of

genes with ChromGene state 1 (corresponding to highly expressed genes with strong signals of active marks such as H3K4me1 and H3K27ac) among the genes expressed in hepatocytes and extraembryonic cells, as well as neutrophils, monocytes, lymphocyte B lineage, venous blood, peripheral blood mononuclear cells, and myeloid cells.

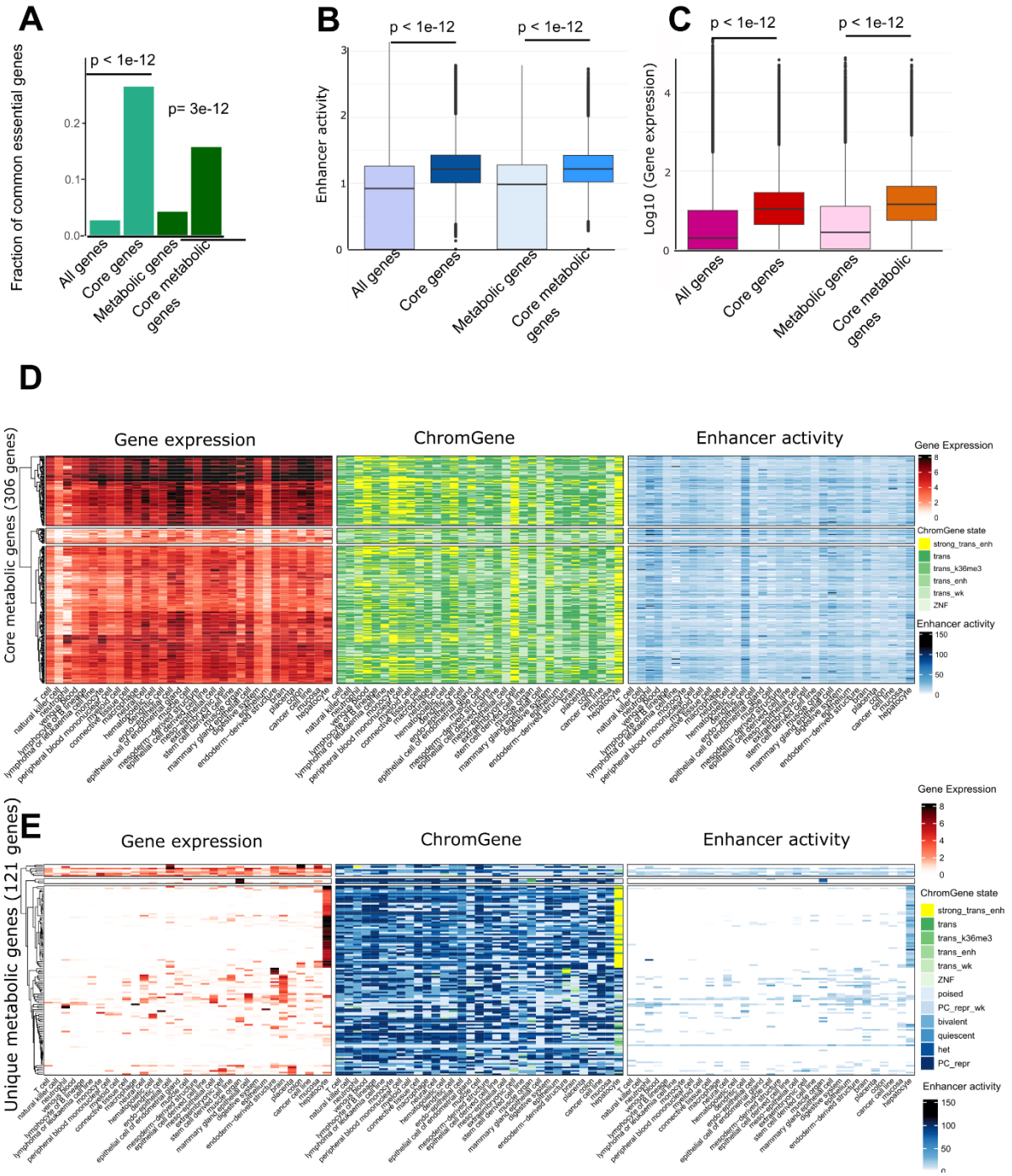


Figure 4: Core and core metabolic genes are essential and exhibit high enhancer activity and high expression. A) Core and core metabolic genes are enriched for essential genes common to most cell types (hypergeometric p-value < 3e-12). **B)** Core and core metabolic genes have higher enhancer activity scores than genes on average (t-test p-value <1e-12) and **C)** higher median expression values (t-test p-value <1e-12). **D)** Core metabolic genes have high expression values ($\log_2$ TPM, left), are linked to active ChromGene states (middle), and have a high enhancer activity across most cell types and tissues (right). Rows contain the 306 core metabolic genes and

columns contain the cell types with at least 3 samples for each data type. **E)** Similar to D but representing 121 unique metabolic genes. Unique genes have a high expression in a single tissue or cell type and have low or no expression in the other tissues. There is also a high number of Polycomb-repressed genes, heterochromatin states, and low enhancer activity scores for these unique genes in other tissues.

For unique genes, we found both higher expression and higher enhancer activity to correspond to an active ChromGene state **(Figure 4E, Supplementary Figure 5)**. Of the 1,013 unique genes, 177, 144, and 109 were specific to hepatocytes, brain, and muscle, respectively **(Supplementary Figure S6A)**. These sets were enriched for gene ontology terms matching their cellular identities (*e.g.*, brain enriched for protein-protein interactions at synapses). Among the 121 unique metabolic genes, as many as 67 were hepatocyte-specific genes and were consistently enriched for tasks and pathways specific to hepatocytes **(Figure 4E, Supplementary Figure S6B, and Table S4-S5)**. The high expression of these unique gene sets matched high enhancer activity scores while low expression and enhancer activity was observed in other tissues. Polycomb-mediated repression was the most common repressive chromatin state at the cell type-specific genes **(Figure 4E)**. However, the number of Polycomb-repressed genes among the 121 unique genes varied between tissues. Muscle cell, lymphoma and leukaemia samples, and cancer cell lines had 62, 60 and 57 Polycomb-repressed unique metabolic genes, respectively. This suggests that besides gene activation in unique cell types, repressive states, and extensive Polycomb-mediated repression, could also contribute to shape metabolic identity.

TFs controlling metabolic identity show cell type-specific activity

Next, we asked whether the high and consistent expression of core genes and core metabolic genes was controlled by the same TFs across tissues, or whether core metabolism is subject to cell type-specific regulation. To this end, we integrated cell type-specific enhancer-gene interactions to obtain sets of regulatory elements linked to each core and core metabolic gene for each cell type **(Figure 5A)**. Many TFs were uniformly enriched at enhancers associated with core genes across different cell types **(Figure 5A).** Conversely, we observed that some TF regulators of core metabolic genes were uniquely enriched in specific cell types and tissues. Overall, 271 non-redundant TF motif modules and tissue-specific regulatory regions were significantly enriched at enhancers associated with core metabolic genes (FDR<0.01), and 40 of these had absolute enrichment scores over 1.2. These TFs control common processes such as growth, proliferation and differentiation (EGR)[29], oncogenic and embryonic development (TFAP2, SPDEF, ZFX, ZIC1, ZBTB7A, PAX, GLIS) [30–40], regulation of chromatin structure (MECP2, HINFP, MBD2, CTCF, KAISO)[41–46], and cell cycle (E2F, RUNX)[47,48]. Moreover, MYB is expressed mainly in hematopoietic tissues, colonic crypts, and neural stem cells. In contrast, RUNX1 is expressed in hematopoietic cells [49], but also plays a role in the innate immune system of non-hematopoietic cells [50]. Nuclear receptors broadly regulate metabolism and inflammation [51]. Motifs of TFs implicated in the maintenance of embryonic development (ZSCAN4) [52], cell development and homeostasis (FOX2), and liver regeneration (ZNF490)[53] were depleted from enhancers associated with core metabolic genes. HINFP, ZNF490, GMEB2 showed the largest variation in enrichment. Furthermore, only 6 TFs were common essential genes: CTCF, NRF1, THAP1, YY1, ZNF143, and ZNF335, consistent with redundant roles among TFs.

Interestingly, we observed an inverse association between TF enrichment values for core metabolic genes and unique metabolic genes **(Figure 5B)**, suggesting distinct upstream regulators of core and tissue-specific metabolic processes. Consistently, some TFs with negative enrichment values in core metabolic genes, such as ZSCAN4 and ZNF24, have high enrichment values for the unique metabolic genes **(Figure 5C)**.

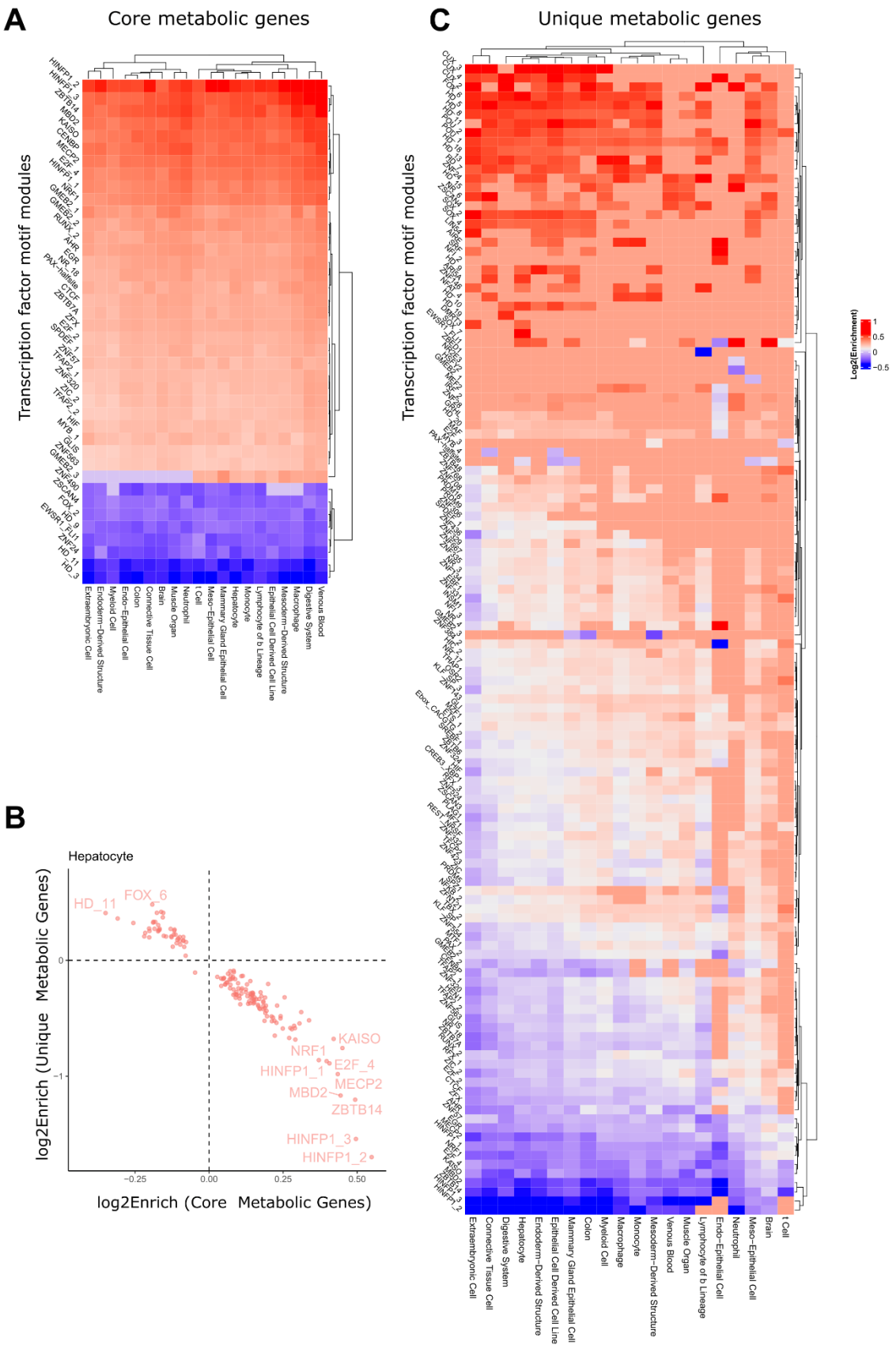


Figure 5: Core metabolic genes and unique metabolic genes are enriched for different TF sets. A) The enhancer-gene interactions were used to obtain a union set of regulatory elements for ten samples randomly chosen from the cell types and tissues with more than 10 samples for the core and core metabolic genes. TFs implicated in

common biological processes such as growth proliferation, and development response to stress or environmental changes were enriched for core metabolic genes. Only TFs with absolute enrichment above 1.2 are depicted. B) Negative correlation between enrichment (log2, FDR<0.01) of TFs for core metabolic genes vs unique metabolic genes. C) Same analysis as A) but for unique genes. Only TFs with an absolute enrichment above 1.3 are depicted.

Other TF groups with high enrichment at enhancers of unique metabolic genes included CUX[54] (implicated in cell cycle control and DNA repair), POU [55](warranting cell identity and differentiation) and SOX (SOX9 plays a role in liver homeostasis and regeneration[56]).

Most of these TFs and motifs have a wide number of target genes and were shown to be dysregulated in numerous cancers [57–59]. A more detailed understanding of the cell type-specific control of core and unique genes could allow targeted interventions in specific cell types such as diseased cells.

Epigenetic control shapes metabolic identity by regulating solute carrier genes in a tissue-dependent manner

Encouraged by our findings that unique genes are marked by Polycomb-mediated repression in other tissues, we next sought to determine if this repression was targeted to genes controlling key nodes in metabolic pathways or broadly across genes in metabolic pathways. In total, 750 metabolic genes were associated with Polycomb-mediated repression in at least one tissue **(Figure 6A, Supplementary Figure S8)**. We observed that leukaemia or lymphoma cell lines (409 metabolic genes), muscle cells (348), natural killer cells (286), and cancer cell lines (282) have the highest number of Polycomb-repressed metabolic genes. Conversely, stem cell-derived cell lines, mammary gland epithelial cell lines, and brain samples have the lowest number (38, 55 and 67). Mammary gland epithelial cells are instead enriched for poised promoters **(Supplementary Figure S9)**, possibly linked to their precise, highly regulated response to hormonal signals. The poised and Polycomb-repressed ChromGene states are mutually exclusive, explaining the lower number of Polycomb-repressed genes in the cells carrying many poised promoters. The 409 Polycomb-repressed genes in leukaemia or lymphoma cell lines are enriched for pathways such as bile secretion, purine metabolism, and steroid hormone biosynthesis, which are mostly active in hepatocytes. Furthermore, in all the tissue types Polycomb-repressed genes are enriched for genes encoding for solute carriers (SLCs) **(Figure 6B)**. The proportion of SLCs among the Polycomb-repressed genes is between 15.8% (1.3-fold enriched) and 25.5% (2.1-fold enriched) for stem-cell derived cell lines and epithelial cells of the endometrial gland, while overall they represent only 12.4% of the metabolic genes **(Figure 6A)**. For lymphoma or leukaemia cell lines, 76 SLCs (out of a total of 229 SLCs, p = 1.e-5) and 8 (ATP)-binding cassette transporters (out of 22, p = 0.08) **(Figure 6A)** are Polycomb-repressed.

To analyse the distribution of Polycomb-mediated repression across metabolic pathways, we compared the binarized ChromGene states and the Polycomb-repression distribution (ChromGene state 12). A core metabolism cluster highlighting mostly the same pathways as in **Figure 3A** is enriched for active chromatin state **(Figure 6C top)** across all tissues and cell types and depleted of any Polycomb-repression **(Figure 6C bottom)**. Several pathways have a presence rate of above 0.6 for ChromGene state 1 in three or more tissues **(Supplementary Table 8**, **Supplementary Figure 10)**. Hepatocytes have a presence rate above 0.6 for 21 pathways indicating once again high metabolic complexity. For Polycomb-mediated repression only a few pathways have a presence rate above 0.3 in more than one cell type (**Supplementary Figure S10**)**.** Leukaemia and lymphoma cell lines had the highest presence rates for Polycomb-repression in butonate metabolism, blood group synthesis, alanine and aspartate metabolism, vitamin D, O- glycan metabolism, glycine, serine, alanine, and threonine metabolism. Unlike what is seen for active chromatin states, Polycomb-repression never

affects every reaction of a pathway, but only some specific branches of pathways (**Supplementary Table S9**).

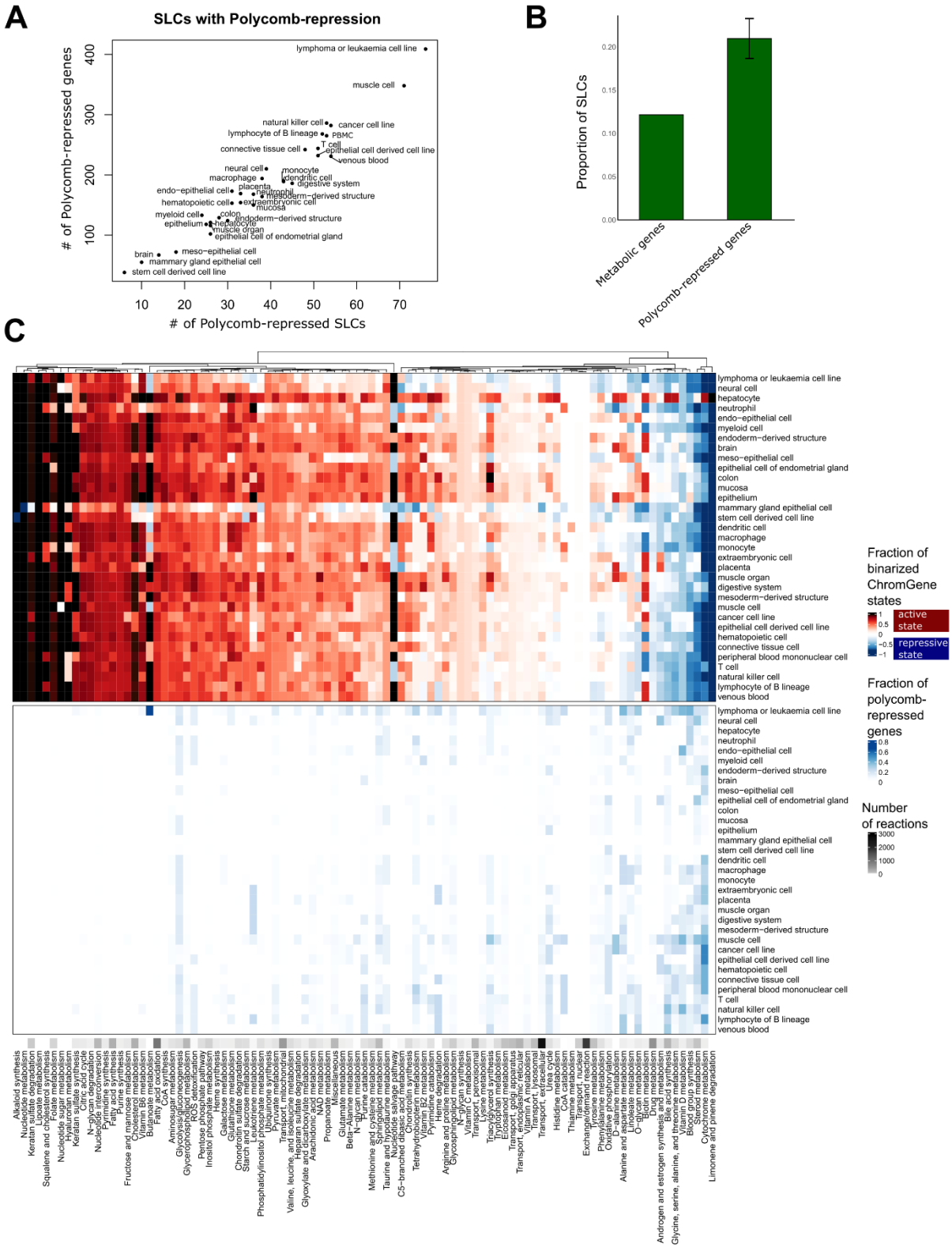


Figure 6: Polycomb repression shapes metabolic networks by repressing solute carriers. A) Number of Polycomb-repressed SLCs over number of Polycomb-repressed genes across cell types. B) Proportion of SLCs in

metabolic genes and Polycomb-repressed genes. C) Presence rate of different ChromGene states for the different metabolic pathways. Top: ChromGene states were binarized for visualization purposes. Active ChromGene states 1-6 were associated with a score of 1, and inactive ChromGene states between 7 and 12 with a score of –1. Gene expression was then mapped using the Gene-Protein-Reaction rules of the metabolic network to obtain a score for each reaction. A mean score for each pathway was computed by summing the individual scores and dividing by the number of reactions (top). Bottom: Same plot but considering ChromGene state 12 (Polycomb-repressed, bottom panel). The black ribbon depicts the number of reactions per pathway.

Transporters and other pathway entry points are key nodes for epigenetic regulation of the metabolic networks

As SLCs are enriched for Polycomb-mediated repression, and hence could serve as switches within the metabolic network, we tested whether SLCs were also enriched among the uniquely expressed metabolic genes **(Figure 7A)**. Within the 121 unique genes, we observed an enrichment for cytochrome P450 (CYP) enzymes (12 out of 48, p = 3e-05), which is consistent with the enrichment for hepatocyte-specific genes. Furthermore, 33 out of 371 genes controlling transport reactions, were among the 121 unique metabolic genes (p = 0.02), (3 ATP binding transporters, 21 SLCs and the transporter genes NPC1L1, SULT2A1, AQP8, PRODH2, RBP2, RBP4, RHCG, ATP1A3, ATP1A4).

To take advantage of the structural information of the metabolic reconstruction, the 121 uniquely active metabolic genes were mapped to the networks to unravel key switches. Among the 5,938 reactions under known gene control in the generic Recon3D reconstruction, 3,744 reactions are associated with transport. Of these, 2,241 are controlled by gene products, while the remaining 1,503 are diffusion reactions.

Among the 269 unique reactions under a selective expression regulation (controlled by one or several of the 121 genes), 151 were transporter reactions controlled by 17 genes. Consequently, an enrichment for unique expression regulation for the transporters was also found at the reaction level (p = 1e-10) **(Figure 7B)**. 97, 28, 6, and 4 out of the 151 transporter reactions are controlled by SLCO1B1, SLC22A10, SLC5A8 and SLC22A2, respectively. Besides, extracellular transporters (151/2652), bile acid synthesis (27/185), drug metabolism (13/573), glycine, serine, alanine, and threonine metabolism (11/47), cytochrome C (8/15) and glycolysis/gluconeogenesis (6/42) have the largest number of reactions controlled by unique genes. These reactions are often importing metabolites into the cell or are the first reaction after a transport reaction to the endoplasmic reticulum or another compartment, as exemplified by the drug metabolism pathway **(Figure 7C)**. Accordingly, drug metabolism is inactivated in other tissues by repressing their import in the cell or repressing enzymes in the endoplasmic reticulum metabolising drugs **(Figure 7C)**. Similarly, we have already shown the importance of regulation of entry points for the bile acid pathway (IHEC flagship paper, under preparation). This pathway has a large fraction of genes and associated reactions with a ChromGene state of 1 (highly expressed in hepatocytes) **(Figure 7D)**. In other tissues, this pathway seems to be mostly inactive due to the Polycomb-repression of genes controlling switches in the network **(Figure 7D)**. Overall, the metabolic network analysis shows an orchestrated epigenetic regulation combining activating and repressing signals to control key nodes in the networks in a cell-type-specific manner.

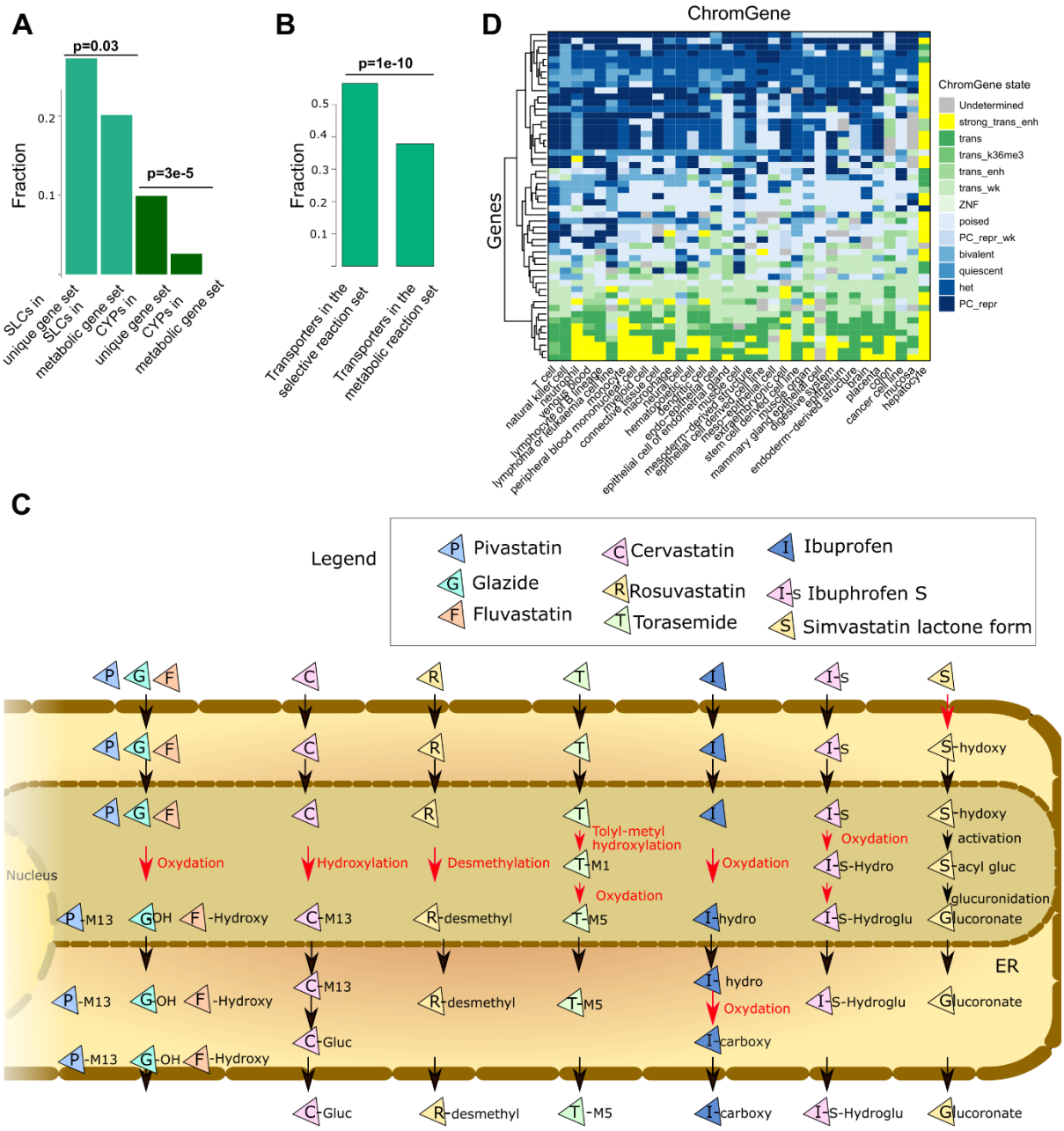


Figure 7: Transporter and CYP genes are enriched among the unique genes and transport reactions are enriched among the unique reactions. A) Among the 121 unique genes, 12 out of 48 genes and 33 out of 271 genes code for CYP proteins and transporter coding genes, respectively (hypergeometric test). B) Among the 256 unique reactions, 152 out of 2,241 are transporter reactions (hypergeometric test). C) Schematic representation of the drug metabolism pathway in hepatocytes. Reactions under unique control are depicted in red. These unique reactions are often located in the endoplasmic reticulum and generally after a transport reaction between two compartments or an import into the cell. D) ChromGene states linked to the bile acid genes are depicted. Some genes implicated in drug metabolism are often uniquely expressed in hepatocytes and control 13 out of 573 reactions. In other tissues, these genes are often Polycomb repressed.

Discussion

In this study, we reconstructed 1,555 metabolic network models capturing core metabolic pathways and processes and the basis of the distinct metabolic identities across cell types. Both core and unique metabolic genes have a higher enhancer activity, and consequently, higher expression values, compared to the remaining metabolic genes, providing robustness to the control of vital processes, consistent with the high number of common essential genes among the core genes. 60% of common essential genes were included among the core genes. We defined core genes based on strict criteria (active ChromGene states for all cell types and tissues), and consequently, many of the common essential genes missing from the core gene set were associated with active chromatin states in most cell types. Gene set enrichment analysis further showed that unique genes (considering both metabolic and non-metabolic unique genes) are sufficient to infer the identity of the tissue and cell types. Unique genes exhibit active ChromGene state in the cell type they are expressed in and strong Polycomb repression in other tissues. Notably, stem cells exhibit fewer Polycomb-mediated repressed genes, suggesting a relationship between differentiation stage and repressive mechanisms. Conversely, we observe more Polycomb-repressed genes in cancer cell lines, consistent with Polycomb dysregulation in cancer [60].

We observed that transport reactions, specifically SLCs, were enriched for unique metabolic genes and often Polycomb-repressed. This is consistent with the strict regulation of transporters, as SLCs control the import of metabolites, and channel them into storage or degradation compartments. While precise transporter regulation is energetically expensive, it enables more robust regulatory systems where regulation is enforced only at key points of the networks. SLCs have diverse expression patterns with potential substrates that range from ions, lipids, vitamins, waste products to drugs [61,62]. Some members of the SLC superfamily are ubiquitously expressed, while others are cell- or even context-specific. Mutations in 20 % of SLC genes are linked to Mendelian [63] diseases. SLCs are also often dysregulated in cancer, as the ability to control non-vital metabolic pathways can facilitate tumor growth. SLCs are often hypermethylated in cancer, and SLC expression has been linked to cancer clinical outcomes [64].

Finally, we showed that core genes are under the regulatory control of TFs linked to essential biological processes such as growth, proliferation, and chromatin structure. However, most of these TFs are not essential, suggesting redundant regulatory systems. Notably, TFs controlling core metabolism were often depleted in the unique gene sets. As an example, binding sites for KAISO were enriched in core metabolism and depleted in the unique gene sets of several tissues. KAISO regulates development in normal tissue and plays a role in tumorigenesis through its impact on cell cycle and methylation. Similarly, binding sites for HINFP, which is critical for late embryogenesis cell cycle control [65], were broadly depleted from unique gene sets despite strong enrichments in the core metabolic set. HINFP is crucial in the control of the cell cycle during late embryogenesis [65].

In summary, our analysis, empowered by the comprehensive EpiATLAS resource, has characterized the metabolic identities across 58 cell types and tissues, delineated core and tissue-unique metabolic components, and, by integrating cell-type-specific gene-level chromatin states and enhancer-gene interactions, identified regulatory elements, upstream transcription factor regulators, and key pathway nodes controlling metabolism. We observed that transporters, specifically SLCs, and first reactions of pathways were enriched for high expression, active chromatin state, and Polycomb-mediated repression. This analysis serves as the foundation for uncovering regulatory sites that govern metabolic identity in human cells, providing a basis for tissue-specific metabolic reprogramming.

Methods

We accessed RNA-seq profiles (1,555 samples across 58 cell types and tissues), ChromGene HMM states (1,698 samples), and enhancer activity data (1565 samples) from the IHEC EpiATLAS resource (https://epigenomesportal.ca/ihec/) [16]. Of these, 965 samples had expression, ChromGene, and enhancer activity data and 938 samples belonged to tissues and cell types with at least 3 samples. RNA-seq and ChromGene data were accessed from the IHEC sftp server in batch mode. The summed enhancer activity on gene-level was taken from the IHEC sftp as well ('gABC_GeneSummedEnhancer_geq2.txt.gz'). This file sums the H3K27ac signal of all enhancers of a gene, based on interactions predicted with the gABC-score[17]. The score of genes without sufficient activity at any of their promoter (≥2 H3K27ac signal ±200 bp around any annotated TSS) was set to zero.

Reconstruction of metabolic models

1,555 sample models were reconstructed for each RNA-seq sample by the rFASTCORMICS [5] using Recon 3D [24] as input reconstruction, as previously described. Briefly, rFASTCORMICS discretizes RNA-seq values into expressed, not expressed, and undetermined expression status. Expressed and not expressed genes are mapped to the generic reconstruction (here Recon3D) to form two sets of reactions: core and inactive reactions. Core reactions are supported by the data and should be included in the model. Inactive reactions are under the control of genes that are not expressed and hence are removed from the network. rFASTCORMICS uses FASTCORE[9] to build a flux-consistent network by adding to the set of core reactions (which are not blocked due to the presence of inactive reactions in the same pathway), a next-to-minimal set of non-core reactions. The models were not medium-constrained.

Jaccard Similarity Index, t-SNE, and Principal Component Analysis

For each pair of models, the Jaccard Similarity Index (JSI) was computed, and the models were then clustered according to the JSI using the Euclidean distance. To verify that the clustering on the models captured metabolic variations, we performed principal component analysis on the metabolic genes retrieved from Recon3D using the *factoextra* package [66].

We generated a t-SNE embedding of the sample models using the *tSNE* package[67] in R. Briefly, we constructed a binary matrix encoding reaction presence across 10,600 Recon 3D reactions in 1,555 samples. We excluded reactions absent from all samples in our dimensionality reduction analysis.

Core, core metabolic, unique, and unique metabolic genes

A consensus ChromGene score is the most represented cell type and tissue gene mark. If the most represented mark accounts for less than 50% of the samples and is an activating mark and the other marks are repressive then a score of zero is assigned. Similarly, if the majority mark accounts for less than 50% of the samples and is repressive and the other marks are activating marks, the samples get a consensus score of zero. We defined a set of core genes as those with active consensus ChromGene states (states below 7) for all 33 cell types in the set of 938 EpiATLAS samples with expression profiles, ChromGene annotations, and enhancer activity data. Conversely, we defined unique genes as those which contained active consensus ChromGene states in only one tissue. We then defined core metabolic and unique metabolic gene sets as the subsets of core and metabolic genes included in Recon3D. The log2 expression values, consensus ChromGene states, as well as the enhancer activity were mapped to the metabolic genes of Recon3D and plotted with the ComplexHeatmap package [68].

Enrichment of common essential genes among core genes and gene set enrichments

To validate the core sets, we computed the enrichment of common essential genes (genes essential across all cell lines in DepMap) downloaded from https://depmap.org/portal/download/all/ (CRISPRInferredCommonEssentials.csv) in the core sets using a hypergeometric test. Gene ontology enrichment terms for all four sets were downloaded from EnrichR [69] (https://maayanlab.cloud/Enrichr/).

Presence rate of ChromGene states

The consensus ChromGene states were mapped to the metabolic genes, and a ChromGene score for each reaction was computed using the Gene-Protein-Reaction rules (GPR rules). GPR rules state which genes control a reaction and if the gene products act as protein complexes (denoted by BOOLEAN AND), or as isozymes (BOOLEAN OR) *e.g.*, Gene 1 AND Gene 2 OR Gene 3. To obtain a single score for each reaction, the maximum ChromGene value was taken for genes linked by a BOOLEAN OR and the minimum for BOOLEAN AND. The ChromGene was binarized to obtain a negative (-1) and a positive score (1) for ChromGene states below or equal to 7 and above 7, respectively. A mean score was then computed for each pathway. A similar approach was used for the genes linked to ChromGene 12 (Polycomb-repression) (Figure 6).

Enrichment of transporter genes and reactions

Hypergeometric tests were performed to assess the enrichment of SLCs and CYPs in unique metabolic genes. Furthermore, genes controlling reactions belonging to the transporters subsystems (pathways) in Recon3D were also assessed for enrichment. Finally, reactions of the transporter reactions were also tested for enrichment among the set of unique reactions (reactions that have an activating consensus ChromGene in a unique cell type and tissue).

Transcription Factor Motif Enrichment Analysis

To identify upstream regulators of shared and tissue-specific metabolic gene regulation, we performed overlap enrichment analysis between transcription factor motif occurrences and metabolic gene-linked regulatory regions. Briefly, we began with the previously described sets for core and unique metabolic genes. For a particular sample, we limited these sets to genes with active ChromGene states in the corresponding tissue type. We used the previously described set of sample-specific enhancer-gene interactions to identify sets of gene-associated regulatory elements. We used the union set of regulatory elements from 10 samples in each of 20 tissue categories in our analysis.

We accessed the human reference genome-wide motif occurrences of 280 non-redundant archetypal transcription factor motif modules [70]. We calculated a tissue-specific binomial-based enrichment between pathway-associated enhancers and motif archetype occurrences [71] using the joint set of all core and unique gene regulatory regions as background regions.

To identify tissue-specific regulators of core and unique metabolic processes, we visualized enrichments across tissue categories. We pruned the resulting heatmaps by retaining motif modules with absolute enrichment of at least 1.2 and 1.3, for modules regulating core genes and unique metabolic genes, respectively.

Availability of data

Transcriptomic and epigenomic data can be obtained from the from the IHEC EpiATLAS resource (https://epigenomesportal.ca/ihec/)[16]. The reconstructed metabolic models can be obtained from https://github.com/sysbiolux/IHEC_metabolism/.

Competing interests

The authors declare no competing interests.

Funding

This research was funded in part by the Luxembourg National Research Fund (FNR), grant reference [PRIDE15/10675146/CANBIO]. D.G. was supported by funding from Fondation du Pélican de Mie et Pierre Hippert-Faber. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number "T32 GM007748". The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. For open access, the authors have applied a Creative Commons Attribution 4.0 International (CC BY 4.0) licence to any Author Accepted Manuscript version arising from this submission.

Acknowledgements

The computational experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg [72], https://hpc.uni.lu). All schematic representations were created with BioRender.com, R, or inkscape.

References

1 Morris SA. The evolving concept of cell identity in the single cell era. *Dev* 2019; **146**. doi:10.1242/dev.169748.

2 Swisa A, Glaser B, Dor Y. Metabolic stress and compromised identity of pancreatic beta cells. *Front Genet* 2017; **8**: 1–11.

3 Rodríguez-Colman MJ, Schewe M, Meerlo M, Stigter E, Gerrits J, Pras-Raves M *et al.* Interplay between metabolic identities in the intestinal crypt supports stem cell function. *Nature* 2017; **543**: 424–427.

4 Bintener T, Pacheco MP, Philippidou D, Margue C, Kishk A, Del Mistro G *et al.* Metabolic modelling-based in silico drug target prediction identifies six novel repurposable drugs for melanoma. *Cell Death Dis* 2023; **14**. doi:10.1038/s41419-023-05955-1.

5 Pacheco MP, Bintener T, Ternes D, Kulms D, Haan S, Letellier E *et al.* Identifying and targeting cancer-specific metabolism with network-based drug target prediction. *EBioMedicine* 2019; **43**: 98–106.

6 Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A *et al.* Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nat Protoc* 2019; : 1.

7 Passi A, Tibocha-Bonilla JD, Kumar M, Tec-Campos D, Zengler K, Zuniga C. Genome-scale metabolic modeling enables in-depth understanding of big data. *Metabolites* 2022; **12**. doi:10.3390/metabo12010014.

8 Moškon M, Režen T. Context-Specific Genome-Scale Metabolic Modelling and Its Application to the Analysis of COVID-19 Metabolic Signatures. *Metabolites* 2023; **13**. doi:10.3390/metabo13010126.

9 Vlassis N, Pacheco MP, Sauter T. Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol* 2014; **10**: e1003424.

10 Li X, Yilmaz LS, Walhout AJM. Compartmentalization of metabolism between cell types in

multicellular organisms: A computational perspective. *Curr Opin Syst Biol* 2022; **29**: 1–13.

11 O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson B. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol* 2013; **9**: 693.

12 Galhardo M, Berninger P, Nguyen TP, Sauter T, Sinkkonen L. Cell type-selective disease-association of genes under high regulatory load. *Nucleic Acids Res* 2015; **43**: 8839–8855.

13 Pacheco MP, John E, Kaoma T, Heinäniemi M, Nicot N, Vallar L *et al.* Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network. *BMC Genomics* 2015; **16**. doi:10.1186/S12864-015-1984-4.

14 Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317–330.

15 Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020; **583**: 699–710.

16 Bujold D, Morais DA de L, Gauthier C, Côté C, Caron M, Kwan T *et al.* The International Human Epigenome Consortium Data Portal. *Cell Syst* 2016; **3**: 496-499.e2.

17 Hecker D, Ardakani FB, Karollus A, Gagneur J, Schulz MH. The adapted Activity-By-Contact model for enhancer–gene assignment and its application to single-cell data. *Bioinformatics* 2023; **39**. doi:10.1093/bioinformatics/btad062.

18 Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010; **28**: 817–825.

19 Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *2013 ACM Conf Bioinformatics, Comput Biol Biomed Informatics, ACM-BCB 2013* 2013; **9**: 813–814.

20 Jaroszewicz A, Ernst J. ChromGene: gene-based modeling of epigenomic data. *Genome Biol* 2023; **24**: 203.

21 Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. Nat. Methods. 2012; **9**: 215–216.

22 Campbell I. 51 ANAESTHESIA AND INTENSIVE CARE MEDICINE 7:2 Liver: metabolic functions Functions of the liver. 2006.

23 Ozougwu JC. Physiology of the liver. 2017.

24 Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Mih N, Gatto F *et al.* Recon3D: A resource enabling a three-dimensional view of gene variation in Human metabolism. *Nat Biotechnol* 2018; **36**: 272–281.

25 Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H *et al.* Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells HHS Public Access Author manuscript. *Nat Genet* 2017; **49**: 1779–1784.

26 Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet. 2019; **51**: 1664–1669.

27 Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;

483: 603–607.

28 Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015; **163**: 1515–1526.

29 Gashler A, Sukhatme VP. Early Growth Response Protein 1 (Egr-1): Prototype of a Zinc-finger Family of Transcription Factors. In: Cohn WE, Moldave K (eds). *Progress in Nucleic Acid Research and Molecular Biology*. Academic Press, 1995, pp 191–224.

30 Brill E, Gobble R, Angeles C, Lagos-Quintana M, Crago A, Laxa B *et al.* ZIC1 overexpression is oncogenic in liposarcoma. *Cancer Res* 2010; **70**: 6891–6901.

31 Harel S, Tu EY, Weisberg S, Esquilin M, Chambers SM, Liu B *et al.* ZFX controls the self-renewal of human embryonic stem cells. *PLoS One* 2012; **7**. doi:10.1371/journal.pone.0042302.

32 Gerard M, Abitbol M, Delezoide A-L, Dufier J-L, Mallet J, Vekemans M. PAX-genes expression during human embryonic development, a preliminary report. *C R Acad Sci III* 1995; **318**: 57–66.

33 Nasimi A, Haddad F, Mirzaei-Damabi N, Rostami B, Hatam M. Another controller system for arterial pressure. AngII-vasopressin neural network of the parvocellular paraventricular nucleus may regulate arterial pressure during hypotension. *Brain Res* 2021; **1769**. doi:10.1016/j.brainres.2021.147618.

34 Merzdorf CS. Emerging roles for zic genes in early development. Dev. Dyn. 2007; **236**: 922–940.

35 Ronima KR, Dey C, Thummer RP. An Insight into the Role of GLIS1 in Embryonic Development, iPSC Generation, and Cancer. Springer International Publishing: Cham, pp 1–17.

36 Shaw T, Barr FG, Üren A. The PAX Genes: Roles in Development, Cancer, and Other Diseases. Cancers (Basel). 2024; **16**. doi:10.3390/cancers16051022.

37 Ye T, Feng J, Wan X, Xie D, Liu J. Double agent: SPDEF gene with both oncogenic and tumor-suppressor functions in breast cancer. *Cancer Manag Res* 2020; **12**: 3891–3902.

38 Yang YL, Zhao LY. AP-2 Family of Transcription Factors: Critical Regulators of Human Development and Cancer. 2021.

39 Song X, Zhu M, Zhang F, Zhang F, Zhang Y, Hu Y *et al.* ZFX Promotes Proliferation and Metastasis of Pancreatic Cancer Cells via the MAPK Pathway. *Cell Physiol Biochem* 2018; **48**: 274–284.

40 Liu XS, Liu Z, Gerarduzzi C, Choi DE, Ganapathy S, Pandolfi PP *et al.* Somatic human ZBTB7A zinc finger mutations promote cancer progression. *Oncogene* 2016; **35**: 3071–3078.

41 Baubec T, Ivánek R, Lienert F, Schübeler D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* 2013; **153**: 480–492.

42 Kaplun D, Starshin A, Sharko F, Ganova K, Filonova G, Zhigalova N *et al.* Kaiso regulates DNA methylation homeostasis. *Int J Mol Sci* 2021; **22**. doi:10.3390/ijms22147587.

43 Marshall AD, Bailey CG, Rasko JEJ. CTCF and BORIS in genome regulation and cancer. Curr. Opin. Genet. Dev. 2014; **24**: 8–15.

44 Mitra P, Xie R-L, Medina R, Hovhannisyan H, Zaidi SK, Wei Y *et al.* Identification of HiNF-P, a Key Activator of Cell Cycle-Controlled Histone H4 Genes at the Onset of S Phase. *Mol Cell Biol*

2003; **23**: 8110–8123.

45 Ong CT, Corces VG. CTCF: An architectural protein bridging genome topology and function. Nat. Rev. Genet. 2014; **15**: 234–246.

46 Vuu YM, Roberts CT, Rastegar M. MeCP2 Is an Epigenetic Factor That Links DNA Methylation with Brain Metabolism. Int. J. Mol. Sci. 2023; **24**. doi:10.3390/ijms24044218.

47 Friedman AD. Cell cycle and developmental control of hematopoiesis by Runx1. J. Cell. Physiol. 2009; **219**: 520–524.

48 Johnson DG and S-BR. Role of E2F in cell cycle control and cancer. *Front Biosci* 1998; **3**: d447-8.

49 De Bruijn M, Dzierzak E. Review Series TRANSCRIPTION FACTORS IN HEMATOPOIESIS AND HEMATOLOGIC DISEASE Runx transcription factors in the development and function of the definitive hematopoietic system. 2017. doi:10.1182/blood.

50 Voon DCC, Hor YT, Ito Y. The RUNX complex: Reaching beyond haematopoiesis into immunity. *Immunology* 2015; **146**: 523–536.

51 Zhao L, Hu H, Gustafsson JÅ, Zhou S. Nuclear Receptors in Cancer Inflammation and Immunity. *Trends Immunol* 2020; **41**: 172–185.

52 Zalzman M, Falco G, Sharova L V, Nishiyama A, Lee S, Stagg CA *et al.* HHS Public Access. 2010; **464**: 858–863.

53 Nygård IE, Mortensen KE, Hedegaard J, Conley LN, Kalstad T, Bendixen C *et al.* The genetic regulation of the terminating phase of liver regeneration. *Comp Hepatol* 2012; **11**: 1–15.

54 Nepveu A. Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development. *Gene* 2001; **270**: 1–15.

55 Malik V, Zimmer D, Jauch R. Diversity among POU transcription factors in chromatin recognition and cell fate reprogramming. *Cell Mol Life Sci* 2018; **75**: 1587–1612.

56 Shang T, Jiang T, Cui X, Pan Y, Feng X, Dong L *et al.* Diverse functions of SOX9 in liver development and homeostasis and hepatobiliary diseases. *Genes Dis* 2024; **11**: 100996.

57 Detilleux D, Spill YG, Balaramane D, Weber M, Bardet AF. Pan-cancer predictions of transcription factors mediating aberrant DNA methylation. *Epigenetics and Chromatin* 2022; **15**. doi:10.1186/s13072-022-00443-w.

58 Laissue P. The forkhead-box family of transcription factors: key molecular players in colorectal cancer pathogenesis. *Mol Cancer* 2019; **18**: 5.

59 Wang J, Liu Q, Shyr Y. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. *BMC Genomics* 2015; **16**. doi:10.1186/1471-2164-16-S7-S5.

60 Piunti A, Shilatifard A. The roles of Polycomb repressive complexes in mammalian development and cancer. Nat. Rev. Mol. Cell Biol. 2021; **22**: 326–345.

61 Dvorak V, Superti-Furga G. Structural and functional annotation of solute carrier transporters: implication for drug discovery. Expert Opin. Drug Discov. 2023; **18**: 1099–1115.

62 Hediger MA, Clémençon B, Burrier RE, Bruford EA. The ABCs of membrane transporters in health and disease (SLC series): Introduction. Mol. Aspects Med. 2013; **34**: 95–107.

63 Lin L, Yee SW, Kim RB, Giacomini KM. SLC transporters as therapeutic targets: Emerging opportunities. Nat. Rev. Drug Discov. 2015; **14**: 543–560.

64 Lavoro A, Falzone L, Tomasello B, Conti GN, Libra M, Candido S. In silico analysis of the solute carrier (SLC) family in cancer indicates a link among DNA methylation, metabolic adaptation, drug response, and immune reactivity. *Front Pharmacol* 2023; **14**. doi:10.3389/fphar.2023.1191262.

65 Xie R, Medina R, Zhang Y, Hussain S, Colby J, Ghule P *et al.* The histone gene activator HINFP is a nonredundant cyclin E/CDK2 effector during early embryonic cell cycles. 2009.

66 Kassambara A, Mundt F. Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]. 2020.

67 Donaldson J. tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE). 2022.

68 Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016; **32**: 2847–2849.

69 Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**: W90–W97.

70 Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D *et al.* Global reference mapping of human transcription factor footprints. *Nature* 2020; **583**: 729–736.

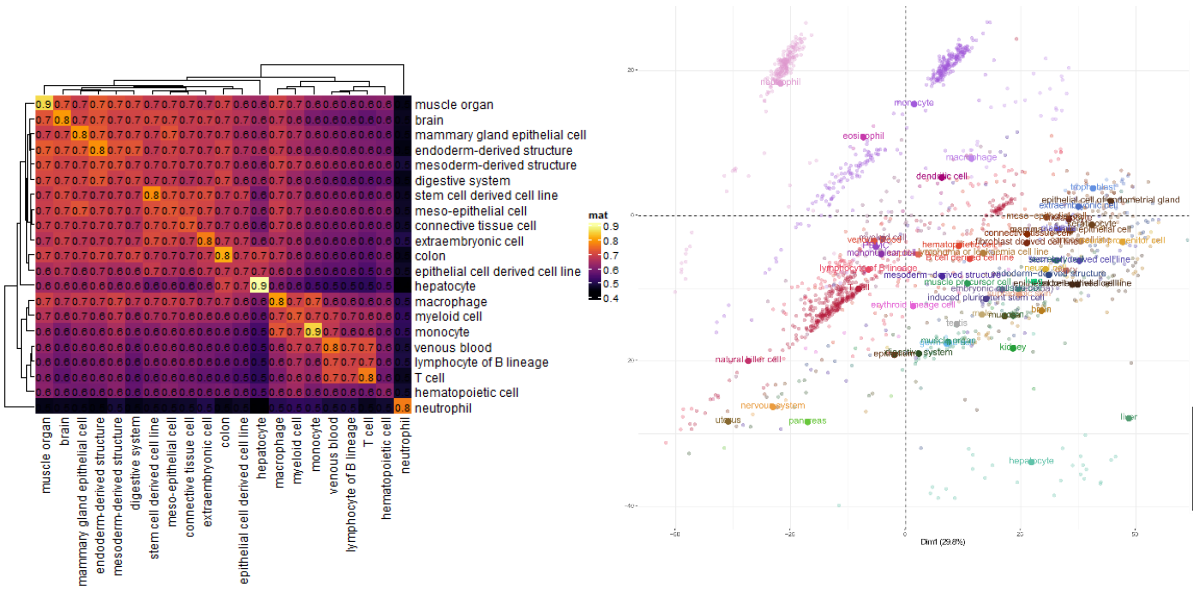
71 Au EH, Fauci C, Luo Y, Mangan RJ, Snellings DA, Shoben CR *et al.* Gonomics: uniting high performance and readability for genomics with Go. *Bioinformatics* 2023; **39**. doi:10.1093/bioinformatics/btad516.

72 Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an academic HPC cluster: The UL experience. In: *2014 International Conference on High Performance Computing & Simulation (HPCS)*. 2014, pp 959–967.

Epigenetic control of metabolic identity across cell types. Supplementary Data

Supplementary Table S1: The ChromGene state integrates active and repressive marks to obtain a gene-specific state instead of a locus state. 1-6 are active states, 7 corresponds to poised and 8-12 are repressive states.

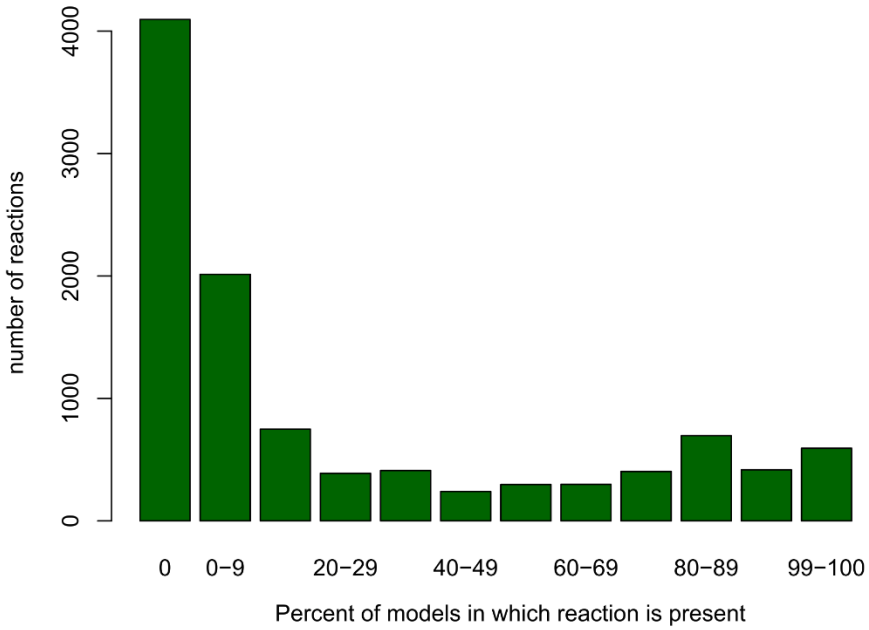
Nb	*Name*	*Description*
1	strong trans enh	Strongly transcribed genes with enhancer marks H3K4me1 and H3K27ac, highest expressed
2	trans	Transcribed genes, highly expressed. Lower levels of H3K36me3 and H3K4me1 than strong trans enh
3	trans k36me3	Transcribed genes, moderately expressed, marked with H3K36me3. Relatively low levels of enhancer marks
4	trans enh	Transcribed genes, moderately expressed, marked with enhancer marks H3K4me1 and to a lesser extent H3K27ac. Relatively low levels of H3K36me3.
5	trans wk	Transcribed genes, lowly expressed, marked with enhancer mark H3K27ac and to a lower extent, H3K4me3. Relatively low levels of H3K36me3
6	ZNF	Zinc Finger-related genes, with a state marked by H3K36me3 and H3K9me3
7	poised	Lowly-expressed genes with high levels of H3K4me1, but low levels of H3K36me3.
8	PC repr wk	Lowly expressed genes with high levels of H3K4me1, weakly marked with polycomb repressive mark H3K27me3.
9	bivalent	Lowly expressed genes, marked with high level of repressive mark H3K27me3, and moderate levels of H3K4me1 and H3K4me3
10	quiescent	Lowly expressed genes, lack of significant histone marks
11	het	Generally unexpressed genes associated with heterochromatin.
12	PC repr	Generally, unexpressed genes, marked with polycomb repressive mark H3K27me3.



Supplementary Figure S1: Metabolic models capture the metabolic identity of tissues and cell types.

A) Hierarchical clustering of metabolic models by median Jaccard Similarity Indexes. A Jaccard Similarity Index (JSI) was computed for each pair of models. Then a median JSI was computed for each combination of cell types and tissues. The similarity within cell types and tissue tended to be larger

than between different cell types and tissues. B) Principal component analysis of the metabolic gene expression data. Neutrophils are separated from the other blood cell samples. Non-blood cells are separated from the blood cells along PC1. The hepatocyte and liver samples like neutrophils also form a distinct cluster but at the other extremity of PC1 and PC2.

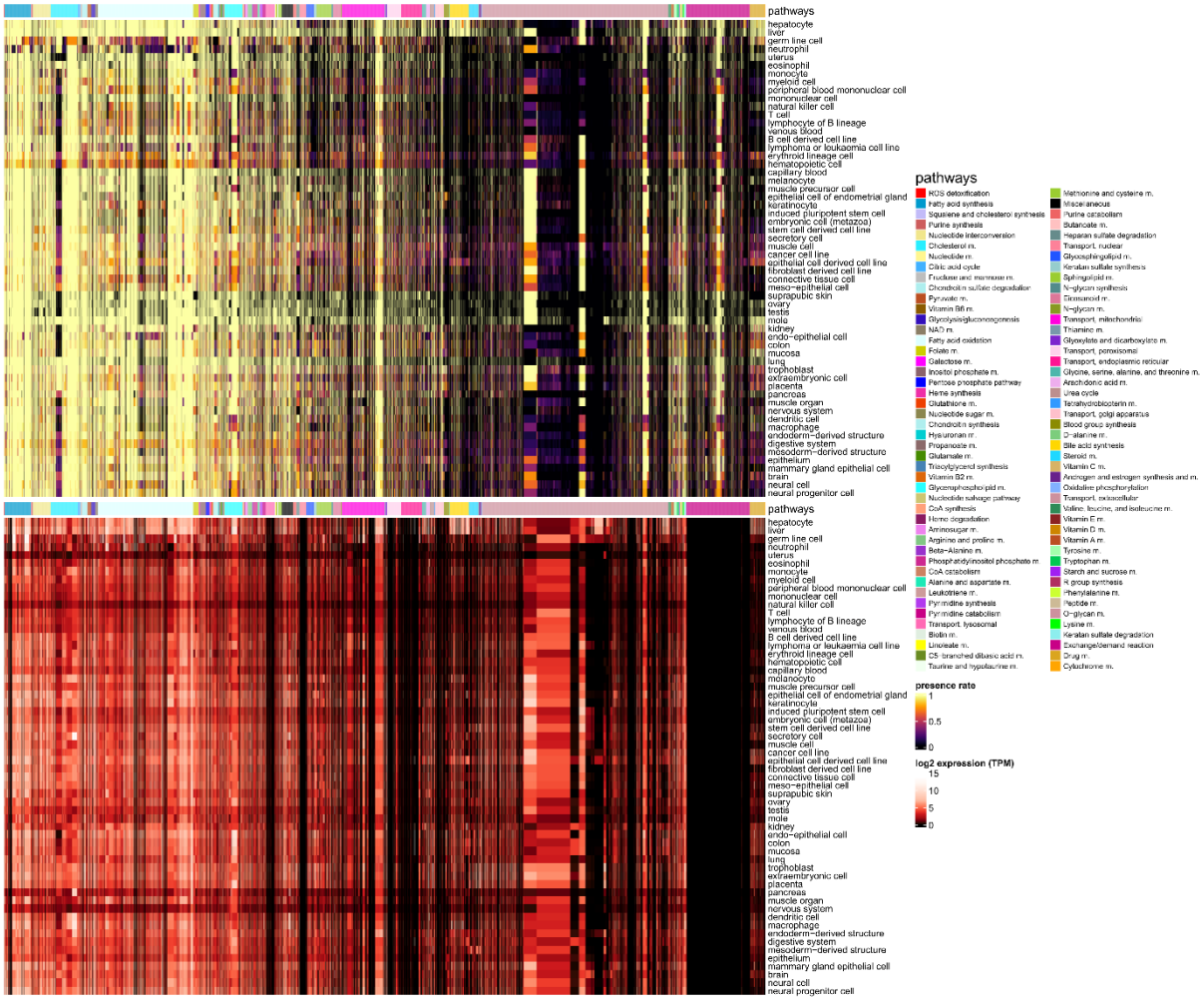


Supplementary Figure S2: Distribution of the number of models in which each reaction is present across the 1,555 metabolic models.

Supplementary Table S2: Number of reactions per pathway that were not included in any model.

Pathways	***Number***	***Pathways***	***Number***	***Pathways***	***Number***
Exchange/demand reaction	1231	*Dietary fiber binding*	12	*Glycolysis/gluconeogenesis*	4
Transport, extracellular	1022	*Methionine and cysteine metabolism*	12	*N-glycan synthesis*	4
Drug metabolism	442	*N-glycan degradation*	12	*Urea cycle*	4
Peptide metabolism	241	*Folate metabolism*	11	*Cholesterol metabolism*	3
Fatty acid oxidation	129	*Nucleotide interconversion*	11	*Glycosphingolipid metabolism*	3
Transport, mitochondrial	74	*O-glycan metabolism*	11	*Limonene and pinene degradation*	3
Tyrosine metabolism	66	*Transport, nuclear*	11	*Oxidative phosphorylation*	3

Miscellaneous	57	*Glycine, serine, alanine, and threonine metabolism*	10	*Pyruvate metabolism*	3
Keratan sulfate degradation	55	*Pentose phosphate pathway*	9	*Vitamin E metabolism*	3
Transport, peroxisomal	34	*Tetrahydrobiopterin metabolism*	9	*Arginine and proline metabolism*	2
Starch and sucrose metabolism	33	*Aminosugar metabolism*	8	*Hippurate metabolism*	2
Transport, endoplasmic reticular	29	*Vitamin D metabolism*	8	*N-glycan metabolism*	2
Transport, lysosomal	28	*Androgen and estrogen synthesis and metabolism*	7	*Protein formation*	2
Tryptophan metabolism	28	*Glycerophospholipid metabolism*	7	*Xenobiotics metabolism*	2
Vitamin A metabolism	28	*Pyrimidine catabolism*	7	*Alanine and aspartate metabolism*	1
Phenylalanine metabolism	23	*Arachidonic acid metabolism*	6	*Alkaloid synthesis*	1
Valine, leucine, and isoleucine metabolism	22	*Galactose metabolism*	6	*Biotin metabolism*	1
Eicosanoid metabolism	21	*Sphingolipid metabolism*	6	*Blood group synthesis*	1
Transport, golgi apparatus	20	*Steroid metabolism*	6	*CoA catabolism*	1
Bile acid synthesis	18	*Fructose and mannose metabolism*	5	*CoA synthesis*	1
Lysine metabolism	18	*NAD metabolism*	5	*Heme synthesis*	1
Ubiquinone synthesis	18	*Purine catabolism*	5	*Lipoate metabolism*	1
Fatty acid synthesis	14	*R group synthesis*	5	*Phosphatidylinositol phosphate metabolism*	1
Histidine metabolism	14	*Vitamin B6 metabolism*	5	*Propanoate metabolism*	1
Inositol phosphate metabolism	13	*Vitamin C metabolism*	5	*Vitamin B12 metabolism*	1
Cytochrome metabolism	12	*C5-branched dibasic acid metabolism*	4		



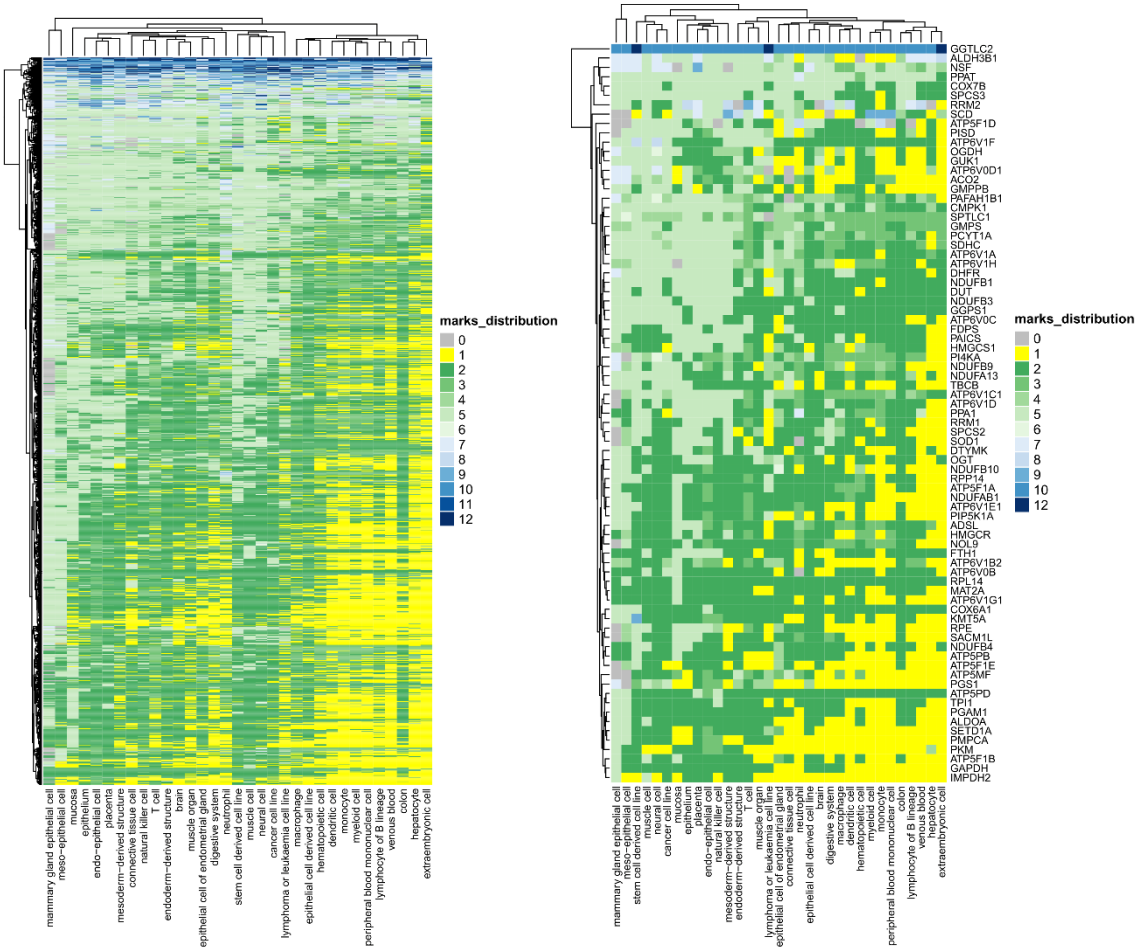
Supplementary Figure S3: The presence rate of the reactions across the 1,555 models corresponds to the gene expression mapped to Recon3D via the Gene Protein Reaction rules.

Supplementary Table S3: Core metabolic genes are enriched for genes implicated in splicing and mRNA processing and energy production, nucleotide, lipid metabolism and ROS metabolism.

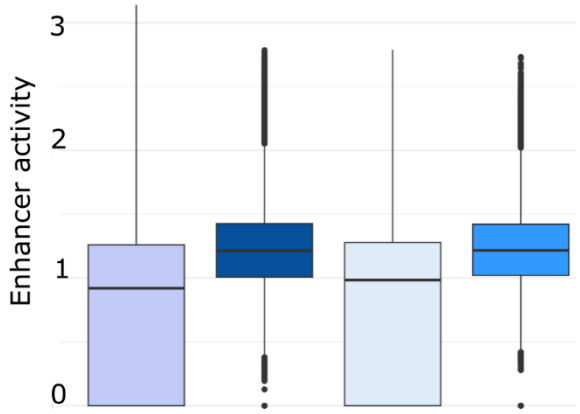
GENE ONTOLOGY	OVERLAP	P-VALUE
CORE GENES		
MRNA SPLICING VIA SPLICEOSOME (GO:0000398)	148/211	3.16E-60
MRNA PROCESSING (GO:0006397)	145/214	3.71E-56
RNA SPLICING VIA TRANSESTERIFICATION REACTIONS WITH BULGED ADENOSINE AS NUCLEOPHILE (GO:0000377)	130/180	1.77E-55
TRANSLATION (GO:0006412)	142/234	2.11E-46
GENE EXPRESSION (GO:0010467)	164/296	2.11E-46
RIBOSOME BIOGENESIS (GO:0042254)	106/155	1.66E-41
RNA PROCESSING (GO:0006396)	113/183	8.30E-38
CYTOPLASMIC TRANSLATION (GO:0002181)	74/93	3.65E-36
RIBONUCLEOPROTEIN COMPLEX BIOGENESIS (GO:0022613)	81/118	1.06E-31

MACROMOLECULE BIOSYNTHETIC PROCESS (GO:0009059)	104/183	2.35E-30
RNA SPLICING (GO:0008380)	71/98	4.03E-30
PEPTIDE BIOSYNTHETIC PROCESS (GO:0043043)	93/158	9.88E-29
PROTEIN-RNA COMPLEX ASSEMBLY (GO:0022618)	89/150	7.60E-28
RIBOSOMAL SMALL SUBUNIT BIOGENESIS (GO:0042274)	61/84	6.70E-26
RRNA PROCESSING (GO:0006364)	68/101	7.91E-26
PROTEASOME-MEDIATED UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PRO.	139/319	4.21E-25
NCRNA PROCESSING (GO:0034470)	63/100	1.63E-21
RRNA METABOLIC PROCESS (GO:0016072)	59/91	4.69E-21
METABOLIC CORE GENES		
METABOLISM R-HSA-1430728	210/2049	6.61E-133
CITRIC ACID (TCA) CYCLE AND RESPIRATORY ELECTRON TRANSPORT	64/163	7.18E-72
RESPIRATORY ELECTRON TRANSPORT, ATP SYNTHESIS BY CHEMIOSMOTIC COUPLING, HEAT PRODUCTION BY UNCOUPLING PROTEINS R-HSA-163200	53/112	1.35E-64
RESPIRATORY ELECTRON TRANSPORT R-HSA-611105	44/90	3.55E-54
COMPLEX I BIOGENESIS R-HSA-6799198	27/51	4.69E-34
METABOLISM OF NUCLEOTIDES R-HSA-15869	24/95	6.30E-21
METABOLISM OF LIPIDS R-HSA-556833	51/732	4.91E-18
CHOLESTEROL BIOSYNTHESIS R-HSA-191273	11/26	3.32E-12
INTERCONVERSION OF NUCLEOTIDE DI- AND TRIPHOSPHATES R-HSA-499943	11/28	7.98E-12
METABOLISM OF CARBOHYDRATES R-HSA-71387	26/285	1.52E-11
FORMATION OF ATP BY CHEMIOSMOTIC COUPLING R-HSA-163210	9/16	1.81E-11
PURINE RIBONUCLEOSIDE MONOPHOSPHATE BIOSYNTHESIS R-HSA-73817	7/10	8.26E-10
PYRUVATE METABOLISM AND CITRIC ACID (TCA) CYCLE R-HSA-71406	12/54	9.25E-10
NUCLEOTIDE BIOSYNTHESIS R-HSA-8956320	7/13	9.74E-09
CRISTAE FORMATION R-HSA-8949613	9/29	9.78E-09
ACTIVATION OF GENE EXPRESSION BY SREBF (SREBP) R-HSA-2426168	10/42	1.66E-08
METABOLISM OF VITAMINS AND COFACTORS R-HSA-196854	18/186	1.66E-08
FATTY ACID METABOLISM R-HSA-8978868	17/173	3.74E-08
PKMTS METHYLATE HISTONE LYSINES R-HSA-3214841	10/47	4.70E-08
METABOLISM OF WATER-SOLUBLE VITAMINS AND COFACTORS R-HSA-196849	14/122	1.30E-07
REGULATION OF CHOLESTEROL BIOSYNTHESIS BY SREBP (SREBF) R-HSA-1655829	10/55	2.16E-07
GLUCONEOGENESIS R-HSA-70263	8/32	4.47E-07
BIOSYNTHESIS OF N-GLYCAN PRECURSOR (DOLICHOL LLO) AND TRANSFER TO PROTEIN R-HSA-446193	11/77	4.99E-07
HEME BIOSYNTHESIS R-HSA-189451	6/14	6.47E-07
SYNTHESIS OF SUBSTRATES IN N-GLYCAN BIOSYTHESIS R-HSA-446219	10/63	7.11E-07
TP53 REGULATES METABOLIC GENES R-HSA-5628897	11/81	7.59E-07
BIOLOGICAL OXIDATIONS R-HSA-211859	17/218	8.13E-07
PHASE II - CONJUGATION OF COMPOUNDS R-HSA-156580	12/107	1.51E-06
MITOCHONDRIAL BIOGENESIS R-HSA-1592230	11/89	1.84E-06
METABOLISM OF PORPHYRINS R-HSA-189445	7/28	2.57E-06
CYTOPROTECTION BY HMOX1 R-HSA-9707564	9/57	2.92E-06
IRON UPTAKE AND TRANSPORT R-HSA-917937	9/59	3.85E-06

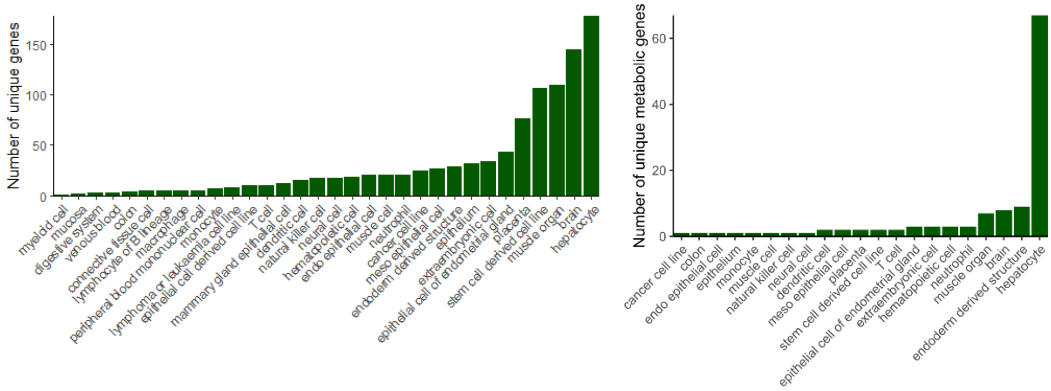
PYRUVATE METABOLISM R-HSA-70268	7/31	5.00E-06
METABOLISM OF STEROIDS R-HSA-8957322	13/153	9.50E-06
MITOCHONDRIAL FATTY ACID BETA-OXIDATION R-HSA-77289	7/36	1.40E-05
INSULIN RECEPTOR RECYCLING R-HSA-77387	6/26	2.83E-05
SYNTHESIS OF UDP-N-ACETYL-GLUCOSAMINE R-HSA-446210	4/8	4.54E-05
VITAMIN B5 (PANTOTHENATE) METABOLISM R-HSA-199220	5/17	5.33E-05
CELLULAR RESPONSE TO CHEMICAL STRESS R-HSA-9711123	13/182	5.80E-05
GLYCOSPHINGOLIPID METABOLISM R-HSA-1660662	7/45	5.92E-05
TRANSFERRIN ENDOCYTOSIS AND RECYCLING R-HSA-917977	6/31	7.45E-05
ASPARAGINE N-LINKED GLYCOSYLATION R-HSA-446203	16/282	8.72E-05
SPHINGOLIPID METABOLISM R-HSA-428157	9/89	9.47E-05
GLUCOSE METABOLISM R-HSA-70326	9/89	9.47E-05



Supplementary Figure S4: Metabolic gene ChromGene States across cell types. 60% of common (left) and metabolic essential genes (right) were in inactive ChromGene states (above seven, in blue) across 33 cell types. The remaining genes were tagged with active ChromGene state (1-6, yellow and green) for most cell types. However, several genes, such as gamma-glutamyl transferase (GGTLC2) were associated with repressive marks in almost all cell types.



Supplementary Figure S5: Unique genes and unique metabolic genes enhancer activities.



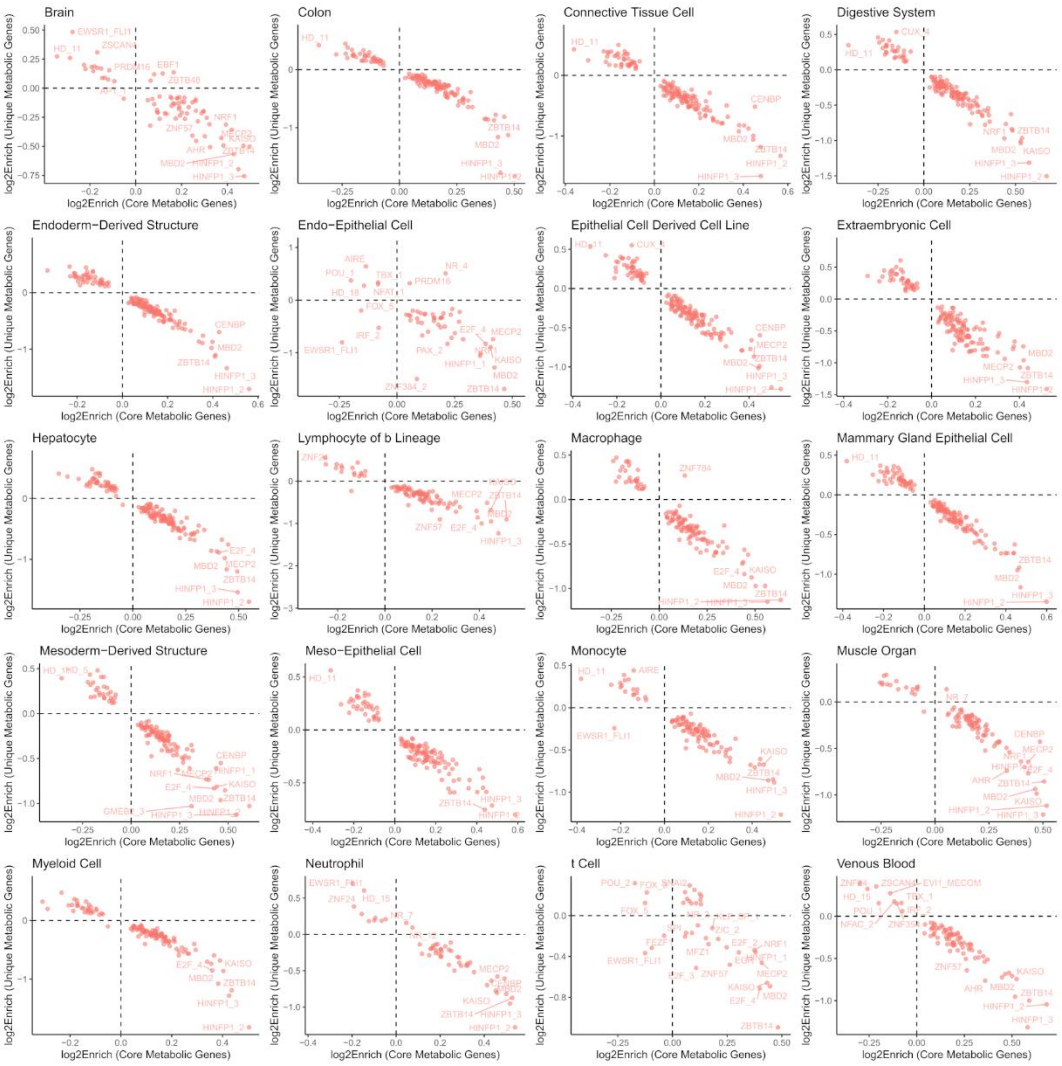
Supplementary Figure S6: Repartition of the unique (left) and unique metabolic genes across tissue and cell types (right)

Supplementary Table S4: <u>Unique genes</u> for hepatocyte, brain, and muscle organs are enriched for organ-matched gene ontology terms. Unique genes were defined by having an active consensus ChromGene state (states below 7) in one single tissue or cell type. No significant enrichments were found for colon and connective tissue cell line due to a limited number of genes (4, 5 genes). Similarly, no enrichments were found for cancer lines due to the substantial tissue heterogeneity of this category.

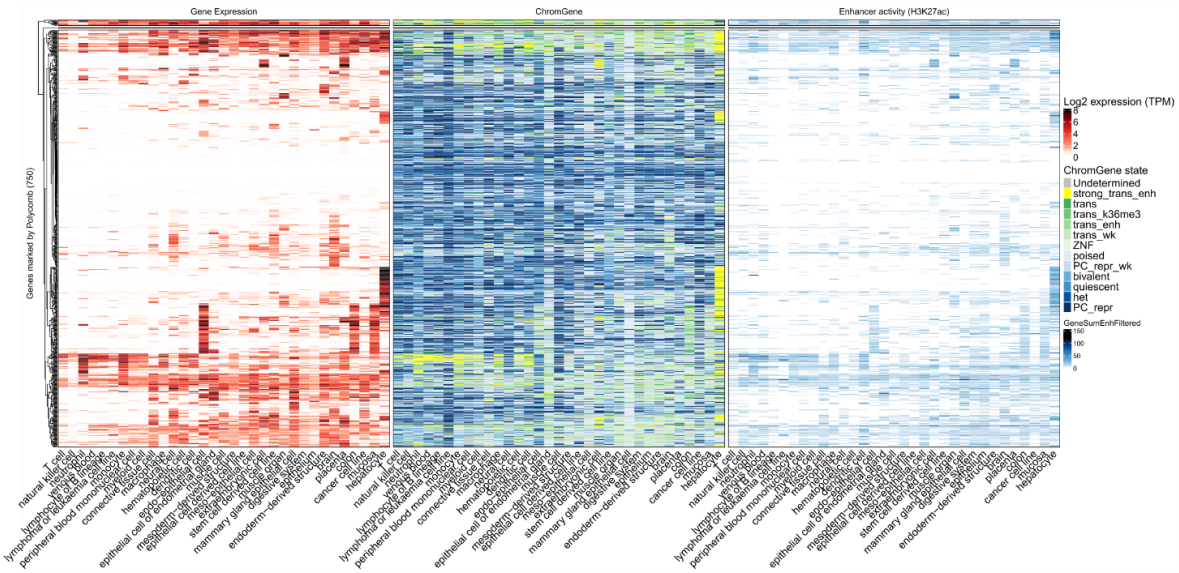
Tissue	unique genes	*Database*	Term	Overlap	Adj. p-value
Hepatocyte	177	KEGG	Steroid hormone biosynthesis	12/61	2.03E-11
			Complement and coagulation cascades	13/85	2.97E-11
			Drug metabolism	13/108	4.64E-10
			Metabolism of xenobiotics by cytochrome P450	11/76	1.80E-09
			Bile secretion	10/90	1.54E-07
			Retinol metabolism	9/68	1.56E-07
			Chemical carcinogenesis	13/239	3.41E-06
			Linoleic acid metabolism	5/29	7.17E-05

			Arachidonic acid metabolism	6/61	2.03E-04
			Maturity onset diabetes of the young	4/26	8.01E-04
			Cholesterol metabolism	5/50	8.01E-04
Brain	144	Reactome	Receptor-type Tyrosine-Protein Phosphatases	5/20	4.87E-05
			Protein-protein Interactions at Synapses	7/88	3.21E-04
			Neuronal System	12/386	1.50E-03
Muscle Organ	109	Reactome	Muscle Contraction	10/196	2.97E-05
			Striated Muscle Contraction	5/33	1.15E-04
			Myogenesis	4/29	0.0014
Placenta	76	Descartes	Trophoblast giant cells in Placenta,	17/89	4.20E-23
			CSH1 CSH2 positive cells in Adrenal	16/272	1.49E-13
Neutrophils	21	Reactome	Neutrophil Degranulation	9/468	1.43E-08
		Reactome	Innate Immune System R-HSA-	9/1035	6.88E-06
		Reactome		10/1943	1.01E-04
Dendritic cell	15	PanglaoDB_Augmented	Dendritic Cells	2/199	0.06

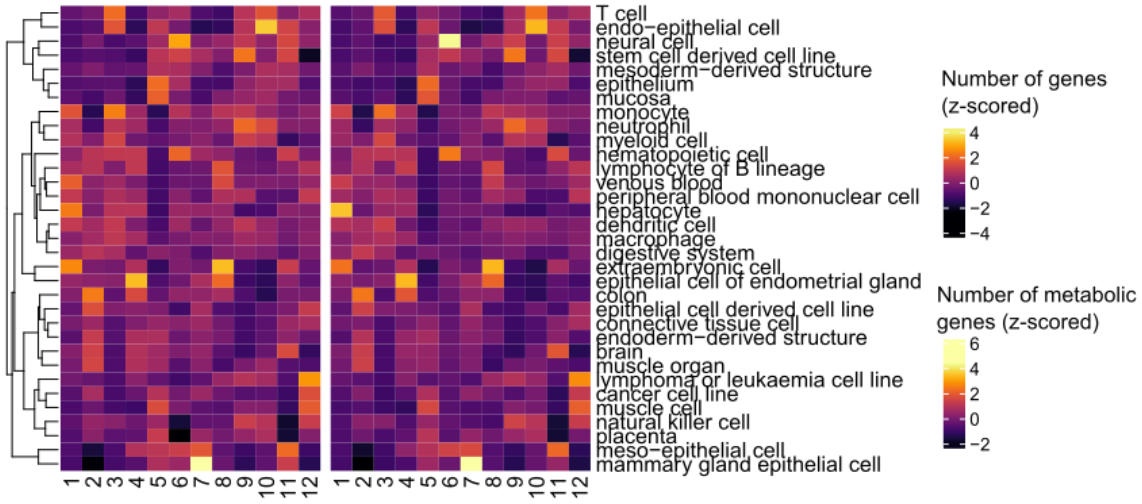
Supplementary Table S5: Unique metabolic genes for hepatocyte, brain, and muscle are enriched for organ-matched gene ontology terms. Unique genes specific to one tissue were defined by having a consensus ChromGene below 7 in one single tissue or cell type.



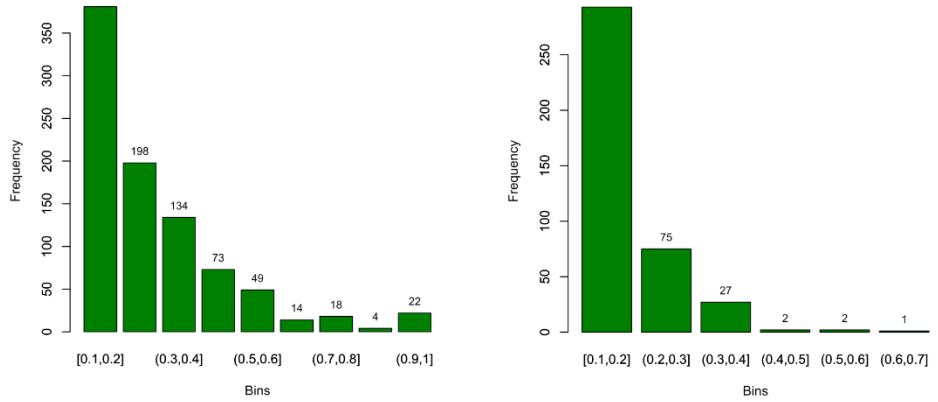
Supplementary Figure S7: Correlation between TF enrichments in enhancers linked to unique metabolic genes vs core metabolic genes.



Supplementary Figure S8: Expression, ChromGene states and sum enhancer and gene interactions for polycomb-associated genes.



Supplementary Figure S9: Number of genes associated with the twelve ChromGene states. Chrome 1-6 are linked to an active state, 7 stand for poised and 8-12 to repressive marks.



Supplementary Figure S10: Number of pathway-cell type combinations having a presence rate between 0.1-0.7 for ChromGene 1 (left) and between 0.1-1 for ChromGene 12 (right).

Supplementary Table S8: Number of cell types with a presence rate above 0.6 for strongly transcribed genes with enhancer marks H3K4me1 and H3K27ac (ChromGene State 1)

Pathways	Number of cell types	Number of reactions
Keratan sulfate synthesis	9	22
Lipoate metabolism	9	1
Alkaloid synthesis	5	1
Keratan sulfate degradation	3	76
Glycolysis/gluconeogenesis	3	42

Fructose and mannose metabolism	3	23
Vitamin B6 metabolism	3	11
Hyaluronan metabolism	3	5
Nucleotide metabolism	3	1
Fatty acid oxidation	1	961
Cholesterol metabolism	1	242
Fatty acid synthesis	1	239
Folate metabolism	1	59
Pyruvate metabolism	1	32
Glutamate metabolism	1	16
Cytochrome metabolism	1	15
Purine synthesis	1	15
Glyoxylate and dicarboxylate metabolism	1	14
Alanine and aspartate metabolism	1	13
Triacylglycerol synthesis	1	9
Leukotriene metabolism	1	8
ROS detoxification	1	7
Squalene and cholesterol synthesis	1	5
Butanoate metabolism	1	3
Limonene and pinene degradation	1	3
Nucleotide sugar metabolism	1	1

Supplementary Table S9: Number of cell types with a presence rate above 0.3 for Polycomb-mediated repressed genes (ChromGene State 12)

Pathways	Number of cell types	Number of reactions
Cytochrome metabolism	12	15
Vitamin D metabolism	4	13
Alanine and aspartate metabolism	3	13
Steroid metabolism	3	90
Bile acid synthesis	2	185
D-alanine metabolism	2	3
Androgen and estrogen synthesis and metabolism	1	26
Blood group synthesis	1	47
Butanoate metabolism	1	3
Glycine, serine, alanine, and threonine metabolism	1	47
O-glycan metabolism	1	18
Triacylglycerol synthesis	1	9