Year : 2020

Integrative Analysis of Gene Expression data from a human cohort

Sönmez Flitman Reyhan

Sönmez Flitman Reyhan, 2020, Integrative Analysis of Gene Expression data from a human cohort

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_3CACAB6F72EA9

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Département de Biologie Computationnelle

Integrative Analysis of Gene Expression data from a human cohort

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

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

par

Reyhan Sönmez Flitman

Master en Statistique de l'Université de Neuchâtel, Suisse

Jury

Prof. Thierry Pedrazzini, Président
Prof. Sven Bergmann, Directeur de thèse
Prof. Marc Robinson-Rechavi, Expert
Prof. Bart Deplancke, Expert

Lausanne
2020

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

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

Président·e	Monsieur	Prof.	Thierry	**Pedrazzini**
Directeur·trice de thèse	Monsieur	Prof.	Sven	**Bergmann**
Expert·e·s	Monsieur	Prof.	Marc	**Robinson-Rechavi**
	Monsieur	Prof.	Bart	**Deplancke**

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

Madame Reyhan Flitman

Master in statistics, Université de Neuchâtel, Suisse

intitulée

Integrative analysis of gene expression data from a human cohort

Lausanne, le 3 juillet 2020

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

Acknowledgements

I am grateful to many people who were with me throughout this long journey and made the journey memorable.

If I remember correctly, Rico was the first colleague I have met in Bugnon 27, and he remains special to this day. He not only mentored me in the first couple of years of my PhD but also became a friend that I like to cherish all my life. Then I met many more amazing people in my CBG days. I will remember Tanguy as an enthusiast sportsman and father whom we could never do the via ferrata, Micha as a tranquil-peaceful soul, Daniel with our lakeside and water sports occasions, David as the other PhD whom I wished to overlap a bit longer, Roger with his trips to home country Spain, Sarvenaz with her smile and easiness, Mirjam with her softness, Tugce my turkish buddy with her eagerness to embrace change, Anneke with her strength, Mattia with his breeze, Bita with her attention & engagement, second Daniel with his quietness and Katarina with our long coffee break chats on motherhood. Thank you all for being part of my life. I would also like to thank Marc for welcoming me to *hub neuchatel* which has been my office in the last bits of this journey, during the Corona days.

And Sven, the best supervisor one can ever have. He has just been incredibly supportive from the beginning till the end of my PhD. I would like to thank him for his availability when I needed it, for his patience and trust with me, and for showing me how to conduct science with integrity. I also greatly appreciate his kind being that allowed me to enjoy other beautiful things that life had brought to me throughout these years.

Ian, my love, the one that is always there next to me. Thank you for sharing my life with me. And finally my girls, Mavi and Defne who are younger than my PhD, thank you for being the most charming things on this planet and making me happy just with your presence.

Summary

Emergence of *Next Generation Sequencing* and other technological advancements that boosted high-throughput measurements, facilitated generation of numerous *-omics* data. As -omics data became more accessible and widespread, integration of them into association studies has become a central challenge in the field. Such synthesis allows investigating the interplay between different organisational layers of a biological system and aids building a more holistic view of the organism.

We had access to various datasets including genomics, metabolomics and gene expression data that were collected under a collaborative effort of the *Cohort Lausannoise* study. We used RNA-Seq data from lymphoblastoid cell lines (LCLs) derived from 555 Caucasian individuals to characterize their transcriptome and in the first part of this work, integrated gene expression data with genotypes to study the genetic variants affecting the gene expression levels, an analysis known as eQTL analysis.

In the second part of the work, we investigated the results of a metabolome- and transcriptome-wide association study to identify genes influencing the human metabolome. As for the metabolome we took an untargeted approach using binned features from 1H nuclear magnetic resonance spectroscopy (NMR) of urine samples from the same subjects allowing for data-driven discovery of associated compounds (rather than working with a limited set of quantified metabolites). We identified 21 study-wide significant associations between metabolome features and gene expression levels. The most significant association was between the gene *ALMS1* and two adjacent metabolome features at 2.0325 and 2.0375 ppm. By using our previously developed metabomatching methodology, we found N-Acetylaspartate (NAA) as the potential underlying metabolite whose urine concentration is correlated with *ALMS1* expression. Indeed, a number of metabolome- and genome-wide association studies (mGWAS) had already suggested the locus of this gene to be involved in regulation of N-acetylated compounds, yet were not able to identify unambiguously the exact metabolite, nor to disambiguate between *ALMS1* and *NAT8*, another gene found in the same locus as the mediator gene. The second highest significant association was observed between *HPS1* and two metabolome features at 2.8575 and 2.8725 ppm. Metabomatching of the association profile of *HPS1* with all metabolite

features pointed at trimethylamine (TMA) as the most likely underlying metabolite. mGWAS had previously implicated a locus containing *HPS1* to be associated with TMA concentrations in urine but could not disambiguate this association signal from *PYROXD2*, a gene in the same locus.

In the third part of the work we studied causality between gene expression levels and metabolite concentrations by Mendelian Randomization analysis. We showed for both *ALMS1* and *HPS1* genes that their expression is causally linked to their associated metabolite concentrations. Our study provided evidence that the integration of metabolomics with gene expression data can support mQTL analysis, helping to identify the most likely gene involved in the modulation of the metabolite concentration.

Résumé

L'émergence du *séquençage de la prochaine génération* et d'autres avancées technologiques qui ont permis les mesures à haut débit, ont facilité la production de nombreuses données de type "omiques". Les données omiques étant ainsi devenues plus accessibles et plus répandues, leur intégration dans les études d'association est devenue un défi central. Une telle synthèse permet d'étudier l'interaction entre les différentes couches organisationnelles d'un système biologique et aide à construire une vision plus globale de l'organisme.

Nous avons eu accès à divers ensembles de données, notamment des données de génomique, de métabolomique et de transcriptomique (expression de gènes), recueillies dans le cadre de l'étude de la *Cohorte Lausannoise*. Nous avons utilisé les données RNA-Seq de lignées de cellules lymphoblastoïdes (LCL) provenant de 555 individus caucasiens pour caractériser leur transcriptome et dans la première partie de ce travail, nous avons intégré les données d'expression des gènes avec les génotypes pour étudier les variantes génétiques affectant les niveaux d'expression des gènes, une analyse connue sous le nom d'analyse eQTL.

Dans la deuxième partie du travail, nous avons étudié les résultats d'une étude d'association à l'échelle du métabolome et du transcriptome pour identifier les gènes influençant le métabolome humain. En ce qui concerne le métabolome, nous avons adopté une approche non ciblée en utilisant directement les données spectrales obtenues par résonance magnétique nucléaire (RMN) d'échantillons d'urine de ces mêmes 555 individus, ce qui a permis de découvrir des métabolites associés à partir de données (plutôt que de travailler avec un ensemble limité de métabolites quantifiés). Nous avons ainsi identifié 21 associations significatives à l'échelle de l'étude entre les caractéristiques des métabolomes et les niveaux d'expression des gènes. L'association la plus significative était entre le gène *ALMS1* et deux caractéristiques métabolomiques adjacentes à 2.0325 et 2.0375 ppm. En utilisant notre méthodologie de metabomatching, nous avons trouvé le N-Acetylaspartate (NAA) comme candidat sous-jacent potentiel dont la concentration urinaire est corrélée avec l'expression de l'*ALMS1*. En effet, un certain nombre d'études d'association à l'échelle du métabolome et du génome (mGWAS) avaient déjà suggéré que le locus de ce gène était impliqué dans la régulation des composés N-acétylés, mais elles n'ont pas pu identifier sans ambiguïté le métabolite exact, ni faire la distinction entre *ALMS1* et *NAT8*, un autre gène trouvé

dans le même locus que le gène médiateur. La deuxième association significative la plus élevée a été observée entre *HPS1* et deux caractéristiques du métabolome à 2,8575 et 2,8725 ppm. Le mGWAS avait précédemment impliqué un locus contenant *HPS1* pour être associé aux concentrations de TMA dans l'urine mais n'a pas pu désambiguïser ce signal d'association de *PYROXD2*, un gène dans le même locus.

Dans la troisième partie du travail, nous avons étudié la causalité entre les niveaux d'expression des gènes et les concentrations de métabolites par analyse de randomisation mendélienne. Nous avons montré pour les gènes *ALMS1* et *HPS1* que leur expression est causalement liée aux concentrations des métabolites qui leur sont associés. Notre étude a fourni la preuve que l'intégration de la métabolomique avec les données d'expression des gènes peut soutenir l'analyse mQTL, aidant à identifier le gène le plus probablement impliqué dans la modulation de la concentration du métabolite.

List of Abbreviations

ACP	Auto-correlation profile
BH-FDR	Benjamini-Hochberg False Discovery Rate
BMRB	Biological Magnetic Resonance Data Bank
CLT	Central Limit Theorem
CNV	Copy Number Variants
CoLaus	Cohorte Lausannoise
CVDs	Cardiovascular diseases
DEGs	Differentially Expressed Genes
DGE	Differential Gene Expression
EBV	Epstein-Barr virus
eQTL	expression Quantitative Trait Loci
eQTS	expression Quantitative Trait Score
FDR	False Discovery Rate
GWAS	Genome-wide association studies
GLM	Generalised Linear Models
GO	Gene Ontology
HMDB	Human Metabolome Database
ISA	Iterative Signature Algorithm
IVs	Instrumental variables
IVW	Inverse Variance Weighted
LCLs	Lymphoblastoid Cell Lines
LD	Linkage disequilibrium
lincRNAs	Long intergenic non-coding RNAs
mGWAS	Metabolome and genome-wide association studies

MAF	Minor Allele Frequency
MR	Mendelian Randomization
mQTLs	Metabolome quantitative trait loci
NAA	N-Acetylaspartate
NAC	N-acetylated compounds
NGS	Next Generation Sequencing
NMR	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
PCA	Principal component analysis
pFDR	positive False Discovery Rate
PGS	Polygenic risk scores
Pro-BNP	pro B-type natriuretic peptide
QC	Quality control
QTL	Quantitative Trait Loci
RNA-Seq	RNA sequencing
RPKM	Reads Per Kilobase of transcript, per Million mapped reads
RT	Reverse transcriptase
SNP	Single nucleotide polymorphism
TAD	Topologically Associated Domain
TM	Transcription Modules
TMA	Trimethylamine
TMM	Trimmed Mean of M-values
TR-lincRNAs	Trait-relevant lincRNAs
TSLS	Two Stage Least Squares
VMH	Virtual metabolic human database

Table of content

List of Figures

List of Tables

1 Introduction

In this chapter I introduce the simultaneous analysis of multi-omics data, the concept of precision medicine, the difference between hypothesis-generating versus hypothesis-driven research, and 'CoLaus', a cross-sectional multi-omics human cohort that we use for the analyses. I also describe various statistical methods and modular approaches I used for the analyses and Mendelian Randomization as a tool of choice to infer causality.

1.1 Integrative analysis of multi-omics data

Understanding the relationship between genotype and phenotype is of fundamental importance in biology. It is also highly relevant in biomedical research in order to better understand the molecular basis of many diseases. It is well known that genetic variants together with the environmental factors influence risk of developing certain diseases and determine complex phenotypic traits [1]. After the completion of *The Human Genome Project* in 2003 [2] millions of DNA sequence variants in the human genome were discovered by the HapMap [3] and other projects. Since around this time *Next Generation Sequencing* (NGS) technologies have been emerging with a fast pace, creating billions of DNA sequences cheaper and faster than previously anticipated [4]. Thanks to these advancements, genome sequence data of humans and other model organisms became widely available and marked the last two decades as the *Genomics era*, signifying the transition from a gene-centric to a genomic view. With the rise of genomic data, genome-wide association studies (GWAS) emerged, where common genetic variants are associated with complex traits and common diseases [5]. As millions of SNPs across the genome can be assayed simultaneously, GWAS represent a promising way to study complex and common diseases in which many genetic variants contribute to disease risk [6-8]. GWAS discoveries therefore provide insights into the contribution of individual variation to the susceptibility of many diseases including multiple sclerosis, Crohn's disease, diabetes, cancer and schizophrenia [9-14] and to continuous traits such as lipids, height and fat mass [15-17]. While GWAS is a powerful tool to connect traits to genetic variability, it is not without limitations. Many trait associated SNPs do not map to protein coding open reading frames and even if they do, they often correspond to synonymous nucleotide changes, not altering the amino acid

sequence of the protein [18]. Moreover, pinpointing causal variants discovered in GWAS remains difficult since the lead variants associated with a trait are often in high linkage disequilibrium (LD) with other variants in the same region with only slightly lower association signals. Such associated LD blocks typically contain several genes or functional elements, preventing the accurate identification of causal variants. For these reasons, it is often difficult to get deeper insight into the biological mechanisms underlying these associations. A criticism of GWAS is on *missing heritability*, a concept referring to the candidate loci reported by GWAS typically having small effect sizes and even jointly explaining only a small fraction of the estimated variance of a trait [19]. However recently it has been shown that explained heritability is getting closer to those of estimated from twin studies [6].

In addition to generating genotypic data, technological advancements have promoted high-throughput measurements also in other fields such as gene expression, epigenetics, metabolomics and proteomics. As high-throughput measurements of these molecular traits become more accessible and widespread, integration of them into association studies has become a central challenge in the field. Such synthesis allows investigating the interplay between different organisational layers of a biological system.

Integration of gene expression with genotypic data, is one of the frequently used methods to derive functional relevance of genetic variants. Associating gene expression levels with genetic variants results in the discovery of expression quantitative trait loci (eQTL) and many studies reported that trait associated genetic variants discovered in GWAS are significantly enriched in eQTLs, suggesting that many trait associated variants affect the phenotype by altering gene expression [20-23]. There is also a growing body of literature highlighting the more pronounced effects of genetic variants on molecular traits compared to phenotypic traits [24-26]. This is not surprising as molecular traits representing fundamental biological processes such as gene expression are intermediates in the genotype to trait causality chain.

Another type of molecular traits that is of interest to GWAS is metabolomics. GWAS with metabolomic traits (mGWAS) search for genetic variants that influence human metabolism. Metabolites are small molecules that reflect various cellular processes taking place in the cells of an organism. Metabolomics techniques, such as mass-spec and NMR allow for estimating the concentrations of large sets of metabolites thus providing snapshots of the physiological states of

a cell. They complement transcriptome and proteome measurements, which reflect events, such as the expression of a particular gene, which may be the cause or the effect of metabolomic changes. In general, metabolite concentrations are influenced by the genetic background and the environment including diet, infections as well as chronic diseases, and interactions between the two. To date more than 150 loci have been identified as modulators of serum and 26 loci for urine metabolites [27]. Having genotype, metabotype and phenotype data, one can start to investigate to what degree the metabolomics data reflect the genotypic background and how informative it is about the phenotype, e.g. disease susceptibility. Changes in metabolite concentrations may be the consequence of the genetic background modulated by the environment, and some of these changes can be causal for developing diseases. Conversely, some changes in metabolites may occur as a results of an organismal dysfunction. Being able to distinguish between these two scenarios, would be of great clinical usefulness, as metabolite changes which are causally upstream are good candidates for developing presymptomatic biomarkers indicating increased disease risk well ahead of the various homeostatic organismal processes leading to disease manifestation.

Despite metabolism and gene expression regulation both being fundamental biological processes that are commonly studied as molecular phenotypes, there are very few studies in humans that focus on the interplay between them. Several studies investigated the relationship between untargeted serum metabolites and whole blood gene expression in humans [28-30], but, to the best of our knowledge no transcriptome- and metabolome-wide association study has been performed using urine metabolome data of healthy human subjects, which we were able to investigate in the context of this work.

1.2 Precision medicine

Personalised medicine seeks tailoring the medical treatments according to the individual characteristics of each patient. The recently emerged term *precision medicine*, while often used interchangeably, is intended to convey a slightly different message. As therapeutics are rarely developed for individuals, precision medicine implies the treatment accuracy among a spectrum of patients, e.g subgroups of patients, instead of single individuals.

The concept was in response to the scientific discoveries that changed the perception of how the unique molecular profile of a person affects its susceptibility to certain diseases. Yet the notion of precision medicine is also in line with many other advancements that took place in recent decades. Having access to more readily available genomics data, improved understanding in the population-level genetic variation, increased digitization of medical records and creative approaches to integrate data, are all contributing to the formation of this notion. With improved prediction, prevention, diagnosis and treatment of the disease, precision medicine has the promise of advancing the traditional methods and endorsing genome-driven medical decision making. Some of the postulated benefits of precision medicine include:

- Screening, diagnostic and prognostic tests to determine the predisposition to common genetically determined diseases and predicting the severity of the disease
- Drug targeting according to disease sub-types and adjusting the effective dose to achieve better therapeutic outcome while reducing the side-effects
- Assessing drug hypersensitivities
- Reduced healthcare costs due to more efficient therapies and reduced side-effect related treatments

Pharmacogenomics is one of the earliest fields that adopted the precision medicine practices by providing support to clinical decision making. Decision support in this sense is making use of genotypic variability among individuals to favor drugs and adjust dosages individually for patients in order to reduce the chances of patients suffering from life threatening side-effects and administering an effective dose for better therapeutic outcomes. Studies associating genetic markers with pharmacogenomic traits, in other terms GWAS with pharmacogenomic traits, have high potential for getting translated into clinical practice and therefore have an immediate impact on human health. GWAS with pharmacogenomic traits are also known to have larger effect sizes compared to the GWAS with complex diseases [31]. One important example of such association study is the investigation of genetic determinants of Hepatitis C virus (HCV) persistence and response to therapy. Several groups, including one at the CHUV, studied patients with HCV infection and found SNPs in the IL28B locus, encoding antiviral cytokine interferon lambda, associating with the progression of chronic HCV infection [32]. Findings of this study facilitated stratification of patients according to their genotypes and revealing their likelihood to respond to

the known therapies. As a result this subgroup of patients were recommended to be prioritized for novel therapeutic strategies in order to avoid drug resistance mutations and treatment failure. Other GWAS applications with pharmacogenomic traits that had sizable effect size estimates include response to tamoxifen in breast cancer and response to statin treatment [31]. Another benefit of GWAS in pharmacogenomic research is the value of disease associated genetic variants as biological targets for drug development.

The role of metabolomics in pharmacogenomic research is important. As metabolism is closely related to cellular processes, any kind of perturbation in cellular physiology can result in altered metabolism and therefore an altered metabolic profile. This shows the great potential of metabolomic profiling of body fluids such as blood and urine in the context of health status monitoring. Due to potential confounding factors, drug response studies are usually designed in vitro, using cell-line based models where different drug doses are administered to the cell lines and in return changes in their gene expression and cell growth are observed. However drug response may not be well reflected when gene expression of a surrogate tissue is used. For instance, gene expression of a cell line might not reflect the metabolic response required to metabolize a drug and likewise blood might not well imitate the drug-organ interaction. Integrating metabolomics into drug response studies can be especially beneficial in this context, where it would be possible to study the drug response in vivo by measuring metabolic response created in return to drug administration and dose adjustment.

Despite the premises of precision medicine it also receives a fair amount of criticism on its validity, applications and consequences. One of the major criticisms it receives is its reliance on algorithms that were developed to tackle cohort or population level genetic variation rather than individual level. Low level of accuracy in a population study translates into a lost opportunity of a discovery whereas it has more destructive consequences in the clinical testing context. Another skepticism of its usefulness is due to the perception of the increased genetic risk by society. Joyner argues that in most cases genetic information will be overvalued and received as deterministic, resulting in devaluation of likely beneficial lifestyle advice [33]. This will either discourage people to make healthier lifestyle choices or in the opposite scenario it will increase the demand for medical surveillance and therefore medical costs, which might be unnecessary. While some see precision medicine as hope for the future generations to come, others see it as a distraction from otherwise simple issues that need attention, "for the sake of a revolution that

might never come" as Joyner states [33-36]. All in all, precision medicine is no doubt an exciting emerging field, yet as it is true for every scientific advancement one should acknowledge its limitations and be cautious about the interpolation of nishe successes for other domains.

1.3 The CoLaus study

Cardiovascular diseases (CVDs) and related risk factors are recognized as the main cause of death in the world, accounting for almost a third of all deaths globally in 2016 [37]. While CVDs are common, the associated risk factors show substantial variation across global regions [38]. The *Cohorte Lausannoise* (short CoLaus) Study belongs to one of the many international efforts aiming to assess population specific prevalence of CVD risk factors, in this case the population of Lausanne. Another aim of this population-based cross-sectional study is to help discovering new genetic determinants of cardiovascular risk factors [39]. Recruitment to the project was done on the basis of a simple, non-stratified random selection of 19,830 Swiss residents from the wider Lausanne area who were aged 35-75 years in 2003. Overall 6,738 subjects participated in the CoLaus Study undergoing extensive phenotyping, but only the subpopulation of 6,188 Caucasian was genotyped. Baseline sampling procedure started in 2003 and ended in 2006 (see Figure 1).

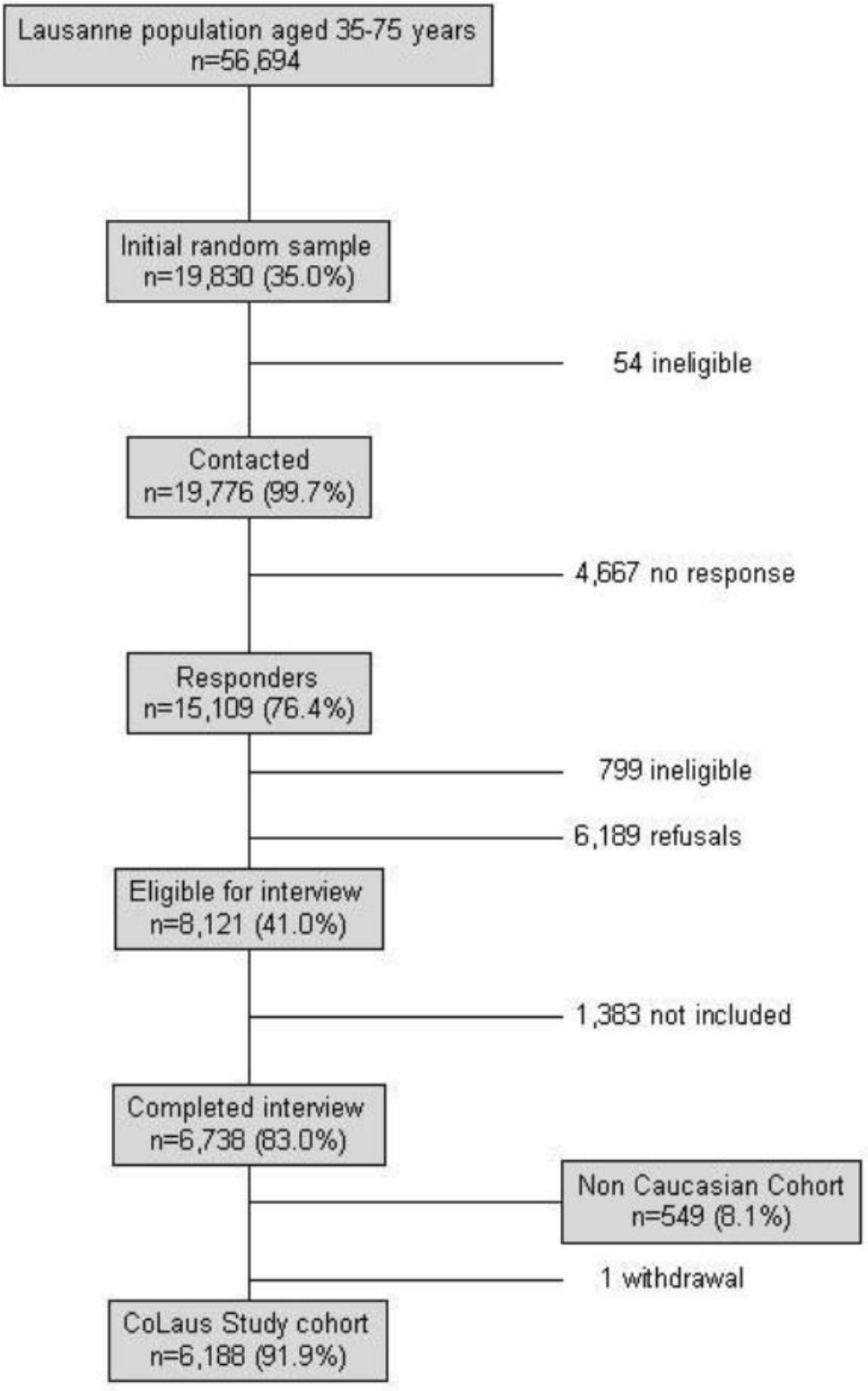


Figure 1: Flowchart of CoLaus study, adapted from Firman et al. 2008 [39].

The clinical phenotypic data were assessed through questionnaires, following face-to-face interviews and medical examinations. The questionnaires and face-to-face interviews recorded information on socio-economic status, lifestyle factors, personal and familial history of CVD and related risk factors, while medical examinations measured anthropomorphic traits and clinical characteristics. An overview of measured phenotypes is listed in Table 1.

Socio-economic status	Country of origin, Family status, Marital status, Education, Work status
Lifestyle factors	Smoking status, Alcohol, Caffeine consumption, Physical activity, Mood
Personal and family history of disease and CV risk factors	Age of occurrence, Number of members affected, Personal medicines used, Reproductive and obstetrical history, Mini-Mental State Evaluation
Anthropomorphic traits	Height, Weight, Body mass index, Hip circumferences, Baldness
Clinical characteristics	Blood pressure (systolic and diastolic) , Hypertension, Family history of diseases, Cholesterol (total, LDL and HDL), Triglicerides, Plasma glucose, Insulin, Leptin, Adiponectin, Creatinin, Other biomarkers

Table 1: Overview of measured phenotypic traits in the CoLaus study.

Genotypic data were acquired using DNA from lymphoblastoid cell lines (LCLs) derived from the subjects' blood cells. Nuclear DNA was extracted from these LCLs for SNP genotyping using the Affymetrix GeneChip Human Mapping 500K array set. Genotypes were called using BRLMM [40]. Next, duplicate individuals and first and second degree relatives were identified through measuring genomic identity-by-descent coefficients by PLINK [41] and the duplicates and the younger individual from the relative pair were removed from the analysis. The 390,631 measured SNPs (with Hardy-Weinberg P-value above 10^{-7} and MAF above 1%) were then imputed to the full set of unmeasured HapMap II SNPs (release 21) by using IMPUTE version 0.2.0 [42]. Lastly, expected allele dosages were computed for 2,557,249 SNPs of 6,188 subjects.

Urine metabolomics profiles were generated for 974 subjects using proton nuclear magnetic resonance (NMR). ^{1}H NMR spectra were acquired at 300 K on a Bruker 16.4 T Avance II 700 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) employing a standard ^{1}H detection pulse sequence with water suppression. Subsequently ^{1}H spectra were binned in chemical shift increments of 0.005 ppm, resulting in metabolic profiles of 2,200 metabolome features. Filtering out features, and then samples with more than 5% of missing values, a dataset composed of 1,276 features for 835 individuals was obtained.

Serum metabolomics profiles of 983 subjects were analyzed by ^{1}H NMR spectroscopy on the same spectrometer. ^{1}H NMR spectra were referenced according to the glucose signal at 5.223 ppm, phase-corrected, and baseline-corrected. No binning was applied. Instead a peak picking approach by using Focus [43] was applied to obtain a reduced set of metabolome features. Later on these features were log-transformed and z-score normalized (both row- and column-wise) to obtain zero mean and unit variance. The final dataset consisted of 388 features for 838 subjects.

Gene expression profiles of LCLs were obtained using the Illumina HiSeq2000 platform. RNA-seq data was produced by the Department of Genetic Medicine and Development at the University of Geneva. Mapping was done onto Genome Reference Consortium Human Build 37 (GRCh37), hg19. Overall 45,470 gene expression profiles were quantified for 555 subjects for whom we also had the metabolomics data.

The systematic follow-up studies of CoLaus were realized in 2009, 2014 and 2018 respectively. During the follow-up information regarding cardiovascular events and deaths were collected. Having access to such longitudinal data will allow for studying the onset and progression of CVD and related risk factors. In particular, having time ordered data opens new ways to investigate causality and search for presymptomatic biomarkers that predict disease risk.

1.4 Hypothesis-driven versus hypothesis-generating research

Hypothesis-driven and hypothesis-generating research are two different approaches to conduct research. In hypothesis-driven research a specific research question is already defined based on a prior groundwork. The goal of the experiments is then to test this hypothesis, in the context of

biomedical research usually through a sequence of assays applied to selected phenotypes of cases and controls.

In contrast, in hypothesis-generating research there is no specific hypothesis *a priori*. Rather, the objective is to explore comprehensive sets of data in order to reveal the patterns or structures within. Eventually, results from a hypothesis-generating research can be used as a basis to formulate concrete testable hypotheses. This is a non-biased way of discovering knowledge where scientists are not constrained by preconceived ideas. The availability of large datasets and advancements in statistical methods to analyze them has facilitated this approach to such an extent that hypothesis-generating research has now also come to be known as *data-driven research*. Genome-wide association studies (GWAS) is an illustration of this type of research, where genetic variants of the entire genome are tested for association with diseases/traits. Resulting discoveries can then be followed up on in a more focussed manner, by translating the results derived from hypothesis-generating research into hypothesis-testing research. Nevertheless, hypothesis-driven and hypothesis-generating research should not be seen as mutually exclusive, while in fact they are complementary to each other. After all, the motivation of hypothesis-generating research is to formulate better and more relevant hypotheses that can be tested.

A more specific example that we accomplished in the scope of this thesis is our CoLaus transcriptome- and metabolome-wide association study. This association study is an example of hypothesis-generating research and it represents the first stage of our work. In the first stage, we were interested in studying the relationship between the two different molecular entities and investigated to what extent their respective features were correlated with each other. By examining the association results we were able to point out metabolite-gene pairs that correlated with each other more than expected by chance. In the second stage, we were then in a position to proceed with some of the interesting results from the first stage and formulate concrete testable hypotheses with them. In our case, we were interested in the causal relationship between specific metabolite-gene pairs and the Mendelian Randomization analysis served to test the specific hypothesis; if the expression of a particular gene was causally upstream to the metabolite concentration it was associated with.

1.5 Some basic statistical concepts

1.5.1 Statistical tests

Parametric vs non-parametric tests

In statistics *parameter* refers to the feature of a population whereas *statistic* refers to a feature of the sampling distribution. A parametric statistical test makes various assumptions on the population parameters and the distribution of the data from which these parameters come from. Student's t-tests and ANOVA tests are among the well known parametric tests.

Non-parametric tests which are also known as *distribution-free tests* do not make any assumptions on the population parameters and are often used when the data from which these parameters come from are non-normally distributed. Mann-Whitney and Krustal-Wallis tests are among the well known non-parametric tests that are used in place of Student's t-tests and ANOVA tests respectively.

Both parametric and non-parametric tests have their advantages and use cases where one should be prefered over the other. Some of the advantages of parametric tests are summarised below:

- They give accurate results when the sample size is large enough and the data is normally distributed. Contrary to common conception, parametric tests can also be used when the data is not normally distributed [44]. In this case however, a higher sample size is required so as to satisfy the central limit theorem, which states that given a sufficiently large sample size, the sampling distribution of the mean will be approximately normally distributed even if the distribution of the population is non-normal. The sample size required to achieve this effect depends on the underlying population distribution of the variable. The further from normal the distribution, the higher the sample size necessary to achieve the normal sampling distribution of the mean.
- When appropriate, parametric tests have greater statistical power to detect true significant effects and they tend to be more accurate.

Some of the advantages of non-parametric tests include:

- They are well suited when the assumptions of parametric tests are not met, typically in the case of small sample sizes.
- Non-parametric tests assessing the median can be favorable in cases where the median is a better measure of central tendency, as in the case of skewed distributions, for example.
- Most of the non-parametric tests compare the distribution of ranks, therefore they are more robust to the presence of outliers.

To summarise, both the distribution and sample size of the data should be considered when making a choice between parametric and non-parametric tests. An evaluation of which central tendency measures suit the data, mean or median, also would help to decide over the two types of tests.

P-values in statistical hypothesis tests

The purpose of statistical tests is to test a hypothesis, where the hypothesis is often an educated guess that can be tested by experiments or observations. In hypothesis-testing terminology, the *alternative hypothesis* corresponds to the proposed relationship between the datasets, whereas the *null hypothesis* corresponds to no relationship between the datasets. Prior to hypothesis-testing, a level of significance, or *alpha level*, is picked. The alpha level represents the probability of rejecting the null hypothesis when it was in fact true, which is known as a *type I error*. Alpha levels should not be seen as definitive in characterizing a hypothesis as valid or invalid; rather, they serve to qualify statistical results in a commonly agreed on manner. For many the consensus value for alpha is 0.05, although lower values have been argued for in favor of reproducibility [45]. A test statistic is calculated on the observed data, then compared to the rejection region defined by alpha in order to support or reject the null hypothesis. A p-value stands for the probability of observing a particular test statistic, given that the null hypothesis chosen for the test statistic is valid. Thus smaller the p-value of an observed statistic, more likely it is that the null hypothesis is not adequately explaining the observed phenomenon.

Multiple hypothesis testing

To every single performed statistical test, there is an associated false discovery rate, defined by the alpha level. When the number of tests performed increases, the number of false discoveries also rises accordingly. The potentially large number of false discoveries resulting from multiple hypotheses testing is known as *multiple hypothesis testing problem*.

There are several different ways to account for the multiple hypothesis testing problem. One of the most traditional methods is the Bonferroni correction, a type of Family-Wise Error Rate (FWER) correction [46]. A Bonferroni-adjusted p-value is calculated by dividing the significance level by the number of tests and it measures the probability of having at least one false positive result. It is a conservative correction method: it focuses on the reduction of false positives, but may in turn also suppress true positives. The False Discovery Rate (FDR) is another popular alternative concept to account for the multiple hypotheses testing problem. Unlike FWER, FDR only controls the number of false discoveries in those tests that result in discovery. Therefore FDR is less conservative than the Bonferroni approach and has greater power to find truly significant discoveries. The FDR concept was first introduced by Hochberg and Benjamini (1990) (called HB-FDR) [47] and then later on improved by Benjamini and Hochberg (1995) (called BH-FDR) [48]. Calculation of FDR simply starts with ranking the p-values from smallest to largest. A BH-critical value is then calculated by multiplying the p-value with $(i/m)Q$, where i is the rank of the p-value, m is the total number of tests and Q is an *a priori* defined false discovery rate. P-values below the BH-critical value are considered significant. Adaptive methods to BH-FDR have been proposed including the one suggested by Storey (2003) called positive FDR (pFDR) [49] which requires at least one significant discovery among all the tests performed. With the introduction of pFDR, Storey also introduces an error measurement called q-value, which is the analogue of p-value related to the false positive rate.

1.5.2 Linear regression analysis

Regression analysis is a well established, commonly used statistical tool in biological, social and behavioral sciences. In the simplest terms, regression analysis models the linear relationship between variables y and x, where y corresponds to observational data (dependent variable) and x

corresponds to explanatory variable (independent variable). In simple linear regression, observational data y only depends on one explanatory variable x, whereas in multivariable linear regression y can be modeled by several explanatory variables $X=(x_1,x_2,...,x_n)$. The general form of multivariable linear regression can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon \,,$$

for $i=(1,...,n)$, where β_0 refers to the intercept, β_i refers to the regression coefficient of x_i and ε refers to the error term.

Regression analysis serves three main purposes that can be summarised as model specification, model fitting and prediction.

- *Model specification* is the process of selecting the explanatory variables $X=(x_1,x_2,...,x_n)$ most relevant to the dependent variable y. Having too few explanatory variables in the model can give rise to underspecified and biased models and too many explanatory variables in the model can give rise to overspecified and less precise models. Model specification will often derive from field expertise. In addition to this, statistical assessments like adjusted R-squared and predicted R-squared can help during the model specification process. These values reflect the improvement of the model with the addition of every new explanatory variable to it. P-values of the regression coefficient estimates is also another way to judge the importance of the variable in the model. Besides these statistics, various regression methods exist that are designed to pick up the most relevant variables for the model in an automated fashion; like stepwise and LASSO regressions. Stepwise regression evaluates the F-test statistics resulting from adding or removing variables to the model, and in the light of these decides which variables should be included in the model. LASSO regression uses shrinkage to shrink data values towards a central point, therefore encouraging sparse models. Nevertheless model specification is by no means an easy task and it indeed causes regression analysis being regarded as an *art*.
- *Model fitting* refers to fitting a regression model to describe the relationship between dependent variable y and the explanatory variables $X=(x_1,x_2,...,x_n)$. When the model is fit, the overall fraction of the variation in Y explained by explanatory variables can be examined as well as the contribution of each explanatory variable, x_i , to explain the variation in Y known as effect size. Ordinary Least Squares method (OLS) is the

standard fitting method used in linear regression. Least squares regression line refers to the line that minimises the distance of data points to the regression line as much as possible, which is also known as minimizing the sum of squared residuals. In simple linear regression, the regression line is fitted in two-dimensional space, to the values of x and y. On the other hand when there are multiple explanatory variables in the model, as in multivariate linear regression, the regression line is replaced by a regression plane and the OLS method still remains valid and applicable. The OLS corresponds to the maximum likelihood criterion when the errors (residuals) of the model are normally distributed. Another least squares method alternative to ordinary least squares, is called regularised least squares. In regularised least squares a penalised version of least squares loss function is minimised as in ridge or LASSO regression.

- *Prediction* is another use of regression where the mean of the dependent variable Y, can be predicted given the values of the explanatory variables $X=(x_1,x_2,...,x_n)$.

Regression models are not limited to linear fits, also non-linear regression alternatives exist. Yet the interpretation of the linear model remains more straightforward when translating the effect of explanatory variables on the dependent variable, as in effect sizes. Even though linear regression models are commonly used and intuitive, they also suffer from numerous drawbacks. First of all it is limited to linear relationships therefore it would not be suitable when the relationship is curved. Second, it only looks at the mean of the dependent variable and when the mean is a poor description of the dependent variable linear regression fails to describe the complete relationship between variables. Third, linear regression is very sensitive to outliers and depending on the nature of the outliers additional effort should be given to account for them.

As mentioned, linear regression models assume the error term to have a normal distribution, meaning a constant change in variable x giving rise to a constant change in y. While this is the case when the variables are normally distributed, certainly it would not hold for all types of relationships. A solution to address this problem is a flexible generalization to OLS, called *Generalised Linear Models* (GLM). GLM can be seen as an extension to OLS, where the dependent variable is allowed to have an error distribution model other than a normal distribution including binomial, poisson and gamma distributions.

1.6 Modular approaches

For GWAS with molecular phenotypes, such as gene expression, methylation or metabolomics data, there is a substantial need to reduce the complexity in the phenotype data. The complexity in such datasets is due to their high dimensional nature and the noise in their measurements. Assigning the data to modules or groups based on a defined criteria and using the weighted average of the group members as a phenotype in the downstream statistical analysis would have several advantages such as; reduction of the complexity of the data as there will generally be fewer groups than individual measurements; reduction of noise as the individual fluctuations of the data would cancel out each other on their average; and finally providing biological focus as the members of a group would share common features.

1.6.1 Principal Component Analysis

Principal component analysis (PCA) takes high-dimensional data as an input and uses the dependencies between variables to project the high-dimensional data into a lower-dimensional space in a way that minimizes the loss of information content of the data. More specifically, PCA uses weighted averages of the original variables to construct new variables, so-called principal components (PCs), that are orthogonal to each other and ordered in such a way that the first component explains the greatest variance in the data, the second component explains the second greatest variance, and so on. It can be shown that the PCs are the eigenvectors of the covariance matrix of the input data. PCA functions usually return two vectors for each component: the component scores and the loadings. The former are the transformed variable values corresponding to a particular data point and the latter are the weights by which each standardized original variable should be multiplied to get the component score. As first leading principal components together typically explain a considerable proportion of the variance, PCA is used to achieve dimensionality reduction in high-dimensional data as well as to inspect the internal variance structure of the data.

In GWAS, PCA is mainly used to identify differences in genotypic information that might be caused by population structure rather than the disease or trait of interest. By performing PCA on genotypic data, the information across millions of SNPS is reduced to a couple of PCs that

typically are governed by differences in studied population. For example PCA on genotypes from studies combining subjects from different regions, countries or ethnicities usually reveals clusters of these structures in the leading PCAs. Since such groups often vary in a given phenotype due to non-genetic factors (such as diet or environmental factors) it is customary to use PCs as covariables in the association model to account for the population stratification of the data. In molecular datasets such as gene expression and metabolomics, PCA can be used for the initial inspection of the data to detect outliers or batch effects that manifest themselves as subgroups of samples whose data received the same bias in some of its features.

1.6.2 Hierarchical Clustering

Hierarchical clustering is a statistical method to discover similar groups in a given dataset. More specifically it uses cluster trees (dendrograms) to represent the data, where every group (node) links two or more successor groups. By definition hierarchical clustering is a nested and organised tree where every node is expected to represent a meaningful group. Hierarchical clustering algorithms are monotonic algorithms that either build by bottom-up or top-down approach. The bottom-up approach, known as hierarchical agglomerative clustering, starts with treating every item as a single cluster and gradually merges two items based on their dissimilarity with other merged items. Pairing continues until all the items are merged into a single cluster. The top-down approach, called divisive clustering, is less frequently used and it starts with one single cluster and splits the clusters into parts based on a defined similarity. The process is repeated until every cluster contains only one item.

Hierarchical clustering is known to be computationally costly and it has high storage requirements. This makes it less convenient for the analysis of large datasets. As the algorithm would find clusters even in the most unsuitable data, the resulting cluster tree can be completely wrong. Also the similarity measures used to define the clusters have a big impact on the appearance of the final tree and it can be puzzling to decide on the most suitable similarity measure to use. Nevertheless hierarchical clustering is a widely used clustering algorithm.

1.6.3 ISA

Iterative signature algorithm (ISA) is a biclustering algorithm that clusters both the rows and the columns of input data into biclusters, or *modules* [50-52]. ISA was developed to find *transcription modules* (TM) in large scale gene expression data. As a soft clustering algorithm, the convenience of ISA over other clustering algorithms is that it allows genes to be involved in multiple clusters which is especially critical when the genes have multiple different functions, a concept known as pleiotropy. Another advantage of ISA, as a biclustering algorithm it allows to study clusters of gene expression under individual experimental conditions unlike clustering algorithms where the expression is generalized under all experimental conditions. This is beneficial as cellular processes are often influenced by a subset of conditions and combining the expression profile over the entire set of conditions would yield an increase in the background noise. By definition, a TM given by ISA contains a group of genes and a group of experimental conditions. Genes in a given TM have more similar expression profiles over the conditions of the TM. Likewise, conditions of the TM are more similar to each other over the genes in the TM. Even though originally the algorithm was designed to work with gene expression over different experimental conditions, it is also possible to apply ISA on gene expression of different samples.

ISA starts with selecting a random subset of genes or conditions, and applies thresholding on these entities in an iterative way. Thresholding is defined as keeping the elements that are certain standard deviation away from the mean, where the standard deviation is specified by the user with *row.threshold* and *col.threshold* parameters. Another parameter called *direction* is used to specify the values to keep if they are significantly higher ('up'), lower ('down') or higher or lower ('updown') than the mean. To give an example; for an input matrix of $E(m \times n)$ with m number of genes and n number of conditions, ISA starts with a random seed vector (r_0), a binary vector of length m, and multiplies it with E'. Next, the result is thresholded and the new thresholded vector c_0 becomes the column signature of r_0. In the next step E is multiplied by c_0 and thresholded to get r_1. This iteration is performed until both r_i and r_{i-1}, c_i and c_{i-1} are similar to each other where the similarity is defined by Pearson correlation. Another parameter, *cor.limit* , allows the modules to converge if they are correlated with each other with specified Pearson correlation coefficient. In the end, ISA outputs modules that consist of similarly expressed genes under a subset of experimental conditions or for a subset of samples. Contribution of each gene or

experimental condition/sample to a module is then summarised by row and column scores. These scores vary between -1 and 1, and further the number is from zero, stronger is the association of the given row/column to the bicluster.

1.7 Mendelian Randomization

Instrumental variable analysis is a statistical method that uses instrumental variables (IVs) to investigate the causal effect of an exposure on an outcome. The principle behind this method is to make use of IVs that affect a particular exposure, and only through this exposure affect an outcome of interest. Effect size estimates of the IV on the exposure and on the outcome are then used to assess the causal effect of the exposure on the outcome. *Mendelian randomization* (MR) is a particular adaptation of instrumental variable analysis into epidemiology, where genetic variants, SNPs, are utilized as IVs [53, 54].

The validity of the causal estimates relies on the three assumptions of MR (see Figure 2). In order to ensure an unbiased causal inference it is crucial to verify that these assumptions are satisfied. The *validity assumption* states that the instruments chosen for the analysis should be strongly related with the exposure. This assumption is satisfied by choosing genetic variants as IVs that are quantitative trait loci (QTLs) of the exposure. The *independence assumption* states that the instruments should not be associated with any confounders of the exposure-outcome relationship. Any known variable that is suspected to confound this relationship should be tested. Typically population stratification can be an example of such a confounding factor. And finally the third *exclusion restriction assumption* states that the instruments should not be linked to the outcome through anything but the exposure, in other words the instruments should not be pleiotropic. This assumption is more difficult to verify, but the plausibility of the assumption can be evaluated in several ways. Pleiotropy can be detected simply by investigating the *heterogeneity* among genetic variants used in MR. A genetic variant is called heterogeneous if its effect on the exposure and the outcome is inconsistent compared to the rest of the genetic variants used in the analysis. This inconsistency is often a sign of horizontal pleiotropy which causes violation of the exclusion restriction assumption. Cochran's Q test or other tests can be used to detect the heterogeneity among the candidate SNPs [55]. Also robust MR analysis methods such as median estimator and MR-Egger regression can be used to evaluate the significance of the causal

estimates [56]. These methods are known to have more relaxed MR assumptions and they can tolerate some degree of heterogeneity among the genetic variants. Agreeing causal estimates given by multiple MR methods are considered as a sign of robust causal estimation [56].

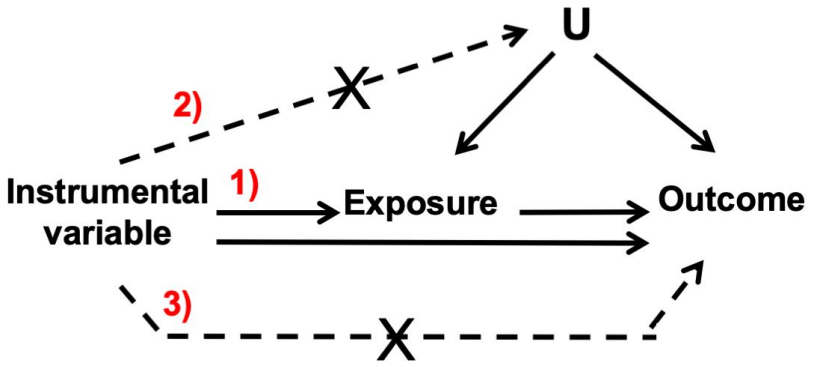


Figure 2: MR assumptions. (1) Validity assumption: the instrumental variable (IV) associates with the exposure; (2) independence assumption: the IV does not associate with any confounding factor U of the exposure-outcome association; (3) exclusion restriction assumption: the IV associates with the outcome only through the exposure.

MR can be applied in two ways depending on the type of the available data. In the case where individual level data is available, effect sizes of SNPs on the exposure and the outcome are estimated by using the same matching samples. However having effect size estimates from matching samples is less common compared to publicly available summary statistics. Fortunately, MR can be applied by using the summary statistics from higher powered studies where effects of SNPs on the exposure and the outcome are estimated by using different samples.

In the same sample MR, causality is studied by using the two stage least squares method (TSLS). In the first stage of TSLS the exposure is regressed on the genetic variants, and in the second stage, the outcome is regressed on the fitted values provided by the first regression. The two stage approach allows to estimate the effect of the exposure on the outcome by using the variability in exposure data driven by genetic variants. Accordingly, the detected causality can be attributed to the causal effect of the exposure on the outcome as the effect of genetic variants can only be upstream of any phenotype. The validity of MR assumptions should be examined prior to the TSLS. Finally, Durbin-Wu-Hausman test of endogeneity [57] can be used to

compare the efficiency and bias of the TSLS estimates against and ordinary least squares (OLS) estimates, in order to justify use of TSLS over OLS estimates.

In two sample MR, summary statistics from higher powered studies are used. Causal estimate is calculated by *Wald Ratio* method where the effect of the genetic variant on the outcome is divided by the effect of the genetic variant on the exposure [58]. Later, individual ratios from different SNPs are often combined by inverse variance weighted method (IVW) to calculate the causal estimate. When using heterogeneous SNPs as IVs, IVW method could give biased causal estimates. On the other hand other meta-analysis methods such as median estimator and MR-Egger are known to be less prone to bias if some of the IVs are not valid.

2 Expression data from CoLaus LCLs

In this chapter I first introduce lymphoblastoid cell lines from which we derived the gene expression profiles, followed by RNA-Seq technology as the method of choice to measure gene expression. Next I describe the initial inspections we performed on CoLaus gene expression data including examining the correlation structure of the data, studying the presence of potential batch effects and detecting outliers via principal component analysis. Finally, I demonstrate a modular analysis on CoLaus gene expression and phenotypic data, where I detect phenotypically relevant subsets of samples for functionally enriched groups of genes.

2.1 Lymphoblastoid cell lines

Lymphoblastoid cell lines (LCLs) are considered *immortal cell lines* that can grow for many generations without turning tumorigenic [59]. Simple preparation of the cell line, effortless and convenient storage in biobanks and the potential to serve as a limitless source of genomic DNA with somatic mutation rate as low as 0.3% are much appreciated in functional and molecular studies [60]. Eliminating the need for resampling is highly beneficial as the discomfort to the donor can be avoided and the concern of unavailable donors due to death or geographic relocation are ruled out. Moreover, having a continuous source of biological material accommodates the growing movement of international biobanking well. Even though LCLs are not the only method to amplify the whole genome, three decades after its discovery it is still considered as the gold standard for long term management of high molecular weight genomic DNA [61].

While there are various ways to generate LCLs, infecting the cell with the Epstein-Barr virus (EBV) remains the most common practice. The method has been used successfully since 1986 and was originally described by Neitzel in detail [62]. Briefly the procedure involves separating the lymphocyte cells from the peripheral whole blood, subsequently infecting the resting B lymphocyte cells with EBV while removing T lymphocyte cells, and finally allowing the EBV infected B cells to proliferate in the growth medium before cryopreservation (see Figure 3) [63].

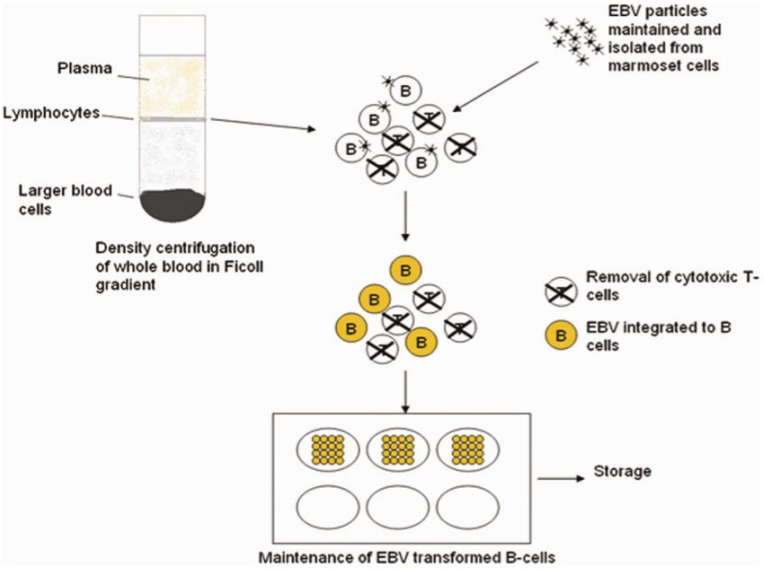


Figure 3: Generation of lymphoblastoid cell line (taken from Sie et al. [63]).

One of the main reasons of LCLs extensive use is due to their ability to maintain an intact genome over many generations [63, 64]. However it is critical to realize to what extent the genetic makeup of LCLs represent that of conventional DNA sources such as whole blood. To date, many studies have reported high correlation between DNA sourced from LCLs and conventional whole blood. For instance, it has been demonstrated that the LCL formation does not cause reproducible change in the structural variation, insertion and deletion polymorphism compared to whole blood [65, 66]. Hence DNA sourced from LCLs recognised as not distinguishable from its genetic counterparts and widely accepted as a faithful copy of their donor's genome.

Besides being used as a renewable source of DNA in genomic screenings, LCLs are also a continuous source of RNA and protein. Thanks to their capacity to produce diverse biomolecules, LCLs are increasingly used in other areas of genomic research such as transcriptomics, epigenetics, proteomics and pharmacological studies. Remarkably, it has been argued that based on the genes expressed in LCLs, these cell lines appear to use a wide range of metabolic pathways in a way that is representative for blood cells but also other cell types of the individuals from whom the cell lines were derived, making them a good model system for use in functional and molecular studies [61]. One of the first publicly available large-scale human

transcriptomics datasets was that of the LCLs derived from HapMap participants [67]. This dataset was first used to identify SNPs that influence inter-individual mRNA variation, known as expression-QTL (eQTL) [68, 69]. Since then, studies based on LCLs have become routine to study inter-population and inter-individual differences in gene expression [70-74]. Differential gene expression studies have become a mainstream tool to understand the etiology of many complex diseases, and have given rise to the discovery of biomarkers [75-81]. LCLs express various proteins found in neuronal cells [82-84]. Among these proteins, amyloid precursor protein (APP) has been associated with Alzheimer's disease and its transcriptional regulation has been found to be similar to those in LCLs [85]. Given the similarity of certain genes' expression and regulation in LCLs and neural cells, LCLs have been suggested as surrogate models for neurological studies [63]. Before the use of LCLs in psychiatric research, where the access to the relevant primary cells were always limited, primary cells such as skin and blood have been traditionally adopted as surrogates of the central nervous system and used to measure peripheral biomarkers. However these cells are confounded by immediate environmental factors such as diet, smoking, alcohol consumption and exposure to toxins and drugs [86]. These primary cells also reflect the acute state of the person, including her circadian rhythm, diet as well as health-related parameters [86]. On the contrary LCLs are less likely to suffer from these environmental and state related changes thus they can be used to study functional aspects of psychological diseases [87]. To summarise, numerous studies showed well use of LCLs in both proteomics and transcriptomics especially in the context of neurological diseases [86-89].

LCLs have been used to study proteome expression response to DNA damage, in particular DNA double-strand break which is a highly cytotoxic event that challenges the genomic integrity of the cell [90]. It is known that the DNA double-strand break plays an important role in the early stages of tumorigenesis and LCLs provide sufficient biological material for the studies to tackle the pathways induced by this damage. Other use of LCLs is in the comparative proteomics field, where a proteome atlas specific to LCLs are build [91, 92] to understand the pathological mechanisms involved in the immortalization process, which is a common process between Epstein-Barr virus (EBV) and many other tumorigenic viruses including papilloma viruses and human T-cell leukemia viruses.

It has been shown that the genetic makeup of a person has a considerable effect on how the person responds to a particular treatment, otherwise known as a hot research topic in precision

medicine. LCLs have been widely used to study the genetics of the drug response in particular drug toxins. They are employed as cost-effective systems to study the effects of drug dosage. Besides the genetic variant discovery they also used in functional genomic studies, where the mechanism of action of the potentially relevant genes are discovered through molecular manipulation of these cell lines [93, 94].

Regardless of LCLs wide use in various fields it is important to be aware of its potential limitations when it comes to its use in genetic and functional studies. As part of the EBV-transformation process the cells go through biological alterations and perturbation in their molecular pathways become inevitable. Also adaptation to culturing, differences in culturing conditions, the age of the cell from newly established to mature, all considered as factors that might affect the appropriate use of LCLs in cell biology. Limitations of LCL model as surrogates of primary B cells are fairly well characterised. As previously mentioned, genetic alterations caused by viral infection is considered negligible thanks to extrachromosomal, circular episome of the viral genome and limited expression of viral genes [62, 95]. Gene expression on the other hand showed difference between transformed and non-transformed B cells, yet it has been shown that the inter-individual differences in gene expression is maintained through the transformation process [96-100]. DNA methylation studies also showed differences between transformed and non-transformed cells, where transformed cells were hypomethylated compared to their non-transformed counterparts [96, 100, 101].

As described there are debates on the use of LCLs as surrogate model systems, however they have proven their worth in genomic studies, they have had great importance in providing essentially inexhaustible supply of DNA and they will likely continue to be a valuable tool to meet the high demands of genetic material in the *biobank-omics era* [102].

2.2 RNA-Seq technology

Individual transcripts were studied as early as the 1980s with low-throughput sequencing methods. *Expressed Sequence Tag* was used to sequence random complementary DNA (cDNA) fragments, and therefore informed about both the sequence and the abundance of RNA [103]. On the other hand, early attempts to study the transcriptome, i.e. the sum of all RNA transcripts of an organism, started with *Serial Analysis of Gene Expression* (SAGE) in the 1990s, where short

tags of cDNA were sequenced [104]. In the mid 1990s these methods were overtaken by a more affordable high-throughput method called *DNA microarray* [105, 106]. The microarray technology relies on hybridization of fluorescently labeled reverse-transcribed DNA to probes that are attached to an array, therefore enabling quantification of predetermined sequences. This technology was widely used until quite recently, and sometimes still is the tool of choice for economic reasons, because it does not require large-scale sequencing, yet allows reasonably accurate quantification of most genes in many samples in a very cost-effective manner. While this revolutionary technology allows the study of genome-wide transcription, it has also important shortfalls. One of the major drawbacks of this technology is its reliance on a priori sequence knowledge for the design of the microarray probes. The need for a reference genome/transcriptome makes this technology not suitable for discovery applications. Other drawbacks of microarrays include limited dynamic detection range caused by both background noise due to cross-hybridization and saturation of signals. By the beginning of 2000s, *Next Generation Sequencing* (NGS), a massive parallel sequencing tool, became available and it changed the transcriptomics field by emerging *RNA-Seq* technology [107-110]. RNA-Seq analysis is a hypothesis-free approach where the transcripts are sequenced individually instead of hybridizing to pre-designed sequences as in microarray. Thanks to this single nucleotide resolution, unlike microarrays, RNA-Seq allows for discovery of novel transcriptional variants. Compared to microarrays it also has a large dynamic range of expression detection and higher reproducibility of the results [111]. To this day RNA-Seq continues to be the method of choice for transcript profiling.

A major application of RNA-Seq is differential gene expression (DGE) analysis and steps involved in this kind of analysis have not changed considerably from the first publications [111, 112]. In a typical RNA-Seq experiment, researchers are interested in the quantification of protein coding and/or long non-coding RNAs. These RNA species bear a polyA tail, and therefore laboratory experiment starts with RNA extraction step followed by an enrichment step which can be achieved by either selection of polyadenylated RNA or depletion of ribosomal RNA (rRNA). Both of these methods serve the same purpose of enriching the polyadenylated RNA while getting rid of the other common RNA species such as rRNA. Next is the fragmentation step, where the RNA samples get fragmented into a certain size range as sequencing platforms usually have size limitations. In the following amplification step, first the RNA samples are

converted to cDNA by reverse transcription (RT) and later these cDNA copies of transcripts get amplified by PCR to overcome the detection limit of sequencers. This step is also useful to allow sequencing of samples with very low input RNA. In the final step, prepared cDNA libraries are sequenced either in one direction (single-end) or both directions (pair-end). Former method is quicker and cheaper compared to the latter method and usually regarded as sufficient to do gene expression analysis. Yet, the latter has benefits of providing more accurate alignments which is especially valuable for novel transcript discovery and gene annotation. As RNA-Seq relies on converting RNA molecules to cDNA before sequencing, the sequencing platforms used for RNA-Seq are the same platforms that are used for high-throughput DNA sequencing. In the final step the cDNA library is sequenced to a read depth of 10-30 million reads per sample, depending on the sequencing platform used [113]. As the accuracy of the RNA-Seq experiment to detect low abundance transcripts is dependent on the number of reads obtained per sample, also known as transcriptome coverage, choosing a sequencer with appropriate transcriptome coverage is of great importance. It has been shown that already 10 million reads per sample is enough for 9 out of 10 genes to be covered by at least 10 reads [113, 114].

Both most recent transcriptomics technologies, microarrays and RNA-Seq, rely on challenging data analysis tasks as part of the experiments. While in microarrays the challenge is handling high resolution images and extracting features from it, in RNA-Seq the challenge is to align millions of short DNA reads into reference genome or reconstruct a de-novo transcriptome without a reference genome. Once the alignment/reconstruction step is achieved, in a typical DGE experiment the next steps include estimation of gene expression by quantifying the reads overlap with transcripts, normalizing the estimates and finally identifying differentially expressed genes (DEGs). There are many tools to identify DEGs such as EdgeR [115], DEseq2 [116], Cuffdiff2 [117], Limma/Voom [118].

2.3 CoLaus RNA-Seq data

In CoLaus, both the genotype and the gene expression data were gathered from the Epstein–Barr-virus-transformed lymphoblastoid cell lines (LCLs) derived from (cryopreserved) whole blood of CoLaus subjects. Total RNA was extracted from these LCLs by following the Illumina TruSeq v2 RNA Sample Preparation protocol (Illumina, Inc., San Diego, CA). Later,

mRNA sequencing was performed on the Illumina HiSeq2000 platform producing 49bp paired-end reads. Paired-end reads were mapped to human genome assembly GRCh37 (hg19) with GEM-Tools using GENCODE v15 as gene annotation [119]. The reads were then filtered for correct orientation of the two ends, a minimum quality score of 150 and allowing 5 mismatches in both ends. Gene level read counts were quantified with in-house script. This resulted in expression profiles of 45,470 genes for 555 individuals, which were quantified as RPKM (Reads Per Kilobase of transcript, per Million mapped reads) values. Above analyses were done at the Department of Genetic Medicine and Development at the University of Geneva.

2.3.1 First look in our RNA-Seq data

We wanted to visualise the RNA-Seq data of CoLaus before using it in downstream statistical analysis. Figure 4 visualizes the entire gene expression matrix of the 45,470 genes across 555 samples, in terms of z-scored $\log_2$ transformed RPKM+1 values on the left and a histogram of $\log_2$ transformed RPKM+1 values on the right. One can see that the majority of RPKM values are small (i.e. < 1), but that some genes (mostly, but not exclusively, protein-coding ones) can obtain much larger values.

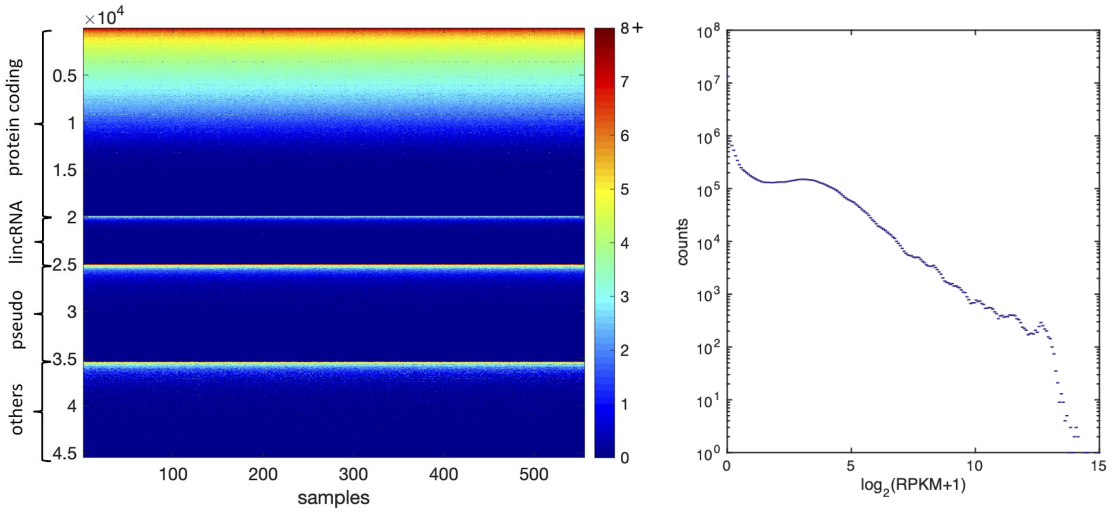


Figure 4: Left: Heatmap of CoLaus RNA-Seq data comprising 45,470 genes of 555 samples. Gene expression is represented as $\log_2$(RPKM+1) values that are z-scored across samples to make genes comparable. Values above 8 are mapped to 8 in order to achieve better visualisation. Protein coding genes, lincRNA and pseudogenes are annotated in the plot, while rest of the genes such as miRNA, processed transcript are under the category called 'others'. Right: Histogram of all $\log_2$(RPKM+1) expression values.

Next we wanted to inspect the correlation structure in the gene expression data. Figure 5 shows the sample-to-sample Pearson correlations across the z-scored $\log_2$(RPKM+1) gene expression values of 45,470 genes (left) and their histogram (right). These correlations ranged from -0.35 to 0.37.

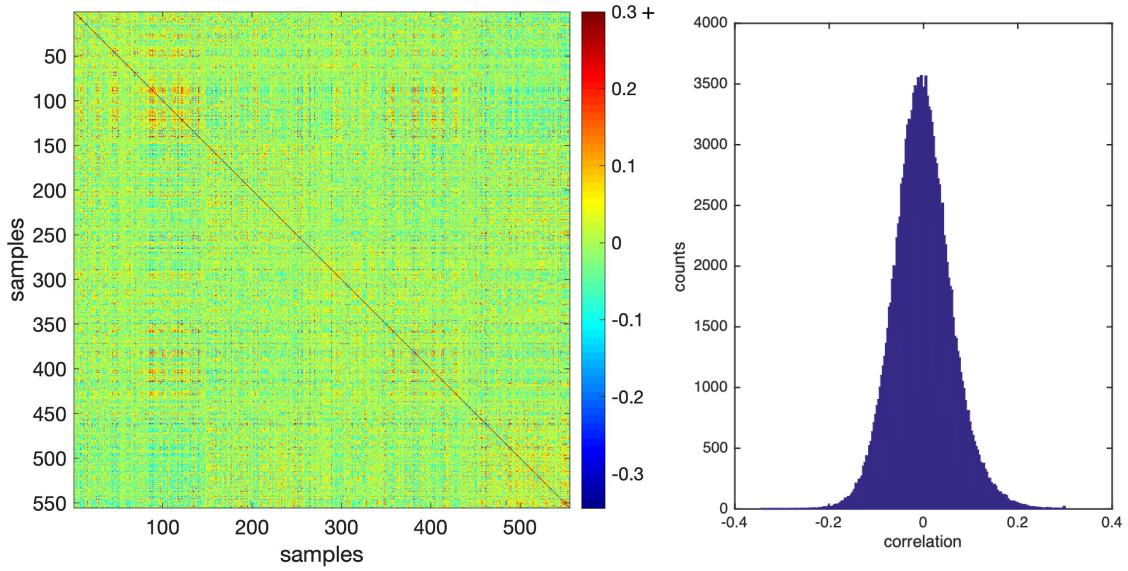


Figure 5: Sample-to-sample correlation plot of 555 samples. Pearson correlation coefficients are calculated on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable. Colour bar represents the correlation coefficients where the coefficients above 0.3 were mapped to 0.3 in order to achieve better visualisation.

2.3.2 Batch effect detection

High-throughput technologies rely on the use of diverse reagents, hardware equipment and technicians to create accurate measurements. During the course of an experiment, varying

conditions in any of these components affect the measurements simultaneously. These effects are known as batch effects, and they manifest as subgroups of samples whose data received the same bias in some of its features. Failing to control for such effects in analyses may result in erroneous biological conclusions. For instance it has been shown that in DGE analysis most of the observed variance is driven by different batches rather than the biological groups [120]. Batch effects can cause serious problems especially when they are correlated with the biological outcome of interest. In such cases not accounting for these batch effects would give rise to misleading and incorrect conclusions. Fortunately, these batch effects can be detected when large amounts of data are available, as it is often the case with high-throughput experiments. Some of the common batch effects in RNA-Seq experiments are driven by DNA preparation, processing groups and processing dates. The protocols used to prepare samples and cDNA libraries might differ among the laboratories, which in turn might also give rise to batch effects when combining data from different labs. The processing date of the samples is also among the other known sources of batch effects and may reflect subtle changes such as temperature or humidity.

Batch effects are more pronounced in studies combining data from different experiments that were generated by different research groups. Even though this is not the case for CoLaus RNA-Seq data, we tested for the potential batch effects that might have an effect on the downstream analysis. We investigated the date of the analysis and the sequencing lane as two potential batch effects. Regarding processing dates; a total of 555 samples were analysed in five days: 192 samples on 20th of June, 192 samples on 27th of June, 84 samples on 4th of July, 84 samples on 10th of July and 3 samples on 13th of November 2014. Regarding sequencing lanes; a total of 555 samples were analysed in eight sequencing lanes. The number of samples analysed in each lane were 75, 72, 72, 72, 72, 72, 72 and 48 respectively.

First we visually inspected the sample-to-sample correlation plot where the samples belonging to the same processing day were grouped together (see Figure 6). Next we examined the correlation between the samples belonging to the same day and samples that were processed in different days (see Figure 7). We tested if the within and between day correlations differed from each other and the results of two sample t-test showed these distributions as significantly different from each other (p-value = 4×10^{-263}).

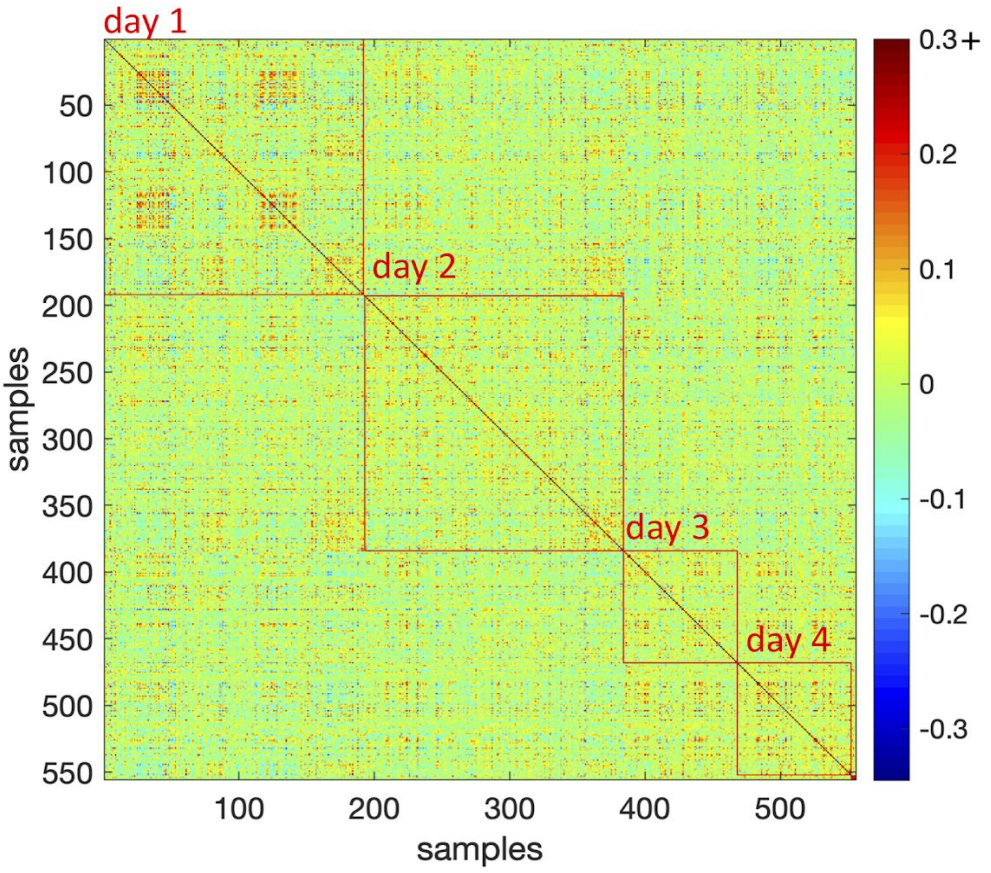


Figure 6: Sample-to-sample correlation plot of 555 samples where samples processed in a given day are grouped together. Heatmap inside the red squares represent the within day correlations of the samples whereas outside the squares represent among day correlations. Day 5 is on the bottom right corner of the plot however as it has three samples it is not visible. Pearson correlation coefficients are calculated on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable. Correlation coefficients above 0.3 were mapped to 0.3 in order to achieve better visualisation.

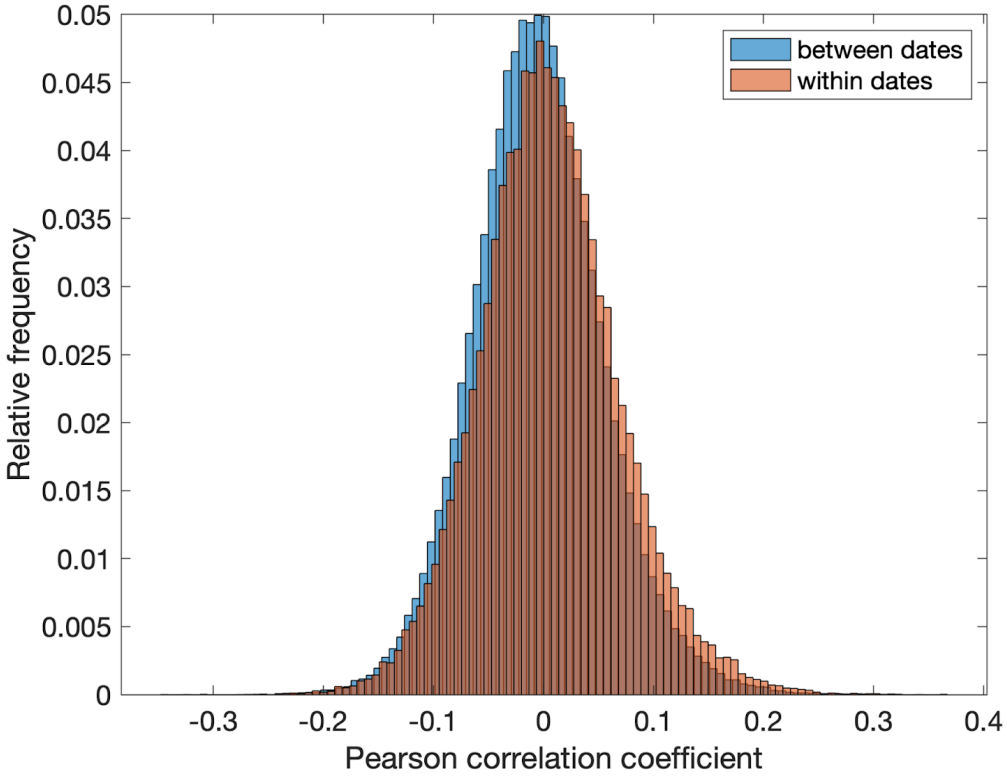


Figure 7: Distribution of sample-to-sample Pearson correlation coefficients given for two groups. Pearson correlation coefficients are calculated on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable. Blue histogram represents 'within days correlations' ; correlations of samples that were processed in the same day summed over five processing days. Red histogram represents 'between days correlations' ; correlations of samples that were processed in different days.

We repeated the same analysis for lanes. First we visually inspected the sample-to-sample correlation plot where the samples analysed in the same lane were grouped together (see Figure 8). Next we examined the correlation between the samples belonging to the same lane and samples that were processed in different lanes (see Figure 9). We tested if the within and between lane correlations differed from each other and the results of two sample t-test showed these distributions as significantly different from each other (p-value = 0.0261).

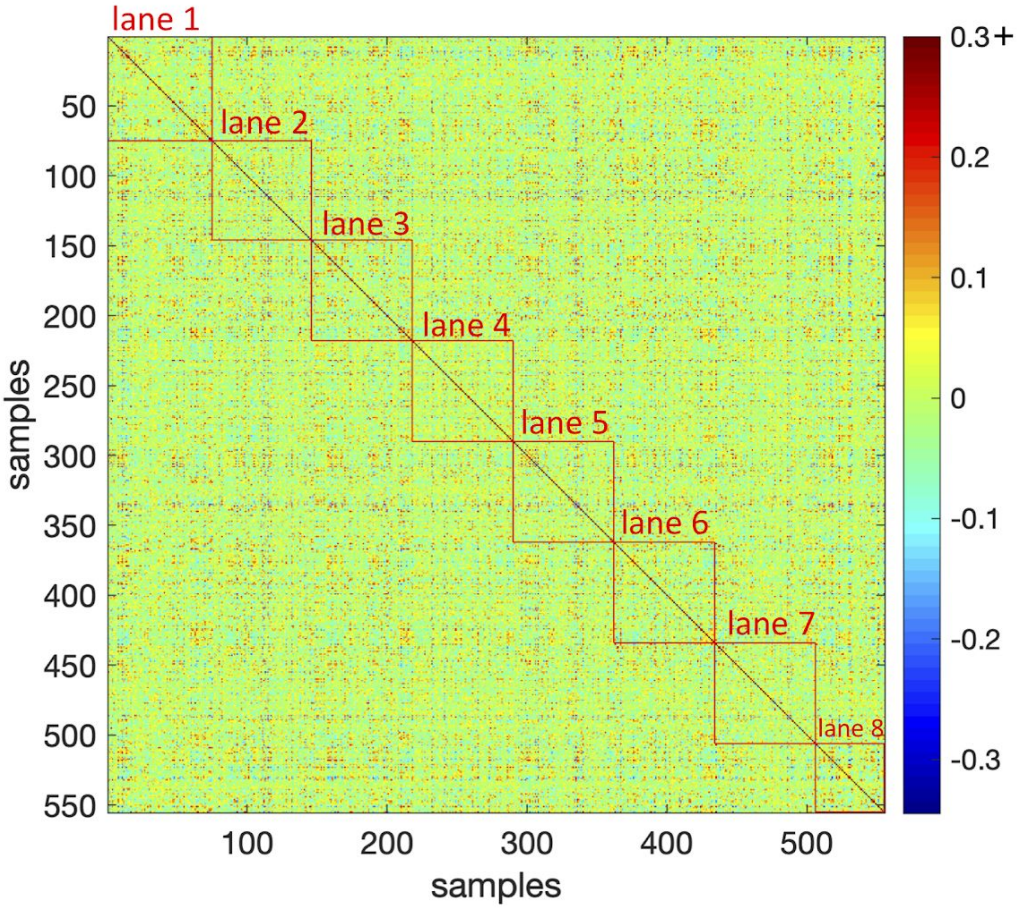


Figure 8: Sample-to-sample correlation plot of 555 samples where samples processed in the same lane are grouped together. Heatmap inside the red squares represent the within lane correlations of the samples whereas outside the squares represent between lane correlations. Pearson correlation coefficients are calculated on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable. Correlation coefficients above 0.3 were mapped to 0.3 in order to achieve better visualisation.

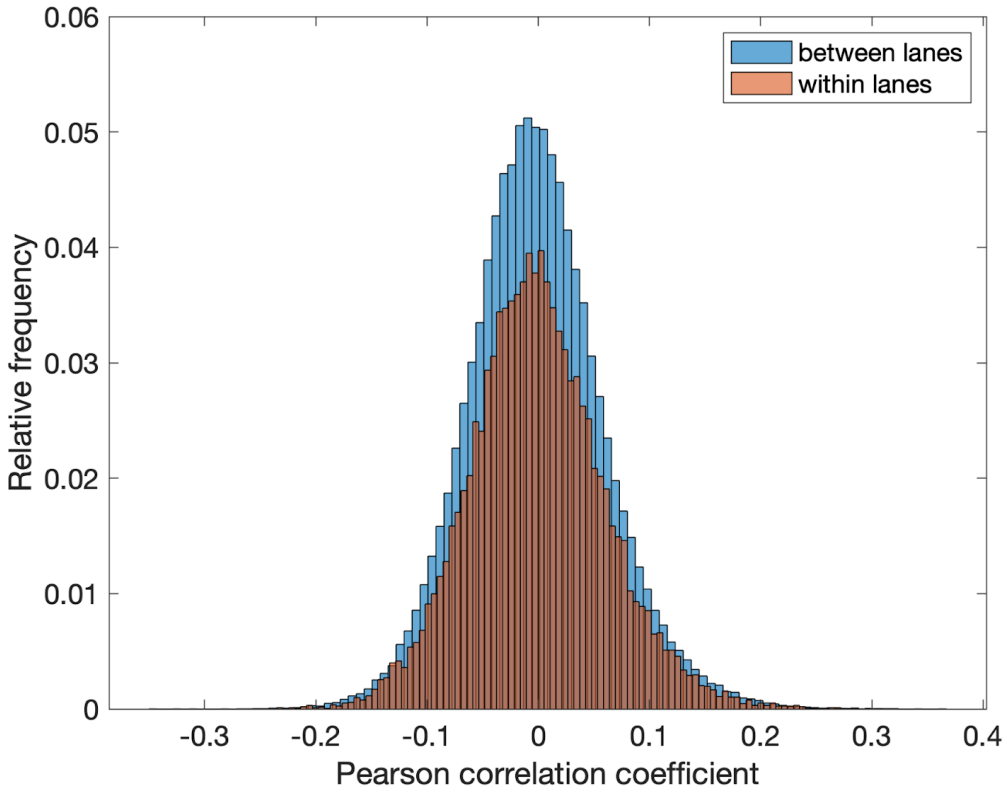


Figure 9 : Distribution of sample-to-sample Pearson correlation coefficients given for two groups. Pearson correlation coefficients are calculated on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable. Blue histogram represents 'within lanes correlations' ; correlations of samples that were processed in the same lane summed over eight lanes. Red histogram represents between lanes correlations; i.e. correlations of samples that were processed in different lanes.

To summarise, when within batch groups and between batch groups correlations were compared we found correlations between samples processed on the same date were significantly smaller than those of samples processed on different dates, therefore suggesting that processing dates incurred batch effects. This was also observed for the sequencing lanes but their effect was much smaller compared to the processing dates.

Principal component analysis

As described in section 1.6.1 one of the uses of principal component analysis (PCA) is to investigate the variability in the data. Therefore anything that would affect the variance structure of the data can be detected by PCA. Here we apply PCA to detect both the outlier samples and the potential batch effects, processing dates and sequencing lanes. Figures 10 and 11 are showing the PCA plots of gene expression data where the samples are color coded based on the day they have been processed and the lane they have been sequenced, respectively. None of the PCA plots show clear outliers, hence we choose not to remove any samples from the gene expression data. Also none of the PCA plots show strong formation of clusters where the samples are enriched for particular processing date or sequencing lane.

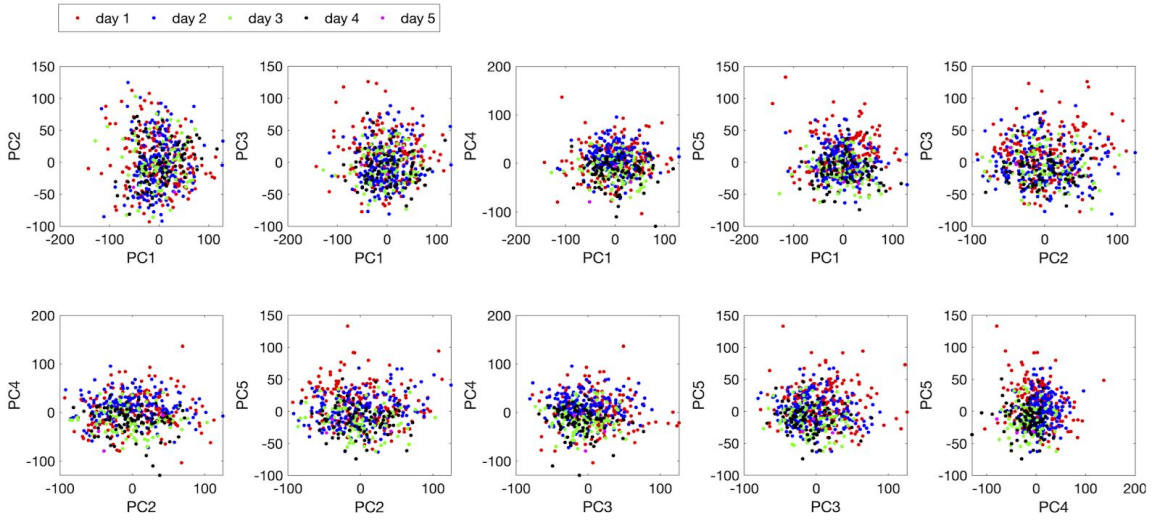


Figure 10: PCA plots showing pairs of principal components from first to fifth. Subjects are color coded based on the processing date. PCA analysis was done on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable.

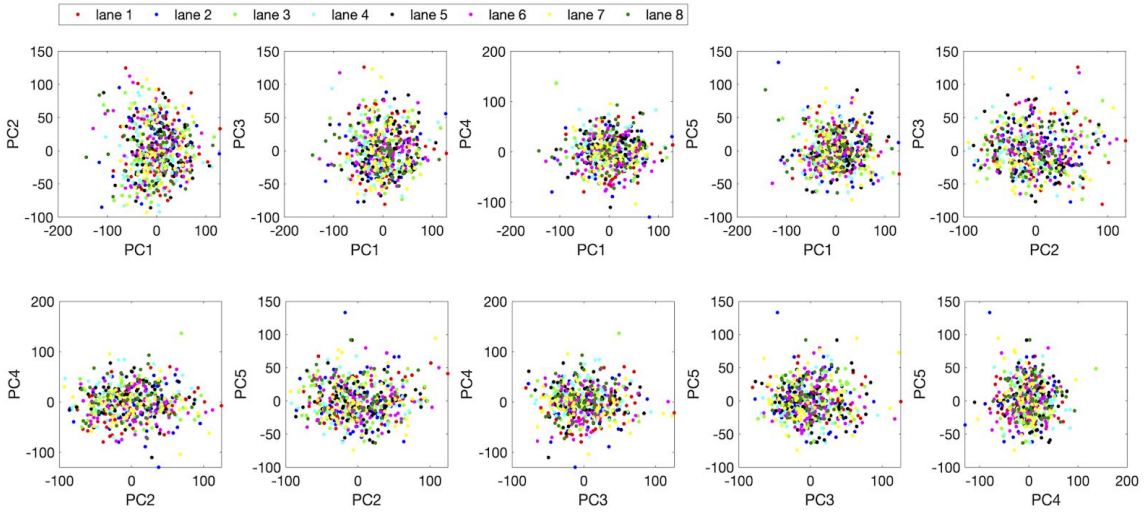


Figure 11: PCA plots showing pairs of principal components from first to fifth. Subjects are color coded based on the sequencing date. PCA analysis was done on $\log_2$(RPKM+1) values of 45,470 genes that were z-scored across samples to make genes comparable.

We also wanted to quantitatively measure whether the principal components were capturing the batch effects. To this end we did regression analyses where each of the first 10 principal components of the gene expression data was used as response variable and the batch effects coded as a design matrix, were the explanatory variables. In the case of processing dates, day 1 was omitted from the coding of the design matrix whereas in the sequencing lanes, lane 1 was omitted. We did remove these variables as they were redundant and not removing them would create a multicollinearity problem for the analysis. P-values resulting from the regressions are shown in Table 2 and 3, for processing dates and lanes, respectively.

	Adjusted R^2 of model	P-values			
		Day2	Day3	Day4	Day5
PC1	-0.01	0.69	0.89	0.98	0.77
PC2	-0.01	0.43	0.69	0.75	0.48
PC3	0.05	6.9×10^{-5}	8.8×10^{-5}	3.0×10^{-6}	0.59
PC4	0.14	0.03	1.3×10^{-6}	3.3×10^{-9}	8.1×10^{-3}
PC5	0.12	1.9×10^{-4}	9.0×10^{-12}	7.3×10^{-12}	0.02
PC6	0.10	7.7×10^{-11}	0.94	1.7×10^{-8}	0.19
PC7	0.02	0.10	0.18	0.05	0.54
PC8	0.02	1.5×10^{-3}	0.02	0.43	0.18
PC9	0.00	0.86	0.10	0.41	0.55
PC10	0.06	8.7×10^{-5}	0.95	5.4×10^{-3}	0.43

Table 2: Linear regression analysis testing whether principal components are associated with processing dates. Day 1 was omitted as it was a redundant variable in the regression. Adjusted R-square is highest for the PC 4 and PC 5, as these PC's also more significantly associated with processing days compared to other PCs.

	Adjusted R^2 of model	P-values						
		Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
PC1	-0.001	0.99	0.45	0.04	0.83	0.30	0.65	0.62
PC2	-0.005	0.65	0.14	0.10	0.32	0.69	0.32	0.29
PC3	-0.008	0.38	0.63	0.26	0.61	0.89	0.31	0.36
PC4	0.031	0.95	0.04	2.0×10^{-3}	4.0×10^{-3}	0.81	0.58	0.02
PC5	0.001	0.74	0.48	0.08	0.20	0.38	0.62	0.41
PC6	0.013	0.19	0.19	0.04	0.04	0.13	0.06	0.32
PC7	-0.006	0.50	0.09	0.24	0.27	0.19	0.20	0.22
PC8	0.014	0.03	0.10	0.28	0.02	0.04	0.05	0.52
PC9	0.014	2.5×10^{-3}	0.12	0.05	0.24	0.03	0.76	0.67
PC10	-0.003	0.16	0.59	0.42	0.25	0.79	0.84	0.44

Table 3: Dummy variable regression to test whether principal components are associated with sequencing lanes. Lane 1 was omitted as it was a redundant variable in the regression. Adjusted R-square is highest for the PC 4, as it is also more significantly associated with processing lanes compared to other PCs.

To summarise, even though PCA plots did not show signs of strong batch effects, the regression analyses where individual principal components were associated with batch effects, showed that some of the PCs significantly associated with dates and lanes. Especially some of the processing dates were strongly associated with principal components 4 and 5, while the association was much weaker for the lanes. This is in agreement with what we observed in the previous section when we compared within and between batch sample-to-sample correlations. Given the above, we decided to use the first 10 principal components of the gene expression data to represent batch effects including processing date, sequencing lane and other batch effects that we are not aware of.

2.4 Modular analysis of CoLaus expression data

We performed a modular analysis of the CoLaus gene expression data by using the *Iterative Signature Algorithm* (ISA) [50-52] (see section 1.6.3 for details). The ISA identifies groups of genes, termed *modules* that are co-expressed across subsets of samples. The goal of our analysis was not to discover novel modules of co-expressed genes, as our data have limited sample size and stem from one type of cells growing under the same experimental condition. Rather, we wanted to see whether despite these limitations our data would nevertheless capture relevant groups of genes enriched in certain functional annotations, and whether the samples in which these genes are co-expressed would also point to groups of subjects with some of their clinical phenotypes being unusual.

For the analysis we used RPKM values of 19,903 protein coding genes for 555 samples. We added a constant of 1 to the RPKM values prior to $\log_2$ transformation. Next we z-score normalised the data first among the samples and then among the genes, therefore making the genes and the samples comparable, respectively. Stringency of the modules depend on the two parameters: *row.threshold* and *column.threshold*. These parameters control how far from the mean, in units of standard deviation, genes and samples, respectively, have to be for inclusion in the module. As a result, the higher these parameters are, the more similar are the module genes or samples to each other. Conversely, the lower these thresholds are, the bigger and less coherent the modules become. We explored a parameter space of these thresholds from 1 to 6, with 0.5 standard deviation increments for both gene and sample selection. We selected the direction

parameter as 'updown' to select genes/samples that are either a certain number of standard deviations higher or lower than the average. By setting the *cor.limit* parameter to 0.5, we allowed only those modules to be kept which were correlated with any other modules by a Pearson correlation coefficient less than 0.5. The analysis yielded 204 modules which consisted of 6 to 3,962 genes and 1 to 106 samples.

Following the module discovery in the gene expression data, we wanted to explore the samples and the genes found in the modules. In particular, we were interested in the enrichment of phenotypes for the module samples in order to see the similarities among subjects that are attributed to the same modules. We were also interested in the enrichment of the pathways/diseases of the module genes, to observe the shared function of the genes belonging to a module. To explore the former, we chose 39 phenotypes, including measurements of diastolic/systolic blood pressure, height, weight, waist circumference, hip circumference, BMI, glucose, cholesterol, HDL, LDL, triglycerides, urinary and serum creatinine, heart rate, bioimpedance, several liver enzymes, alcohol consumption, metabolic syndrome, heart failure and CVD risk markers (see Table 4).

To study the enrichment of modules with these phenotypes, we first stratified module samples based on their sex in order to prevent faulty associations due to sex confounding. Next, we used the sample score of each module and tested its correlation with each of the 39 phenotypes. The module score of samples is represented by a vector consisting of numbers between -1 and 1, for each sample, where the absolute value of its magnitude corresponds to the strength of its association with the bicluster. Overall 19 module-phenotype pairs had a Pearson correlation p-value lower than 0.001, involving 13 unique modules.

Next, for these 13 phenotype enriched modules we did gene enrichment analysis using the following libraries in enrichR [121]: GO molecular function, GO biological process, KEGG 2016, OMIM disease, Disease signatures from GEO and WikiPathways 2016. Six out of 13 phenotype-enriched modules showed gene- and phenotype enrichment that are functionally coherent.

Variable name	Transformation	Variable full name	Physiology
age		Age in Years	
sex		Gender (0=male,1=female)	
ht		Height (cm)	
wt	log(p)	Weight (kg)	
bdmsix	log(p)	BMI (kg/m*2)	Body mass index
wst	log(p)	Waist (cm)	
hipcr		Hip (cm)	
bmpsc		Bioimpedance	Body's response to applied electrical current; estimates body fat
hrtrte1	log(p)	Heart Rate 1 (beats/min)	
hrtrte2	log(p)	Heart Rate 2 (beats/min)	Repetetive heart beat measurements
hrtrte3	log(p)	Heart Rate 3 (beats/min)	
bpd1		Diastolic BP (mm Hg) - 1	
bpd2		Diastolic BP (mm Hg) - 2	Repetetive diastolic blood pressure measurements
bpd3		Diastolic BP (mm Hg) - 3	
bps1		Systolic BP (mm Hg) - 1	
bps2		Systolic BP (mm Hg) - 2	Repetetive systolic blood pressure measurements
bps3		Systolic BP (mm Hg) - 3	
adtrn	log(1+p)	ADTRN (%)	Alcohol consumption marker
alb		Albumin (g/L)	Plasma protein produced by the liver; decreased in cases of malnutrition, liver cirrhosis, renal losses, gastro-intestinal losses
alkp	log(p)	Alkaline phosphatase (U/L)	Liver enzyme; marker of biliary obstruction, intra/extran hepatic cholestasis, chronic renal failure, congestive heart failure, infection/inflammation
alt	log(p)	ALT (U/L)	Liver enzyme; marker of liver damage
ast	log(p)	AST (U/L)	Liver enzyme; marker of liver damage
chol		Total cholesterol (mmol/L)	Cholesterol; cardiovascular risk factor
clhdl	log(p)	Total to HDL cholesterol (ratio)	Ratio of total to HDL-cholesterol; marker of cardiovascular risk
cr		CR (micromol/L)	Serum creatinine; marker of renal function
cru	log(p)	CRU (micromol/L)	Urinary creatinine; mainly used to adjust other urinary markers
gamgt	log(p)	Gamma GT (U/L)	Liver enzyme; marker of liver damage
gluc	log(p)	Glycaemia (mmol/L)	Marker of diabetes and metabolic syndrome
hctn	log(p)	HCTN (micromol/L)	Homologue of the amino acid cysteine; non-classical cardiovascular risk factor
hdlch		HDL cholesterol (mmol/L)	Blood lipid, high-density lipoprotein; classical cardiovascular risk factor
ldlch		LDL cholesterol (mmol/L)	Blood lipid, low-density lipoprotein; classical cardiovascular risk factor
trig	log(p)	Triglycerides (mmol/L)	Blood lipid; risk factor for atherosclerosis, heart disease and stroke
uric	log(p)	Uricaemia (micromol/L)	Marker of metabolic syndrome and inflammation; has been associated with the incidence of hypertension and cardiovascular disease and with cardiovascular mortality
conso_hebdo	log(1+p)	Alcohol consumption (units/week)	
insulin	log(p)	Insulin (microU/mL)	Hormone controlling blood glucose levels
ldlsize		LDL size (Angstrom)	Size of LDL particles; small LDL promote the formation of fatty deposits in the arteries
adiponectin	log(p)	Adiponectin (ng/mL)	Adipose tissue hormone; negatively correlated with BMI
leptin	log(p)	Leptin (ng(mL)	Adipose tissue hormone; positively correlated with BMI
probnp	log(p)	Pro B natriuretic peptide	Marker of heart failure; decreases cardiac output and blood volume

Table 4: 39 CoLaus phenotypes that were used in phenotypic enrichment of discovered modules by ISA.

The first module contains samples that were phenotypically similar in their Pro-BNP profiles, a heart failure marker; and genes enriched for amyotrophic lateral sclerosis (in GWAS catalog), a multi-system neurodegenerative disorder that has implications on cardiac function. The second module was phenotypically enriched for glucose; and its genes were enriched for hemochromatosis (in OMIM), which is a disease related to high iron accumulation in the body. It has been reported that more than 50% of the patients diagnosed with hemochromatosis suffer from diabetes due to beta-cell damage caused by high iron levels [122]. The third module contains samples similar in their LDL profile; and genes enriched for multiple sclerosis (in OMIM). It has been shown that patients with relapsing-remitting multiple sclerosis patients have smaller LDL compared to healthy people [123]. The fourth module contains samples enriched for total homocysteine, a non-classical cardiovascular risk factor; and genes enriched for atrial fibrillation (in GWAS catalog) and fibrosis (in OMIM). The fifth module contains samples similar in their heart rate profile; and genes enriched for cardiovascular diseases (in GWAS catalog). Finally, the sixth module contains samples similar in their levels of Gamma GT, a liver enzyme and alcohol consumption marker; and genes enriched in high alcohol use (in GWAS catalog).

To summarize, in this study we demonstrated an application of ISA using gene expression data from LCLs (see Section 1.5.2 for details of the algorithm). Thanks to the biclustering nature of the algorithm, we could study clusters of gene expression for a subset of samples, unlike clustering algorithms where co-expression is always computed over all samples, which for large sets of samples often fails to identify genes whose co-expression is only observed in small subsets of samples, as this is masked by the background noise. In addition to investigating the gene enrichment of the cluster to infer the functional relevance of it, we used the sample specificity of each module to study the phenotypic enrichment of the subjects corresponding to its samples which resulted in some interesting matches. These results not only underline the capacity of ISA to detect functionally relevant biclusters from our LCL expression data but also demonstrate that these data are rich in information reflecting the phenotypic state of the subjects from whom the LCLs were derived from, which was one of our main motivations to perform this analysis.

3 Integration with genotypes

In this chapter I give a brief overview of the studies that investigated the genetic variation in the human genome sequence and the genotyping technologies used for this purpose. Next I introduce genome-wide association studies (GWAS) - a line of research that has emerged thanks to the advancements in the genomics field. I also discuss expression quantitaive trait loci (eQTL) analysis and present the cis-eQTL analysis I performed on CoLaus data. The final two sections of the chapter are dedicated to two collaborations that resulted in publications. The first focused on characterizing the regulatory roles of biologically and trait-relevant lincRNAs (TR-lincRNAs) in human LCLs. My role in this project was to perform genome-wide lincRNA *cis*-eQTLs analysis on CoLaus, which served as a replication study. The second aimed to gain further insights on how genetic variants influence genes and converge on pathways that are essential for complex traits. To this end, the study uses trans-eQTLs and significant polygenic risk score (PGS) - gene expression associations. My role in the project was to perform various QTL analysis on CoLaus data, which served as one of the LCL replication cohorts.

3.1 Genotypes and their measurements in cohorts

Among the 3 billion nucleotides present in the human genome [2], only few vary across a population; while millions of DNA sequence variants were discovered, these only amount to a small fraction of the human genome. Indeed, it has been estimated that common genetic sequence variants, that variants present in over 1% of the population, called *single nucleotide polymorphisms* (SNPs) account for close to 90% of the genetic variation among humans [124]. Other sequence variations of the human genome include rare variants, copy number variants (CNVs) and structural variants. Genotyping is a method to detect the differences in the genetic make-up of an individual, so-called genotype, by screening different genetic variants they carry in their genome.

To date, many efforts have been made to study genetic variation in genome sequence with a particular focus on SNPs. The international HapMap project [67] was a pioneering project that not only studied the genetic variation across human populations but also defined high-resolution haplotype maps. A haplotype block defines a set of SNPs inherited together due to linkage

disequilibrium (LD) and proved to be useful when choosing the most informative subsets of SNPs, known as *tagging* SNPs, to design arrays used in genetic screening of the populations [125]. A more recent initiative, the 1000 Genomes project [126], aimed to more deeply characterise the human genome variation by extending genotyping to rare variants (allele frequencies below 1%).

In the past decades, a great deal of effort has been made to develop accurate, fast and cost effective genotyping methods to achieve population level screenings. Several genotyping methods exist, each with their own benefits and drawbacks yet hybridization to a solid array, known as DNA microarrays, remains as one of the simplest techniques for high-throughput genotyping. DNA microarrays, also known as SNP arrays, rely on hybridization of single-stranded DNA fragments of samples to SNP arrays that contain thousands of unique probe sequences. Marker densities of SNP arrays have been increasing throughout the last decades and nowadays a wide range of SNP arrays are available in the market that are custom built for different human populations and have different coverages to better suit the needs of the studies. Meanwhile, *next generation sequencing* (NGS) technologies have been emerging making the acquisition of billions of DNA sequences cheaper and faster than previously anticipated [4]. NGS offers all the premises of SNP arrays with the added benefits of greater resolution and accuracy. Use of paired-end reads in NGS allows to discover additional CNVs that would otherwise be missed, and to discover structural variants such as inversions and translocations that would be entirely missed by SNP arrays [5].

To be able to detect more loci, while NGS requires deeper sequencing, SNP microarrays require denser probes. To compensate for the limited number of tag-SNPs present in microarrays, *genotype imputation* is used to predict unmeasured genotypes based on the reference haplotype panels. Currently many national large-scale whole-genome sequencing projects are realized all over the world including UK [127], Japan [128], Netherlands [129] and Singapore [130], discovering more population specific genetic variants and forming more accurate reference haplotype panels.

Regardless of the choice of genotyping platform, genotypes are at the center of biomedical research and there is no reason to believe that they will lose their popularity in the near future, given that we describe the present-day as the *genomics era*, where genomic information is not a limiting factor for discoveries anymore.

Genome-wide association studies (GWAS) is a line of research that makes extensive use of these genetic variants. A GWAS employs genomics data acquired for a large collection of samples to search for SNPs that are associated with certain traits or diseases [5]. Due to their abundance and high density, SNPs have been ideal polymorphic markers for these association studies. To date, GWAS improved our understanding of the genetic basis of many complex diseases including multiple sclerosis [12], Crohn's disease [9], diabetes [14], cancer [10, 11]; and schizophrenia [13]. Several common variants influencing the continuous traits such as lipids, height and fat mass have also been found [15-17]. While GWAS have identified thousands of common variants that are associated with complex traits [8], the regulatory mechanisms behind these associations mostly remain poorly understood. Pinpointing causal variants is difficult, since the lead variants associated with a trait are often in high linkage disequilibrium (LD) with other variants in the same region with only slightly lower association signal. Such associated LD blocks typically contain several genes or functional elements, preventing the accurate identification of causal variants. Furthermore, some trait associated variants fall into intergenic regions of the genome with no obvious functional role at all [18]. Nevertheless the knowledge on the LD structure around the SNPs has been crucial to fine-map the trait associated SNPs and predict disease related genes [131].

In the meantime, a number of studies reported that trait associated genetic variants are significantly enriched in expression quantitative trait loci (eQTLs), suggesting that many trait associated variants affect the phenotype by altering gene expression [20-23]. There is also a growing body of literature highlighting the more pronounced effects of genetic variants on molecular traits compared to phenotypic traits [24-26]. This is not surprising as molecular traits representing fundamental biological processes such as gene expression are intermediate in the genotype to trait causality chain.

Genetic variants can affect gene expression in two ways: they could affect the expression of the gene which is nearby (*cis*-eQTL) or expression of the gene that is further away or is even on different chromosome (*trans*-eQTL). The *cis*-eQTL is potentially useful to pinpoint the true disease gene from an associated locus implicated by GWAS. The *trans*-eQTL however, allows us to identify the downstream affected disease associated genes which were not implicated by GWAS studies at all, thereby potentially revealing previously unknown pathways.

3.2 *Cis*-eQTL analysis of CoLaus

Genotypes of CoLaus were measured by using the Affymetrix GeneChip Human Mapping 500 K array set. The full set of unmeasured HapMap II SNPs (release 21) was imputed and expected allele dosages were computed for 2,557,249 SNPs. Gene expression data was generated by Illumina HiSeq2000 platform and mapped onto human genome 19 (Genome Reference Consortium Human Build 37 (GRCh37)), resulting in RNA-Seq profiles of 45,470 genes. eQTL analysis was done on 555 subjects whom we had genotype and gene expression data.

For the *cis*-eQTL analysis we considered 19,903 protein coding genes and further removed genes which had RPKM value smaller than 1 for 10% or more of the samples. Motivation in doing so was to eliminate problematically distributed gene expression values, which then could give rise to faulty association signals. By applying this criteria, 49% of the protein coding genes were removed from the analysis. Next we removed the genes that did not distribute continuously between 2.5th and 97.5th percentiles, in order to exclude ill-distributed genes that can cause biased inferences in the downstream statistical analysis. This criteria removed an additional 6% of the protein coding genes. Overall 8,924 protein coding genes were selected for the analysis.

For each gene, we defined a *cis*-window ranging 500kb out from the gene mid-point, and calculated the Spearman rank correlation between the gene expression value and the SNPs within the corresponding gene's *cis*-window. For each gene only the SNP with minimum Spearman rank correlation p-value was recorded (p_{obs}). Subsequently, we accounted for multiple hypothesis testing by controlling Benjamini-Hochberg false discovery rate at 5% and found that almost all genes had an eQTL. As shown in Figure 12 for chromosome 19, we see that the p-value distribution of the *cis*-SNPs - gene expression correlations significantly deviates from the standard uniform distribution while we do not observe this for the non-cis-SNPs - gene expression corrleations. When the performed statistical tests are independent, p-values are expected to follow a standard uniform distribution. Yet, in the eQTL analysis this assumption does not hold due to the correlation structure in both genotype (LD) and gene expression data (co-expression). To account for this we used a type of non-parametric randomization test called *permutation test* where the expected null distribution of the p-values are calculated and the likelihood of observed p-values originating from this computed null distribution is assessed

along with the multiple testing correction method of choice. In our case, we estimated the null distribution of the p-values by randomly permuting the sample labels of the expression data 1,000 times and recording the most significant SNP - gene pair Spearman rank correlation p-value each time (p_{exp}). For a given gene, we considered the highest correlated *cis*-SNP a *cis*-eQTL, if p_{obs} was smaller than 95% of the p_{exp} values calculated for the given gene. As described the significance of the *cis*-eQTLs are therefore reported on the basis of permutation adjusted BH-FDR of 5%.

Figure 13 shows the distance between gene-midpoint and the discovered *cis*-eQTLs. We found that the majority of the *cis*-eQTL SNPs map within 100 kb of the gene-midpoint.

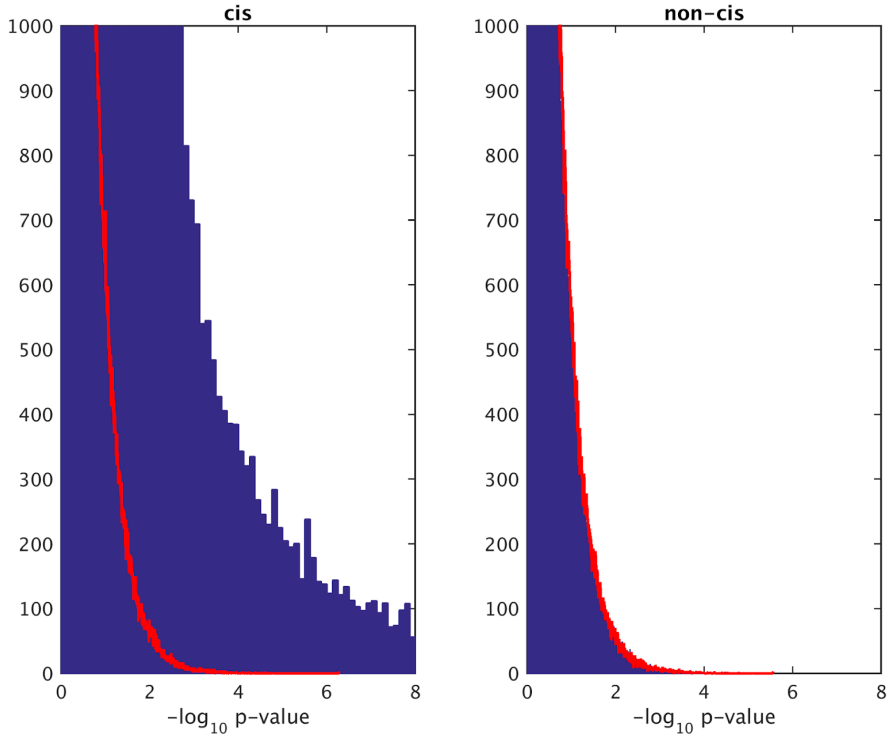


Figure 12: Distribution of $-\log_{10}$ p-values coming from SNPs association with gene expression values. Left panel shows all chromosome 19 genes' Spearman rank correlation p-values with their respective *cis*-SNPs. Right panel shows the same genes' Spearman rank correlation p-values with randomly selected SNPs outside their *cis*-windows. Red line in both panels represents the $-\log_{10}$ of uniform distribution, the expected distribution of p-values under null hypothesis.

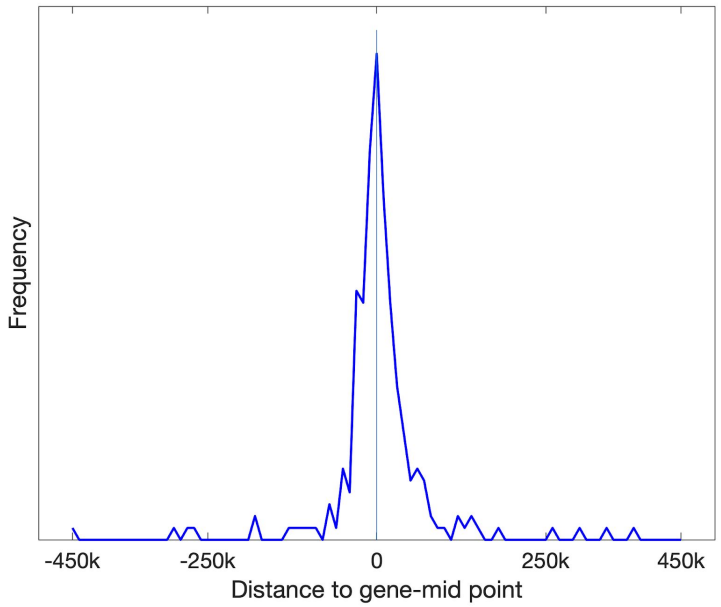


Figure 13: Distribution of *cis*-eQTLs distance to gene-midpoint.

Out of 8,643 genes analysed 3,811 were found to have a significant eQTL at permutation-adjusted BH-FDR of 5%. Figure 14a shows the proportion of the genes that have eQTL by chromosomes. Chromosome 19 had the lowest number of genes with a *cis*-eQTL (34%), whereas chromosome 21 had the highest rate (59%). On average we found 44% of the genes having an eQTL in CoLaus LCL derived gene expression data. Out of 1,655,025 SNPs analysed, 165,944 SNPs found to be an eQTL at permutation adjusted BH-FDR of 5%. Figure 14b shows the proportion of the eQTLs by chromosomes. Chromosome 13 had the lowest eQTL rate (5%), whereas chromosome 19 had the highest eQTL rate 18%. On average we observed 10% of the analysed SNPs being an eQTL.

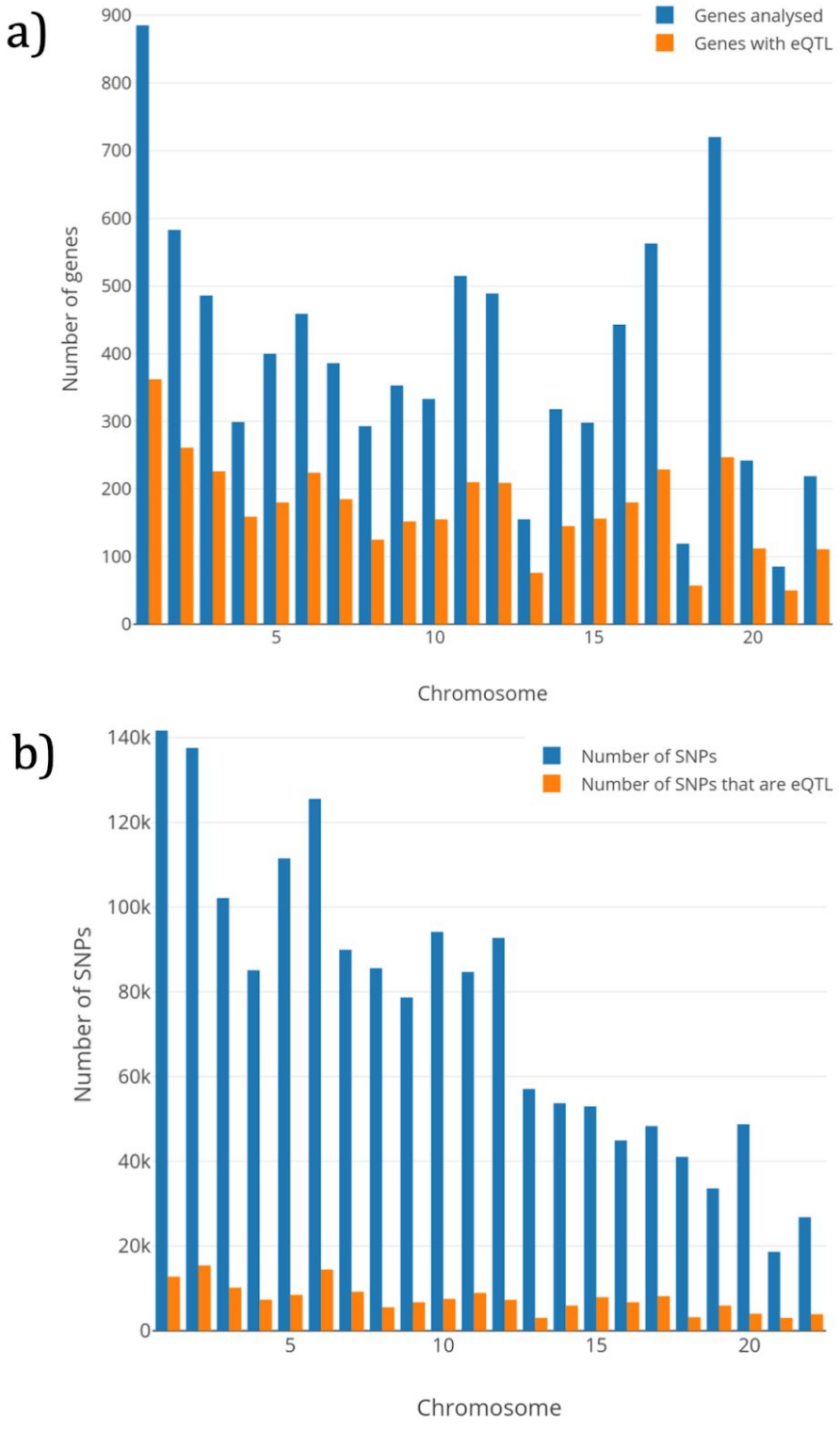


Figure 14: a) Number of genes with an eQTL reported by chromosome b) Number of eQTLs reported by chromosome.

We also investigated whether highly or lowly expressed genes tend to have more *cis*-eQTLs. To investigate this, we split genes on chromosome 19 into five quantiles according to their mean RPKM values. As shown in Figure 15, we found lowly expressed genes to be more likely to have a *cis*-eQTLs.

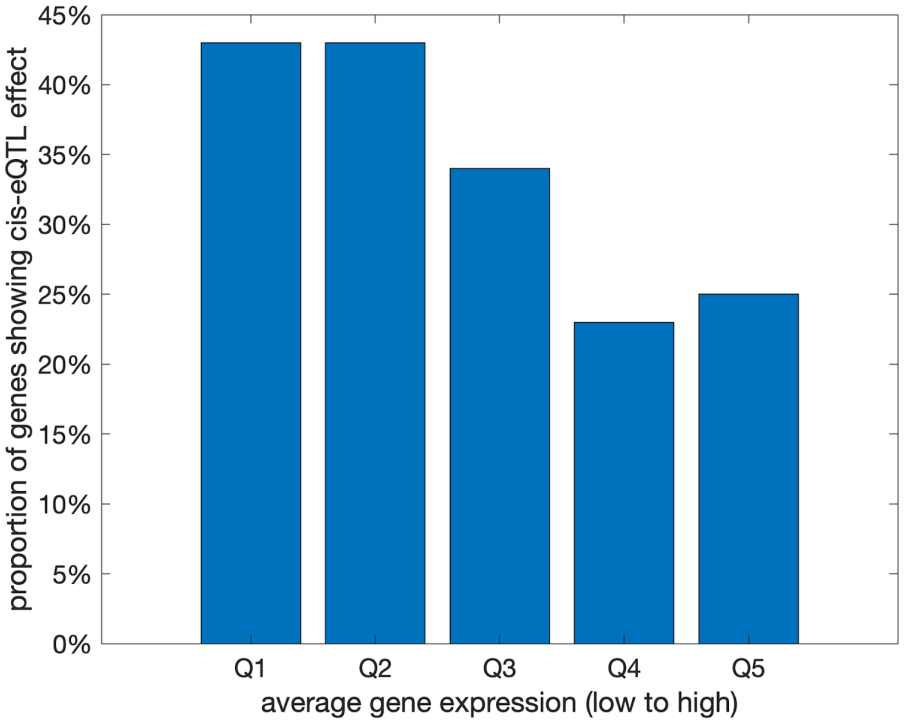


Figure 15: Percentage of genes with *cis*-eQTLs, stratified into five quantiles by mean RPKM values (lowest quantile Q1 to highest quantile Q5).

3.2.1 Comparison with other studies

We compared CoLaus *cis*-eQTL results with two other studies. The first study is Blood eQTL Browser [132] which is an eQTL meta-analysis of non-transformed peripheral blood samples based on microarray gene expression data. For the comparison of discovered eQTLs, first we wanted to define the set of SNPs and the genes analysed in both Blood eQTL Browser and CoLaus studies. As genotypes were imputed to Hapmap Phase 2 in both studies, there was no need to identify the overlap in SNP space. On the other hand, the sets of genes analysed in the two of the studies differed, with an intersection counting 8,643 genes. Among these common

genes, Blood eQTL Browser identified eQTLs for 3,562 genes, while CoLaus identified 3,786 genes (see Figure 16). 1,806 of these genes with an eQTL overlapped between two studies. This overlap is significant, with a Fisher's exact test p-value (over representation) of 1.81 x 10^{-27}).

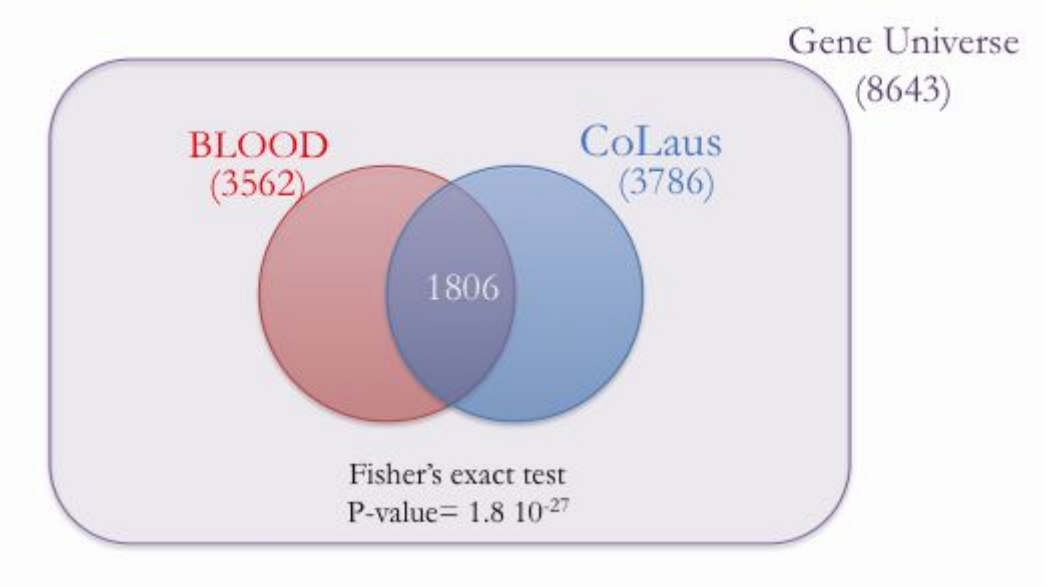


Figure 16: Overlap between BLOOD and CoLaus; BLOOD has 3562 and CoLaus has 3786 unique genes that have a significant eQTL, while 1806 of these eQTL overlap.. The gene universe refers to the 8643 genes that were included in the *cis*-eQTL analysis of both CoLaus and BLOOD.

The second study we compared our results to is GEUVADIS [70], an eQTL analysis reported in LCLs using RNA-Seq technology. We restricted our analysis to genes and SNPs present in both studies resulting in 1,515 genes in GEUVADIS and 3,811 genes in CoLaus with at least one significant eQTL. We found 1,343 genes overlapping between two studies and the enrichment of the overlap was very significant (Fisher's Exact Test, right tail 'over representation' p-value = numerically zero). Figure 17 shows the Venn diagram representation of the *cis*-eQTL comparison. Given the difference in batch effect treatment in GEUVADIS and CoLaus, where GEUVADIS uses a sophisticated batch effect detection method called PEER [133] as opposed to CoLaus not attempting to remove any batch effects, a remarkable 89% of GEUVADIS eQTLs were replicated in CoLaus.

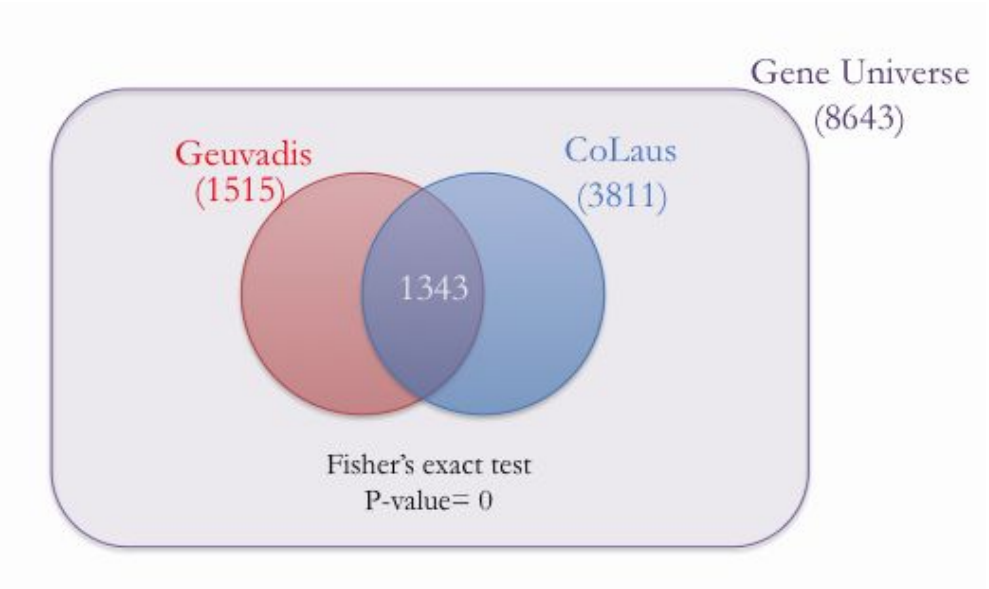


Figure 17: Overlap between GEUVADIS and CoLaus; GEUVADIS has 1,515 and CoLaus has 3,811 unique genes that have a significant eQTL, while 1343 of these eQTL overlap. The gene universe refers to the 8643 genes that were included in the *cis*-eQTL analysis of both CoLaus and GEUVADIS.

The fact that GEUVADIS and CoLaus both use LCLs and Illumina RNA-seq, is the likely reason why we observed a larger overlap with GEUVADIS compared to Blood eQTL Browser which uses blood tissue and microarray technology.

3.2.2 Conclusions

We selected a rather small set of protein coding genes (8,643 genes) that passed strict QC criteria and ran a *cis*-eQTL discovery analysis for these genes. We found an eQTL for 44% of these genes in CoLaus LCL derived gene expression data (BH-FDR of 5% against a permutation derived null distribution). The proportion of protein coding genes with *cis*-eQTLs reported as 27% in the literature [134], lower than what we observe in our study. It is very likely that we observed higher percentage of genes having eQTL as the set of genes considered for the *cis*-eQTL analysis were less to begin with, due to stringent QC we used in our analysis. On average we observed 10% of the analysed SNPs being an eQTL, however this number largely affected by the density of genotype imputation and the definition of *cis*-window. We found the majority of the *cis*-eQTLs to be in the 100 kb neighborhood of the gene-midpoint, in aggreement

with published results [135]. We observed lowly expressed genes having more *cis*-eQTLs compared to highly expressed genes. This result is contradictory to what has been reported in the literature by Westra et al. [132], yet we think the observed disagreement might be due to the different gene expression technologies used in two studies. While Westra et al. uses microarray technology where the low expression values are known to be noisy, CoLaus uses RNA-seq technology which is known to detect lowly expressed genes more accurately.

When we compared our *cis*-eQTL results with two studies, Blood eQTL browser and GEUVADIS, we found that the CoLaus *cis*-eQTLs are significantly overrepresented in the other studies, demonstrating the robustnes of our *cis*-eQTL analysis.

3.3 *cis*-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture

In this chapter, I describe the above titled project I have been involved in.

3.3.1 Background

It has been shown that despite 75% of the human genome being transcribed, only a small proportion of these transcripts are translated into proteins [136]. The transcripts without protein-coding capacity are called non-coding RNA and they are classified into two categories based on their length. RNAs that are shorter than 200 nucleotides are known as small non-coding RNAs and they mainly consist of microRNAs, small nucleolar RNAs. RNAs that are longer than 200 nucleotides are known as long intergenic non-coding RNAs (lincRNAs). The long nature of the lincRNAs serves as an added functionality, by allowing them folding upon themselves and creating complex structures. As a result their interactions are not solely based on base-pairing but also the tertiary structure [137]. Due to their lack of protein coding ability, lincRNAs were formerly seen as 'junk RNA'. However this perception has been drastically changed after many studies showing their regulatory role in disease and normal phenotypes [138-140]. A large proportion of GWAS variants fall into non-coding regions of the genome and they are enriched for eQTLs [141]. To date, eQTL analysis of many protein-coding genes helped us to better understand how these genetic variants affect the human complex traits [76, 142].

Recently similar studies have been done on lincRNAs and like their protein-coding counterparts, lincRNAs were also found to contribute to phenotypes [70, 134, 143, 144].

3.3.2 Scope of the project

Although many studies associate lincRNAs with complex traits, the underlying mechanism of action of these associations remain largely unknown. So far the functional role of lincRNAs has been shown for regulation of epigenetic markers and gene expression; in particular in transcriptional and post-transcriptional regulation [145-147]. Marques et al. [148] reported chromatin signature at lincRNA transcription start sites exhibiting differences depending on the functional role of lincRNA. More specifically they show the lincRNAs with enhancer-like chromatin signatures correlating more often with neighboring protein-coding genes, thus pointing to local acting regulatory function of lincRNAs. They also remark the trait associated eQTL variants being enriched for enhancer regions[149], and accordingly suggest a link between enhancer associated lincRNAs and complex human traits.

Aim of the current study is to extensively characterize the regulatory roles of biologically and trait-relevant lincRNAs (TR-lincRNAs) in human LCLs. Biological relevance of these TR-lincRNA was decided upon them being conserved in recent human history and genetic interactions with other trait-associated loci.

3.3.3 My contribution

My role in this project was to do genome-wide lincRNA *cis*-eQTLs analysis on CoLaus LCL gene expression data, which served as a replication study.

For the analysis we considered two types of lincRNA: lincRNAs that were in GENCODE version 19 and de-novo LCL expressed lincRNAs that were detected in the discovery cohort. First we used HTSeq [150] to quantify RNA-Seq reads that overlap lincRNAs (GENCODE version 19 + de-novo LCL expressed lincRNA) and protein coding genes (GENCODE version 19). Expression level of each gene for each sample was subsequently estimated as RPKM by mapping the total number of exonic reads of the gene. Next, the genes that were unexpressed across the population were removed from the analysis by discarding the genes that had zero RPKM values upto half of the samples. Later we performed PEER normalization to account for

potential technical variation across samples [151]. PEER-corrected expression values were then transformed to standard normal distribution.

The *cis*-eQTL analysis was performed for genome-wide significant trait-related autosomal SNPs ($p< 5\times10^{-8}$; [152]) located within two MB window centered on the transcription start site (TSS) of each expressed lincRNA and protein-coding gene. For the analysis we used Pearson's correlation coefficient, rho (r), to estimate the association between PEER-corrected and transformed expression values and trait-associated SNPs. Global significance of the associations were assessed by using permutation adjusted false discovery rate. More specifically we permuted the gene expression values of each gene 1,000 times and recorded the maximum absolute rho value in each permutation (r_{exp}). Then only the *cis*-eQTLS with r_{obs} higher than 95% of all r_{exp} values were considered to be significant (FDR controlled at 5%).

3.3.4 Results and conclusions

In the discovery cohort 111 and 1,479 *cis*-eQTLs were detected for 73 lincRNAs and 756 protein coding genes respectively. Despite the number of lincRNA *cis*-eQTLs being relatively low compared to protein-coding gene *cis*-eQTLs, authors showed when the differences in gene lengths and expression level in two groups are taken into account, both groups have indinguishable proportion of eQTLs ($p=0.68$, two-tailed χ^2 test) suggesting that lincRNA properties limit the power to detect eQTLs.

Overall 68% of the identified lincRNA *cis*-eQTLs and 71% of the identified protein-coding *cis*-eQTLs were replicated in CoLaus. The proportion of eQTLs in both groups remained similar ($p=0.69$, two-tailed Fisher's exact test). Additionally, authors adopted *regulatory trait concordance* (RTC) method to reduce false positive eQTL detections by taking into account local LD structure [153]. RTC method assesses the likelihood of the identified *cis*-eQTL to be most likely driven by the complex trait associated genetic variant and not due to LD with another SNP. The high confidence lincRNAs and protein-coding genes resulting from this analysis are likely to be true trait-relevant gene candidates and they are called trait-relevant lincRNAs (TR-lincRNA) and trait-relevant protein-coding genes (TR-pcgenes). Interestingly, when the TR-lincRNA and the TR-pcgenes were considered together, 73% of the GWAS *cis*-eQTLs were replicated in CoLaus.

Authors did additional analysis and illustrated following: i) TR-lincRNA are enriched for enhancer-like chromatin signatures ii) TR-lincRNA interact with nearby TR-pcgenes iii) TR-lincRNA are enriched at topologically associated domain (TAD) boundaries. Taken together, they suggest TR-lincRNAs are likely to regulate proximal trait-ralated gene expression in *cis* by modulating local chromosomal architechture.

The paper entitled, *cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture*, was published in Cell Reports in February 2017. The manuscript is accessible in Appendix 1 and also online: http://dx.doi.org/10.1016/j.celrep.2017.02.009.

3.4 Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis

In this chapter, I describe the above titled project I have been involved in.

3.4.1 Background

Despite the widespread use of *cis*-eQTLs to interpret the regulatory mechanisms of GWAS variants, they explain only a modest proportion of the disease heritability [132]. *Trans*-eQTLs however used to provide insights of the effects of a single genetic variant on many genes and they have been already successfully used to elucidate putative key driver genes involved in diseases [154]. Yet the discovery of *trans*-eQTLs require larger samples as their effects tend to be much weaker compared to those of *cis*-eQTLs.

While identifying the downstream effect of a genetic variant with *trans*-eQTL analysis is quite straightforward, other approaches needed to combine the consequences of trait-associated variants. Polygenic risk scores (PGS) have been proposed to sum up the effects of individual variants contributing to a disease and used to comprehensively characterise the overall genome-wide risk of a disease [155]. As much as PGS being useful, the way polygenic effects manifest themselves still remain largely unknown.

3.4.2 Scope of the project

This paper systematically investigates the *trans*-eQTLs along with associations of PGS with gene expression (expression quantitative trait score, eQTS). The aim of the work is to gain further insights on how genetic variants influence genes and converge on pathways that are essential for complex traits. To maximise the power to detect *trans*-eQTL and eQTS effects, study combines data from 37 cohorts reaching to 31,684 blood samples in the context of eQTLGen consortium. The meta-analysis achieves a six-fold increase in size over the previous large-scale studies [132].

3.4.3 My contribution

CoLaus was one of the LCL replication cohorts along with ALSPAC [156], MuTHER [157] and Geuvadis [70], which were meta-analysed together. I used the recommended RNA-Seq pipeline to replicate their various QTL analysis, which consisted following steps:

Step 1 - Preparation of expression data: The suitability of the CoLaus RNA-Seq quantification for the replication analysis was discussed with the authors. We concluded that the quality of the sequencing, performed sample quality control and mapping parameters were well suited for the purpose of the replication study.

Step 2 - Preparation of genotype data: I first computed 4 MDS components of the non-imputed autosomal SNPs of CoLaus, to use for correction of expression data in the downstream statistical analysis. Genotype Harmonizer [158] was used to harmonize, filter and convert the genotype data to match the GIANT release of 1000G.

Step 3 - Formatting expression data: First I investigated the outliers of the expression data by performing a PCA on the data without prior centring or scaling. As seen in Figure 18, by plotting the first two principal components, I haven't detected any clear outliers, therefore did not remove any samples from the analysis.

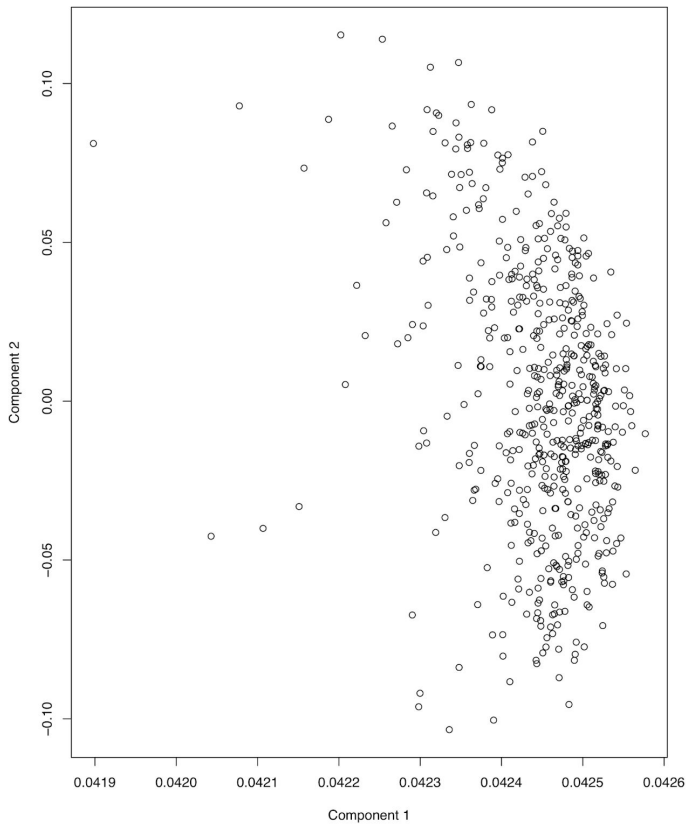


Figure 18: Principal components of gene expression data prior to transformations. Plotting components 1 vs 2 suggesting none of the samples as clear outliers.

In the next step I applied a series of normalizations to the expression data. I started with TMM (Trimmed Mean of M-values) normalization as implemented in edgeR package [115], removed genes with no variance, applied logarithm 2 and Z-score transformation (centering and scaling the genes). Finally the first 4 MDS of the genotype data was removed from expression data to account for possible population stratification.

Step 4 - MixupMapper: As demonstrated in Westra et al. [159], sample mix-ups often occur in genomic datasets. By analysing five publicly available human genomics datasets they found 3% to 23% of the samples being assigned to wrong expression phenotypes. MixupMapper calculates the predicted gene expression values solely based on on the genotype of the *cis*-SNPs and compares the predicted gene expession values with the observed gene expression values to

detect anomalies. Using the distance measures the method is able to detect and correct sample mix-ups with high specificity and sensitivity [159].

I used MixupMapper on CoLaus genotype and gene expression data and found five problematic samples. Figure 19 shows the best matching expression file per genotype file. The strength of original genotype-expression matching samples (on the x-axis) is plotted against the strength of best-matching genotype-expression samples (on the y-axis), where the smaller numbers represent more likely matches. Examining Figure 19, I detected one obvious mix-up, where the best matching trait score was much smaller than the original linked trait score, indicating a strong probability of one-to-one mixup (indicated as empty red circles). Indeed, looking into sample identifiers I realised they differed from each other only in one digit, in one sample it being 'O' (AA02JWO) and in the other one it being '0' (AA02JW). I fixed this one-to-one mix-up. There were some other genotype-expression pairs which were flagged as mix-ups, in the upper right corner of Figure 19, where some expression dataset seemed to match slightly better to the given genotype file, but the difference was not dramatic. For those expression files, I checked the strength of the link with original, assigned genotype files (highlighted as red dots in the diagonal) and as they had relatively smaller scores I concluded that the original assigned matches for those genotype files being more trustworthy than the match proposed by the tool. Thus, I did not remove these samples.

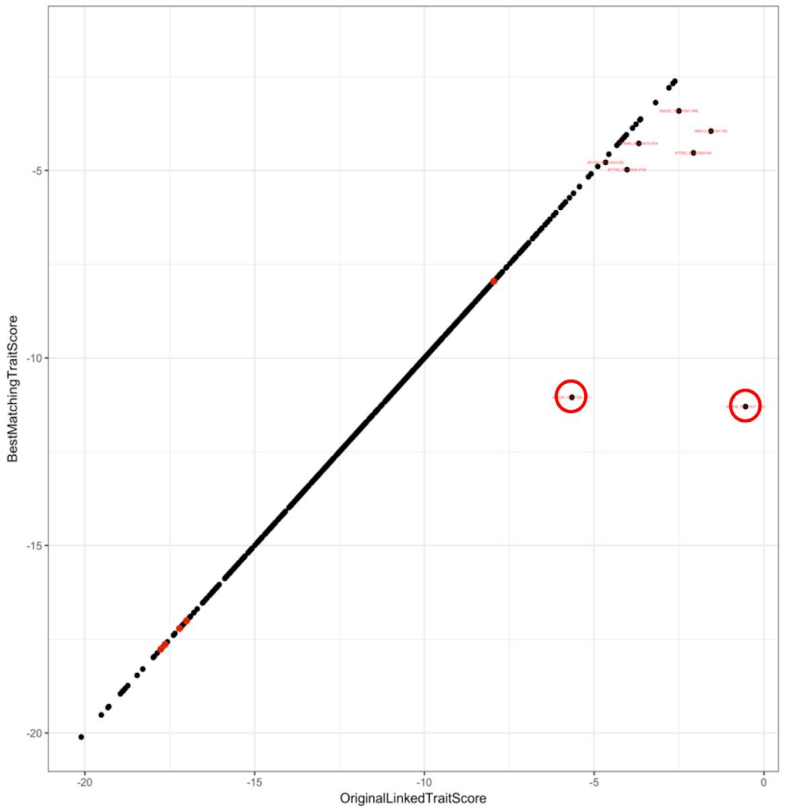


Figure 19: Best-matching expression file per genotype file. X-axis denotes the strength of original genotype-expression matching and y-axis denotes the strength of best-matching genotype-expression samples. Two samples detected as one-to-one mix-ups, are marked with empty red circles.

Next I investigated a similar plot, where this time the best matching genotype file was given per expression file (see Figure 20). I could detect the previously mentioned very strong and obvious mix-ups (empty red circles), however based on Figure 20 there were three additional mix-ups where the originally assigned genotype file was not the best-matching one (empty blue circles). I removed these three samples from the analysis. Additionally, there were some samples flagged as mix-ups in the upper right corner of Figure 20, but the difference was not very big so I did not remove these samples.

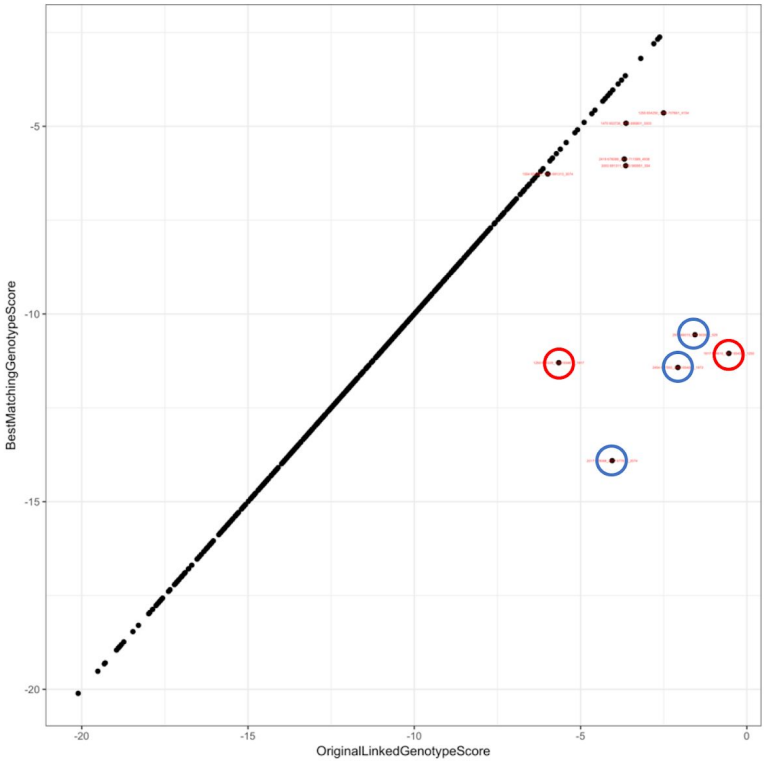


Figure 20: Best-matching genotype file per expression file. X-axis denotes the strength of original expression-genotype matching and y-axis denotes the strength of best-matching expression-genotype samples. Two samples detected as one-to-one mix-ups are marked with empty red circles while additional three samples detected as mix-ups are marked with blue circles.

Step 5 - Removing principal components: Before the QTL analysis, removal of first 20 principal components of expression was recommended to increase the power to detect *cis*- and *trans*-eQTLs. However we wanted to avoid removing principal components of expression which explained genetic variation. To this end, we associated the first 20 principal components of expression with genotypes and found the principal component 17 and 20 as significantly associating with genotype at FDR of zero. We removed all but 17th and 20th principal components of expression from the gene expression data.

Step 6 - Performing the *cis*-eQTL analysis: For genotypes I used 1000 Genomes imputed SNPs with the following criteria: Hardy-Weinberg equilibrium P-value $>10^{-4}$, MAF > 0.01 and call rate > 0.95. *Cis*-eQTL analysis was performed on CoLaus RNA-Seq gene-level quantified and normalized data. For *cis*-eQTL mapping, the maximum distance between the SNP and the

middle of the gene was 1,000,000 bp (=1 Mbp). To control for multiple testing I performed 10 permutations, thereby shuffling sample labels to calculate the FDR controlled at 0.05.

Step 7 - Performing conditional *cis*-eQTL analysis: Authors noted that removing the *cis* effects greatly enhances the power to detect trans effects. To maximize the power to identify *trans* effects, I wanted to regress out all the independent *cis*-eQTL effects. To do so, I performed an iterative conditional *cis*-eQTL analysis which enabled the identification of secondary/tertiary/quaternary/... *cis*-eQTLs. In the first round, a *cis*-eQTL analysis is performed as in the discovery *cis*-eQTL analysis (Step 6). For all significant (FDR<0.05) *cis*-eQTL effects, the most significant SNP effect for each gene was regressed out from the gene expression matrix. The next round of *cis*-eQTL mapping analysis was conducted on the adjusted expression matrix while testing only genes that had any significant *cis*-eQTL effect prior regression. The analysis was performed iteratively until no significant (FDR<0.05) effects remained for a given gene. In CoLaus gene expression data it took nine iterations and resulted in removing the effects of 13,934 SNPs in order to conditionaly remove entire *cis*-eQTL effects from the gene expression data.

Step 8 - Performing the *trans*-eQTL analysis: Two different versions of the expression data were used in *trans*-eQTL analysis in order to observe the effect of principal component correction on eQTL discoveries. Used gene expression data were: The gene expression data where the first 18 non-genetic expression principal components were removed and the expression data where no principal components were removed. In both cases, the effect of *cis*-eQTLs found in Step 7 were used to correct the expression matrices before *trans*-eQTL mapping. Standard settings for the *trans*-eQTL mapping were: Hardy-Weinberg equilibrium P-value > 10^{-4}, MAF > 0.01, and call rate > 0.95. Due to the limitations on the computational power and file sizes I tested only a preselected list of SNPs in this phase, consisting of EBI GWAS Catalogue and Immunobase (all GWAS catalog SNPs with p-value < 5×10^{-8}). For *trans*-eQTL mapping, the minimum distance between the SNP and the middle of the probe was 5,000,000bp (=5 Mbp). To control for multiple testing, I performed 10 permutations by shuffling sample labels, to calculate the false discovery rate (FDR) at 0.05.

Step 9 - Performing the Polygenic Risk Score eQTL analysis: I calculated polygenic risk scores (PRS) for all the individuals in the eQTL dataset, using publicly available GWAS summary

statistics for traits specified by the authors. Those risk scores were then correlated with expression levels of all genes to identify novel trait-associated "hub" genes and pathways. PRS - *trans*-eQTL analysis was done for two different expression matrices similar to Step 8, principal component corrected expression matrix and not corrected expression matrix, again with the motivation to observe the effect of principal component correction on eQTL discoveries.

3.4.4 Results and conclusions

64% of the identified *trans*-eQTL SNPs in the discovery cohort have been previously associated with blood composition phenotypes [160]. This was expected as SNPs that regulate the abundance of a specific blood cell type were also expected to have *trans*-eQTL effects on genes, especially expressed in that cell type. To disentangle this, authors wanted to distinguish the blood cell type specific *trans*-eQTLs from the *trans*-eQTLs caused by intracellular mechanisms. To this end they acquired eQTL data from LCLs, induced pluripotent cells, several purified blood cell types and blood DNA methylation data. CoLaus data was meta-analysed with two other cohorts that had LCL gene expression data to study the profoundness of this cell type specific effect.

Replication efforts resulted in 3,853 significant *trans*-eQTLs that are replicated in at least one of the methylation or cell type specific data including LCLs. Replication rate corresponded to 6.4% of the entire set of discovered *trans*-eQTLs. Authors denoted this set of *trans*-eQTLs as *intracellular eQTLs*, and suggested them being less likely to be driven by cell type composition. They also acknowledged the replication effort as conservative due to limited sample size of the replication study (N=1,460). More specifically when only LCL data was considered, the replication rate of the *trans*-eQTLs was as low as 0.6%, where the 88% of the replications agreed on the direction of the effect (see Figure 21).

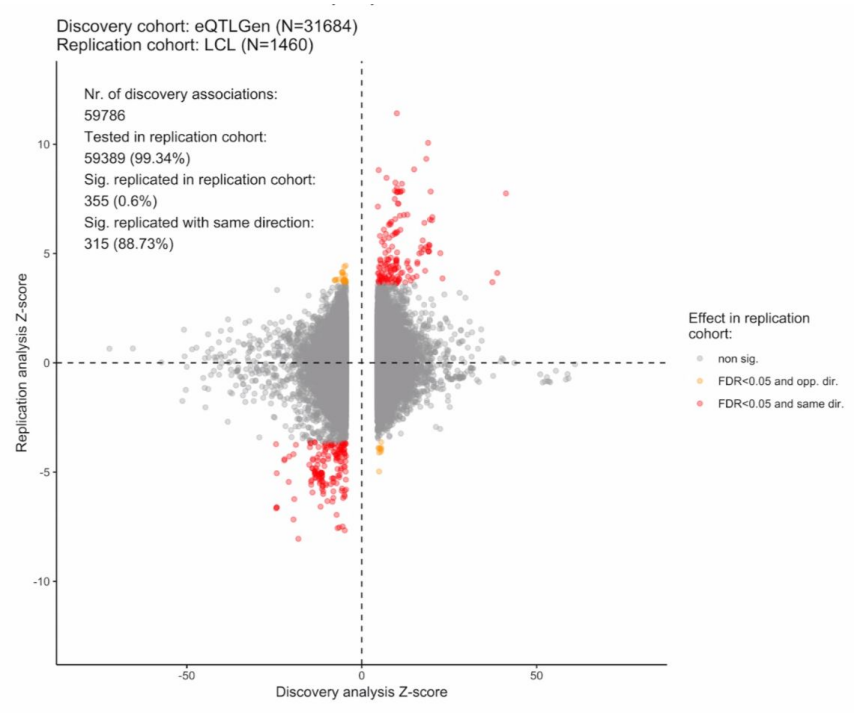


Figure 21: *Trans*-eQTL replication in purified cell type *LCLs*. Z-score comparisons between discovery and replication dataset and corresponding statistics are shown (taken from Vosa et al.[135], Extended Data Figure 11).

To further validate the discovered blood *trans*-eQTLs in purified cell types, similar analysis was done on GTEx data where 32 cell types and cell lines were analysed in total. Replication rate of blood *trans*-eQTLs in non-blood tissues of GTEx data was also very low ranging from 0% to 0.03%. The enrichment of *trans*-eQTLs in LCLs in particular, ranked as the lowest among all the cell types analysed, with the replication rate of 0%. Liver, kidney and non-sun exposed skin however, were among the tissues that were most enriched for blood *trans*-eQTLs.

Another focus of the paper was performing an association analysis between PGS and gene expression levels, so-called PGS - *trans*-eQTL analysis, in order to find eQTS (expression quantitative trait score). Rationale behind this approach was that when the expression level of a gene is associated with PGS of a certain trait/disease, downstream *trans*-eQTL effects of the individual genetic variants would converge on the very same gene, therefore highlighting the gene as the driver of the trait/disease. The meta-analysis resulted in the discovery of 18,210 eQTS effects (FDR<0.05) of which only 10 replicated in LCL replication cohorts (see Figure 22). Out

of 10 replicated eQTS, nine agreed on the direction of the effect between blood tissue and LCL. All things considered the very low replication rate of blood eQTS in non-blood tissues, lead to the conclusion of these effects being highly cell-type specific.

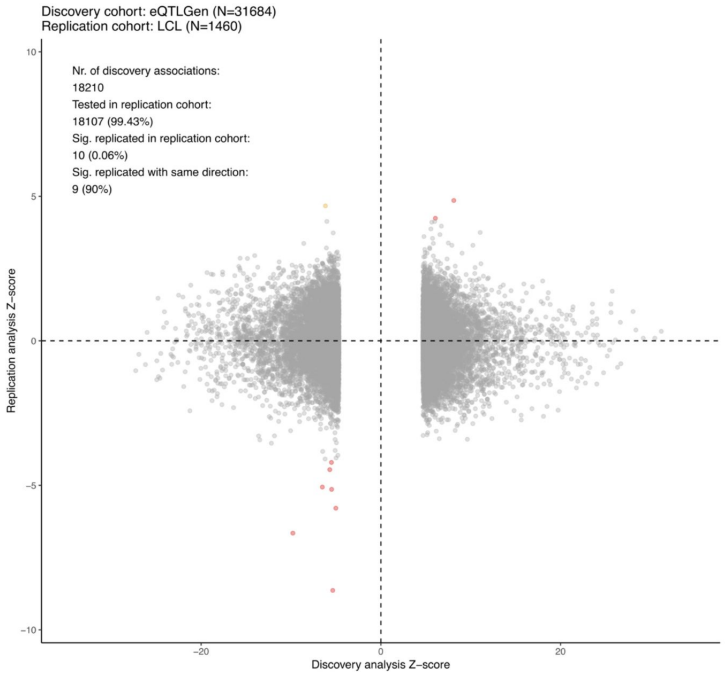


Figure 22: Replication of eQTS results in LCLs. Z-score comparisons between discovery and replication dataset and corresponding statistics are shown (taken form Vosa et al.[135], Extended Data Figure 16A).

The undertaken work has many other remarkable results in addition to the cell-type specific effects of QTLs which the CoLaus data have contributed. Authors found *cis*-eQTLs for 16,989 genes, *trans*-eQTLs for 6,298 genes and eQTS effects for 2,568 genes out of 19,960 genes studied. When specifically looking at the genes expressed in blood, 88% of the genes had significant *cis*-eQTL and 32% of the genes had significnat *trans*-eQTL effects. They observed 84% of the strong *cis*-eQTL SNPs mapping within 20 kb of the transcription start or end site of the genes, suggesting variants close to start or end of the transcripts driving *cis*-eQTL effects. Yet the 80% trait-associated variants mapped within 33 kb from the gene, which lead them to conclude that trait-associated variants having a different genetic architechture compared to those of *cis*-eQTLs,

therore limiting the ability to use *cis*-eQTLs to pinpoint casual genes from susceptibility loci suggested by GWAS.

In contrast, they suggested *trans*-eQTLs being more informative than *cis*-eQTLs on disease etiology. They showed multiple unlinked genetic variants which are associated with the same trait, frequently converging on *trans*-genes that are known to be important for the disease. This was also observed when PGS were associated with expression levels of the genes, where many of the significantly associated genes were known to drive these traits.

The paper entitled, *Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis*, is submitted to Nature in October 2018 and currently under revision. The manuscript is accessible in Appendix 2 and also online: https://www.biorxiv.org/content/10.1101/447367v1.

4 Integration with metabotypes

I start this chapter by introducing CoLaus metabolome data and explaining 'metabomatching', a method developed to identify metabolites underlying novel SNP associations with NMR metabolomic features. I briefly describe metabolome and genome-wide association studies (mGWAS) and report mGWAS results of CoLaus. Next I point to the importance of simultaneous analysis of cross-sectional multi-omics data from large population studies and describe the metabolome and transcriptome-wide association study of CoLaus. In the last section I present a paper published by our group that uses the metabomatching method to generate metabolic signatures from large-scale NMR data in an unsupervised fashion.

4.1 CoLaus metabolome data

Urinary metabolic profiles were generated using one-dimensional proton nuclear magnetic resonance (NMR) spectroscopy. NMR spectra were acquired at 300 K on a Bruker 16.4 T Avance II 700 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) using a standard ^{1}H detection pulse sequence with water suppression. The spectra were referenced to the TSP signal and phase- and baseline corrected. We binned the spectra into chemical shift increments of 0.005 ppm, obtaining metabolome profiles of 2,200 metabolome features, of which 1,276 remain after filtering for missing values [161]. Lastly, the dataset was log10-transformed and standardised across features then samples, to make samples and feature intensities comparable.

4.2 Metabomatching

Metabomatching is a method to identify metabolites underlying novel SNP associations with metabolome features [161]. It compares the significance profile of the associations with a given variable with all metabolome features across the full chemical shift range, the so-called pseudospectrum, with NMR spectra of pure metabolites available in public databases such as HMDB [162] and BMRB [163]. For each metabolite m, metabomatching defines a set of features $F_{\delta}(m)$ that contains all the features f that fall within a δ ppm vicinity of any NMR spectrum peak of m listed in the database. Metabomatching then computes the sum

$$(1)\ s(F_{\delta}(m)) = \sum_{f \in F_{\delta}(m)} \left(\frac{\hat{\beta}}{\widehat{SE}_f} \right)^{\alpha},$$

where $\hat{\beta}_f$ and $\widehat{SE}_f$ are the point estimates of the effect size and standard error of the association between f and m. Assuming a χ^2 -distribution if $\alpha = 2$, or a Z-distribution if $\alpha = 1$, for the sum with $\left|F_{\delta}(m)\right|$ degrees of freedom, metabomatching defines a score for each m as the negative logarithm of the nominal p-value corresponding to the observed sum. These scores are calculated for all the metabolites with NMR spectra in the database, allowing to rank them with regard to their likelihood to underlie the association of the variable with the metabolomic features.

Although metabomatching was originally developed to use pseudospectra from SNP - metabolome associations, we have recently shown that it can also use co-varying features of metabolome data itself to identify metabolites [164]. In our metabolome- and transcriptome-wide association study we used metabomatching to identify metabolites underlying gene expression - metabolome associations.

4.3 Metabolome and genome-wide association studies (mGWAS)

GWAS with metabolic traits (mGWAS) search for genetic variants that influence human metabolism. Metabolites are unique chemical fingerprints that reflect various cellular processes taking place in the cells of a subject, thus serving as snapshots of the physiological states of a cell. Together with transcriptome and proteome they serve as molecular phenotypes that reflect the functional status of events occurring further upstream. To date more than 150 independent genetic variants have been identified as modulators of serum metabolites, and 26 for urine metabolites [27]. Once having the genotype, metabotype and phenotype data, one can start to investigate to what degree the metabolomics data reflect the genotypic background and how informative it is about the phenotype, e.g. disease susceptibility. Changes in metabolite concentrations may be the direct consequence of the genetic background modulated by the environment, and some of these changes can be causal for developing or progressing a disease. Conversely, some changes in metabolites may occur as a results of an organismal dysfunction. Importantly, being able to distinguish between these two scenarios, would be of great clinical

usefulness, as metabolite changes which are causally upstream are good candidates for developing presymptomatic biomarkers indicating increased disease risk well ahead of the various homeostatic organismal processes leading to disease manifestation.

4.4 mQTLs of CoLaus

In his work, Rico et al. [161] performed a metabolome-wide genome-wide association study (mGWAS) of CoLaus untargeted urine NMR data. He discovered 139 independent genome-wide significant associations of which 56 replicated in an independent cohort (p-value < 5×10^{-8}). He designed a method called *metabomatching* to identify the metabolites underlying the observed genotype - metabolome features associations (see section 4.2 for details). When metabomatching was applied to the 56 replicated SNP-feature associations, he found 11 locus-metabolite associations that are reported in Table 5, SNPs mapping to *ALMS1*, *ADXT2*, *PSMD9* and *PYROXD2* genes have been previously reported to associate with the respective metabolites listed in the table, thus the current study served as a replication. SNPs in *NAT2* and *PYROXD2* genes on the other hand had robust metabomatching with several metabolome features which did not overlap with those of reported metabolites in the literature. SNPs in *ACADL*, *ABO* and *ACADS* genes have been previously associated with serum metabolites yet the current study could not distinguish if the associations of these genes in urine corresponded to the urine analogs of those in serum or they were novel discoveries. Two novel findings of the study were *FUT2*'s association with fucose and *SLC7A9*'s association with lysine. Urine fucose levels concentration has been linked to gut microbial health and Crohn's disease, whereas lysine has been linked to kidney function and kidney failure, hence two novel findings of the study showing clinical relevance.

Gene	Chr	SNP	Associated Metabolite
ALMS1	2	rs11884776	N-acetylated compounds
ACADL	2	rs3764913	Unknown
ADXT2	5	rs37370	3-Aminoisobutyrate
NAT2	8	rs4921914	Unknown
ABO	9	rs579459	Unknown
PYROXD2	10	rs2147896	Trimethylamine
PYROXD2	10	rs4345897	Unknown
ACADS	12	rs3916	Unknown
PSMD9	12	rs7314056	2-Hydroxyisobutyrate
SLC7A9	19	rs8101881	Lysine
FUT2	19	rs492602	Fucose

Table 5: Association results of mGWAS (Adapted from Rueedi et al. 2014 [161]).

This study reported a high level of replication rate in an independent cohort where the experimental conditions were considerably different. Also most of the significant associations found in this study were previously reported in the literature. This demonstrates the reliability and robustness of feature-based NMR metabolomics.

4.5 Associating metobotypes with gene expression levels

Genome-wide association studies (GWAS) have identified thousands of common variants that are associated with complex traits [8], but the regulatory mechanisms behind these associations mostly remain poorly understood. Pinpointing causal variants is difficult, since the lead variants associated with a trait are often in high linkage disequilibrium (LD) with other variants in the same region with only a slightly lower association signal. Such associated LD blocks typically contain several genes or functional elements, preventing the accurate identification of causal genes. Furthermore, some trait associated variants fall into intergenic regions of the genome with no obvious functional role at all [18].

A number of studies reported that trait associated genetic variants are significantly enriched in expression quantitative trait loci (eQTLs), suggesting that many trait associated variants affect the phenotype by altering gene expression [20-23]. There is also a growing body of literature highlighting the more pronounced effects of genetic variants on molecular traits compared to

phenotypic traits [24-26, 165]. This is not surprising, since molecular traits representing fundamental biological processes such as gene expression and metabolism are intermediates in the genotype to trait causality chain.

With high-throughput measurements becoming more accessible and widespread, integration of molecular traits into association studies has become a central challenge in the field. Such synthesis allows investigating the interplay between different organisational layers of a biological system. Despite metabolism and gene expression regulation both being fundamental biological processes that are commonly studied as molecular phenotypes, there are very few studies in humans that focus on the interplay between them. Several studies investigated the relationship between untargeted serum metabolites and whole blood gene expression in humans [28-30], but, to the best of our knowledge no transcriptome- and metabolome-wide association study has been performed using urine metabolome data of healthy human subjects.

Most metabolome- and genome-wide association studies (mGWAS) reporting metabolite quantitative trait loci (mQTL) use targeted approaches where the concentrations of a limited number of metabolites are estimated from the metabolome data generated by mass spectrometry or NMR spectroscopy. This targeted approach is limited to the number of known quantifiable metabolites in the biofluid under study. In the current study we adopted an untargeted approach, making use of the entire metabolomic data captured by binned ^{1}H NMR spectra as our molecular traits. So here we present an untargeted metabolome- and transcriptome-wide association study using the entire NMR spectral information to characterize the urine metabolomes of 555 subjects and RNA-Seq data of lymphoblastoid cell lines (LCLs) derived from the same set of individuals. LCL have been widely used in genomic studies and proven their worth as faithful surrogates of primary tissues for studying both gene expression variation among individuals and the genetic architecture underlying regulatory variation of gene expression[96, 97, 99, 166]. LCLs thus present an interesting system whose genetic variance in expression resembles that of the cell types affecting the urine metabolome, with the added advantage of not being influenced by immediate environmental factors such as recent changes in the diet or exposure to a drug. Despite having limited statistical power and using surrogate tissue, we identified two strong associations between gene expression levels and urine metabolome features, which allowed us to refine previous links between the corresponding genes and metabolites.

4.5.1 Association analysis

We performed an untargeted metabolome- and transcriptome-wide association study by pairwise linear regression of log-transformed expression levels of each of the 43,614 genes (as response variable) onto each of the 1,276 metabolome features (as explanatory variable). The metabolome features resulted from binning the raw urinary NMR spectra with a bin-size of 0.005 ppm, and rank normalizing each bin passing QC (see Section 4.1 for details). The gene expression levels, quantified as RPKM, were measured using RNA-Seq on lymphoblastoid cell lines derived from the same set of 555 subjects (see Section 2.3 for details). The model also included the following common confounding factors: age, sex, the first four principal components of the genotypic data (correcting for population stratification) and the first 10 principal components of the gene expression data (correcting for potential batch effects). For the completeness of the analysis we did not apply any exclusion criteria to remove genes from the analysis. As a consequence, the significant associations need to be further evaluated in order to remove problematically distributed genes that could give rise to inaccurate regression estimates. We applied a nominal Bonferroni threshold for multiple testing $p_{max} = 0.05/(125\times1109) = 3.6\times10^{-7}$ by taking into account the effective number of tests which we estimated to be 125 for metabolome features and 1109 for genes (i.e. the number of principal components explaining more than 95% of the data [167]). Only associations with a p-value below p_{max} were considered significant. All statistical analyses were performed using Matlab [168].

Figure 23 shows the qq-plot of all pairwise associations. It is well calibrated, and only two association p-values (both involving the *ALMS1* gene, see below) are highly significant (FDR<0.05). Yet, applying an adjusted Bonferroni threshold of 3.6×10^{-7} to account for the effective number of independent variables, we identified 25 additional marginally significant feature-gene associations. The 27 association pairs involved 22 unique genes and 25 unique features.

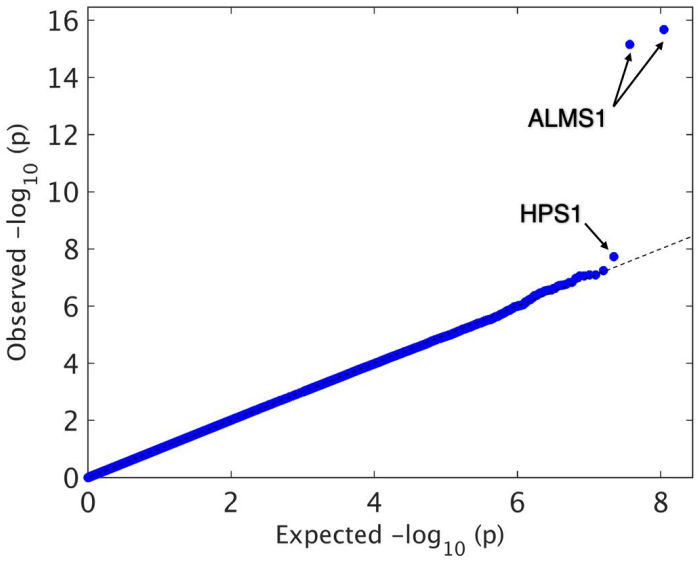


Figure 23: QQ-plot showing -$\log_{10}$(p)-values of metabolome- and transcriptome-wide association analysis. The features that significantly associate with *ALMS1* expression are ranking as 1st , 2nd and 8th; the features associated with *HPS1* expression are ranking as 3rd and 5th and the features associated with *ALMS1P* expression are ranking as 6th and 7th.

As we did not apply any a-priori exclusion criteria to remove genes from the analysis, we inspected the expression value distributions of these 22 significant genes in order to identify cases in which the small p-value may be due to a problematic distribution of the expression values. Indeed, we observed that some of the genes had zero expression values for a sizable fraction of the samples and very low expression values otherwise. Based on the distributions we filtered out all genes with zero RPKM values in more than 95% of the samples if all these expression values were below 1. Applying this rather mild filtering removed 11,547 out of the 43,614 all autosomal genes (26%) and 1,994 out of 19,123 protein-coding genes (10%). Amongst the 22 marginally significant associations five (23%) were removed. Scatter plots of gene expression and associated metabolite feature can be seen in Figure 24 and 25 for the discarded and valid associations respectively. We report the remaining 21 significant associations corresponding to 17 unique genes and 19 unique features in Table 6.

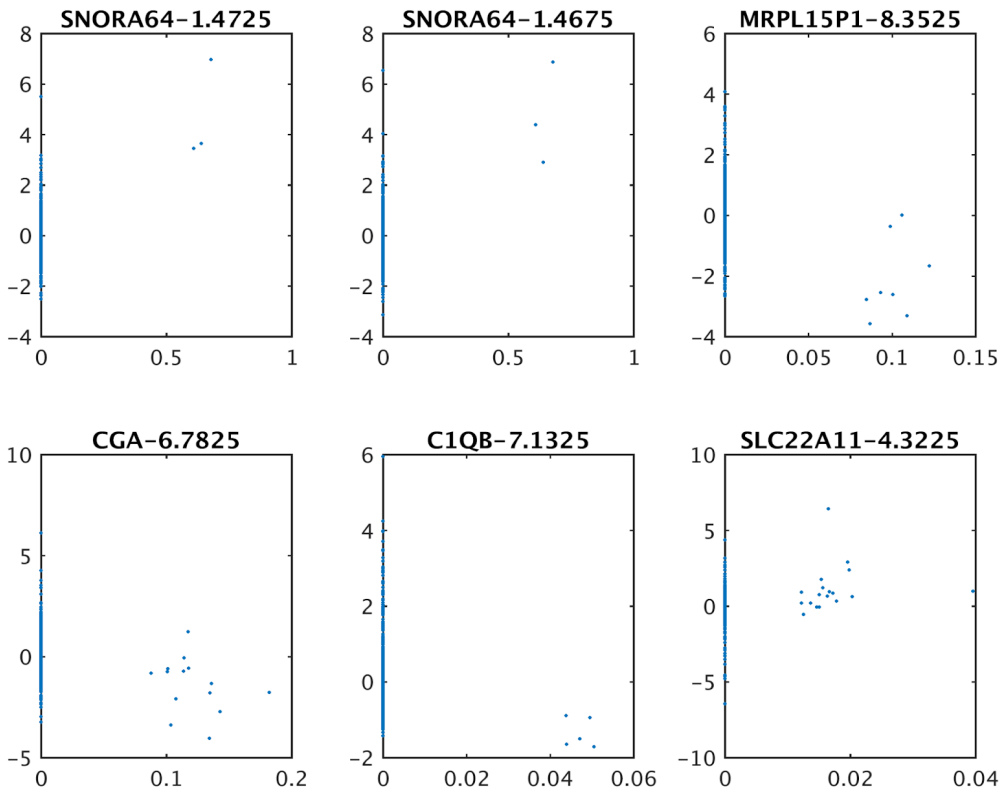


Figure 24: Scatter plots of six removed study wide significant metabolite feature - gene expression associations. X-axis shows the gene's RPKM values and the Y-axis shows the $\log_{10}$ transformed and Z-scored metabolite feature that is associated with the respective gene.

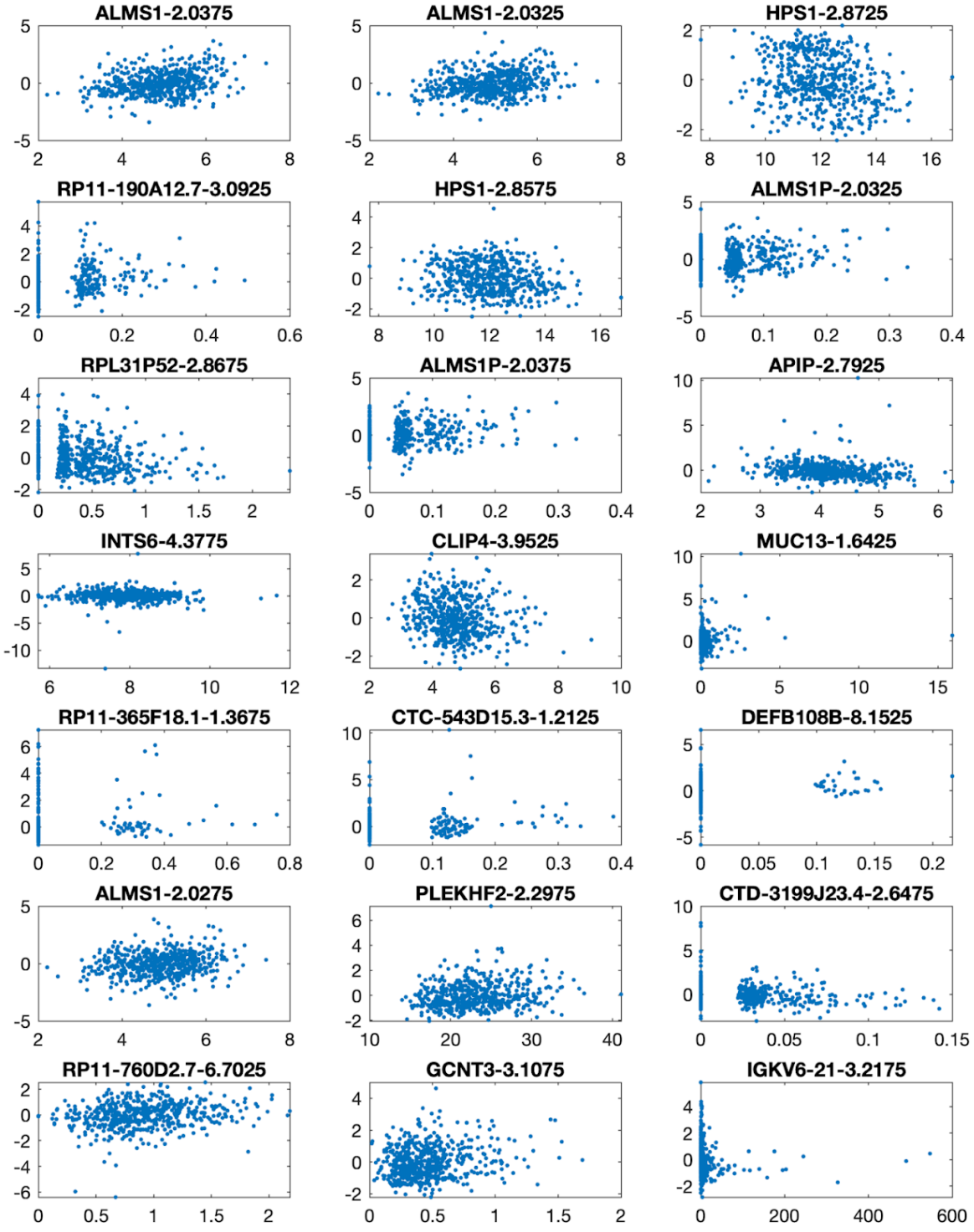


Figure 25: Scatter plots of 21 study wide significant metabolite feature - gene expression associations. X-axis shows the gene's RPKM values and the Y-axis shows the log10 transformed and Z-scored metabolite feature that is associated with the respective gene.

Genes			Metabolite	Association		Published as mGWAS
Gene ID	Chr	Gene symbol	Feature(s)	*X*	*P*	Body fluid
ENSG00000116127	2	ALMS1	2.0375, 2.0325, 2.0275	0.51, 0.50, 0.32	2.1×10^{-16}, 7.0×10^{-16}, 2.7×10^{-7}	Serum [185,165] Urine[169,172]
ENSG00000107521	10	HPS1	2.8725, 2.8575	-0.31, -0.29	1.9×10^{-8}, 8.0×10^{-8}	Serum [185,165] Urine[169]
ENSG00000256029	1	RP11-190A12.7	3.0925	0.24	5.6×10^{-8}	Serum [165]
ENSG00000163016	2	ALMS1P	2.0325, 2.0375	0.24, 0.23	8.1×10^{-8}, 1.5×10^{-7}	Serum [185,165] Urine[169]
ENSG00000219355	12	RPL31P52	2.8675	-0.23	1.0×10^{-7}	
ENSG00000149089	11	APIP	2.7925	-0.27	1.5×10^{-7}	Serum [185,165] Urine[169]
ENSG00000102786	13	INTS6	4.3775	-0.34	1.9×10^{-7}	Serum [185,165]
ENSG00000115295	2	CLIP4	3.9525	-0.31	1.9×10^{-7}	Serum [185,165]
ENSG00000173702	3	MUC13	1.6425	0.24	2.1×10^{-7}	Serum [165]
ENSG00000228360	7	RP11-365F18.1	1.3675	0.22	2.4×10^{-7}	
ENSG00000267273	19	CTC-543D15.3	1.2125	0.22	2.4×10^{-7}	
ENSG00000184276	11	DEFB108B	8.1525	0.23	2.7×10^{-7}	
ENSG00000175895	8	PLEKHF2	2.2975	0.31	2.8×10^{-7}	Serum [185]
ENSG00000267267	17	CTD-3199J23.4	2.6475	-0.22	2.9×10^{-7}	
ENSG00000213650	7	RP11-760D2.7	6.7025	0.23	3.0×10^{-7}	
ENSG00000140297	15	GCNT3	3.1075	0.27	3.3×10^{-7}	Serum [185,165]
ENSG00000211611	2	IGKV6-21	3.2175	-0.22	3.5×10^{-7}	

Table 6: 21 study-wide significant associations from metabolome- and transcriptome-wide association analysis, corresponding to 17 unique genes and 19 unique features. Abbreviations: GeneID - Ensembl Gene ID (NCBI build 37), Chr - chromosome, X - effect size, P - P-value.

4.5.2 Metabolite discovery

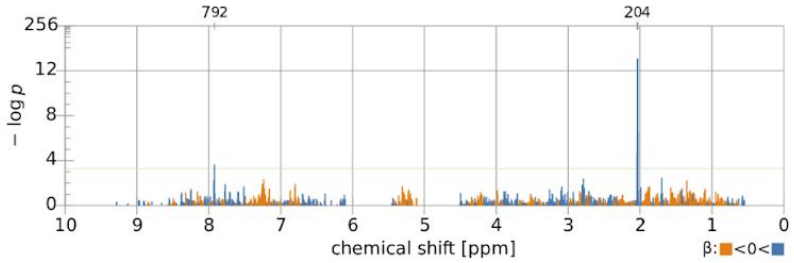
To find the metabolites underlying these significant associations between gene expression levels and metabolome features we used metabomatching. Metabomatching has been previously established as an effective tool for prioritizing candidate metabolites underlying SNP-metabolome features association profiles, so-called pseudospectra [161, 169]. In this study we used association profiles of genes which had at least one significantly associated metabolite feature as input to metabomatching and found that the pseudospectra of *ALMS1* and *ALMS1P* matched well with the N-Acetylaspartate (NAA) NMR spectrum and that the pseudospectrum of *HPS1* matched well with the trimethylamine (TMA) NMR spectrum (see Figure 26).

A

Metabomatching Settings

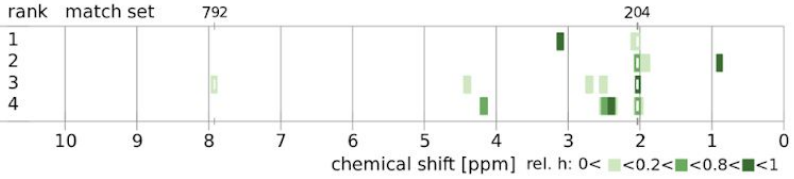
mode peak, δ = 0.030
scoring χ^2
database UMDB

Pseudospectrum of ENSG00000116127 ALMS1



Candidate Metabolites

score	name	rank
26.5	1,3-diaminopropane	1
25.5	isovalerate	2
20.4	N-acetyl-L-aspartate	3
17.8	pyroglutamate	4

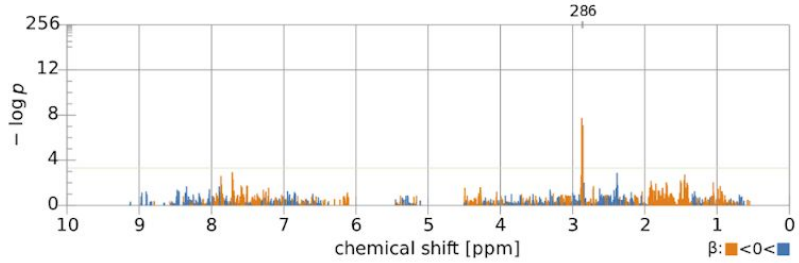


B

Metabomatching Settings

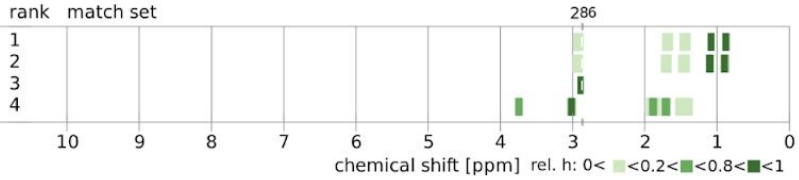
mode peak, δ = 0.030
scoring χ^2
database UMDB

Pseudospectrum of ENSG00000107521 HPS1



Candidate Metabolites

score	name	rank
13.5	2-methyl-3-ketovalerate	1
12.9	3-methyl-2-oxovalerate	2
10.6	trimethylamine	3
9.0	lysine	4

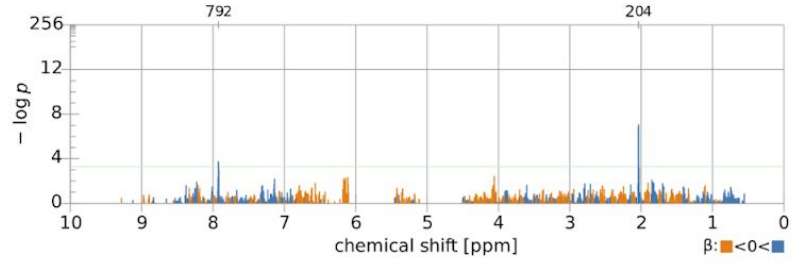


C

Metabomatching Settings

mode peak, δ = 0.030
scoring χ^2
database UMDB

Pseudospectrum of ENSG00000163016 ALMS1P



Candidate Metabolites

score	name	rank
9.9	aminoadipate	1
8.7	1,3-diaminopropane	2
8.3	2-hydroxyglutarate	3
7.7	N-acetyl-L-aspartate	4

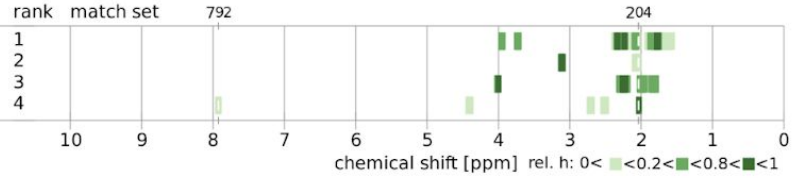


Figure 26: Metabomatching figures showing the pseudospectra derived from gene expression - metabolome features associations [170]. The features in each pseudospectrum are color-coded by the sign of the effect size and the four highest ranking candidate metabolites are listed on the lower left with their reference NMR spectra shown on the right (color coding indicating their relative peak intensities). A) CoLaus urine metabolome-*ALMS1* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.03 ppm and 7.92 ppm regions which match well with the highest intensity peak of NAA and one of the lower intensity peaks of the NAA NMR spectrum, respectively. B) CoLaus urine metabolome - *HPS1* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.87 and 2.86 ppm which match well with TMA singlet. C) CoLaus urine metabolome-*ALMS1P* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.03 ppm and 7.92 ppm regions which match well with the highest intensity peak of NAA and one of the lower intensity peaks of the NAA NMR spectrum respectively.

As shown in Table 6, the expression of *ALMS1* significantly associates with three neighboring features at 2.0375 ppm (p-value=2×10^{-16}), 2.0325 ppm (p-value=7×10^{-16}) and 2.0275 ppm (p-value=3×10^{-7}). There are few metabolites with resonances in this region and usually a singlet signal in this area is interpreted as the N-acetylated resonance detected in the ^{1}H NMR spectrum of N-acetylated compounds [171]. As illustrated in Figure 26A, among the top three metabolites suggested by metabomatching that have a peak at 2.03 ppm, the only one with the highest intensity peak at this position is NAA. Also the presence of a secondary peak in the pseudospectrum at 7.9225 ppm matches well with one of the lower intensity peaks of the NMR spectrum of NAA reported at 7.92 ppm in HMDB, even though the association p-value of this feature is below the Bonferroni threshold (p-value=2×10^{-4}). Similarly, metabomatching the pseudospectrum of *ALMS1P* (*ALMS1* pseudogene) points to NAA as the most likely matching N-acetylated compound (Figure 26C). The metabolome features pointing to NAA are the same features as in *ALMS1* but with lower association p-values (2.0375 ppm with p-value=1×10^{-7}, 2.0325 ppm with p-value= 8×10^{-8} , 7.9225 ppm with p-value= 2×10^{-4}).

The third and fifth strongest associations in Table 6 are between *HPS1* gene expression and two neighboring metabolome features at 2.8725 ppm (p-value=2×10^{-8}) and 2.8575 ppm (p-value= 8×10^{-8}), respectively. Figure 26B shows the metabomatching result of the *HPS1* pseudospectrum. Among the top three metabolites suggested by metabomatching,

trimethylamine (TMA) is the most plausible metabolite driving the association pattern, as it is the only metabolite with its highest intensity NMR peak at 2.86 ppm region and no missing peaks. Schematic representation of the match between pseudospectra and the NMR spectra for both *ALMS1* and *HPS1* can be seen in Figure 27.

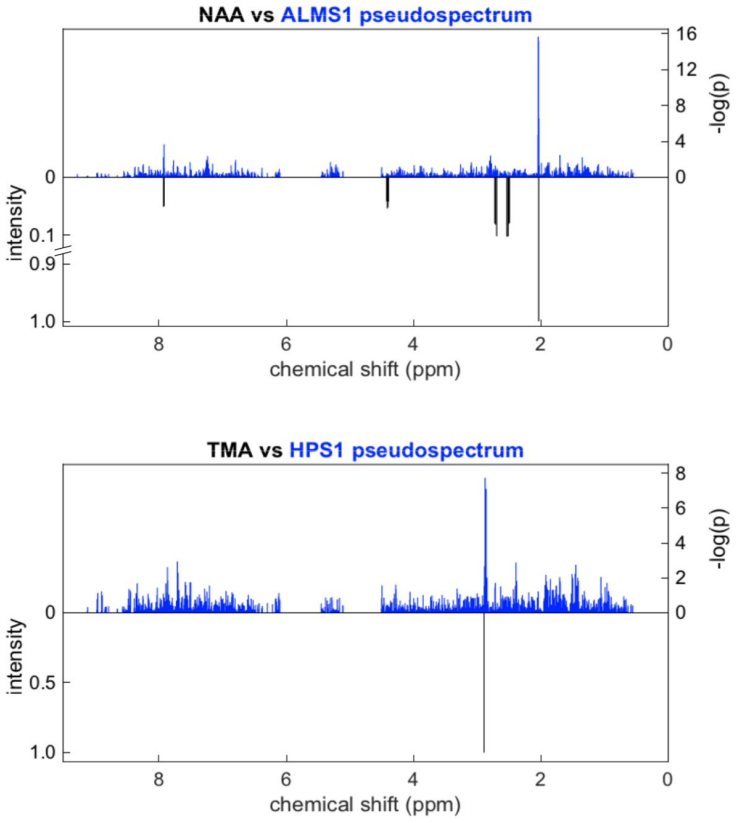


Figure 27: Schematic representation of the match between *ALMS1* association pseudospectrum and NAA NMR spectrum(top plot) and the match between *HPS1* association pseudospectrum and TMA NMR spectrum (bottom plot). Each figure shows -log10 transformed gene expression - metabolome features association p-values on the top and the reference NMR spectrum of the matching metabolite on the bottom. In the NAA - *ALMS1* match, leading features allowing metabolite identification are at 2.03 ppm and 7.92 ppm regions which match well with the highest intensity peak of NAA and one of the lower intensity peaks of the NAA NMR spectrum respectively. In the TMA - *HPS1* match, leading features allowing metabolite identification are 2.87 and 2.86 ppm which match well with TMA singlet.

The reference spectrum of NAA in the Urinary Metabolome Database (UMDB) that we used for metabomatching was recorded in water. In order to verify that the peaks of this spectrum are comparable to those of NAA in urine, we spiked NAA into pooled urine samples from our collection at a concentration of 1 and 10 mM and recorded their ^{1}H NMR spectrum. Inspecting the 5 multiplet regions of NAA, we concluded that the NAA peak positions are very similar in both solvents (see Figure 28).

To further investigate if a better match exists among all the N-acetylated family of compounds, we built a library consisting of all N-acetylated compounds proton NMR spectra available in HMDB and the Biological Magnetic Resonance Data Bank (BMRB). NAA remained the best metabomatching hit for the *ALMS1* pseudospectrum (see Figure 29). Figure 30 illustrates the relationship between *ALMS1* gene expression level and the NAA metabolite concentration where every point in the plot represents a study sample and each of the samples are color coded according to the genotype at rs7566315 SNP, that is an eQTL of *ALMS1* and mQTL of NAA.

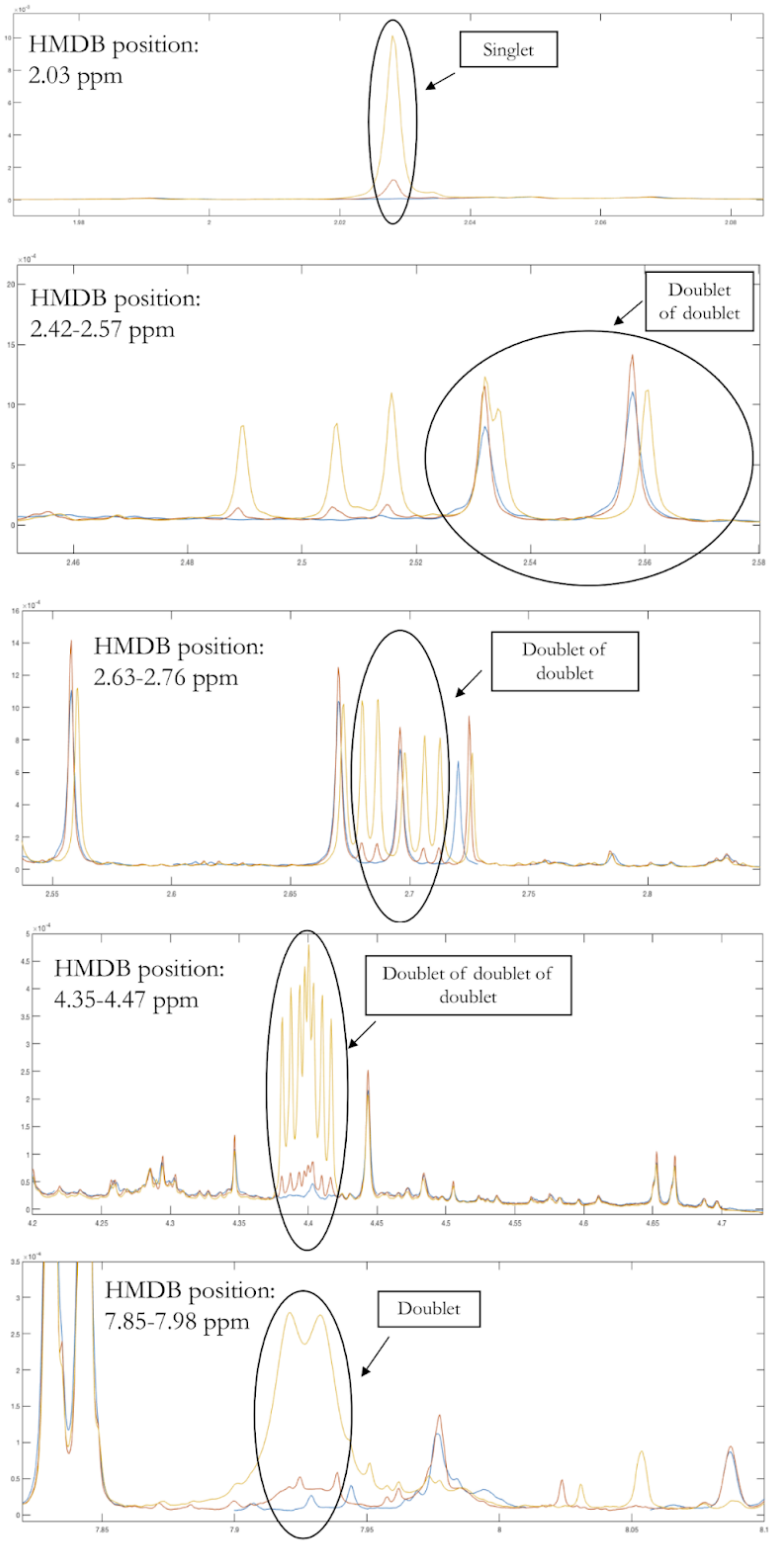


Figure 28: Combined NMR profile of 3 different NMR experiments. Blue spectrum: randomly selected and pooled urine samples. Red spectrum: NAA spike-in into pooled samples where the NAA concentration in the solution is 1 mM. Yellow spectrum: NAA spike-in into pooled samples where the NAA concentration in the solution is 10 mM.

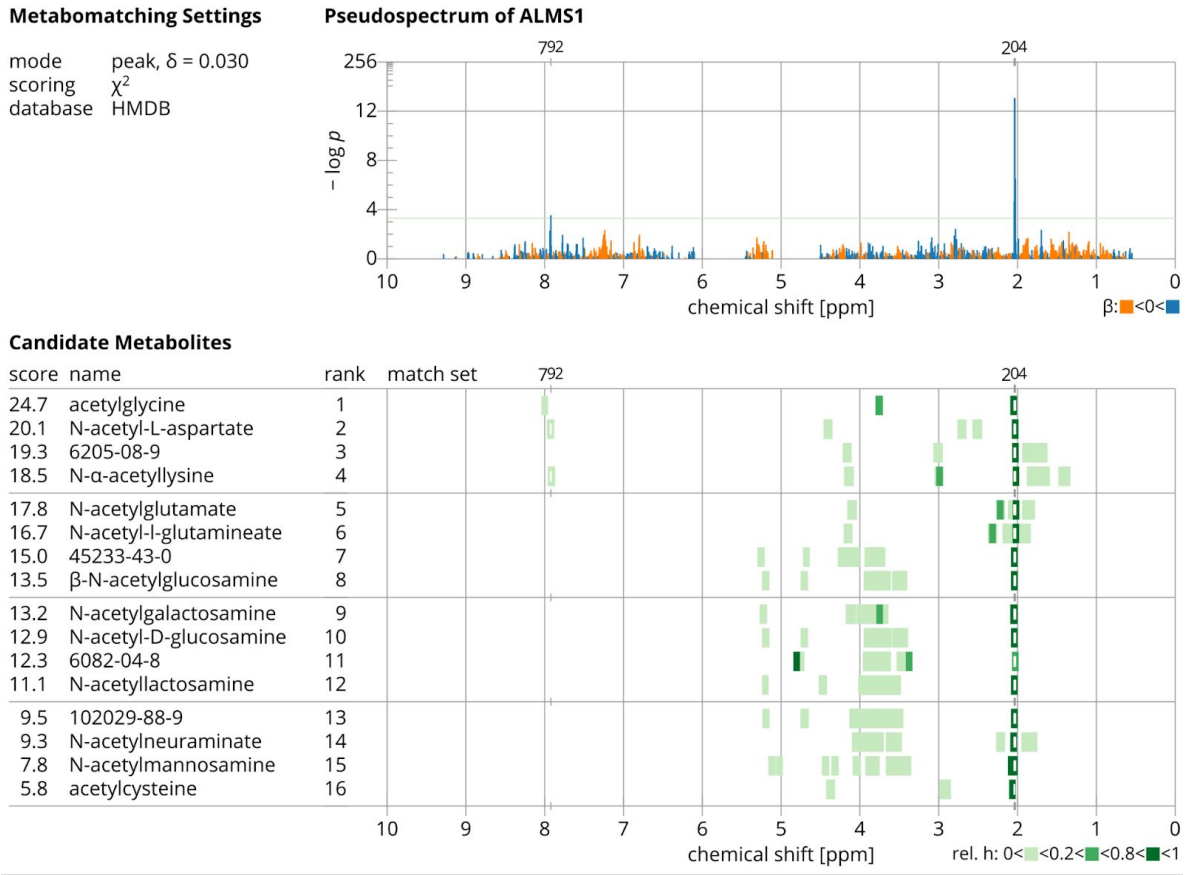


Figure 29: CoLaus urine metabolome- *ALMS1P* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.0375 ppm, 2.0325 and 7.9225 ppm regions, respectively, which match well with the highest intensity peak of NAA and one of the lower intensity peaks of NAA NMR spectrum, respectively.

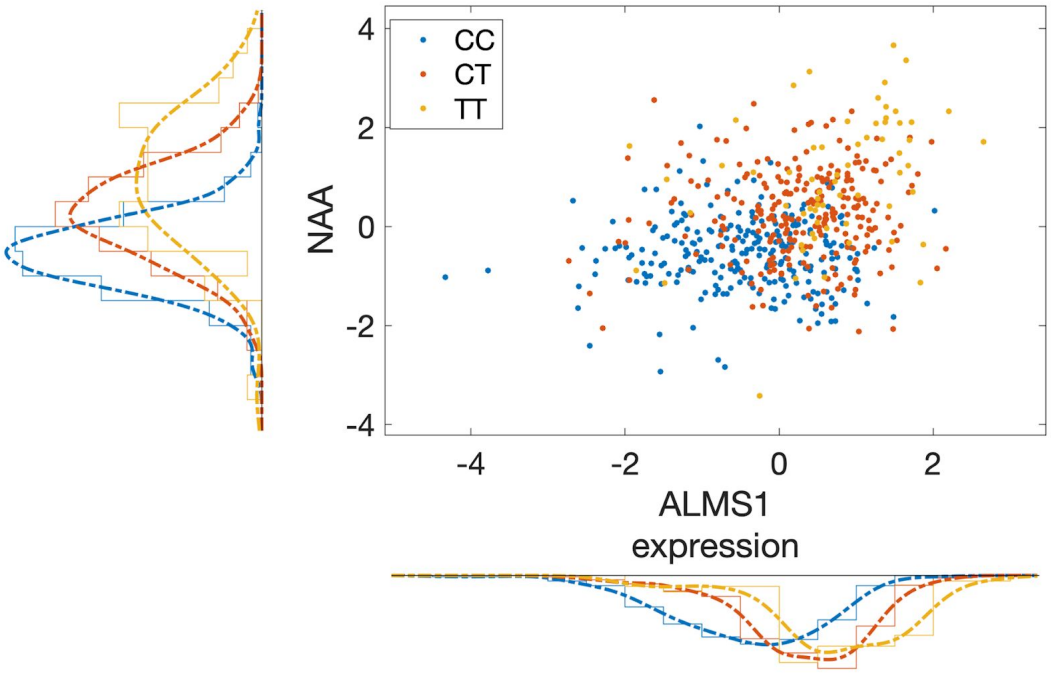


Figure 30: SNP rs7566315, showing a mQTL effect on NAA and an eQTL effect on *ALMS1* gene expression. Each point represents a study sample. NAA concentration is approximated by the feature at 2.0375 ppm that is log10 transformed after feature- and sample-wise z-scoring (y-axis). *ALMS1* expression is quantified as $\log_2$ transformed RPKM+1 values (x-axis). Color code represents the genotype of rs7566315 (legend).

4.5.3 Validation of *ALMS1*, *HPS1* and *ALMS1P* associations

To the best of our knowledge, there is no other study with urine NMR spectra and expression data of LCLs derived from the same subjects that is of comparable or larger sample size, precluding proper out-of-sample replication of our results. We have, however, access to additional urine NMR spectra from samples collected for a subset of 301 CoLaus subjects in a follow-up study conducted five years after the baseline data collection. We note that the follow-up NMR data are not independent from the baseline data, yet they were obtained from physically different samples collected at a significantly later time and processed in a different NMR spectrometer and facility. As for the expression data, we only have those from LCLs derived from blood taken at baseline, so we could only test whether the associations we observed between baseline metabolomics and baseline transcriptomics measurements would persist as associations between follow-up metabolomics and baseline transcriptomics data.

We thus asked whether our significant and marginally significant results can be confirmed also using the follow up metabolomics data. We focused on the *ALMS1* and *ALMS1P* gene expression association with NAA and the *HPS1* gene expression association with TMA. As baseline and follow-up urine NMR data were each processed and binned individually, the features did not correspond one-to-one between the studies. To test the association of these three genes with relevant features, we selected all features within +/- 0.03 ppm neighborhood of top features associated with these genes from baseline dataset; i.e. 2.0375 ppm for *ALMS1* and *ALMS1P*, and 2.8575 ppm for *HPS1*. This resulted in 12 features to test for each of the genes. We used a Bonferroni multiple testing corrected p-value threshold of 0.05/(12 features x 3 genes) = 1.4×10^{-3}.

In the follow-up, *ALMS1* gene expression level significantly associated with three neighboring features at 2.042 ppm (p-value=5.1×10^{-7}), 2.037 ppm (p-value=3.7×10^{-6}) and 2.032 ppm (p-value=3.9×10^{-4}), likely corresponding to the features at 2.0375 and 2.0325 ppm in the baseline association study. *HPS1* gene expression level significantly associated with 2 features at 2.869 ppm (p-value=2.2×10^{-5}) and 2.859 ppm (p-value=1.3×10^{-3}) that likely correspond to the features at 2.8725 and 2.8575 ppm in the baseline dataset. *ALMS1P* however did not show any significant association with candidate features in the follow-up study. Table 7 summarises our validation results.

ALMS1	shifts	2.067	2.062	2.057	2.052	2.047	2.042	2.037	2.032	2.027	2.022	2.017	2.011
follow-up	beta	0.0266	0.0625	0.0988	0.0126	0.1019	0.2943	0.2733	0.2094	0.0375	0.0655	0.0126	0.0296
	se	0.0588	0.0599	0.0596	0.0582	0.0593	0.0572	0.0579	0.0584	0.0585	0.0580	0.0577	0.0582
	p-value	6.51E-01	2.97E-01	9.83E-02	8.29E-01	8.71E-02	5.12E-07	3.73E-06	3.92E-04	5.21E-01	2.60E-01	8.27E-01	6.12E-01
	r-sq	0.0704	0.0397	0.0486	0.0874	0.0594	0.1275	0.1087	0.0943	0.0882	0.1030	0.1111	0.0962
HPS1	shifts	2.884	2.879	2.874	2.869	2.864	2.859	2.854	2.848	2.843	2.838	2.833	2.828
follow-up	beta	-0.0285	-0.0530	-0.1917	-0.2492	-0.1903	-0.1978	-0.1290	-0.0040	-0.0202	-0.1004	0.0500	0.0681
	se	0.0625	0.0607	0.0606	0.0578	0.0611	0.0610	0.0613	0.0614	0.0610	0.0602	0.0630	0.0618
	p-value	6.49E-01	3.83E-01	1.74E-03	2.24E-05	2.02E-03	1.33E-03	3.63E-02	9.48E-01	7.41E-01	9.67E-02	4.28E-01	2.72E-01
	r-sq	0.0561	0.1129	0.1174	0.1968	0.0924	0.1028	0.0866	0.0875	0.0989	0.1213	0.0481	0.0825
ALMS1P	shifts	2.067	2.062	2.057	2.052	2.047	2.042	2.037	2.032	2.027	2.022	2.017	2.011
follow-up	beta	0.0055	-0.0274	-0.0097	0.0265	0.0225	0.1166	0.1060	0.1237	0.0189	0.0420	0.0331	0.0053
	se	0.0643	0.0656	0.0655	0.0636	0.0652	0.0651	0.0655	0.0649	0.0640	0.0635	0.0631	0.0637
	p-value	9.31E-01	6.76E-01	8.82E-01	6.78E-01	7.30E-01	7.45E-02	1.07E-01	5.75E-02	7.67E-01	5.09E-01	6.00E-01	9.34E-01
	r-sq	0.0697	0.0366	0.0393	0.0879	0.0499	0.0559	0.0467	0.0648	0.0871	0.1003	0.1118	0.0954

Table 7: Validation of three essential associations discovered in CoLaus baseline. Association statistics coming from associating CoLaus follow-up urine NMR data with the expression levels of *ALMS1*, *HPS1* and *ALMS1P*.

4.5.4 Comparison with mGWAS results

We performed a mGWAS study with metabolome features in NAA and TMA NMR peak regions using data from 826 individuals of the CoLaus cohort for whom the urinary NMR spectra are available (similar to [161]). Figure 31A shows the locuszoom figure of SNPs in loci surrounding *ALMS1/NAT8* locus with significant association p-values with metabolome feature at 2.0375 ppm. The SNPs most strongly associated with this metabolome feature are correlated with each other and lie within a locus containing *ALMS1*, *ALMS1-IT1*, *NAT8* and *ALMS1P* genes (r^2>0.8). In Figure 31B we show the p-values for association of expression values from 15 genes with five different metabolome features that represent all multiplet regions of NAA. *ALMS1* and *ALMS1P* have the most significant association results with 2.0375 ppm feature, compared to the rest of the genes. *ALMS1* and *ALMS1P* have the most significant association results with the 2.0375 ppm feature, compared to the rest of the genes. Concordantly, *ALMS1* and *ALMS1P* gene expression levels are associated more significantly to the feature at 7.9225 ppm, the secondary feature in our NAA identification, compared to the other genes at the locus. Figure 32A shows the significant association pattern of SNPs in the loci surrounding *HPS1/PYROXD2* locus with metabolome feature at 2.8725 ppm and Figure 32B shows the significance level for association of expression values from seven genes with the same metabolome feature. Even though the SNPs with the most significant association with feature 2.8725 are physically located closer to *PYROXD2* gene rather than *HPS1* gene, the expression level of *PYROXD2* does not show significant association with this feature. Inspecting the list of published mGWAS in humans [27], we found that the SNPs in both *ALMS1* and *HPS1* loci have been previously reported to associate with a number of metabolic traits. The *ALMS1* locus has previously been associated with a number of N-acetylated compounds, while *HPS1* locus has been associated with various metabolites including trimethylamine and dimethylamine (see Table 8) [161, 169, 172]. In mGWAS studies determining the mediator genes is not a straightforward procedure, as mQTL SNPs are indistinguishable from neighboring SNPs in LD, and mediator genes of the mQTLs are often inferred based on their physical proximity to the SNPs or functional relevance. Consequently, published mGWAS studies were not able to distinguish between *NAT8* and *ALMS1* or *HPS1* and *PYROXD2* as mediator genes of NAA and TMA, respectively. In contrast, in the current association study we use gene expression data allowing us to pinpoint *ALMS1* and *HPS1* as mediator genes.

Reference	Platform	Biofluid	Locus	Reported mGWAS results
Nicholson *et al.* 2011	MS + NMR	Urine + Plasma	ALMS1, NAT8	N-acetylated compounds
Montoliu et al. 2013	NMR	Urine	ALMS1	N-acetylated compounds
Rueedi *et al.* 2014	NMR	Urine	ALMS1	2.0375 (suggested as N-acetylated compounds)
Raffler et al. 2015	NMR	Urine	NAT8	2.031 (metabomatching: N-acetyl L-aspartate)
Suhre et al.2011	MS	Serum	NAT8	N-acetylornithine
Yu *et al.* 2014	MS	Serum	NAT8	N-acetylornithine
Shin *et al.* 2014	MS	Serum	NAT8	N-acetyllysine, Unknown compounds
Nicholson *et al.* 2011	MS + NMR	Urine + Plasma	HPS1, PYROXD2	Trimethylamine (urine), Dimethylamine (plasma)
Rueedi *et al.* 2014	NMR	Urine	PYROXD2	Trimethylamine, unknown compound, 1.8025
Raffler et al. 2015	NMR	Urine	PYROXD2	2.854 (metabomatching: trimethylamine)
Raffler et al. 2013	NMR	Plasma	PYROXD2	2.757
Rhee *et al.* 2013	MS	Plasma	HPS1	Asymmetric dimethylarginine
Krumsiek *et al.* 2012	MS	Serum	HPS1, PYROXD2	Multiple compunds, Unknown compounds
Hong *et al.* 2013	MS	Serum	HPS1	Caprolactam
Shin *et al.* 2014	MS	Serum	PYROXD2	Unknown compounds

Table 8: List of published mGWAS results in humans concerning *ALMS1/NAT8* and *HPS1/PYROXD2* loci. MS:Mass Spectrometry, numbers in reported mGWAS results section refer to NMR spectral shift positions in ppm.

To further evaluate the possible regulation of NAA and TMA by other genes suggested by published mGWAS studies, we investigated the metabomatching plots of these genes in order to see if they pointed to any N-acetylated compounds/TMA. The investigated genes either (a) were the target of an eQTL SNP that is mQTL of NAA/TMA, or (b) were within 500kb of *ALMS1/HPS1*. However none of these candidate genes produced a pseudospectrum containing even a single nominally significant signal pointing to NAA/TMA.

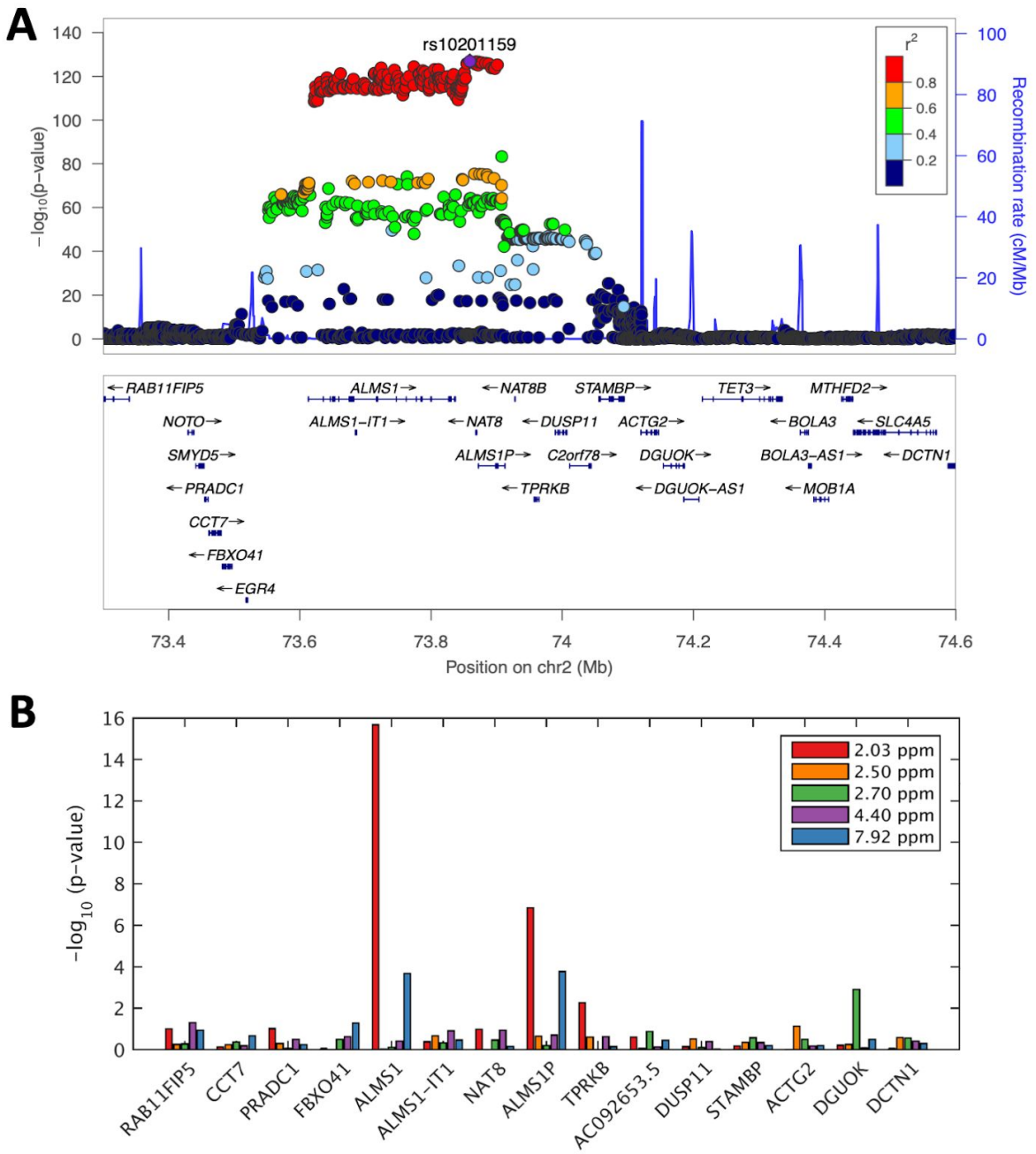


Figure 31: A) LocusZoom plot for *ALMS1/NAT8* locus, where the SNPs are associated with metabolome feature at 2.0375 ppm, LD colored with respect to lead mQTL. B) Bar plot shows -log10 transformed p-values from associating expression value of 15 genes in the locus with the five NAA features.

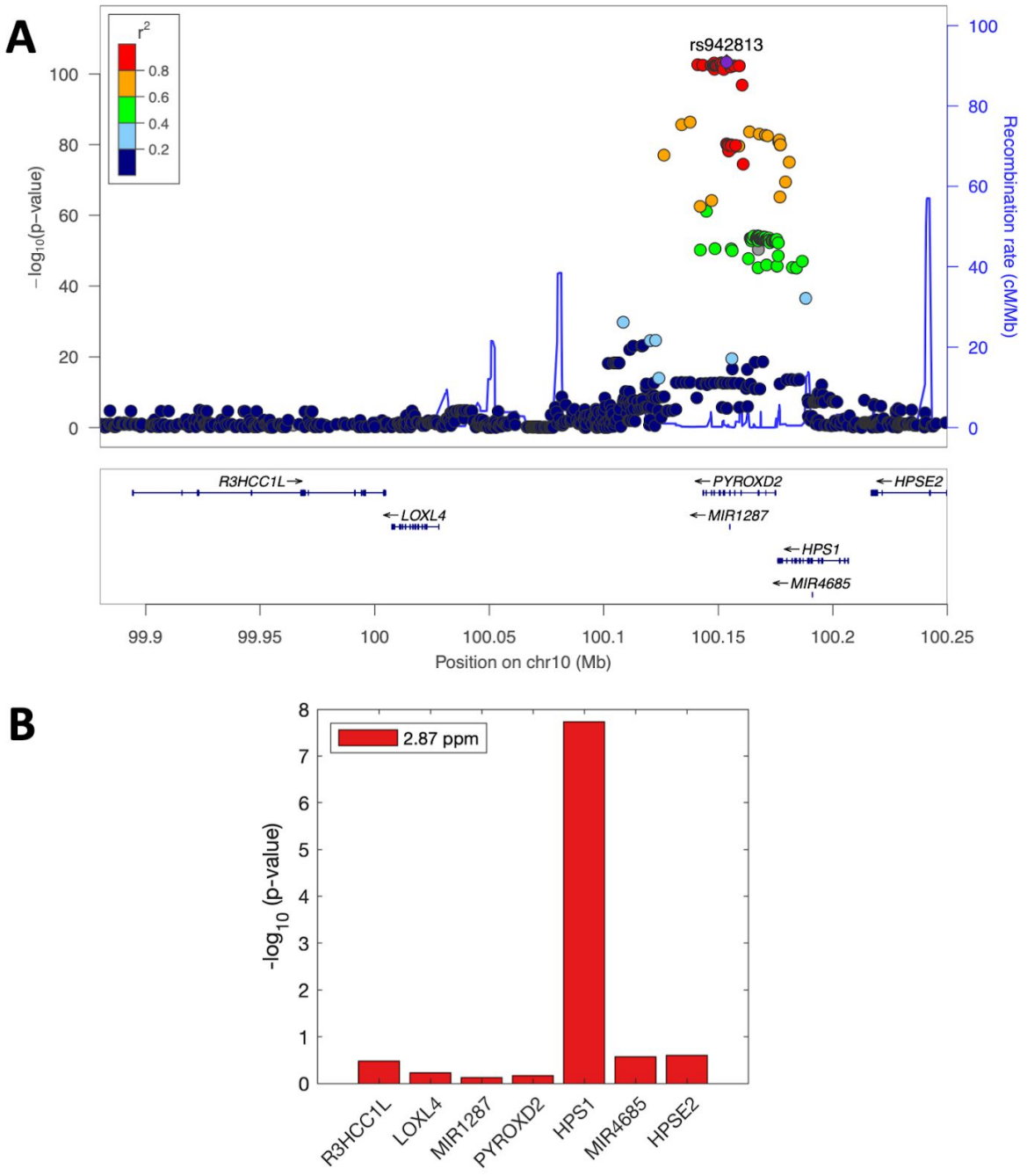


Figure 32: A) LocusZoom plot for *HPS1/PYROXD2* locus, where the SNPs are associated with metabolome feature at 2.8725 ppm, LD colored with respect to lead metaboliteQTL. B) Bar plot shows -$\log_{10}$ transformed p-values from associating expression values of seven genes in the locus with the TMA feature at 2.87 ppm.

4.5.5 Discussions & Conclusion

In this study, we present a metabolome- and transcriptome-wide association study using matching RNA-Seq and NMR urine profiles from 555 subjects of the CoLaus cohort. This is the first time such a study is performed on untargeted urine metabolome of healthy individuals. In contrast to targeted approaches that are restricted to a limited set of urine metabolites, our association study uses the binned features of the entire ^{1}H NMR spectra as metabolic traits. We identified one gene (*ALMS1*) whose association with two adjacent NMR features around 2.03 ppm is highly significant, surviving even the most conservative correction for multiple hypotheses testing. 16 additional genes are associated with metabolic features with marginal significance of p-value below an adjusted threshold accounting for the estimated number of independent variables (see Table 6). Among the top 17 genes, 12 are in loci with SNPs that have been previously reported as mQTLs. This shows the sensitivity of our study to extract likely candidates of metabolically relevant genes, despite its small sample size and low power.

We used metabomatching to search for promising metabolite candidates underlying gene expression-metabolome features associations. This approach was particularly insightful for our top hit *ALMS1*, as well as the strongest marginally significant association involving *HPS1*: Both genes had previously been implicated by mGWAS linking their loci to compound families. However, in both cases the reported mQTL also harbored other genes, leaving the exact gene-metabolite association ambiguous. Figures 33 and 34 are summarising what has been previously reported about the loci and the findings of this study.

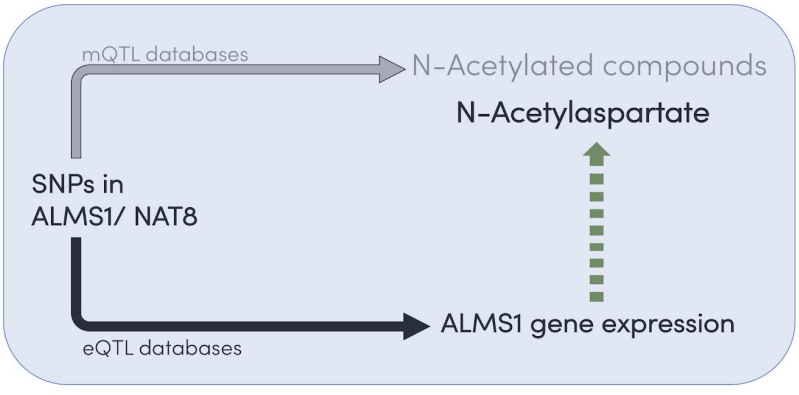


Figure 33: Previously reported mQTL SNPs associated with N-acetylated compounds (in grey). These mQTLs are also eQTLs of *ALMS1* expression in GTex eQTL database (in black). In the current study we found *ALMS1* gene expression being casual on N-Acetyl L-Aspartate (green dashed line).

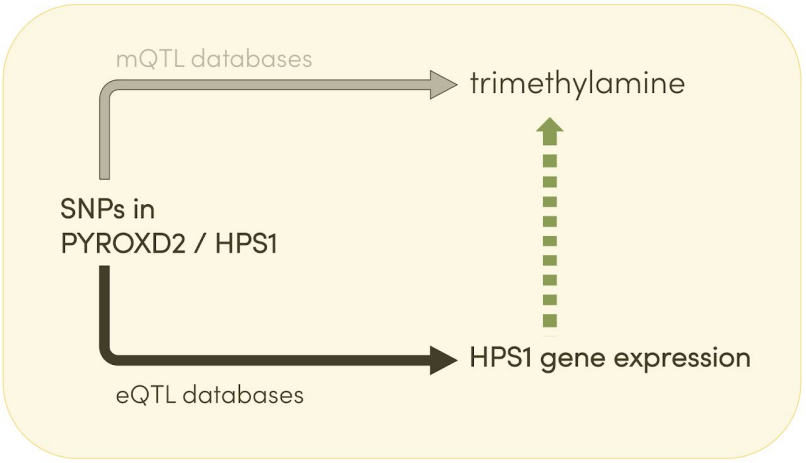


Figure 34: Previously reported mQTL SNPs pointing to TMA in urine. In the current study we found *HPS1* gene expression correlated with TMA in agreement with previous studies.

Specifically, the locus associated through mGWAS with N-acetylated compounds includes both *ALMS1* and the *NAT8* gene [161, 169, 172, 173], and the latter seemed to be the more likely candidate due to its known N-acetyltransferase activity. Yet, our association study using transcriptomics data only implicates *ALMS1* and not *NAT8*. Thus, while we cannot rule out a functional role of *NAT8*, the mQTLs of this locus likely act, at least predominantly, as eQTLs through *ALMS1*, pointing to its regulatory role in modulating the compound concentration. This metabolic role of *ALMS1* is also supported through its known role in Alström syndrome characterised by metabolic deficits (PMC6327082) and kidney health disorder phenotypes [174]. Interestingly, in the mGWAS reported by Montoliu et al. using data from a Brazilian cohort, the authors observed the association between N-acetylated compounds and the SNPs located in *ALMS1/NAT8* locus with stronger SNP associations in the *ALMS1* gene rather than *NAT8* [173]. They argued that the high ethnic diversity of their study population might have been responsible for breaking down the linkage disequilibrium in the *ALMS1/NAT8* region of the genome, resulting in a stronger association for SNPs close or in the *ALMS1* gene compared to other studies.

Our study also sheds more light onto the involved compound: Applying metabomatching on the pseudospectrum from association of all NMR features with the *ALMS1* expression level using a database composed of all N-acetylated compounds NMR spectra, suggested NAA as the best matching metabolite due to the presence of a secondary peak at 7.92 ppm and not missing any

high intensity peaks unlike other N-acetylated compounds (see Figure 29). Interestingly, NAA is the second most abundant metabolite in the brain and involved in neural signalling by serving as a source of acetate for lipid and myelin synthesis in oligodendrocytes [175]. NAA can be detected in urine of both healthy and unhealthy individuals in low concentrations [176] and it has a long history of being a surrogate marker of neural health and a broad measure of cognitive performance [177, 178]. Recently it has been shown that NAA correlates with time measures of neuropsychological performance [179]. The signals of SNPs in *ALMS1* by GWAS with intellectual phenotypes such as self-reported ability in mathematics [180, 181] might therefore be due to its role in modulating NAA. This conjecture of course assumes that NAA levels in relevant brain tissues reflect those in urine and that the *ALMS1* expression variation, and in particular its genetic component, in LCLs or blood, can serve as a proxy for brain tissue. As for *HPS1*, our second strongest association of a gene expression level with urine NMR features, we note that mGWAS previously associated its locus with TMA levels [161, 169, 172]. Yet, most of these studies, including the aforementioned GWAS using a Brazilian cohort [173] considered the *PYROXD2* gene, which is in the same locus, as the most likely modulator of TMA concentrations due to its known function as pyridine nucleotide-disulfide oxidoreductase. While we cannot rule out that this gene is indeed involved in TMA metabolism, in contrast to *HPS1* we have no evidence for association of *PYROXD2* expression levels with TMA. Thus, our data indicates that the mQTLs of this locus act predominantly as eQTLs through *HPS1*, pointing to its regulatory role in modulating TMA.

Our work illustrates the potential of metabolome- and transcriptome-wide association studies for deciphering gene-metabolite relationships. In particular, even with our modest sample size of 555 matched profiles we already had enough power to detect one significant and several marginally significant associations. Moreover, our two strongest associations pinpointed genes in loci implicated by mGWAS as the most likely candidates for transcriptional metabolite regulation. We also showed the possibility of extending correlative work and studying the causal relationship between gene expression levels and metabolite concentrations. Furthermore, this work demonstrated that our metabomatching tool, whose usefulness for elucidating candidate metabolites from mGWAS association profiles [161, 170] as well as auto-correlation signals in NMR data [164] was demonstrated previously, performs equally well on pseudospectra generated by association with gene expression levels. In our two examples the compounds implicated by

previous mGWAS were amongst the top metabomatching candidates and in the case of *ALMS1* restriction of the search space to the relevant compound family clearly favored a particular compound. This suggests that a future version of metabomatching could profit from implementing feature weighting (since in the case of *ALMS1* the lead feature clearly pointed to N-acetylated compounds).

Our study has many limitations: First, we only had access to gene expression levels of LCLs. While blood and such blood-derived cells are the easiest samples one can obtain from healthy subjects, their expression levels in many cases may only reflect poorly those of the relevant cells and tissues. Furthermore, metabolic reactions are of course driven by enzymes whose protein concentration determines the metabolic rate, and variation in gene expression levels is only one source of variation in active enzyme concentration (next to post-transcriptional and post-translational modifications, as well as their decay rate). Second, metabolite concentrations in urine correspond to excess that is cleared from the body, which depends on food intake and provide a poor proxy for many metabolite concentrations in their relevant location. Nevertheless, our study shows the promise of co-analyzing two or more distinct molecular traits observed in the same cohort.

The preprint entitled, *Untargeted metabolome- and transcriptome-wide association study identifies causal genes modulating metabolite concentrations in urine*, is submitted to bioRxiv in May 2020. The manuscript is accessible in Appendix 3 and also online: https://doi.org/10.1101/2020.05.22.110197 .

4.6 Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites

In this chapter, I describe the above titled project I have been involved in.

4.6.1 Background

Nuclear magnetic resonance (NMR) spectroscopy is a widely used method for metabolomic profiling, thanks to its accuracy and reproducibility. Single proton NMR spectroscopy (^{1}H NMR)

allows generating high throughput spectral data at low cost for different biofluids. The first step in NMR analysis usually involves identification of metabolites that are giving rise to the spectrum. However identification of the metabolites is usually not a straightforward procedure as human biofluids typically contain a large number of metabolites and often their corresponding peak positions overlap. Therefore when sample size permits expert annotation is still the most accurate way to identify metabolites. Nonetheless for the large dataset this approach is not scalable due to time and cost involved as well as potential compromise on reproducibility. It has been shown that analyzing the co-varying features in NMR data can facilitate identification of metabolites simply because the features belonging to the same metabolite are expected to be significantly correlated in large datasets [182]. Once the metabolites are identified, often the next task is to quantify the identified metabolites. Many publicly available tools exist to serve this purpose yet they also require expert refinement. There are multiple reasons why it remains challenging to achieve a fully automated metabolite quantification. First, human biofluids contain large numbers of metabolites whose concentrations vary across multiple orders of magnitude. This makes it difficult to gauge the contribution of the metabolites with low concentrations, especially if their peaks overlap with other metabolites' peaks. Second, the exact position of the peaks depends on the biofluid, which does not necessarily match to the one used while acquiring the reference spectra. Third, the peak positions are very sensitive to pH, ionic strength and the protein content of the sample. And lastly, the reference databases are certainly growing but they are far from being complete.

In recent studies, colleagues demonstrated that the limitations of the tragated NMR metabolomics can be addressed specifically in the context of metabolome-GWAS, where the genetic determinants of metabolites are studied [161, 170]. Rueedi et al. observed that the effect of a genetic variant on the concentration of a metabolite often translates into associations with many features in the NMR spectrum. To identify the metabolites underlying the significant associations he developed a method called *metabomatching*, which uses association results of NMR features with a genetic variant to suggest the most likely metabolites underlying this association [161, 170].

4.6.2 Scope of the project

In this project we used metabomatching method to generate metabolic signatures from large-scale NMR data in an unsupervised fashion. More specifically we identified metabolites from untargated urine NMR data by using covarying features of the NMR data among a large number of samples, instead of associating them with external variables. Covarying spectral features were selected based on three different methods including the iterative signature algorithm (ISA), averaged correlation profiles (ACP) and principal component analysis (PCA). Summary of the workflow is shown in Figure 35. ISA is a biclustering method that is designed to find coherent subsets in the data as described in Section 2.4. Parameters of the algorithm were set in a way that the discovered modules corresponded to NMR features that on average have higher or lower intensities in the selected samples compared to the remaining samples. ACP on the other hand is a greedy approach that we used for generating correlation profiles of feature pairs f_i and f_j that are at least 0.1 ppm further away from each other. We started with the most highly correlated feature pair and consecutively added other pairs if they had no other feature pair that is already selected in 0.1 ppm neighborhood. In the final step we calculated the averaged correlation profiles of feature pairs f_i and f_j: $c_k = (C_{ik} + C_{jk})/2$. Standard PCA was implemented to compute loadings of all features onto eigenvectors of the sample-sample correlation matrix of all features.

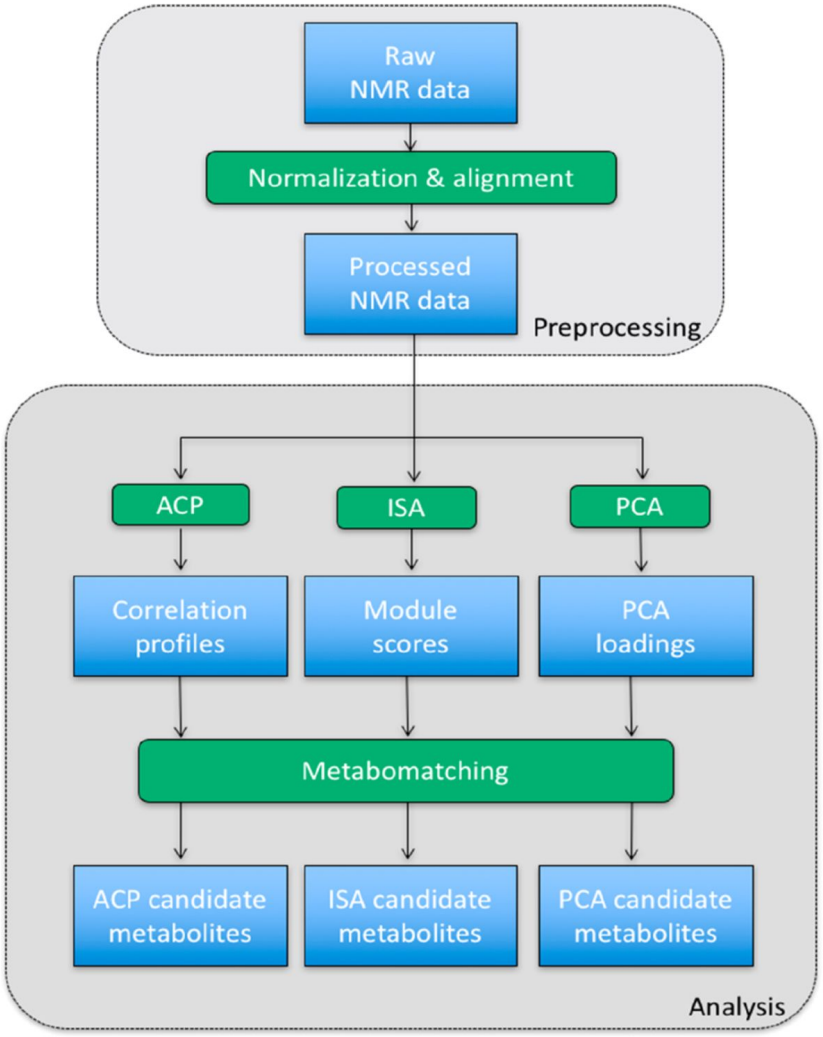


Figure 35: Workflow of unsupervised analysis of NMR data.

4.6.3 My contribution

My main contribution to the study was to perform association analysis of the 14 metabolites discovered in the study, with gene expression profiles. First I employed a candidate gene approach to study the potential links between the metabolites and the genes. I investigated whether or not the 14 discovered metabolites were linked to a gene in the virtual metabolic human database (VMH) and found five metabolites including oxoglutarate, creatine, ethanol, lactate and citrate linked to various genes. Later I performed association analysis between these genes and pseudo-quantified metabolite concentrations however results were not significant.

Next I performed a transcriptome-wide association with 43,614 genes instead of focusing on a set of candidate genes. In the transcriptome-wide association analysis I also investigated the association results gathered for different pseudo-quantifications; pseudo-quantification of the metabolites only considering the NMR spectral peaks that are captured as the signatures in the ACP and ISA method, and the pseudo-quantification of the metabolites based on all the NMR spectral peaks of the metabolite as reported in HMDB. Results showed no strong association between any of the 14 metabolites concentrations and gene expression levels, regardless of the method of pseudo-quantification. Nonetheless, I observed a suggestively significant association between citrate (pseudo-quantified based on the peaks suggested by ACP) and SLC29A4P2 gene which is a solute carrier pseudogene (see Figure 36). Interestingly in the first candidate gene approach I found citrate being associated with two genes SLC25A1 and SLC13A5, which are both solute carrier genes that code for transporter proteins that help the mobility of citrate in and out of the cell. However the SLC29A4P2 gene being a pseudogene, we could not conclude the functional relevance of the observed association. Despite considering different covariables in the association models, the associations failed to point to any other promising metabolite-gene associations.

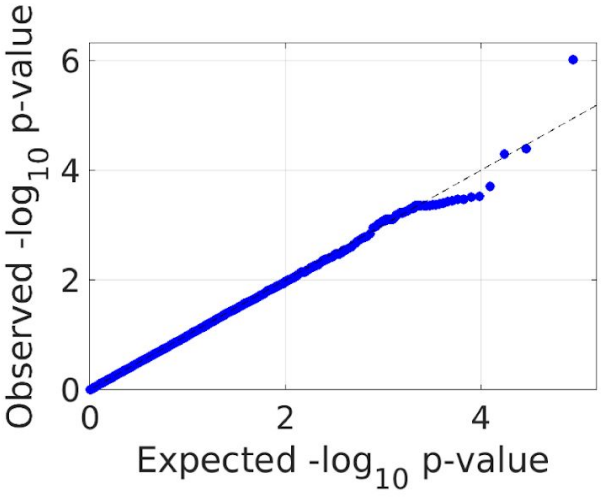


Figure 36: QQ-plot showing expected versus observed p-values of citrate concentration's association with 43,614 genes.

My other contributions to this paper includes creating some of the main figures, discussing the results and implications, and contribution to the manuscript.

4.6.4 Results and conclusions

In this study we showed that metabomatching can be used to identify metabolites based on the internal structure of the large-scale NMR data. We suggest that in a large collection of NMR samples there is enough power to identify metabolites based on the coherent features present across the samples.

Using ACP and ISA driven pseudospectra metabomatching identified a number of metabolites that are present in human urine. Five metabolites identified by both methods included citrate, ethanol, P-hydroxyphenylacetate, D-glucose and hippurate; whereas five metabolites identified only by ISA were 3-aminoisobutyrate, 3-methylhistidine, creatinine, α-lactose and lactate; and four metabolites identified only by ACP were taurine, creatine, oxoglutarate and 3-hydroxyisovalerate (see Figure 37). These compounds are all urinary metabolites that are known to be present in high concentrations in urine. In contrast to ACP and ISA, PCA did not generate pseudospectra robustly matching to a metabolite. We hypothesize that this is due to leading principal components possessing variation signatures that are driven by many metabolites.

By design of the ISA, it finds subsets of data where many features show coherent variation only over a subset of samples. We think this property of ISA is very suitable for integrating data from heterogeneous samples such as diseased or medicated subpopulations. Current implementation of metabomatching allows simultaneous identification upto two compounds. We observed this when an ISA module captured both ethanol and its specific product ethyl glucuronide, demonstrating the power of identifying compounds in the same pathway. Extending metabomatching beyond two compound identification is challenging due to the high number of combinations, yet future work can address this issue by limiting the tested metabolites to those belonging to a particular pathway.

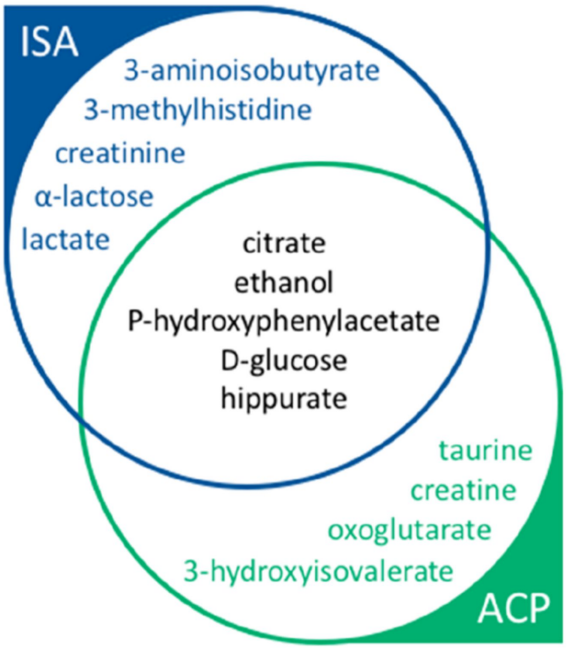


Figure 37: Metabolites robustly matched by metabomatching to pseudospectra driven by iterative signature algorithm (ISA, in blue), average correlation profile (ACP, in green), or both methods (black).

To conclude we believe this work shows the potential for large-scale automated analysis of NMR, and increased sample size shall allow identification of further metabolites. The paper entitled, *Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites*, was published in the Journal of proteome research in July 2019. The manuscript is accessible in Appendix 4 and also online:

https://pubs.acs.org/doi/pdf/10.1021/acs.jproteome.9b00295 .

5 Causality analysis - Integration of genotype, gene expression & metabolomics data

In this chapter I further study the association between the metabolome and the gene expression by exploring the causality between them. To this end I use SNPs as instrumental variables and analyze the direction of the causality between gene-metabolite pairs that were significantly associated in the metabolome-transcriptome association study.

5.1 Methodology

We performed Mendelian randomization (MR) analysis [53, 54] to assess the causal relationship between gene expression and metabolite concentration. While we used SNPs as instrumental variables (IVs), gene expression and metabolome features were interchangeably used as exposure and outcome to determine the direction of causality. For the MR analysis, we used summary statistics from mQTL/eQTL studies with higher statistical power [135, 169]. Causal effects were estimated by using Wald method where the effect of a genetic variant on the outcome is divided by the effect of the same genetic variant on the exposure [58]. Next, ratio estimates from different instruments (SNPs) were combined by the inverse variance weighted method (IVW) to calculate the causal estimate [183].

IVs were selected based on them being significant eQTL/mQTL in the relevant databases. To detect the independent SNPs we used a stepwise pruning approach where first we selected the strongest lead eQTL/mQTL and stepwise pruned the rest of the SNPs if they were correlated with the lead SNP ($r^2 > 0.2$). We repeated the pruning process with the next available SNP until there were no SNPs left to prune. We used Cochran's Q test to determine heterogeneity among the candidate instruments [55]. The SNPs were pruned in a stepwise manner from the model until the model did not show any more signs of heterogeneity (Cochran's Q statistic p-value > 0.05/#of original instruments). We also applied more robust MR analysis methods than IVW, such as the median estimator and MR-Egger regression to evaluate the significance of the causal estimates [56]. These methods are known to have more relaxed MR assumptions and they can

tolerate the violation of the exclusion-restriction assumption for some instruments. For all MR analysis we used the Mendelian Randomization package implemented in R [184].

5.2 Results & Conclusions

We performed MR analysis using summary statistics from the eQTLGen Consortium [135] and Raffler et al. [169] for eQTL and untargeted mQTL results, respectively. We investigated both the causal effect of the gene expression on the metabolite concentration and vice versa for the *ALMS1*-NAA and *HPS1*-TMA gene-metabolite pairs.

In the MR analysis where we investigated the causal effect of *ALMS1* gene on NAA concentration, instrumental variables (IVs) were selected among the SNPs that were reported as significant eQTLs (FDR<0.05) in eQTLGen and that were also measured in Raffler et al., resulting in 86 SNPs. By applying the stepwise pruning approach (see Methods) we found 14 independent SNPs as candidate IVs. Next, we performed Cochran's Q test to detect heterogeneity among these 14 SNPs and removed a further three of those, resulting in 11 SNPs as potentially valid IVs to use in the MR analysis (see Methods). As for the outcome, we used NMR peak intensities as proxies for the concentration of NAA as there were no targeted studies reporting summary statistics explicitly for NAA concentration. To this end we used the peak at 2.0308 ppm reported in Raffler et al., as this peak is the highest peak in the NAA spectrum and often used to estimate the concentration of N-acetylated compounds (NAC) [172,173]. NAA has other NMR peaks in its spectrum, yet the observed intensities at these peaks are much lower and therefore difficult to detect robustly by NMR spectroscopy. Indeed these peaks were only weakly correlated amongst themselves and with the main peak at 2.03 ppm region (Pearson correlation coefficient<0.5), so they were too noisy to define a more robust estimate of the NAA concentration than the main peak on its own. For these reasons we decided to perform our MR analysis using only the intensity measure at 2.03 ppm as outcome, which implies therefore that we studied the causality of any NAC rather than NAA specifically. Causal effect estimates given by different meta-analysis methods are reported in Table 9. All methods agreed on *ALMS1* expression level being causal for NAC concentrations.

For the completeness of the analysis, we also tested the causal effect of NAC on *ALMS1* gene expression level. IVs were selected among the SNPs that were reported as significant mQTLs

(p-value < 1×10-6) in Raffler et al. [27]. Amongst the *cis*-eQTLs of *ALMS1* from eQTLGen, most candidate IVs seemed to have direct pleiotropic effect on *ALMS1* expression in *cis*, reflected by the strong heterogeneity between their expected and observed effects. To overcome this problem we sought to use also *trans*-eQTLs of *ALMS1*, however none of the candidate IVs were measured in the *trans*-eQTL study of eQTLGen. As an alternative, we performed an association study between the candidate IVs and *ALMS1* gene expression level as measured in CoLaus and used these eQTL results in the MR analysis. Overall, we identified 26 significant mQTLs for the 2.03 ppm feature in Raffler et al. (p-value < 1×10-6) which corresponded to six independent SNPs. Two of the six candidate IVs exhibited pleiotropic effects and they were removed from the analysis. Finally, we had four SNPs as potentially valid IVs to use in the MR analysis (see Methods). Causal effect estimates given by different meta-analysis methods are reported in Table 9. None of the methods found NAC concentration to be causal for *ALMS1* gene expression level. However, it should be noted that due to low sample size of *trans*-eQTL study, this particular MR analysis was underpowered.

	Method	Causal Effect Size Estimate	Std. Error	95% CI	P-value	Cochran's Q-statistic p-value
ALMS1 -> NAC	Inverse variance weighted	0.967	0.061	0.847 - 1.087	$< 2\times10^{-16}$	0.2323
	Weighted median	1.111	0.075	0.965 - 1.257	$< 2\times10^{-16}$	NA
	MR - Egger	0.994	0.092	0.812 - 1.175	$< 2\times10^{-16}$	0.1776
	Maximum-likelihood	0.999	0.065	0.872 - 1.126	$< 2\times10^{-16}$	0.249
NAC -> ALMS1	Inverse variance weighted	-0.015	0.264	-0.532 - 0.502	0.955	0.7443
	Weighted median	0.122	0.321	-0.507 - 0.751	0.704	NA
	MR - Egger	1.495	1.976	-2.377 - 5.368	0.449	0.7256
	Maximum-likelihood	-0.015	0.266	-0.535 - 0.505	0.955	0.7443

Table 9: MR results for testing causal effect of *ALMS1* gene expression levels on N-acetylated compounds (*ALMS1* -> NAC) and MR results for testing causal effect of N-acetylated compounds on *ALMS1* gene expression levels (NAC -> *ALMS1*) using summary statistics data.

For the MR analysis of the *HPS1* gene, IVs were selected among the SNPs that were reported as significant eQTLs (FDR<0.05) in eQTLGen and that were also measured in Raffler et al. [27]. As for the outcome, similarly to NAA, there were no studies reporting targeted summary statistics for TMA concentration, therefore we used the NMR peak intensities to estimate the concentration of TMA. According to HMDB, TMA has one singlet at 2.89 ppm where the peak position ranges from 2.79 to 2.99 ppm. In the Raffler et al. dataset we used the intensity of feature at 2.8541 ppm as a proxy of TMA concentration. For the MR analysis we had 77

candidate SNPs six of which were selected as valid IVs as they were independent and did not exhibit heterogeneity (see Methods). Causal effects estimated by using different meta-analysis methods are reported in Table 10. All of the methods agreed on *HPS1* gene expression having a causal effect on TMA concentration.

We also explored the causal effect in the other direction, testing the causal effect of TMA concentration on *HPS1* gene expression. There were 87 significant mQTLs in Raffler et al. [27] that were also measured in eQTLGen. By applying the stepwise pruning approach and removing the SNPs showing heterogeneity (see Methods) we had 18 SNPs to use as IVs in the MR analysis. Causal effects estimated by using different meta-analysis methods are reported in Table 10. All of the methods agreed on TMA concentration being causal on *HPS1* expression. To sum up, the estimated causal effect size of *HPS1* on TMA ranged from 0.27 to 0.37 depending on the method, while the causal effect size of TMA on *HPS1* was around -0.09, pointing to the existence of a negative feedback loop.

	Method	Causal Effect Size Estimate	Std. Error	95% CI	P-value	Cochran's Q-statistic p-value
HPS1 -> TMA	Inverse variance weighted	0.266	0.094	0.082 - 0.450	0.005	0.0803
	Weighted median	0.311	0.072	0.170 - 0.453	$< 2\times10^{-16}$	NA
	MR - Egger	0.37	0.126	0.123 - 0.617	0.003	0.0852
	Maximum-likelihood	0.267	0.094	0.083 - 0.452	0.004	0.0829
TMA -> HPS1	Inverse variance weighted	-0.089	0.012	-0.113 - -0.065	< 2×10-16	0.0958
	Weighted median	-0.09	0.011	-0.111 - -0.068	$< 2\times10^{-16}$	NA
	MR - Egger	-0.086	0.013	-0.111 - -0.061	$< 2\times10^{-16}$	0.0758
	Maximum-likelihood	-0.09	0.012	-0.114 - -0.066	$< 2\times10^{-16}$	0.1258

Table 10: MR results for testing causal effect of *HPS1* gene expression level on TMA (*HPS1* -> TMA) and MR results for testing causal effect of TMA on *HPS1* gene expression level (TMA -> *HPS1*) using summary statistics data.

6 Conclusions

6.1 Summary of findings

The main findings of this thesis relate to gene expression data; either about the patterns we observed within these data (via modular analysis) or the association of these data with other molecular phenotypes and genotypes.

For the modular analysis of CoLaus gene expression data, we used the biclustering algorithm *Iterative Signature Algorithm*, to identify clusters of genes whose expression levels were similar over a subset of samples. Reducing the expression data to modules lowers both the complexity and the noise of the data. This modular analysis of gene expression provides particular insight when genes belonging to a module are enriched for a certain biological function, characterized for example by a gene ontology (GO) term, a phenotype in the GWAS catalog, or a disease in OMIM (Online Mendelian Inheritance in Man). As the enriched modules would potentially involve genes that are poorly annotated or not annotated at all, these unannotated genes become plausible candidates to further study their relevance for the enriched GO term or disease. By working with data from the phenotype rich cohort CoLaus, we also had the chance to investigate phenotypic enrichment of the module samples. Concordance in phenotype and gene enrichment then indicates modules of particular biological relevance. However, the low sample size of our gene expression data did constitute a limiting factor, so that the modular approach was unlikely to provide novel results. Our focus for the modular analysis was therefore to verify the integrity of the gene expression data and of the processing procedure we applied. Overall we found six modules whose phenotype and gene enrichment were concordant. The phenotype and the corresponding gene enrichment of the six modules are as follows: 1) Pro-BNP, a heart failure marker; and amyotrophic lateral sclerosis, a multi-system neurodegenerative disorder that has implications on cardiac function 2) glucose; and hemochromatosis, a disease related to high iron accumulation in the body where the patients are reported to suffer from diabetes 3) LDL size; and the multiple sclerosis where it has been reported that relapsing-remitting multiple sclerosis patients having smaller LDL compared to healthy people 4) homocysteine, a non-classical

cardiovascular risk factor; and atrial fibrillation and fibrosis 5) heart rate; and cardiovascular diseases 6) Gamma GT, a liver enzyme and alcohol consumption marker; and high alcohol use.

By integrating CoLaus gene expression data with genotypes we performed a *cis*-eQTL analysis. Even though we considered a limited subset of protein coding genes for the analysis, had a low sample size and did not apply preprocessing steps to boost the *cis*-eQTL discovery, we found that the *cis*-eQTLs of Colaus were overrepresented in other eQTL databases representing blood (Blood eQTL browser) and LCL (GEUVADIS) eQTLs. Even though Blood eQTL browser is a more powerful study with 5,311 samples compared to Geuvadis with 373 samples, CoLaus had a larger overlap with GEUVADIS. The fact that GEUVADIS and CoLaus both use LCLs and Illumina RNA-seq, is the likely reason why we observed a larger overlap with GEUVADIS compared to Blood eQTL Browser which uses blood tissue and microarray technology.

We investigated the association between CoLaus gene expression and urine metabolome data to identify genes influencing the human metabolome. To the best of our knowledge, this was the first time such a study was performed on the untargeted urine metabolome of healthy individuals. We identified one gene, *ALMS1*, whose association with NMR features was highly significant, surviving even the most conservative correction for multiple hypothesis testing. We also identified other genes including *ALMS1P* and *HPS1*, that were associated with metabolome features with marginal significance with p-values below an adjusted threshold accounting for the estimated number of independent variables. We also observed that among the top genes we discovered through this transcriptome-metabolome association analysis, many were in loci with SNPs that have been previously reported as mQTLs. This shows the sensitivity of our study to extract likely candidates of metabolically relevant genes, despite its small sample size and low power. We used metabomatching to search for promising metabolite candidates underlying gene expression-metabolome features associations. This approach was particularly insightful for our top hit *ALMS1*, as well as the strongest marginally significant association involving *HPS1*: both genes had previously been implicated by mGWAS linking their loci to compound families. However, in both cases the reported locus also harbored other genes, leaving the exact gene-metabolite association ambiguous. We found N-Acetylaspartate (NAA) as the potential underlying metabolite whose urine concentration is correlated with *ALMS1* expression. Indeed, a number of metabolome- and genome-wide association studies (mGWAS) had already suggested the locus of this gene to be involved in regulation of N-acetylated compounds (NAC),

yet were not able to identify unambiguously the exact metabolite, nor to disambiguate between *ALMS1* and *NAT8*, another gene found in the same locus as the mediator gene. We also found *HPS1* associating with trimethylamine (TMA). mGWAS had previously implicated a locus containing *HPS1* to be associated with TMA concentrations in urine but could not disambiguate this association signal from *PYROXD2*, a gene in the same locus. Finally we used Mendelian Randomization (MR) to study the direction of causality between the *ALMS1* and *HPS1* genes and their respective associated metabolites. For *ALMS1*- NAA association, the MR results suggested *ALMS1* gene expression levels being causal on NAC. For the *HPS1* - TMA association causal estimates were significant in both directions yet the causal effect size was much smaller for the effect of TMA on *HPS1*. In addition the causal effect sizes had opposite signs, thus showing presence of negative feedback loop between *HPS1* gene expression and TMA concentration. Our study provides evidence that the integration of metabolomics with gene expression data can support mQTL analysis, helping to identify the most likely gene involved in the modulation of the metabolite concentration.

We also performed metabolome-genome wide association study on Colaus data and discovered mQTLs. When we compared some of the results from this study to the metabolome-transcriptome association study of CoLaus, we observed at times using gene expression data instead of genetic variants were more effective for the metabolite identification; in particular metabomatching plots had more precise matches as the association signals were less noisy.

Throughout these projects I had the opportunity to take different steps involved in scientific research. I started with a broad view, an open question on how two molecular datasets of matching samples relate to each other. I employed a data-driven approach by performing an association study. The findings pointed to a handful of associations allowing me to construct more concrete testable hypotheses to further study them in detail as in the case of causality analysis between the entities. In a way we were fortunate enough to have a reasonable number of results that allowed me to do more focused analysis. If I were to have many more associations I would probably pursue different approaches such as pathway enrichment / annotation analyses to study the relevance of the association results.

6.2 Outlook

The results presented in this thesis are derived from a relatively small subset of a larger cohort; with 555 samples of matching gene expression and metabolome data out of 6,187 samples genotyped. Having a small sample size limits the statistical power to detect variants with small effect sizes. Nowadays it is not uncommon to have cohorts as large as half million participants and of course the discovery capacity of those cohorts is not comparable to CoLaus. Although our analysis using CoLaus data showed the premise of integrating molecular phenotypes, it would certainly benefit from having a larger sample size.

The gene expression dataset we used was gathered from lymphoblastoid cell lines. While such cell lines have many advantages, including a reduction of environmental perturbations that affect in-vivo tissues, these cells are not very similar to those that directly affect the urine metabolome. Indeed, having access to gene expression data from more relevant tissue such as the kidney, could have yielded more associations. Yet, given the design of the cohort this was not an option, since CoLaus participants were randomly selected people from the Lausanne population and therefore not a viable source to give tissue samples or biopsies. If we used another cohort that kidney samples were available for the gene expression analysis then most likely they would not be healthy individuals. Therefore it was always a tradeoff between having more appropriate tissue to couple with urine metabolism and having the chance to study the transcriptome-metabolome link in healthy individuals. Nevertheless it would be interesting to see to what extent the discovered associations would persist if gene expression of a more relevant tissue was used in the association analysis.

We used untargeted urine metabolomics data of CoLaus and used metabomatching method to identify the metabolites underlying the association signals. We found two metabolites in particular that are strongly associated with gene expression levels, namely N-acetylaspartate and trimethylamine. It would be interesting, if we were to do targeted metabolomics and quantify these two metabolites specifically in order to see if we continue to observe the associations between these metabolites and the genes.

We took a data-driven approach and generated hypotheses that could potentially be further tested. A good way to validate the hypothesis we generated by integrating gene expression and

metabolome data, would be to set up a knockdown experiment in mice where the homologs of human genes *ALMS1* and *HPS1*, *Alms1* and *Hps1* respectively, are inactivated. And indeed if the metabolite concentrations of N-acetylaspartate and trimethylamine were to respond to the silencing of these genes. On the other hand, extrapolating results gathered from rodent knockdown experiments to humans can be questionable at times. A more straightforward way to evaluate the validity of the discovered gene-metabolite links would be to reach out to human patients suffering from the malfunctioning of the mentioned genes. For instance it is known that mutations in the *ALMS1* gene can give rise to Alstrom syndrome, a rare genetic disorder affecting multiple systems of the patient including hearing/vision abnormalities, obesity, heart disease, diabetes, kidney and liver problems. If there were a possibility to design a study with the people affected by this syndrome and acquire their urine samples, we could analyse their urine to look for abnormalities in their urine N-acetylaspartate concentrations.

To the best of our knowledge, there were no other studies with urine NMR spectra and gene expression data of LCLs derived from the same subjects that are of comparable or larger sample size to the data we analysed. As a result, we could not perform proper out-of-sample replication of our results. It would be intriguing to perform such a replication study in the future when such a replication cohort becomes available.

CoLaus being a longitudinal study, we had access to follow-ups of clinical phenotypes and urine metabolomics data, whereas gene expression data was only acquired in the baseline sampling. Availability of the longitudinal phenotypic and metabolomics data would allow us to perform prospective studies, with the aim to identify whether any metabolic markers measured at baseline have predictive power for new incidences (during the follow-up period) or for significant changes of risk factors or disease incidents. The prospective design of metabolomics would also be useful as the metabolic markers would have been measured years before the clinical event manifested itself, as its treatment and comorbidities would have confounded the analysis.

Even though we had the chance to investigate the association of baseline gene expression with follow-up metabolomics data, it would have also been interesting to see to what extent the association between metabolome and gene expression would be persisted throughout the matching follow-ups of both datasets. Whether a change in expression correlates with a change in the metabolome across several years or decades is a very interesting question to examine the

relatedness of the two molecular entities and also their robustness to changing conditions. Another curious question would be if the baseline gene expression values could predict the follow-up urine metabolome or vice versa; yet given the not so high correlation of gene expression values with metabolome, more powerful studies with higher sample sizes would be required to investigate this.

We performed pathway enrichment analysis for the genes we found in our modular analysis of gene expression data alone. We could have also performed a pathway enrichment analysis for the genes that reached to study-wise significance with their association with metabolome features, even though their association profile did not point to clear metabolite matches in metabomatching. It might as well be that the genes associating with metabolome features are enriched for metabolism related functions. Complementary to this we could have also performed metabolic pathway enrichment analysis for the candidate metabolites pointed by metabomatching. However with our untargeted urine metabolome data we would be always limited to a small number of metabolites which would have a limited value for pathway analysis.

We could also quantify gene expression in transcript level and study the effect of isoforms in our association studies. Lastly, having access to epigenomics data of CoLaus participants would have been of great use given its relatedness with gene expression data.

References

1. King, R.A., J.I. Rotter, and A.G. Motulsky, *The genetic basis of common diseases*. Vol. 44. 2002: Oxford university press.
2. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome*. Nature, 2004. **431**(7011): p. 931.
3. Consortium, I.H., *Integrating common and rare genetic variation in diverse human populations*. Nature, 2010. **467**(7311): p. 52.
4. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies*. Nature Reviews Genetics, 2016. **17**(6): p. 333.
5. LaFramboise, T., *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances*. Nucleic acids research, 2009. **37**(13): p. 4181-4193.
6. Tam, V., et al., *Benefits and limitations of genome-wide association studies*. Nature Reviews Genetics, 2019. **20**(8): p. 467-484.
7. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey*. Human molecular genetics, 2008. **17**(R2): p. R156-R165.
8. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019*. Nucleic acids research, 2018. **47**(D1): p. D1005-D1012.
9. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci*. Nature genetics, 2010. **42**(12): p. 1118.
10. Eeles, R.A., et al., *Multiple newly identified loci associated with prostate cancer susceptibility*. Nature genetics, 2008. **40**(3): p. 316.
11. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci*. Nature genetics, 2010. **42**(6): p. 504.
12. Beecham, A.H., et al., *Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis*. Nature genetics, 2013. **45**(11): p. 1353.
13. Ripke, S., et al., *Genome-wide association analysis identifies 13 new risk loci for schizophrenia*. Nature genetics, 2013. **45**(10): p. 1150.
14. Billings, L.K. and J.C. Florez, *The genetics of type 2 diabetes: what have we learned from GWAS?* Annals of the New York Academy of Sciences, 2010. **1212**: p. 59.
15. Wang, K., et al., *A genome-wide association study on obesity and obesity-related traits*. PloS one, 2011. **6**(4): p. e18939.
16. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height*. Nature genetics, 2014. **46**(11): p. 1173.
17. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease*. Nature genetics, 2008. **40**(2): p. 161.
18. Hardy, J. and A. Singleton, *Genomewide association studies and human disease*. New England Journal of Medicine, 2009. **360**(17): p. 1759-1768.
19. Manolio, T.A., et al., *Finding the missing heritability of complex diseases*. Nature, 2009. **461**(7265): p. 747-753.
20. Consortium, G., *The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans*. Science, 2015. **348**(6235): p. 648-660.
21. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA*. Science, 2012. **337**(6099): p. 1190-1195.

22. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS*. PLoS genetics, 2010. **6**(4): p. e1000888.
23. Ward, L.D. and M. Kellis, *Interpreting non-coding variation in complex disease genetics*. Nature biotechnology, 2012. **30**(11): p. 1095.
24. Lloyd-Jones, L.R., et al., *The genetic architecture of gene expression in peripheral blood*. The American Journal of Human Genetics, 2017. **100**(2): p. 228-237.
25. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population*. Nature, 2010. **464**(7289): p. 773.
26. Wright, F.A., et al., *Heritability and genomics of gene expression in peripheral blood*. Nature genetics, 2014. **46**(5): p. 430.
27. Kastenmüller, G., et al., *Genetics of human metabolism: an update*. Human molecular genetics, 2015. **24**(R1): p. R93-R101.
28. Bartel, J., et al., *The Human Blood Metabolome-Transcriptome Interface*. PLoS Genet, 2015. **11**(6): p. e1005274.
29. Burkhardt, R., et al., *Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood*. PLoS Genet, 2015. **11**(9): p. e1005510.
30. Inouye, M., et al., *Metabonomic, transcriptomic, and genomic variation of a population cohort*. Molecular systems biology, 2010. **6**(1).
31. Ritchie, M.D., *The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era*. Human genetics, 2012. **131**(10): p. 1615-1626.
32. Rauch, A., et al., *Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study*. Gastroenterology, 2010. **138**(4): p. 1338-1345. e7.
33. Joyner, M.J., *Precision medicine, cardiovascular disease and hunting elephants*. Progress in cardiovascular diseases, 2016. **58**(6): p. 651-660.
34. Ashley, E.A., *Towards precision medicine*. Nature Reviews Genetics, 2016. **17**(9): p. 507.
35. Collins, F.S. and H. Varmus, *A new initiative on precision medicine*. New England journal of medicine, 2015. **372**(9): p. 793-795.
36. Jameson, J.L. and D.L. Longo, *Precision medicine—personalized, problematic, and promising*. Obstetrical & gynecological survey, 2015. **70**(10): p. 612-614.
37. Naghavi, M., et al., *Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016*. The Lancet, 2017. **390**(10100): p. 1151-1210.
38. Kaptoge, S., et al., *World Health Organization cardiovascular disease risk charts: revised prediction models to estimate risk in 21 global regions*. 2019.
39. Firmann, M., et al., *The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome*. BMC cardiovascular disorders, 2008. **8**(1): p. 6.
40. Affymetrix, *BRLMM: an improved genotype calling method for the genechip human mapping 500k array set*. 2006, Affymetrix Santa Clara, CA.
41. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. The American journal of human genetics, 2007. **81**(3): p. 559-575.
42. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes*. Nature genetics, 2007. **39**(7): p. 906-913.
43. Alonso, A., et al., *Focus: a robust workflow for one-dimensional NMR spectral analysis*. Analytical chemistry, 2014. **86**(2): p. 1160-1169.

44. Lumley, T., et al., *The importance of the normality assumption in large public health data sets.* Annual review of public health, 2002. **23**(1): p. 151-169.
45. Benjamin, D.J., et al., *Redefine statistical significance.* Nature Human Behaviour, 2018. **2**(1): p. 6.
46. Dunn, O.J., *Multiple comparisons among means.* Journal of the American statistical association, 1961. **56**(293): p. 52-64.
47. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Statistics in medicine, 1990. **9**(7): p. 811-818.
48. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
49. Storey, J.D., *The positive false discovery rate: a Bayesian interpretation and the q-value.* The Annals of Statistics, 2003. **31**(6): p. 2013-2035.
50. Bergmann, S., J. Ihmels, and N. Barkai, *Iterative signature algorithm for the analysis of large-scale gene expression data.* Physical review E, 2003. **67**(3): p. 031902.
51. Ihmels, J., S. Bergmann, and N. Barkai, *Defining transcription modules using large-scale gene expression data.* Bioinformatics, 2004. **20**(13): p. 1993-2003.
52. Ihmels, J., et al., *Revealing modular organization in the yeast transcriptional network.* Nature genetics, 2002. **31**(4): p. 370-377.
53. Burgess, S., D.S. Small, and S.G. Thompson, *A review of instrumental variable estimators for Mendelian randomization.* Statistical methods in medical research, 2017. **26**(5): p. 2333-2355.
54. Davey Smith, G. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International journal of epidemiology, 2003. **32**(1): p. 1-22.
55. Greco M, F.D., et al., *Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome.* Statistics in medicine, 2015. **34**(21): p. 2926-2940.
56. Staley, O.Y.J., *MendelianRandomization: Mendelian Randomization Package. R package version 0.4.1.* 2019.
57. Baum, C.F., M.E. Schaffer, and S. Stillman, *Instrumental variables and GMM: Estimation and testing.* The Stata Journal, 2003. **3**(1): p. 1-31.
58. Wald, A., *The fitting of straight lines if both variables are subject to error.* The Annals of Mathematical Statistics, 1940. **11**(3): p. 284-300.
59. Sugimoto, M., et al., *Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein-Barr virus.* Cancer research, 2004. **64**(10): p. 3361-3364.
60. Mohyuddin, A., et al., *Genetic instability in EBV-transformed lymphoblastoid cell lines.* Biochimica et Biophysica Acta (BBA)-General Subjects, 2004. **1670**(1): p. 81-83.
61. Amoli, M., et al., *EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood.* International journal of epidemiology, 2008. **37**(suppl_1): p. i41-i45.
62. Neitzel, H., *A routine method for the establishment of permanent growing lymphoblastoid cell lines.* Human genetics, 1986. **73**(4): p. 320-326.
63. Sie, L., S. Loong, and E. Tan, *Utility of lymphoblastoid cell lines.* Journal of neuroscience research, 2009. **87**(9): p. 1953-1959.
64. Thorley-Lawson, D.A. and A. Gross, *Persistence of the Epstein–Barr virus and the origins of associated lymphomas.* New England Journal of Medicine, 2004. **350**(13): p. 1328-1337.
65. Nickles, D., et al., *In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing.* BMC genomics, 2012. **13**(1): p. 477.

66. Redon, R., et al., *Global variation in copy number in the human genome.* nature, 2006. **444**(7118): p. 444.
67. Consortium, I.H., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299.
68. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743.
69. Stranger, B.E., et al., *Population genomics of human gene expression.* Nature genetics, 2007. **39**(10): p. 1217.
70. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506.
71. Li, J.-W., et al., *Transcriptome sequencing of Chinese and Caucasian population identifies ethnic-associated differential transcript abundance of heterogeneous nuclear ribonucleoprotein K (hnRNPK).* Genomics, 2014. **103**(1): p. 56-64.
72. Martin, A.R., et al., *Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture.* PLoS genetics, 2014. **10**(8): p. e1004549.
73. Storey, J.D., et al., *Gene-expression variation within and among human populations.* The American Journal of Human Genetics, 2007. **80**(3): p. 502-509.
74. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-853.
75. Chen, Y., et al., *Variations in DNA elucidate molecular networks that cause disease.* Nature, 2008. **452**(7186): p. 429-435.
76. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-428.
77. Cookson, W., et al., *Mapping complex disease traits with global gene expression.* Nature Reviews Genetics, 2009. **10**(3): p. 184-194.
78. Hu, V.W., et al., *Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes.* BMC genomics, 2006. **7**(1): p. 118.
79. Baron, C.A., et al., *Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism.* Journal of autism and developmental disorders, 2006. **36**(8): p. 973-982.
80. Kakiuchi, C., et al., *Up-regulation of ADM and SEPX1 in the lymphoblastoid cells of patients in monozygotic twins discordant for schizophrenia.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2008. **147**(5): p. 557-564.
81. Joehanes, R., et al., *Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study.* Physiological genomics, 2012. **44**(1): p. 59-75.
82. Abe, K., et al., *Induction of amyloid precursor protein mRNA after heat shock in cultured human lymphoblastoid cells.* Neuroscience letters, 1991. **125**(2): p. 169-171.
83. Gutekunst, C.-A., et al., *Identification and localization of huntingtin in brain and human lymphoblastoid cell lines with anti-fusion protein antibodies.* Proceedings of the National Academy of Sciences, 1995. **92**(19): p. 8710-8714.
84. Kobayashi, H., et al., *Haploinsufficiency at the α-synuclein gene underlies phenotypic severity in familial Parkinson's disease.* Brain, 2003. **126**(1): p. 32-42.
85. Arosio, B., et al., *Fibroblasts from Alzheimer's disease donors do not differ from controls in response to heat shock.* Neuroscience letters, 1998. **256**(1): p. 25-28.
86. Hayashi-Takagi, A., M.P. Vawter, and K. Iwamoto, *Peripheral biomarkers revisited: integrative profiling of peripheral samples for psychiatric research.* Biological psychiatry, 2014. **75**(12): p. 920-928.

87. Sanders, A.R., et al., *Transcriptome study of differential expression in schizophrenia.* Human molecular genetics, 2013. **22**(24): p. 5001-5014.
88. Yoshimi, A., et al., *Proteomic analysis of lymphoblastoid cell lines from schizophrenic patients.* Translational psychiatry, 2019. **9**(1): p. 126.
89. Kitchen, R.R., et al., *Decoding neuroproteomics: integrating the genome, translatome and functional anatomy.* Nature neuroscience, 2014. **17**(11): p. 1491.
90. Dirksen, E.H., et al., *Human lymphoblastoid proteome analysis reveals a role for the inhibitor of acetyltransferases complex in DNA double-strand break response.* Cancer research, 2006. **66**(3): p. 1473-1480.
91. Toda, T. and M. Sugimoto, *Proteome analysis of Epstein–Barr virus-transformed B-lymphoblasts and the proteome database.* Journal of Chromatography B, 2003. **787**(1): p. 197-206.
92. Caron, M., et al., *Proteomic map and database of lymphoblastoid proteins.* Journal of Chromatography B, 2002. **771**(1-2): p. 197-209.
93. Welsh, M., et al., *Pharmacogenomic discovery using cell-based models.* Pharmacological reviews, 2009. **61**(4): p. 413-429.
94. Wheeler, H.E. and M.E. Dolan, *Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation.* Pharmacogenomics, 2012. **13**(1): p. 55-70.
95. Farrell, P.J., *Epstein-Barr virus immortalizing genes.* Trends in microbiology, 1995. **3**(3): p. 105-109.
96. Çalışkan, M., et al., *The effects of EBV transformation on gene expression levels and methylation profiles.* Human molecular genetics, 2011. **20**(8): p. 1643-1652.
97. Bullaughey, K., et al., *Expression quantitative trait loci detected in cell lines are often present in primary tissues.* Human molecular genetics, 2009. **18**(22): p. 4296-4303.
98. Mazzei, F., et al., *8-Oxoguanine DNA-glycosylase repair activity and expression: a comparison between cryopreserved isolated lymphocytes and EBV-derived lymphoblastoid cell lines.* Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 2011. **718**(1-2): p. 62-67.
99. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type–dependent manner.* Science, 2009. **325**(5945): p. 1246-1250.
100. Ozgyin, L., et al., *Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research.* Scientific reports, 2019. **9**(1): p. 1-16.
101. Grafodatskaya, D., et al., *EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines.* Genomics, 2010. **95**(2): p. 73-83.
102. Nam, H.-Y., et al., *Human lymphoblastoid cell lines: a goldmine for the biobankomics era.* Pharmacogenomics, 2011. **12**(6): p. 907-917.
103. Sutcliffe, J.G., et al., *Common 82-nucleotide sequence unique to brain RNA.* Proceedings of the National Academy of Sciences, 1982. **79**(16): p. 4942-4946.
104. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-487.
105. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nature biotechnology, 1996. **14**(13): p. 1675-1680.
106. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-470.
107. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-536.
108. Emrich, S.J., et al., *Gene discovery and annotation using LCM-454 transcriptome sequencing.* Genome research, 2007. **17**(1): p. 69-73.

109. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31-46.
110. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621.
111. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews genetics, 2009. **10**(1): p. 57-63.
112. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome biology, 2010. **11**(12): p. 220.
113. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome biology, 2016. **17**(1): p. 13.
114. Hart, S.N., et al., *Calculating sample size estimates for RNA sequencing data.* Journal of computational biology, 2013. **20**(12): p. 970-978.
115. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
116. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.
117. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature biotechnology, 2013. **31**(1): p. 46.
118. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.
119. Klei, L., et al., *GemTools: a fast and efficient approach to estimating genetic ancestry*. arXiv preprint arXiv:1104.1162, 2011.
120. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-883.
121. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC bioinformatics, 2013. **14**(1): p. 128.
122. Raju, K. and S.M. Venkataramappa, *Primary hemochromatosis presenting as Type 2 diabetes mellitus: A case report with review of literature.* International Journal of Applied and Basic Medical Research, 2018. **8**(1): p. 57.
123. Jorissen, W., et al., *Relapsing-remitting multiple sclerosis patients display an altered lipoprotein profile with dysfunctional HDL.* Scientific reports, 2017. **7**: p. 43410.
124. Kruglyak, L. and D.A. Nickerson, *Variation is the spice of life.* Nature genetics, 2001. **27**(3): p. 234-236.
125. Johnson, G.C., et al., *Haplotype tagging for the identification of common disease genes.* Nature genetics, 2001. **29**(2): p. 233-237.
126. Consortium, G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061.
127. consortium, U.K., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**(7571): p. 82-90.
128. Nagasaki, M., et al., *Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals.* Nature communications, 2015. **6**: p. 8018.
129. Francioli, L.C., et al., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nature genetics, 2014. **46**(8): p. 818.
130. Wong, L.-P., et al., *Deep whole-genome sequencing of 100 southeast Asian Malays.* The American Journal of Human Genetics, 2013. **92**(1): p. 52-66.
131. Schaid, D.J., W. Chen, and N.B. Larson, *From genome-wide associations to candidate causal variants by statistical fine-mapping.* Nature Reviews Genetics, 2018. **19**(8): p. 491-504.

132. Westra, H.-J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nature genetics, 2013. **45**(10): p. 1238-1243.
133. Stegle, O., et al., *A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.* PLoS computational biology, 2010. **6**(5).
134. Popadin, K., et al., *Genetic and epigenetic regulation of human lincRNA gene expression.* The American Journal of Human Genetics, 2013. **93**(6): p. 1015-1026.
135. Võsa, U., et al., *Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis.* BioRxiv, 2018: p. 447367.
136. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-108.
137. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annual review of biochemistry, 2012. **81**: p. 145-166.
138. Esteller, M., *Non-coding RNAs in human disease.* Nature reviews genetics, 2011. **12**(12): p. 861.
139. Batista, P.J. and H.Y. Chang, *Long noncoding RNAs: cellular address codes in development and disease.* Cell, 2013. **152**(6): p. 1298-1307.
140. Yan, X., et al., *Comprehensive genomic characterization of long non-coding RNAs across human cancers.* Cancer cell, 2015. **28**(4): p. 529-540.
141. Edwards, S.L., et al., *Beyond GWASs: illuminating the dark road from association to function.* The American Journal of Human Genetics, 2013. **93**(5): p. 779-797.
142. Gilad, Y., S.A. Rifkin, and J.K. Pritchard, *Revealing the architecture of gene regulation: the promise of eQTL studies.* Trends in genetics, 2008. **24**(8): p. 408-415.
143. Kumar, V., et al., *Human disease-associated genetic variation impacts large intergenic non-coding RNA expression.* PLoS genetics, 2013. **9**(1).
144. McDowell, I., et al., *Many long intergenic non-coding RNAs distally regulate mRNA gene expression levels.* BioRxiv, 2016: p. 044719.
145. Ponting, C.P., P.L. Oliver, and W. Reik, *Evolution and functions of long noncoding RNAs.* Cell, 2009. **136**(4): p. 629-641.
146. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nature reviews genetics, 2009. **10**(3): p. 155-159.
147. Vance, K.W. and C.P. Ponting, *Transcriptional regulatory functions of nuclear long noncoding RNAs.* Trends in Genetics, 2014. **30**(8): p. 348-355.
148. Marques, A.C., et al., *Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs.* Genome biology, 2013. **14**(11): p. R131.
149. Schaub, M.A., et al., *Linking disease associations with regulatory information in the human genome.* Genome research, 2012. **22**(9): p. 1748-1759.
150. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-169.
151. Stegle, O., et al., *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nature protocols, 2012. **7**(3): p. 500.
152. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic acids research, 2014. **42**(D1): p. D1001-D1006.
153. Nica, A.C., et al., *Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations.* PLoS Genet, 2010. **6**(4): p. e1000895.
154. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An expanded view of complex traits: from polygenic to omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.

155. Lewis, C.M. and E. Vassos, *Prospects for using risk scores in polygenic medicine.* Genome medicine, 2017. **9**(1): p. 96.
156. Boyd, A., et al., *Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children.* International journal of epidemiology, 2013. **42**(1): p. 111-127.
157. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS genetics, 2011. **7**(2).
158. Deelen, P., et al., *Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration.* BMC research notes, 2014. **7**(1): p. 901.
159. Westra, H.-J., et al., *MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects.* Bioinformatics, 2011. **27**(15): p. 2104-2111.
160. Astle, W.J., et al., *The allelic landscape of human blood cell trait variation and links to common complex disease.* Cell, 2016. **167**(5): p. 1415-1429. e19.
161. Rueedi, R., et al., *Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links.* PLoS genetics, 2014. **10**(2): p. e1004132.
162. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018*. Nucleic acids research, 2018. **46**(D1): p. D608-D617.
163. Ulrich, E.L., et al., *BioMagResBank.* Nucleic acids research, 2007. **36**(suppl_1): p. D402-D408.
164. Khalili, B., et al., *Automated analysis of large-scale NMR data generates metabolomic signatures and links them to candidate metabolites.* bioRxiv, 2019: p. 613935.
165. Suhre, K., et al., *A genome-wide association study of metabolic traits in human urine.* Nature Genetics, 2011. **43**(6): p. 565-569.
166. Ding, J., et al., *Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals.* The American Journal of Human Genetics, 2010. **87**(6): p. 779-789.
167. Gao, X., J. Starmer, and E.R. Martin, *A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2008. **32**(4): p. 361-369.
168. MATLAB, *8.5.0.197613 (R2015a)*. 2015, The MathWorks Inc.: Natick, Massachusetts.
169. Raffler, J., et al., *Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality.* PLoS genetics, 2015. **11**(9): p. e1005487.
170. Rueedi, R., et al., *Metabomatching: Using genetic association to identify metabolites in proton NMR spectroscopy.* PLoS computational biology, 2017. **13**(12): p. e1005839.
171. Engelke, U.F., et al., *N-acetylated metabolites in urine: proton nuclear magnetic resonance spectroscopic study on patients with inborn errors of metabolism.* Clinical chemistry, 2004. **50**(1): p. 58-66.
172. Nicholson, G., et al., *A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection.* PLoS genetics, 2011. **7**(9): p. e1002270.
173. Montoliu, I., et al., *Current status on genome–metabolome-wide associations: an opportunity in nutrition research.* Genes & nutrition, 2013. **8**(1): p. 19.
174. Chambers, J.C., et al., *Genetic loci influencing kidney function and chronic kidney disease.* Nature genetics, 2010. **42**(5): p. 373-375.
175. Simmons, M., C. Frondoza, and J. Coyle, *Immunocytochemical localization of N-acetyl-aspartate with monoclonal antibodies.* Neuroscience, 1991. **45**(1): p. 37-45.
176. Masaharu, M., et al., *N-acetyl-l-aspartic acid, N-acetyl-a-l-aspartyl-l-glutamic acid and β-citryl-l-glutamic acid in human urine.* Clinica Chimica Acta, 1982. **120**(1): p. 119-126.

177. Barker, P.B., *N-acetyl aspartate—a neuronal marker?* Annals of neurology, 2001. **49**(4): p. 423-424.
178. Jung, R.E., et al., *Biochemical markers of intelligence: a proton MR spectroscopy study of normal human brain.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 1999. **266**(1426): p. 1375-1379.
179. Patel, T. and J.B. Talcott, *Moderate relationships between NAA and cognitive ability in healthy adults: implications for cognitive spectroscopy.* Frontiers in human neuroscience, 2014. **8**: p. 39.
180. Davies, G., et al., *Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function.* Nature communications, 2018. **9**(1): p. 1-16.
181. Lee, J.J., et al., *Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals.* Nature genetics, 2018. **50**(8): p. 1112-1121.
182. Cloarec, O., et al., *Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets.* Analytical chemistry, 2005. **77**(5): p. 1282-1289.
183. Hartung, J., G. Knapp, and B.K. Sinha, *Statistical meta-analysis with applications*. Vol. 738. 2011: John Wiley & Sons.
184. Staley, O.Y.J., *MendelianRandomization: Mendelian Randomization Package*. 2019, <james.staley@bristol.ac.uk>.
185. Shin, S.-Y., et al., *An atlas of genetic influences on human blood metabolites*. Nature genetics, 2014. 46(6): p. 543.

Appendices

Appendix 1: cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture

Appendix 2: Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis

Appendix 3: Untargeted metabolome- and transcriptome-wide association study identifies causal genes modulating metabolite concentrations in urine

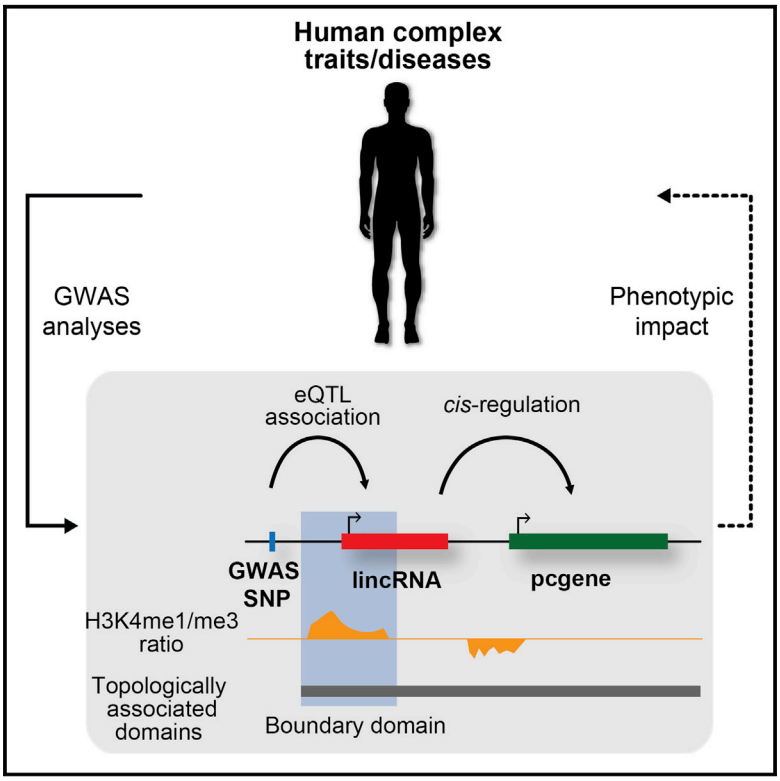
Appendix 4: Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites

Cell Reports

Resource

cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture

Graphical Abstract



Authors

Jennifer Yihong Tan, Adam Alexander Thil Smith, Maria Ferreira da Silva, ..., Zoltán Kutalik, Sven Bergmann, Ana Claudia Marques

Correspondence

jennifer.tan@unil.ch (J.Y.T.), anaclaudia.marques@unil.ch (A.C.M.)

In Brief

Tan et al. identify and characterize 69 human complex trait/disease-associated lincRNAs in LCLs. They show that these loci are often associated with *cis*-regulation of gene expression and tend to be localized at TAD boundaries, suggesting that these lincRNAs may influence chromosomal architecture.

Highlights

- We identify 69 lincRNAs associated with human complex traits (TR-lincRNAs)
- TR-lincRNAs are conserved in humans and interact with other disease-relevant loci
- TR-lincRNAs often associate with *cis*-regulation of proximal protein-coding gene expression
- TR-lincRNAs are enriched at TAD boundaries and may modulate chromatin architecture

Tan et al., 2017, Cell Reports *18*, 2280–2288
February 28, 2017
http://dx.doi.org/10.1016/j.celrep.2017.02.009

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture

Jennifer Yihong Tan,[1,2,*] Adam Alexander Thil Smith,[1,2] Maria Ferreira da Silva,[1,2] Cyril Matthey-Doret,[1,2] Rico Rueedi,[2,3] Reyhan Sönmez,[2,3] David Ding,[4] Zoltán Kutalik,[3,5] Sven Bergmann,[2,3] and Ana Claudia Marques[1,2,6,*]
[1]Department of Physiology, University of Lausanne, 1015 Lausanne, Switzerland
[2]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[5]Institute of Social and Preventive Medicine, University Hospital Lausanne (CHUV), 1011 Lausanne, Switzerland
[6]Lead Contact
*Correspondence: jennifer.tan@unil.ch (J.Y.T.), anaclaudia.marques@unil.ch (A.C.M.)
http://dx.doi.org/10.1016/j.celrep.2017.02.009

SUMMARY

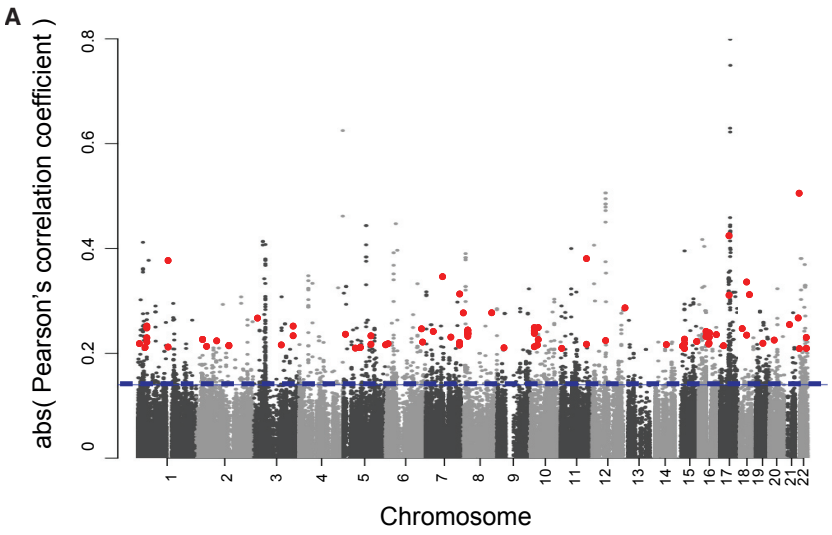
Intergenic long noncoding RNAs (lincRNAs) are the largest class of transcripts in the human genome. Although many have recently been linked to complex human traits, the underlying mechanisms for most of these transcripts remain undetermined. We investigated the regulatory roles of a high-confidence and reproducible set of 69 trait-relevant lincRNAs (TR-lincRNAs) in human lymphoblastoid cells whose biological relevance is supported by their evolutionary conservation during recent human history and genetic interactions with other trait-associated loci. Their enrichment in enhancer-like chromatin signatures, interactions with nearby trait-relevant protein-coding loci, and preferential location at topologically associated domain (TAD) boundaries provide evidence that TR-lincRNAs likely regulate proximal trait-relevant gene expression in *cis* by modulating local chromosomal architecture. This is consistent with the positive and significant correlation found between TR-lincRNA abundance and intra-TAD DNA-DNA contacts. Our results provide insights into the molecular mode of action by which TR-lincRNAs contribute to complex human traits.

INTRODUCTION

An increasing number of reports suggest that long intergenic noncoding RNAs (lincRNAs), which were previously regarded as "junk RNA" (Hüttenhofer et al., 2005), can contribute to normal and disease phenotypes in humans (Esteller, 2011). For example, candidate screens followed by detailed functional characterization of a few individual trait-associated lincRNAs illustrate how genetic variants affecting the lincRNA sequence can underlie human complex traits (Ishii et al., 2006; Zheng et al., 2016). Recently, RNA capture followed by sequencing in multiple disease-associated protein-coding gene deserts led to the identification of lowly and tissue-specifically expressed lincRNA loci (Mercer et al., 2014). Detailed experimental analysis of these lincRNA candidates is now required to establish whether and how these loci contribute to disease.

Although thousands of common genetic variants have been associated with complex human traits through genome-wide association studies (GWASs), only a small proportion fall within exonic coding sequences (Hindorff et al., 2009; Maurano et al., 2012). Instead, most GWAS variants map within noncoding regulatory regions that are enriched in population and tissue-specific expression quantitative trait loci (eQTLs) (Edwards et al., 2013). eQTL analysis has previously led to the identification of protein-coding genes and pathways that are disrupted in human complex traits (for example, Emilsson et al., 2008; Fairfax et al., 2012; Gilad et al., 2008). Recently, lincRNAs whose expression correlate with GWAS variants were also identified using this approach (Kumar et al., 2013; Lappalainen et al., 2013; McDowell et al., 2016; Popadin et al., 2013), suggesting that the transcription or the transcripts arising from lincRNA loci in eQTLs with GWAS variants may similarly contribute to phenotypes. Although a handful of studies have investigated the relationship between individual lincRNAs with risk-variant-associated expression and their linked traits (for example, Ishii et al., 2006; Jendrzejewski et al., 2012), the underlying mechanism of action for most remains undetermined.

So far, functionally characterized lincRNAs have been implicated in both transcriptional and post-transcriptional regulation of local or distal genes (Vance and Ponting, 2014). We have previously shown that chromatin signatures at lincRNA transcriptional start sites allow the distinction between these two regulatory classes (Marques et al., 2013). Specifically, the expression of lincRNAs arising from regulatory elements that carry enhancer-like chromatin signatures correlates with neighboring protein-coding gene abundance, suggesting that transcription at these loci contributes to local regulation of expression (Marques et al., 2013). Interestingly, eQTL GWAS variants are enriched within enhancer regions (Ernst et al., 2011; Schaub



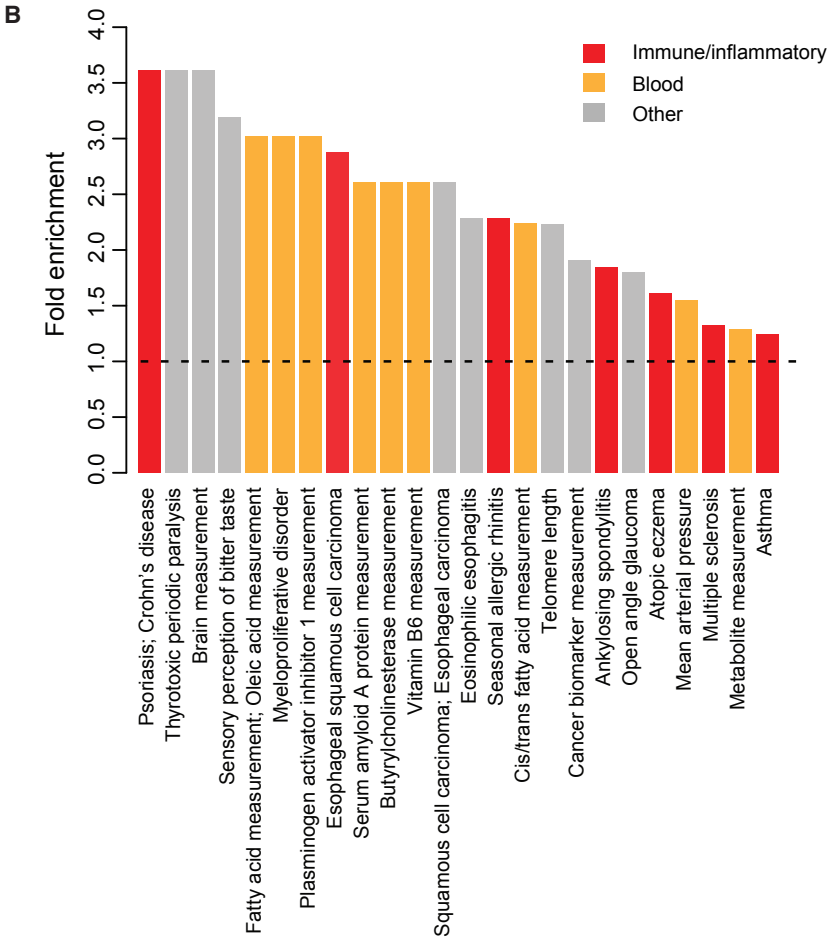


Figure 1. Identification of GWAS *cis*-eQTLs for lincRNAs and Protein-Coding Genes

(A) Manhattan plot showing absolute Pearson's correlation coefficient (*r*) calculated for all possible GWAS *cis*-eQTL associations with LCL-expressed lincRNAs (TR-lincRNAs) and protein-coding genes (TR-pcgenes) across human autosomes. Significance cutoff is represented by a horizontal dashed line (absolute *r* of 0.145). Significant TR-lincRNA *cis*-eQTLs are highlighted in red.

(B) The GWAS human complex traits that are significantly enriched (fold-enrichment, $p < 0.05$, hypergeometric test) within genome-wide significant *cis*-eQTLs (TR-lincRNAs + TR-pcgenes), relative to all possible GWAS *cis*-eQTL associations. Traits are grouped into immune/inflammatory responses (red), blood-related traits (orange), and others (gray).

See also Figure S1 and Tables S1 and S2.

et al., 2012), suggesting a link between enhancer-associated lincRNAs and complex human traits.

Here, we used functional, evolutionary, and population genomics to extensively characterize the regulatory interactions between a high-confidence set of trait-associated lincRNAs and protein-coding genes identified through GWAS *cis*-eQTL analysis. Our results demonstrate that most human complex-trait-associated lincRNAs arise from enhancer-like regions and are frequently located at the boundaries of topologically associated domains (TADs), which have been previously shown to contribute to chromosomal architecture and gene transcription regulation (Rao et al., 2014). Together, these findings support that the transcription of trait-relevant lincRNAs contributes to chromosomal architecture and thereby the regulation of nearby trait-associated protein-coding gene expression levels.

RESULTS

Identification of Trait-Relevant lincRNAs and Protein-Coding Genes

We considered all lymphoblastoid cell line (LCL)-expressed de novo (Experimental Procedures) and GENCODE-annotated loci with at least one genome-wide significant ($p < 5 \times 10^{-8}$) GWAS SNP (7,451 GWAS SNPs) (Welter et al., 2014) in their vicinity (Experimental Procedures). We calculated the Pearson's correlation between the expression of these coding and noncoding loci and the corresponding genotype of their neighboring GWAS SNPs in a panel of 373 LCLs derived from individuals of European descent (Lappalainen et al., 2013). This led to the identification of 111 and 1,479 GWAS *cis*-eQTLs significantly correlated (false discovery rate [FDR] < 5%; Experimental Procedures) with the expression levels of 73 lincRNAs and 756 protein-coding genes, respectively (Figure 1A). We asked whether differences in length and expression level (Figure S1) between lincRNAs and mRNAs would account for

the relatively lower number of eQTL-lincRNAs. After restricting our analysis to length- and expression-matched mRNAs, we found that the proportion of eQTL-lincRNAs (2.9%) is statistically indistinguishable from that of eQTL-mRNAs (3.2% of size- and expression level-matched mRNAs; $p = 0.68$, two-tailed χ^2 test), suggesting that lincRNA properties indeed limit the power to identify lincRNA-eQTLs. Despite the restricted power in lincRNA *cis*-eQTL detection, most of the identified GWAS lincRNA *cis*-eQTLs (68%; Table S1) could be replicated using data from an independent set of LCLs, derived from 555 individuals of European descent from the Lausanne population (Cohorte Lausannoise [CoLaus]; Firmann et al., 2008). The proportion of replicated lincRNA associations is similar to what was found for mRNA *cis*-eQTLs (71%, $p = 0.69$, two-tailed Fisher's exact test), corroborating the robustness of our *cis*-eQTL findings.

Evidence that these GWAS *cis*-eQTLs are enriched in immune/inflammatory response and blood-related traits, including metabolite levels (Figure 1B), suggests that despite known limitations (Choy et al., 2008), lymphoblastoid cells are suitable to investigate the contributions of lincRNA loci to human complex traits.

Genetic variants do not segregate randomly in the human population and SNPs found within the same linkage disequilibrium (LD) block are likely to correlate, to some extent, with the expression levels of all gene loci within the same LD block, leading to false-positive *cis*-eQTL associations between GWAS SNPs and gene expression (Stranger et al., 2007). To address this issue, we used regulatory trait concordance (RTC), an empirical method that accounts for local LD structure (Nica et al., 2010). We estimated the rank of the identified GWAS *cis*-eQTL among all nearby common SNPs based on decreasing absolute correlation with gene expression, thus assessing the likelihood that the identified *cis*-eQTL is most likely driven by the complex-trait-associated genetic variant and not due to local LD with another SNP. This approach does not exclude, however, that the expression of the coding or noncoding loci could be under the influence of an unknown variant in linkage with the GWAS *cis*-eQTL. After applying a previously tested RTC threshold (0.9) to identify high-confidence eQTL associations (Nica et al., 2010), we obtained 69 lincRNAs that are likely true trait-relevant gene candidates (trait-relevant lincRNAs [TR-lincRNAs]), as well as 723 protein-coding genes (TR-pcgenes; Table S1). Importantly, 73% of the GWAS *cis*-eQTLs associated with TR-lincRNAs and TR-pcgenes were validated in CoLaus, a significant 11% increase in replication rate from all identified *cis*-eQTLs ($p < 0.05$, two-tailed Fisher's exact test), reinforcing the reliability of this set.

TR-lincRNAs are likely involved in pathways relevant to their associated traits. Specifically, we asked whether the expression levels of trait-relevant loci are correlated with those of other genes associated with the same trait, as would be expected if they contribute to the same phenotype. For each trait-relevant loci, we used the pathway scoring algorithm "Pascal" (Lamparter et al., 2016) to identify all loci located within LD blocks containing other significant GWAS ($p < 5 \times 10^{-8}$) variants for that trait, and we tested for their co-expression with the *cis*-eQTL loci candidates, a surrogate for genetic interaction. We found that 83% of TR-lincRNAs (57/69) are significantly co-expressed ($p < 0.05$, permutation test; Experimental Procedures) with genes associated with the same trait, a proportion similar to that found for TR-pcgenes (89% [642/723], $p = 0.17$, two-tailed Fisher's exact test; Table S2).

Trait-Relevant lincRNAs Are Conserved in Humans

The biological relevance of lincRNA transcription is generally unclear, and there is ongoing debate as to whether it is the transcript or the act of transcription that underlies the function of most noncoding loci (Wilusz et al., 2009). Evolutionary analyses can provide initial insights into this question, as selective constraint at exons would not be required if it is the act of transcription and not the transcript sequence that underlies function.

We investigated the evolution of TR-lincRNAs' exons in humans and found that they exhibit a significantly higher proportion of low-frequency alleles (derived allele frequency [DAF] < 0.1) compared to local neutrally evolving sequences (ancestral repeats [ARs]), TR-lincRNA intronic regions, and other LCL-expressed lincRNA exons ($p < 0.05$, two-tailed Fisher's exact test; Figure 2A). The proportion of SNPs with DAF < 0.1 found within TR-lincRNA and protein-coding gene exons is statistically indistinguishable ($p = 0.56$, two-tailed Fisher's exact test; Figure 2A). This is in contrast to exons of all LCL-expressed lincRNAs, which have a similar proportion of low derived allele frequency polymorphic sites as local ARs ($p = 0.15$, two-tailed Fisher's exact test; Figure S2A), consistent with previous analyses (Haerty and Ponting, 2013). No statistically significant difference in derived allele frequency was observed between introns and exons of all LCL-expressed lincRNAs ($p = 0.89$, two-tailed Fisher's exact test; Figure S2A). Our results indicate that purifying selection has acted to remove deleterious mutations within TR-lincRNA exons during recent human evolution, which reinforces the functional relevance of these noncoding transcripts in humans. Surprisingly, analysis of putative promoters of TR-lincRNAs suggests that these regions evolved neutrally or nearly neutrally (Figure S2B). The difference in evolutionary constraint between the promoter and exon sequences can likely be explained by inaccurate prediction of proximal promoter regions, which would result in reduced power to infer their constraint. Despite limitations, our analysis of exonic sequence evolution supports that TR-lincRNA transcripts were preserved during recent human evolution.

Unexpectedly, the higher selective constraint observed for TR-lincRNAs relative to other LCL-expressed lincRNAs appears to be an evolutionary signature specific to recent human evolution, as we found no significant differences in their sequence conservation during either mammalian or primate evolution, estimated using phastCons scores, a measure of nucleotide conservation (Siepel et al., 2005) (Figures 2B and S3). Specifically, relative to other LCL-expressed lincRNAs, TR-lincRNA exons, introns, and promoters exhibit statistically indistinguishable median phastCons scores (Figure S3). This observation could be the result of rapidly evolving repetitive elements within TR-lincRNAs (Kapusta et al., 2013; Kelley and Rinn, 2012). Indeed, we found that TR-lincRNA exons and promoters are enriched in long terminal repeat (LTR)-derived transposable elements relative to other LCL-expressed lincRNAs (3.8- to 7.9-fold enrichment, $p < 0.05$). In particular, TR-lincRNAs exons and promoters are enriched in human endogenous retrovirus K (ERVK) LTRs (1.6- to 2.2-fold enrichment,

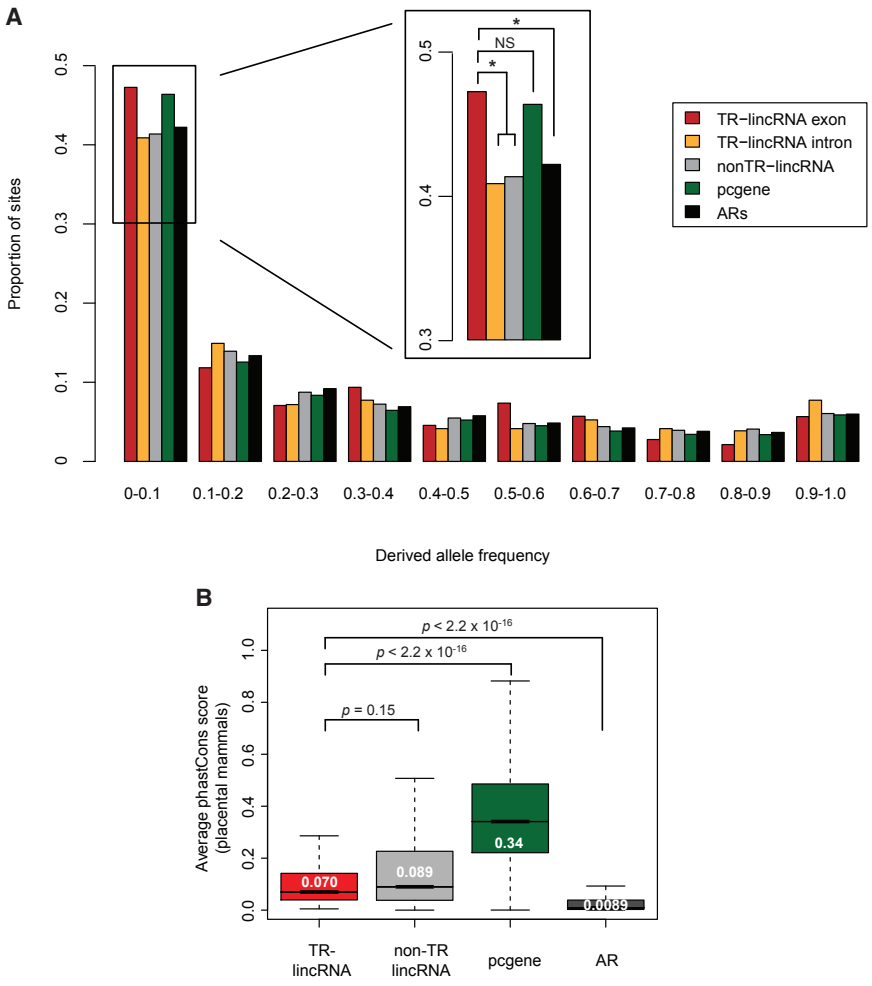


Figure 2. TR-lincRNAs Evolved under Purifying Selection during Recent Human History

(A) Distribution of derived allele frequency (DAF) for variants within exons (red) and introns (yellow) of TR-lincRNA, LCL-expressed lincRNA exons (gray), protein-coding gene exons (green), and ancestral repeats (ARs; black). Low-frequency polymorphic sites (DAF < 0.1) for all classes of genes are depicted in the insert. Asterisks indicate levels of significance in the comparison (*$p < 0.05$; NS, not significant [$p > 0.05$]; two-tailed Fisher's exact test).

(B) Distribution of sequence conservation, as estimated using phastCons scores across placental mammals (y axis), within the exonic sequence of TR-lincRNAs (red), other LCL-expressed lincRNAs (light gray), protein-coding genes (green), and ancestral repeats (dark gray). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

See also Figures S2 and S3 and Table S3.

$p < 0.05$; Table S3; Experimental Procedures), whose transcription was previously shown to be elevated upon immune system stimulation (Manghera and Douville, 2013).

Trait-Relevant lincRNA Transcription Is Associated with *cis* Regulation

lincRNAs can regulate the expression levels of local and distal targets (Vance and Ponting, 2014). To gain insights into the molecular mode of action of TR-lincRNAs, we examined their relationship with TR-pcgenes. For each protein-coding gene, we defined its territory as the genomic region containing all nucleotides that are closer to the gene than they are to its most proximal up- and downstream protein-coding genes. We found that TR-lincRNAs are significantly more likely than expected to reside within TR-protein-coding gene territories (fold enrichment = 2.4, $p < 1 \times 10^{-3}$; Experimental Procedures).

Next, we estimated the median co-expression (Pearson's correlation) in LCLs between pairs of TR-lincRNAs and protein-coding genes in their vicinity (within <20 kb, 20–100 kb, 100–500 kb, and >500 kb of each other). Consistent with their proposed regulatory interactions, we found TR-lincRNAs to be significantly more highly correlated in expression with nearby protein-coding genes than other LCL-expressed lincRNAs (Figure 3A). Furthermore, TR-lincRNAs are over 2.5 times more likely to share an eQTL with at least one nearby protein-coding gene (43/69 [62.3%]) compared to other LCL-expressed lincRNAs (592/2441 [24.3%]), a significantly higher proportion ($p < 1 \times 10^{-3}$, two-tailed Fisher's exact test; Experimental Procedures), suggesting that TR-lincRNAs are more likely than other transcripts to affect the expression of nearby loci.

To dissect the regulatory interaction between TR-lincRNAs and their nearby co-expressed TR-pcgenes, we focused on the 30 trait-relevant lincRNAs with nearby TR-pcgenes that share the same GWAS *cis*-eQTL (Table S4; Experimental Procedures), hereafter referred to as *cis*TR-lincRNAs. We tested, using hierarchical linear regression, whether adding the expression levels of the *cis*TR-lincRNA strengthens the *cis*-eQTL association of its linked TR-pcgene (Experimental Procedures). 87% (26/30) of *cis*TR-lincRNAs significantly improves the association between the expression levels of the nearby TR-pcgenes and their trait-associated variants (Table S5). Furthermore, *cis*TR-lincRNA associations with GWAS *cis*-eQTLs relative to common SNPs in the region (median RTC = 0.97) are significantly higher than those for TR-pcgene associations (median RTC = 0.95, $p < 0.05$, two-tailed Mann-Whitney paired *U*-test; Table S6).

To assess how changes in *cis*TR-lincRNA or TR-pcgene copies impact the expression levels of their nearby associated loci, we identified copy-number variants (CNVs; 1000 Genomes Project Consortium et al., 2012) that uniquely encompass either *cis*TR-lincRNAs or TR-pcgenes (Table S7). CNVs that overlap the shared GWAS *cis*-eQTL or those that contain both the linked *cis*TR-lincRNA and TR-pcgene were excluded. We estimated the absolute fold difference in *cis*TR-lincRNA or TR-pcgene

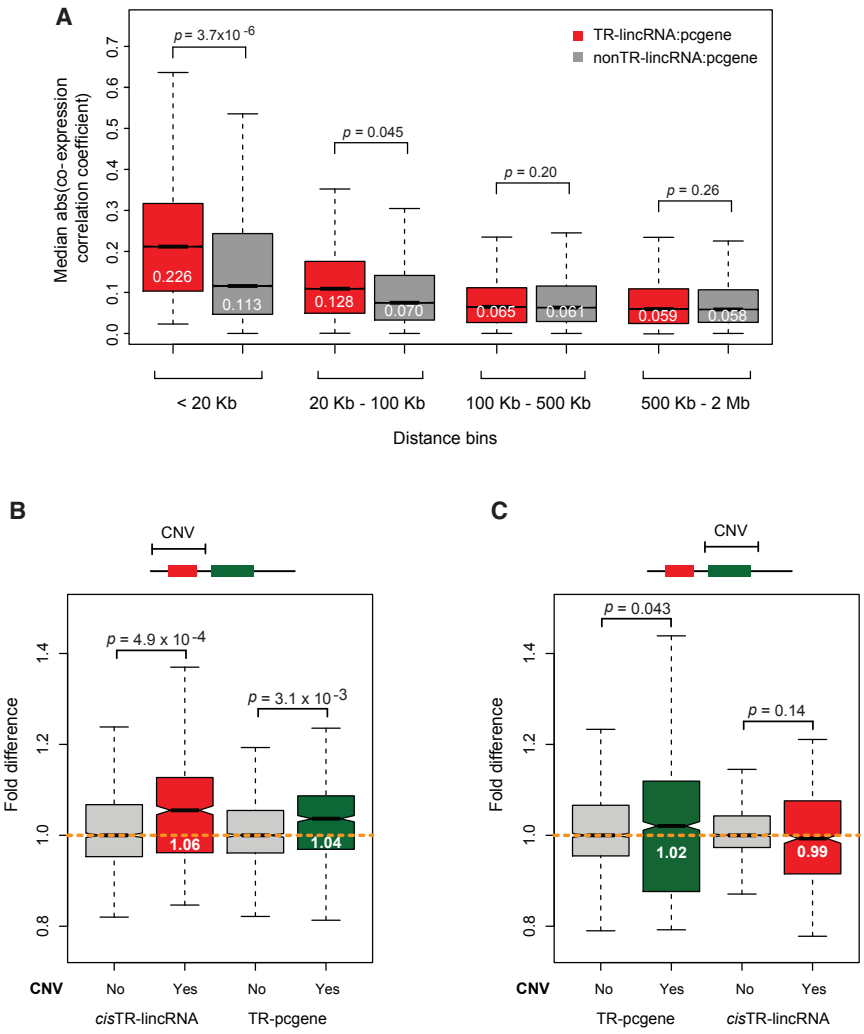


Figure 3. TR-lincRNAs Are Enriched at TAD Boundaries and Regulate Proximal TR-pcgenes in *cis*, Likely by Modulating Chromatin Architecture

(A) Distribution of median absolute correlation coefficient between expression levels in LCLs of TR-lincRNAs (red) or other LCL-expressed lincRNAs (gray) and nearby protein-coding genes. Pairs are split into bins based on their genomic distance (<20 kb, 20–100 kb, 100–500 kb, and 500 kb to 2 Mb).

(B and C) Absolute fold difference in expression levels across individuals that carry copy-number variants (CNVs) (1000 Genomes Project Consortium et al., 2012) that encompass (B) *cis*TR-lincRNAs (red) or (C) TR-pcgenes (green) and that of the nearby trait-relevant protein-coding genes or lincRNAs, respectively, relative to the expression of the loci in individuals without CNVs (gray). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

See also Tables S3, S4, S5, S6, and S7.

expression between individuals with or without CNVs and found that variations in *cis*TR-lincRNA copy number are associated with significant changes in the levels of TR-pcgenes ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure 3B). In contrast, no significant difference in the levels of *cis*TR-lincRNAs was observed when CNVs encompassed TR-pcgenes ($p = 0.14$, two-tailed Mann-Whitney *U* test; Figure 3C). Together, these observations provide preliminary evidence that *cis*TR-lincRNAs contribute to the regulation of the levels of TR-pcgenes in their vicinities.

Trait-Relevant lincRNAs Are Associated with Local Chromosomal Architecture

TADs are genomic regions where DNA-DNA interactions are frequent (Dixon et al., 2012). These genomic structures have been proposed to modulate gene transcription through increased accessibility to shared local regulatory elements (Nora et al., 2013). This hypothesis is supported by evidence of frequent co-expression between genes within the same TAD (Le Dily et al., 2014; Neems et al., 2016). We investigated whether frequent localization within the same TAD would explain the co-expression between pairs of trait-relevant coding and noncoding loci. First, we found that cisTR-lincRNAs are enriched within LCL TADs that also contain TR-pcgenes (fold enrichment = 3.2, $p < 1 \times 10^{-3}$; Experimental Procedures). Interestingly, when we analyzed the location of *cis*TR-lincRNAs within sub-compartments of TADs, we found them to be significantly enriched at the boundaries and depleted at the center of these genomic units (Figure 4A). Such enrichment at TAD boundaries is specific to *cis*TR-lincRNAs, as no preferential location was found when we analyzed the distribution of other LCL-expressed lincRNAs. To assess the relevance of *cis*TR-lincRNAs to local chromosomal architecture, we investigated the correlation between their expression levels and intra-TAD DNA-DNA contact density (Experimental Procedures). We found that the density of chromosomal contacts is significantly higher for TADs containing *cis*TR-lincRNAs (9.1 times, $p < 5 \times 10^{-3}$, two-tailed Mann-Whitney *U* test; Figure 4B) relative to those containing other LCL-expressed lincRNAs. Interestingly, this difference appears to be specific to LCLs, supporting cell-type-specific functions of *cis*TR-lincRNAs ($p > 0.05$, two-tailed Mann-Whitney *U* test; Figure S4A). Strikingly, we found a significant positive correlation between the levels of *cis*TR-lincRNAs and DNA-DNA contacts within their associated TADs relative to other LCL-expressed lincRNAs ($r = 0.163$, Spearman's correlation, $p < 0.05$; Figure 4C). Importantly, this association is also cell-type-specific and restricted to TR-lincRNAs (Figures S4B–S4D), strongly supporting the role of these loci in the modulation of chromosomal architecture.

Previous studies have demonstrated that active enhancer-like regulatory elements are enriched at the boundaries of TADs (Huang et al., 2015). Interestingly, transcription at these

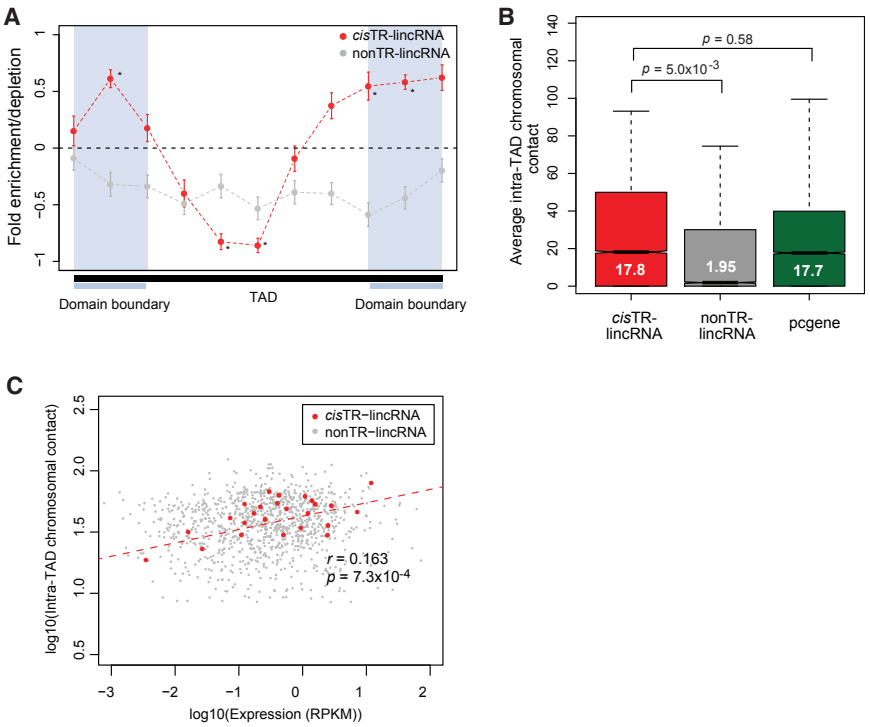


Figure 4. TR-lincRNAs Are Enriched at TAD Boundaries and Regulate Proximal TR-pcgenes in *cis*, Likely by Modulating Chromatin Architecture

(A) Fold enrichment or depletion of *cis*TR-lincRNA (red) and other LCL-expressed lincRNAs (gray) at fractional positions within LCL TADs (GM12878, black bar; Rao et al., 2014) and at TAD boundaries (light blue bar, area shaded in light blue). Significant fold differences are denoted with an asterisk, and SD is shown with error bars ($p < 0.05$, permutation test).

(B) Average chromosomal contacts within TAD that contain *cis*TR-lincRNAs (red), other LCL-expressed lincRNAs (gray), and pcgenes (green) in LCLs (GM12878; ENCODE Project Consortium, 2012). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

(C) Correlation (Spearman's) between expression levels of *cis*TR-lincRNAs ($r = 0.163$, $p = 7.3 \times 10^{-4}$, red) and other LCL-expressed lincRNAs ($r = 0.105$, $p = 0.53$, gray) with the average chromosomal contacts within their residing TADs in LCLs (GM12878; ENCODE Project Consortium, 2012).

See also Figure S4 and Tables S3, S4, S5, and S6.

enhancers is widespread in humans (Andersson et al., 2014), and a large fraction of lincRNA transcription has been previously shown to originate at enhancers (Marques et al., 2013). We investigated whether TR-lincRNAs were enhancer associated. We found that relative to other LCL-expressed lincRNAs, the promoters of *cis*TR-lincRNAs are enriched in mono- versus trimethylation of histone H3K4, a well-established signature of enhancer elements ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figures 5, S5A, and S5B), indicating their likely enhancer origin. Interestingly, we found that the syntenic regions in mouse of our *cis*TR-lincRNA putative promoters are also significantly enriched in enhancer-associated chromatin marks (murine LCLs [CH12 cells]; Mouse ENCODE Consortium et al., 2012) relative to other LCL-expressed lincRNAs ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure S5C), suggesting their associated enhancer activity is conserved between species at some of these loci. These *cis*TR-lincRNAs are also more enriched in the nucleus versus the cytoplasm relative to other LCL-expressed lincRNAs ($p < 0.05$, two-tailed Mann-Whitney *U* test; Figure S5D), which is as expected and consistent with their role in transcriptional regulation.

The cohesin protein complex, known to be enriched at active enhancer elements and TAD boundaries, has been previously shown to be important for intra-TAD gene regulation in a cell-type-specific manner (Merkenschlager and Odom, 2013). For example, cohesin depletion is associated with disrupted promoter-enhancer interactions within TADs (Kagey et al., 2010; Seitan et al., 2011). Another central player in the regulation of chromatin architecture and gene expression is the CTCF transcription factor (reviewed in Merkenschlager and Odom, 2013). Unlike cohesin, which is involved in cell-specific intra-TAD interactions, CTCF is important for the spatial segregation of topological domains (Zuin et al., 2014) with binding sites that are often conserved and shared across different species and cell types (Kim et al., 2007). We observed that cohesin binding sites are significantly enriched at *cis*TR-lincRNAs loci (fold enrichment = 1.43, $p < 0.05$). In contrast, CTCF binding sites are depleted at these noncoding RNA loci (fold depletion = −0.86, $p < 0.05$; Experimental Procedures) relative to intergenic regions of the human genome. These observations suggest that rather than acting to establish TAD architecture, TR-lincRNAs are more likely to be involved in cell-type-specific regulation of enhancer-promoter interactions within TADs.

Taken together, (1) the positive co-expression of a large proportion of trait-relevant lincRNAs with their proximal TR-pcgenes, (2) the contribution to their nearby TR-pcgene GWAS *cis*-eQTL, (3) enrichment at TAD boundaries and cohesin binding sites, and (4) enrichment in enhancer-like RNA properties are all compatible with enhancer origins and local regulatory roles of TR-lincRNAs.

DISCUSSION

Since the discovery of pervasive lincRNA transcription in humans (Carninci et al., 2005), extensive research efforts have strived to establish what might be their contribution, if any, to organismal phenotypes (Marx, 2014). Previous studies (Kumar et al., 2013; Lappalainen et al., 2013; McDowell et al., 2016; Popadin et al., 2013) have led to the identification of lincRNAs associated with complex human traits and diseases, often through *cis*-eQTL analysis. This wealth of information comes with a new and challenging question: what might be the functions of

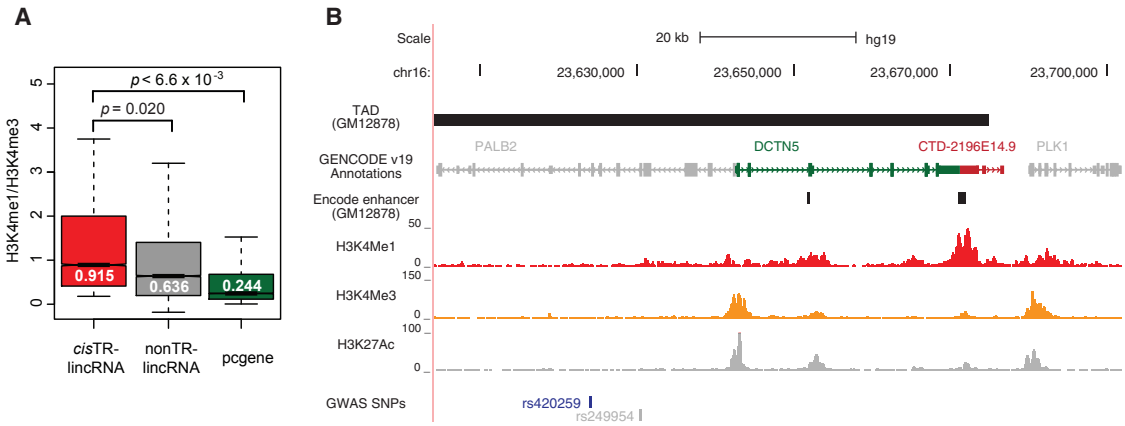


Figure 5. TR-lincRNA Promoter Regions Are Enriched in Enhancer-Associated Chromatin Marks

(A) Ratio of the number of H3K4me1 to H3K4me3 sequencing reads mapped to the putative promoter regions (1 kb upstream and downstream of the TSS) in LCLs (GM12878; ENCODE Project Consortium, 2012) for *cis*TR-lincRNAs (red), other LCL-expressed lincRNAs (gray), and protein-coding genes (green). Differences between groups were tested using a two-tailed Mann-Whitney *U* test, and p values are indicated.

(B) UCSC genome browser view of one *cis*TR-lincRNA, CTD-2196E14.9 (ENSG00000260482, chr16: 23,681,332–23,684,448, red), and a neighboring TR-pcgene, DCTN5 (ENSG00000166847, green), which is associated with the same GWAS *cis*-eQTL (rs420259, blue). Non-trait-associated protein-coding genes between CTD-2196E14.9 and COG7 are colored in gray. Arrows within introns indicate direction of transcription. CTD-2196E14.9 overlaps predicted enhancer elements in a lymphoblastoid cell line (GM12878, vertical black bars; ENCODE Project Consortium, 2012) at the boundary of a TAD (GM12878, horizontal dark gray bar; Rao et al., 2014), and its transcription start site has a high H3K4me1 (red track) over H3K4me3 (yellow track) ratio.

See also Figure S5 and Tables S3, S4, S5, and S6.

these candidates, and how might they contribute to phenotype? Given the heterogeneity of the known molecular mechanisms underlying lincRNA functions and the current lack of approaches to predict them, genetic dissection of these trait-associated candidates is challenging and has only been achieved for a handful of transcripts thus far (for example, Ishii et al., 2006; Jendrzejewski et al., 2012).

Our genome-wide analysis of a stringent set of TR-lincRNAs suggests that these loci often associate with *cis* regulation of nearby trait-associated protein-coding genes and provides a working hypothesis for how lincRNAs can contribute to human complex traits. While co-expression between loci in close genomic proximity is common (McDowell et al., 2016), we show this phenomenon is stronger between TR-lincRNAs and protein-coding genes in their vicinity than between pairs of non-trait-associated loci. Furthermore, we provide evidence that changes in TR-lincRNA copy number are specifically associated with changes in the levels of nearby TR-pcgenes, consistent with the roles of these lincRNAs in the regulation of proximal TR-pcgene expression levels. Recent studies have shown that boundary elements are key to maintaining TAD organization and that mutations in these boundary elements disrupt regulatory interactions and influence phenotypes, specifically during development (Guo et al., 2015; Lupiáñez et al., 2015). The preferential location of TR-lincRNAs at TAD boundaries and their frequent and evolutionarily conserved enhancer origin suggest that TR-lincRNA transcription affects the levels of trait-relevant genes in their vicinity, likely by modulating local chromosomal organization, thus impacting complex normal and disease phenotypes in humans. The correlation observed between TR-lincRNA expression and intra-TAD DNA-DNA interactions in LCLs provides genome-wide support for this hypothesis.

Our results suggest that lincRNAs are generally lowly expressed (Cabili et al., 2011), which is likely to limit their ability to regulate the expression of mRNAs in *trans*. In contrast, regulation of gene expression in *cis* through the modulation of chromosomal architecture is likely to require fewer transcript copies or merely the act of transcription. Therefore, we propose that this mechanism of enhancer-associated lincRNA transcription is likely not restricted to trait-relevant lincRNAs.

While further work is still required to dissect the biological role of individual TR-lincRNAs, our genome-wide results provide the much needed mechanistic insights into their functions, furthering the understanding of the intricate genetic networks underlying complex human traits and diseases.

EXPERIMENTAL PROCEDURES

cis-eQTL Analysis

Mapped RNA-sequencing reads of Epstein-Barr virus (EBV)-transformed LCLs derived from 373 individuals of European descent (Utah Residents with Northern and Western Ancestry [CEU], British in England and Scotland [GBR], Finnish in Finland [FIN], and Toscani in Italy [TSI]) and the corresponding processed genotypes were downloaded from EBI ArrayExpress (EBI: E-GEUV-1) (Lappalainen et al., 2013).

eQTL analysis was performed for genome-wide significant ($p < 5 \times 10^{-8}$; Welter et al., 2014) trait-associated autosomal SNPs located within a 2-Mb window centered on the predicted transcription start site (TSS) of each expressed lincRNA and protein-coding gene. We estimated Pearson's correlation (r_{obs}) between corrected and transformed gene expression levels and trait-associated SNP genotypes. A detailed description of the *cis*-eQTL identification process is provided in Supplemental Experimental Procedures.

Enhancer-Associated TR-lincRNAs

Coordinates of ENCODE-predicted enhancer elements and H3K4me1 and H3K4me3 chromatin immunoprecipitation (ChIP) sequencing reads in human

GM12878 and mouse CH12 LCLs (ENCODE Project Consortium, 2012; Mouse ENCODE Consortium et al., 2012) were downloaded from the UCSC database (Rosenbloom et al., 2015). We estimated the ratio of H3K4me1 to H3K4me3 reads mapping to putative promoter regions of lincRNAs (using HTseq version 0.6.1; Anders et al., 2015). Details on defining putative promoter regions of TR-lincRNAs in human and mouse LCLs are provided in Supplemental Experimental Procedures.

Spatial Chromosomal Architecture Analysis

Intra-chromosomal interactions were calculated using Hi-C contact matrices for four ENCODE cell lines (GM12878, K562, HUVEC, and NHEK; Rao et al., 2014). All computations were performed on 5-kb-resolution matrices with a Mapping Quality (MAPQ) score above 30. Spearman's correlation was estimated between gene expression levels and the average density of contacts within the TAD where the gene resides. Comparisons between Spearman's correlations was performed using the two-sided Fisher's *z* test (1925) based on independent groups implemented in the "cocor" R package (Diedenhofen and Musch, 2015). Details on data normalization and estimation of average intra-TAD contacts are described in Supplemental Experimental Procedures.

Additional materials and methods are described in Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.02.009.

AUTHOR CONTRIBUTIONS

J.Y.T. and A.C.M. designed the study. J.Y.T., A.A.T.S., M.F.d.S., C.M.-D., R.R., R.S., and D.D. performed analyses. J.Y.T., Z.K., S.B., and A.C.M. conceived methods and discussed the results. A.C.M. supervised the analysis. J.Y.T. and A.C.M. wrote the manuscript. All authors approved the manuscript.

ACKNOWLEDGMENTS

We thank Chris P. Ponting and members of the Marques group, Dario Bottinelli and Adriano Biasini for valuable comments and discussion. We thank Wilfried Haerty and Chris Rands for discussion on DAF analysis and Mathieu Heulot for discussion on experimental design. This work was funded by the Swiss National Science Foundation (grant PP00P3_150667 to A.C.M., grant FN 31003A-143914 to Z.K., and grant FN 310030_152724/1 to S.B.) and the NCCR in RNA & Disease.

Received: September 17, 2016
Revised: December 16, 2016
Accepted: January 30, 2017
Published: February 28, 2017

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

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

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C., et al. (2008). Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet. *4*, e1000287.

Diedenhofen, B., and Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. PLoS ONE *10*, e0121945.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat. Genet. *44*, 502–510.

Firmann, M., Mayor, V., Vidal, P.M., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K.S., Yuan, X., et al. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc. Disord. *8*, 6.

Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell *162*, 900–910.

Haerty, W., and Ponting, C.P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome Biol. *14*, R49.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Huang, J., Marco, E., Pinello, L., and Yuan, G.C. (2015). Predicting chromatin organization using histone marks. Genome Biol. *16*, 162.

Hüttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? Trends Genet. *21*, 289–297.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., et al. (2006). Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J. Hum. Genet. *51*, 1087–1099.

Jendrzejewski, J., He, H., Radomska, H.S., Li, W., Tomsic, J., Liyanarachchi, S., Davuluri, R.V., Nagy, R., and de la Chapelle, A. (2012). The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc. Natl. Acad. Sci. USA *109*, 8646–8651.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. *9*, e1003470.

Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. *13*, R107.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kumar, V., Westra, H.J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Võsa, U., et al. (2013). Human disease-associated genetic variation impacts large intergenic noncoding RNA expression. PLoS Genet. *9*, e1003201.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput. Biol. *12*, e1004714.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Le Dily, F., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. *28*, 2151–2162.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Manghera, M., and Douville, R.N. (2013). Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology *10*, 16.

Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol. *14*, R131.

Marx, V. (2014). A blooming genomic desert. Nat. Methods *11*, 135–138.

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.

McDowell, I.C., Pai, A.A., Guo, C., Vockley, C.M., Brown, C.D., Reddy, T.E., and Engelhardt, B.E. (2016). Many long intergenic non-coding RNAs distally regulate mRNA gene expression levels. bioRxiv. Published online March 19, 2016. http://dx.doi.org/10.1101/044719.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Neems, D.S., Garza-Gongora, A.G., Smith, E.D., and Kosak, S.T. (2016). Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. Proc. Natl. Acad. Sci. USA *113*, E1691–E1700.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.

Nora, E.P., Dekker, J., and Heard, E. (2013). Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? BioEssays *35*, 818–828.

Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E.T., and Antonarakis, S.E. (2013). Genetic and epigenetic regulation of human lincRNA gene expression. Am. J. Hum. Genet. *93*, 1015–1026.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. *43*, D670–D681.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–471.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

Vance, K.W., and Ponting, C.P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends Genet. *30*, 348–355.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Zheng, J., Huang, X., Tan, W., Yu, D., Du, Z., Chang, J., Wei, L., Han, Y., Wang, C., Che, X., et al. (2016). Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. Nat. Genet. *48*, 747–757.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van IJcken, W.F., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. USA *111*, 996–1001.

Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis

Urmo Võsa*#[1,2], Annique Claringbould*#[1], Harm-Jan Westra**[1], Marc Jan Bonder**[1], Patrick Deelen**[1,3], Biao Zeng[4], Holger Kirsten[5], Ashis Saha[6], Roman Kreuzhuber[7,8], Silva Kasela[2], Natalia Pervjakova[2], Isabel Alvaes[9], Marie-Julie Fave[9], Mawusse Agbessi[9], Mark Christiansen[10], Rick Jansen[11], Ilkka Seppälä[12], Lin Tong[13], Alexander Teumer[14], Katharina Schramm[15,16], Gibran Hemani[17], Joost Verlouw[18], Hanieh Yaghootkar[19], Reyhan Sönmez[20,21], Andrew Brown[22,23,21], Viktorija Kukushkina[2], Anette Kalnapenkis[2], Sina Rüeger[24], Eleonora Porcu[24], Jaanika Kronberg-Guzman[2], Johannes Kettunen[25], Joseph Powell[26], Bernett Lee[27], Futao Zhang[28], Wibowo Arindrarto[29], Frank Beutner[30], BIOS Consortium, Harm Brugge[1], i2QTL Consortium, Julia Dmitreva[31], Mahmoud Elansary[31], Benjamin P. Fairfax[32], Michel Georges[31], Bastiaan T. Heijmans[29], Mika Kähönen[33], Yungil Kim[34,35], Julian C. Knight[32], Peter Kovacs[36], Knut Krohn[37], Shuang Li[1], Markus Loeffler[5], Urko M. Marigorta[4], Hailang Mei[38], Yukihide Momozawa[31,39], Martina Müller-Nurasyid[15,16,40], Matthias Nauck[41], Michel Nivard[42], Brenda Penninx[11], Jonathan Pritchard[43], Olli Raitakari[44], Olaf Rotzchke[27], Eline P. Slagboom[29], Coen D.A. Stehouwer[45], Michael Stumvoll[46], Patrick Sullivan[47], Peter A.C. 't Hoen[48], Joachim Thiery[49], Anke Tönjes[46], Jenny van Dongen[11], Maarten van Iterson[29], Jan Veldink[50], Uwe Völker[51], Cisca Wijmenga[1], Morris Swertz[3], Anand Andiappan[27], Grant W. Montgomery[52], Samuli Ripatti[53], Markus Perola[54], Zoltan Kutalik[24], Emmanouil Dermitzakis[22,23,21], Sven Bergmann[20,21], Timothy Frayling[19], Joyce van Meurs[18], Holger Prokisch[55,56], Habibul Ahsan[13], Brandon Pierce[13], Terho Lehtimäki[12], Dorret Boomsma[11], Bruce M. Psaty[10,57], Sina A. Gharib[58,10], Philip Awadalla[9], Lili Milani[2], Willem Ouwehand[7,59], Kate Downes[7], Oliver Stegle[8,60,61], Alexis Battle[62], Jian Yang[28,63], Peter M. Visscher[28], Markus Scholz[5], Gregory Gibson[4], Tõnu Esko[2], Lude Franke#[1]

* These authors contributed equally to this work.

** These authors contributed equally to this work.

1. Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands
2. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu 51010, Estonia
3. Genomics Coordination Center, University Medical Centre Groningen, Groningen, The Netherlands
4. School of Biological Sciences, Georgia Tech, Atlanta, United States of America
5. Institut für Medizinische InformatiK, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
6. Department of Computer Science, Johns Hopkins University, Baltimore, United States of America
7. Department of Haematology, University of Cambridge and NHS Blood and Transplant Cambridge Biomedical Campus, Cambridge, United Kingdom
8. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
9. Computational Biology, Ontario Institute for Cancer Research, Toronto, Canada
10. Cardiovascular Health Research Unit, University of Washington, Seattle, United States of America
11. Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
12. Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
13. Department of Public Health Sciences, University of Chicago, Chicago, United States of America
14. Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
15. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
16. Department of Medicine I, University Hospital Munich, Ludwig Maximilian's University, München, Germany

17. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
18. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands
19. Exeter Medical School, University of Exeter, Exeter, United Kingdom
20. Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
21. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
22. Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
23. Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland
24. Lausanne University Hospital, Lausanne, Switzerland
25. University of Helsinki, Helsinki, Finland
26. Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, Australia
27. Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore
28. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
29. Leiden University Medical Center, Leiden, The Netherlands
30. Heart Center Leipzig, Universität Leipzig, Leipzig, Germany
31. Unit of Animal Genomics, WELBIO, GIGA-R & Faculty of Veterinary Medicine, University of Liege, 1 Avenue de l'Hôpital, Liège 4000, Belgium
32. Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom
33. Department of Clinical Physiology, Tampere University Hospital and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
34. Department of Computer Science, Johns Hopkins University, Baltimore, United States of America
35. Genetics and Genomic Science Department, Icahn School of Medicine at Mount Sinai, New York, United States of America
36. IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany
37. Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany
38. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
39. Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Kanagawa 230-0045, Japan
40. DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany
41. Institute of Clinical Chemistry and Laboratory Medicine, Greifswald University Hospital, Greifswald, Germany
42. Faculty of Genes, Behavior and Health, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
43. Stanford University, Stanford, United States of America
44. Turku University Hospital and University of Turku, Turku, Finland
45. Department of Internal Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands
46. Department of Medicine, Universität Leipzig, Leipzig, Germany
47. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
48. Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands
49. Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
50. University Medical Center Utrecht, Utrecht, The Netherlands
51. Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
52. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
53. Statistical and Translational Genetics, University of Helsinki, Helsinki, Finland
54. National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland
55. Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
56. Institute of Human Genetics, Technical University Munich, Munich, Germany.
57. Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States of America
58. Department of Medicine, University of Washington, Seattle, United States of America
59. Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton Cambridge, United Kingdom
60. Genome Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
61. Division of Computational Genomics and Systems Genetics, German Cancer Research Center, 69120 Heidelberg, Germany
62. Departments of Biomedical Engineering and Computer Science, Johns Hopkins University, Baltimore, United States of America
63. Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

[#] Correspondence can be addressed to
Urmo Võsa (urmo.vosa@gmail.com)
Annique Claringbould (anniqueclaringbould@gmail.com)
Lude Franke (ludefranke@gmail.com)

Summary

While many disease-associated variants have been identified through genome-wide association studies, their downstream molecular consequences remain unclear.

To identify these effects, we performed *cis*- and *trans*-expression quantitative trait locus (eQTL) analysis in blood from 31,684 individuals through the eQTLGen Consortium.

We observed that *cis*-eQTLs can be detected for 88% of the studied genes, but that they have a different genetic architecture compared to disease-associated variants, limiting our ability to use *cis*-eQTLs to pinpoint causal genes within susceptibility loci.

In contrast, *trans*-eQTLs (detected for 37% of 10,317 studied trait-associated variants) were more informative. Multiple unlinked variants, associated to the same complex trait, often converged on *trans*-genes that are known to play central roles in disease etiology.

We observed the same when ascertaining the effect of polygenic scores calculated for 1,263 genome-wide association study (GWAS) traits. Expression levels of 13% of the studied genes correlated with polygenic scores, and many resulting genes are known to drive these traits.

Main text

Expression quantitative trait loci (eQTLs) have become a common tool to interpret the regulatory mechanisms of the variants associated with complex traits through genome-wide association studies (GWAS). C*is*-eQTLs, where gene expression levels are affected by a nearby single nucleotide polymorphism (SNP) (<1 megabases; Mb), in particular, have been widely used for this purpose. However, *cis*-eQTLs from the genome tissue expression project (GTEx) explain only a modest proportion of disease heritability[1].

In contrast, *trans*-eQTLs, where the SNP is located distal to the gene (>5Mb) or on other chromosomes, can provide insight into the effects of a single variant on many genes. *Trans*-eQTLs identified before[1–7] have already been used to identify putative key driver genes that contribute to disease[8]. However, *trans*-eQTL effects are generally much weaker than those of *cis*-eQTLs, requiring a larger sample size for detection.

While *trans*-eQTLs are useful for the identification of the downstream effects of a single variant, a different approach is required to determine the combined consequences of trait-associated variants. Polygenic scores (PGS) have been recently applied to sum genome-wide risk for several diseases and likely will improve clinical care[9,10]. However, the exact consequences of different PGS at the molecular level, and thus the contexts in which a polygenic effects manifest themselves, are largely unknown. Here, we systematically investigate *trans*-eQTLs as well as associations between PGS and gene expression (expression quantitative trait score, eQTS) to determine how genetic effects influence and converge on genes and pathways that are important for complex traits.

To maximize the statistical power to detect eQTL and eQTS effects, we performed a large-scale meta-analysis in 31,684 blood samples from 37 cohorts (assayed using three gene expression

platforms) in the context of the eQTLGen Consortium. This allowed us to identify significant *cis*-eQTLs for 16,989 genes, *trans*-eQTLs for 6,298 genes and eQTS effects for 2,568 genes (**Figure 1A**), revealing complex regulatory effects of trait-associated variants. We combine these results with additional data layers and highlight a number of examples where we leverage this resource to infer novel biological insights into mechanisms of complex traits. We hypothesize that analyses identifying genes further downstream are more cell-type specific and more relevant for understanding disease (**Figure 1B**).

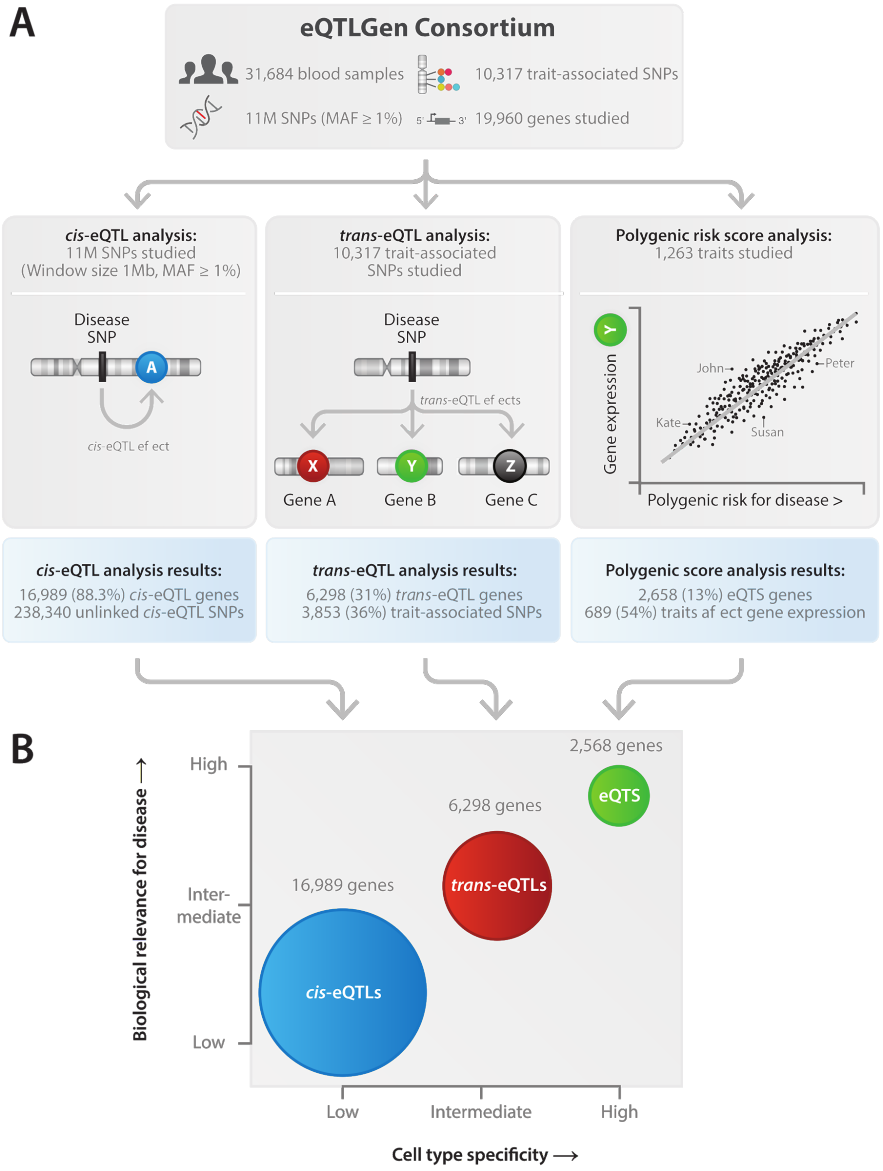


Figure 1. Overview of the study. (**A**) Overview of main analyses and their results. (**B**) Model of genetic effects on gene expression. *Cis*-eQTL are common and widely replicable in different tissues and cell types, whereas *trans*-eQTLs and eQTS are more cell type specific. The biological insight derived from our *cis*-eQTL results are usually not well interpretable in the context of complex traits, suggesting that weaker distal effects give additional insight about biological mechanisms leading to complex traits.

Local genetic effects on gene expression in blood are widespread and replicable in other tissues

Using eQTLGen consortium data from 31,684 individuals, we performed *cis*-eQTL, *trans*-eQTL and eQTS meta-analyses (**Figure 1A**, **Supplementary Table 1**). Different expression profiling platforms were integrated using a data-driven method (**Online Methods**). To ensure the robustness of the identified eQTLs, we performed eQTL discovery per platform and replicated resulting eQTLs in the other platforms, observing excellent replication rates and consistency of allelic directions (**Online Methods**, **Supplementary Note**, **Extended Data Figure 1A-C**). We identified significant *cis*-eQTLs (SNP-gene distance <1Mb, gene-level False Discovery Rate (FDR)<0.05; **Online Methods**) for 16,989 unique genes (88.3% of autosomal genes expressed in blood and tested in *cis*-eQTL analysis; **Figure 1A**). Out of 10,317 trait-associated SNPs tested, 1,568 (15.2%) were in high linkage disequilibrium (LD) with the lead eQTL SNP showing the strongest association for a *cis*-eQTL gene, (R^2>0.8; 1kG p1v3 EUR; **Supplementary Table 2**; **Online Methods**). Genes highly expressed in blood but not under any detectable *cis*-eQTL effect were more likely ($P=2\times10^{-6}$; Wilcoxon two-sided test; **Figure 2A**) to be intolerant to loss-of-function mutations in their coding region[11], suggesting that eQTLs on such gene would interfere with the normal functioning of the organism.

We observed that 92% of the lead *cis*-eQTL SNPs map within 100kb of the gene (**Figure 2D**), and this increased to 97.2% when only looking at the 20% of the genes with the strongest lead *cis*-eQTL effects. Of these strong *cis*-eQTLs, 84.1% of the lead eQTL SNPs map within 20kb of the gene. GWAS simulations[12] indicate that lead GWAS signals map within 33.5kb from the causal variant in 80% of cases, which suggests that our top SNPs usually tag causal variants that map

directly into either the promoter region, the transcription start site (TSS), the gene body, or the transcription end site (TES). For strong *cis*-eQTLs we observed that lead *cis*-eQTL SNPs located >100kb from the TSS or TES overlap capture Hi-C contacts (37%; **Figure 2E**) more often than short-range *cis*-eQTL effects (16%; Chi^2 test $P = 2\times10^{-5}$), indicating that, for long-range *cis*-eQTLs the SNP and gene often physically interact to cause the *cis*-eQTL effect. For instance, a capture Hi-C contact for *IRS1* overlapped the lead eQTL SNP, mapping 630kb downstream from *IRS1* (**Figure 2F**).

We observed that our sample-size improved fine-mapping: for 5,440 protein-coding *cis*-eQTL genes that we had previously identified in 5,311 samples[1] we now observe that the lead SNP typically map closer to the *cis*-eQTL gene (**Extended Data Figure 4**).

Cis-eQTLs showed directional consistency across tissues: in 47 postmortem tissues (GTEx v7[13]) we observed an average of 14.8% replication rate (replication FDR<0.05 in GTEx; median 15.1%; range 3.6-29.7%; whole blood tissue excluded) and on average a 95.0% concordance in allelic directions (median 95.3%, range 86.7-99.3%; whole blood tissue excluded) among the *cis*-eQTLs that significantly replicated in GTEx (**Extended Data Figure 5**, **Supplementary Note** and **Supplementary Table 3**).

However, our lead *cis*-eQTL SNPs show significantly different epigenetic histone mark characteristics, as compared to 3,668 SNPs identified in GWAS (and associated to blood related traits or immune-mediated diseases to minimize potential confounding). We observed significant differences for 20 out of 32 tested histone marks with H3K36me3, H3K27me3, H3K79me1 and H2BK20ac showing the strongest difference (Wilcoxon $P = 10^{-39}$, 10^{-21}, 10^{-19} and 10^{-18},

respectively), suggesting that *cis*-eQTLs have a different genetic architecture, as compared to complex traits and diseases.

We tested this for 16 well-powered complex traits (Supplementary Table 20) and observed that genes prioritized by combining *cis*-eQTL and GWAS data using summary statistics based Mendelian randomization (SMR[14]**; Online Methods**) did not overlap significantly more with genes prioritized through an alternative method (DEPICT) that does not use any *cis*-eQTL information[15]. While the genes prioritized with SMR were informative, and enriched for relevant pathways for several immune traits (**Supplementary Table 20**)**,** non-blood-trait-prioritized genes were difficult to interpret in the context of disease. Moreover, the lack of enriched overlap between DEPICT and SMR indicates that employing *cis*-eQTL information does not necessarily clarify which genes are causal for a given susceptibility locus. As such, some caution is warranted when using a single *cis*-eQTL repository for interpretation of GWAS.

One third of trait-associated variants have *trans*-eQTL effects

An alternative strategy for gaining insight into the molecular functional consequences of disease-associated genetic variants is to ascertain *trans*-eQTL effects. We tested 10,317 trait-associated SNPs ($P \leq 5\times10^{-8}$; **Online Methods**, **Supplementary Table 2**) for *trans*-eQTL effects (SNP-gene distance >5Mb, FDR < 0.05) to better understand their downstream consequences. We identified a total of 59,786 significant *trans*-eQTLs (FDR<0.05; **Supplementary Table 4**, **Extended Data Figure 6**), representing 3,853 unique SNPs (37% of tested GWAS SNPs) and 6,298 unique genes (32% of tested genes; **Figure 1A**). When compared to the previous largest *trans*-eQTL meta-analysis[1] (N=5,311; 8% of trait-associated SNPs with a significant *trans*-eQTL), these results

indicate that a large sample size is critical for identifying downstream effects. Colocalization analyses in a subset of samples (n=4,339; **Supplementary Note**) using COLOC[16] estimated that 52% of *trans*-eQTL signals colocalize with at least one *cis*-eQTL signal (posterior probability > 0.8; **Extended Data Figure 7A-B**). Corresponding colocalizing *cis*-eQTL genes were enriched for transcription factor activity ("regulation of transcription from RNA polymerase II promoter"; $P < 1.3\times10^{-9}$; **Extended Data Figure 7C**). Finally, highly expressed genes without a detectable *trans*-eQTL effect were more likely to be intolerant to loss-of-function variants ($P=6.4\times10^{-7}$; Wilcoxon test, **Figure 2B**), similar to what we observed for *cis*-eQTLs.

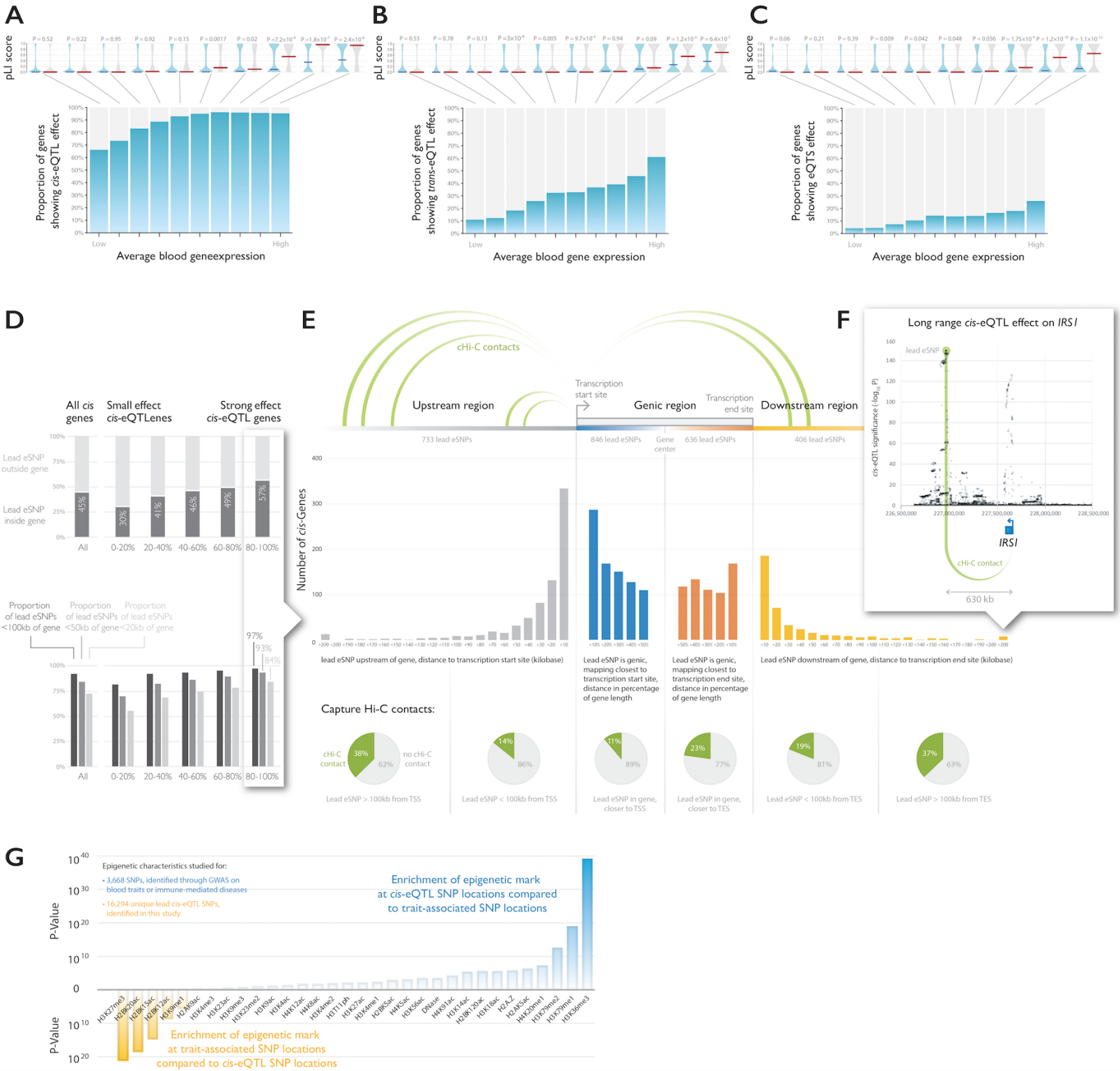


Figure 2. Results of the *cis*- and *trans*-eQTL analysis. All genes tested in (**A**) *cis*-eQTL analysis, (**B**) *trans*-eQTL analysis, and (**C**) eQTS analysis were divided into 10 bins based on average expression levels of the genes in blood. Highly expressed genes without any eQTL effect (grey bars) were less tolerant to loss-of-function variants (Wilcoxon test on pLI scores). Indicated are medians per bin. (**D**) Genes with strong effect sizes are more likely to have a lead SNP within (top panel) or close to the gene (bottom panel) (**E**) Top *cis*-eQTL SNPs positioning further from transcription start site (TSS) and transcription end site (TES) are more likely to overlap capture Hi-C contacts with TSS. (**F**) Enrichment analyses on epigenetic marks of *cis*-eQTL lead SNPs, compared to SNPs identified through GWAS and associated to blood-related or immune-mediated diseases, reveal significant differences in epigenetic characteristics.

In order to study the biological nature of the *trans*-eQTLs we identified, we conducted several enrichment analyses (**Supplementary Note, Extended Data Figure 8, Figure 3**). We observed 2.2 fold enrichment for known transcription factor (TF) - target gene pairs[17] (Fisher's exact test P = 10^{-62}; **Supplementary Note**), with the fold enrichment increasing to 3.2 (Fisher's exact test P < 10^{-300}) when co-expressed genes were included to TF targets. Those genes are potentially further downstream of respective TF targets in the molecular network. Similarly, we observed 1.19 fold enrichment of protein-protein interactions[18] among *trans*-eQTL gene-gene pairs (Fisher's exact test P=0.05). Some of these *cis-trans* gene pairs encode subunits of the same protein complex (e.g. *POLR3H* and *POLR1C*). While significant *cis-trans* gene pairs were enriched for gene pairs showing co-expression (Pearson R > 0.4; Fisher's exact test P=10^{-35}), we did not observe any enrichment of chromatin-chromatin contacts[19] (0.99 fold enrichment; Fisher's exact test P=0.3). Using the subset of 3,831 samples from BIOS, we also ascertained whether the *trans*-eQTL effect was mediated through a gene that mapped within 100kb from the *trans*-eSNP (i.e. using the *cis*-gene as G × E term). We observed significant interaction effects for 523 SNP-*cis*-*trans*-gene

combinations (FDR < 0.05; **Supplementary Table 5**), reflecting a 5.3 fold enrichment compared to what is expected by chance (Fisher's exact $P = 7 \times 10^{-67}$). For instance, for rs7045087 (associated to red blood cell counts) we observed that the expression of interferon gene *DDX58* (mapping 38bp downstream from rs7045087) significantly interacted with *trans*-eQTL effects on interferon genes *HERC5, OAS1, OAS3, MX1, IFIT1, IFIT2, IFIT5, IFI44, IFI44L, RSAD2* and *SAMD9* (**Extended Data Figure 9**).

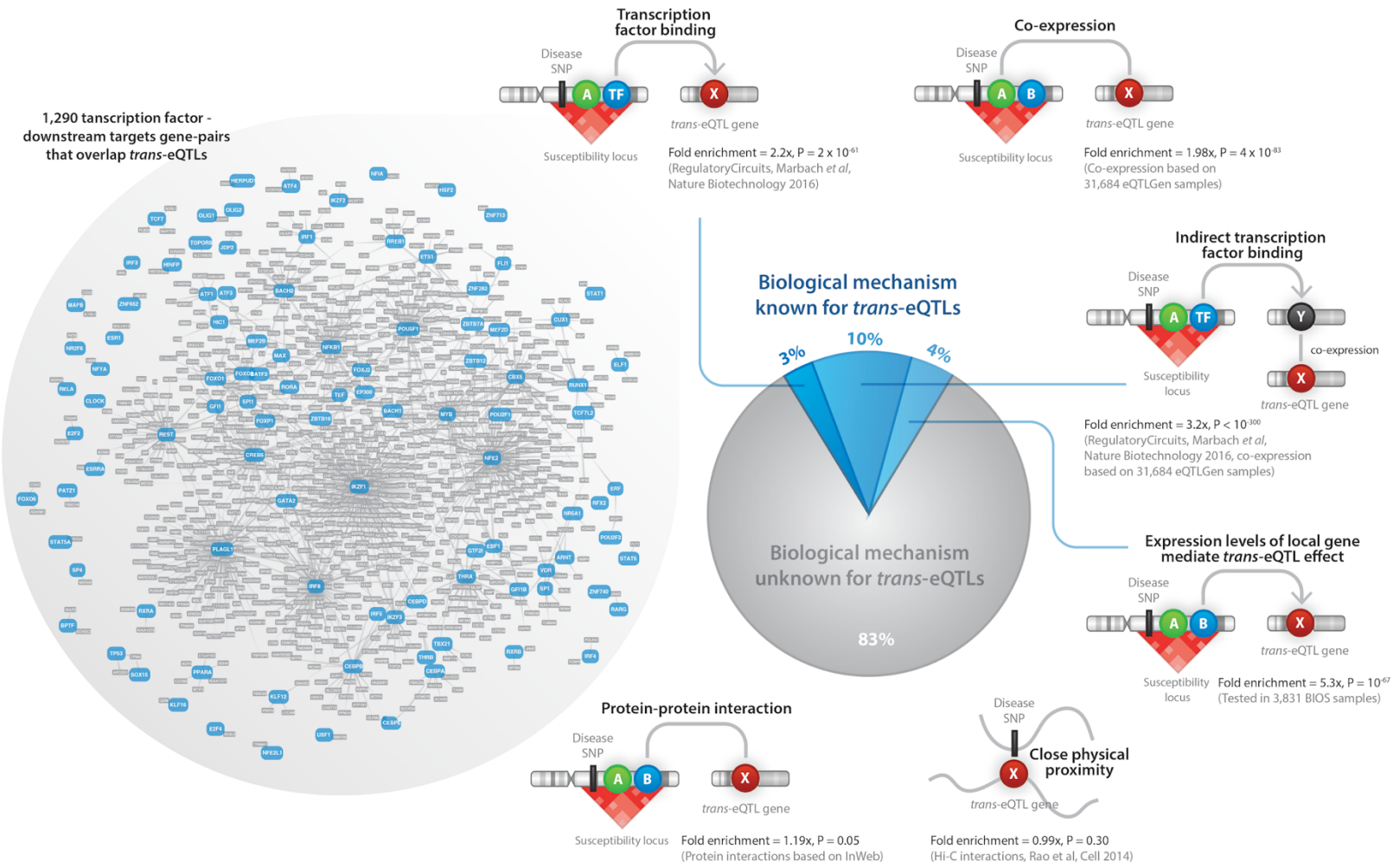


Figure 3. Mechanisms leading to *trans*-eQTLs. Shown are the results of enrichment analyses for known TF associations, HiC contacts, protein-protein interactions, gene co-expression and mediation analyses.

We estimate that 17.4% of the identified *trans*-eQTLs are explainable by (indirect) TF binding or mediation by *cis*-genes (**Supplementary Note**). This leaves 82.6% of the observed *trans*-eQTL effects unexplained. While it is likely that many of these *trans*-eQTLs reflect unknown (indirect) effects of TFs, we speculate that novel and unknown regulatory mechanisms could also play a role. By making all *trans*-eQTL results (irrespective of their statistical significance) publicly available, we envision this dataset will help to yield such insight in the future.

To estimate the proportion of loci where the trait-associated SNP explained the *trans*-eQTL signal in the locus, we performed locus-wide conditional *trans*-eQTL analysis in a subset of 4,339 samples for 12,991 *trans*-eQTL loci (**Online Methods; Extended Data Figure 10; Supplementary Table 6**). In 43% of these loci, we observed that the trait-associated SNP was in high LD with the *trans*-eQTL SNP having the strongest association in the locus ($R^2 > 0.8$, 1kG p1v3 EUR; **Supplementary Table 7**). For 95 cases, the strongest *cis*- and *trans*-eQTL SNPs were both in high LD with GWAS SNP ($R^2 > 0.8$ between top SNPs, 1kG p1v3 EUR; **Supplementary Table 7**).

The majority (64%) of *trans*-eQTL SNPs have previously been associated with blood composition phenotypes, such as platelet count, white blood cell count and mean corpuscular volume[20]. In comparison, blood cell composition SNPs from the same study comprised only 20.7% od all the tested trait-associated SNPs. This was expected, since SNPs that regulate the abundance of a specific blood cell type would result in *trans*-eQTL effects on genes, specifically expressed in that cell type.

Therefore, we aimed to distinguish *trans*-eQTLs caused by intracellular molecular mechanisms from blood cell type QTLs using eQTL data from lymphoblastoid cell line (LCL), induced pluripotent cells (iPSCs), several purified blood cell types (CD4+, CD8+, CD14+, CD15+, CD19+, monocytes and platelets) and blood DNA methylation QTL data. In total, 3,853 (6.4%) of *trans*-eQTLs showed significant replication in at least one cell type or in the methylation data (**Extended Data Figure 11**, **Supplementary Table 11A**). While this set of *trans*-eQTLs (denoted as the “intracellular eQTLs”) is less likely to be driven by cell type composition, we acknowledge that the limited sample size of the available *trans*-eQTL replication datasets make our replication effort very conservative. Furthermore, *trans*-eQTLs caused by variants associated with cell type proportions may be

informative for understanding the biology of a trait. Therefore, we did not remove these kinds of *trans*-eQTLs from our interpretative analyses.

Next, we aimed to replicate the identified *trans*-eQTLs in the tissues from GTEx[13]. Although the replication rate was very low (0-0.03% of *trans*-eQTLs replicated in non-blood tissues, FDR < 0.05, same allelic direction; **Supplementary Table 11B**), we did observe an inflation of signal (median chi-squared statistic) for identified *trans*-eQTLs in several GTEx tissues (**Extended Data Figure 12**). Non-blood tissues showing the strongest inflation were liver, heart atrial appendage and non-sun-exposed skin.

Trans-eQTLs are effective for discerning the genetic basis of complex traits

As described above, *trans*-eQTLs can arise due to *cis*-eQTL effects on TFs, whose target genes show *trans*-eQTL effects. We describe below such examples, but also highlight *trans*-eQTLs where the eQTL SNP works through a different mechanism. **Combining *cis*- and *trans*-eQTL effects can pinpoint the genes acting as drivers of *trans*-eQTL effects.** For example, the age-of-menarche-associated SNP rs1532331[21] is in high LD with the top *cis*-eQTL effect for transcription factor *ZNF131* (R^2 > 0.8, 1kG p1v3 EUR). *Cis*-eQTL and *trans*-eQTL effects for this locus co-localized for 25 out of the 75 downstream genes (**Figure 4A**). In a recent short hairpin RNA knockdown experiment of *ZNF131*[22], three separate cell isolates showed downregulation of four genes that we identified as *trans*-eQTL genes: *HAUS5, TMEM237, MIF4GD* and *AASDH* (**Figure 4A**). *ZNF131* has been hypothesized to inhibit estrogen signaling[23], which may explain how the SNP in this locus contributes to altering the age of menarche.

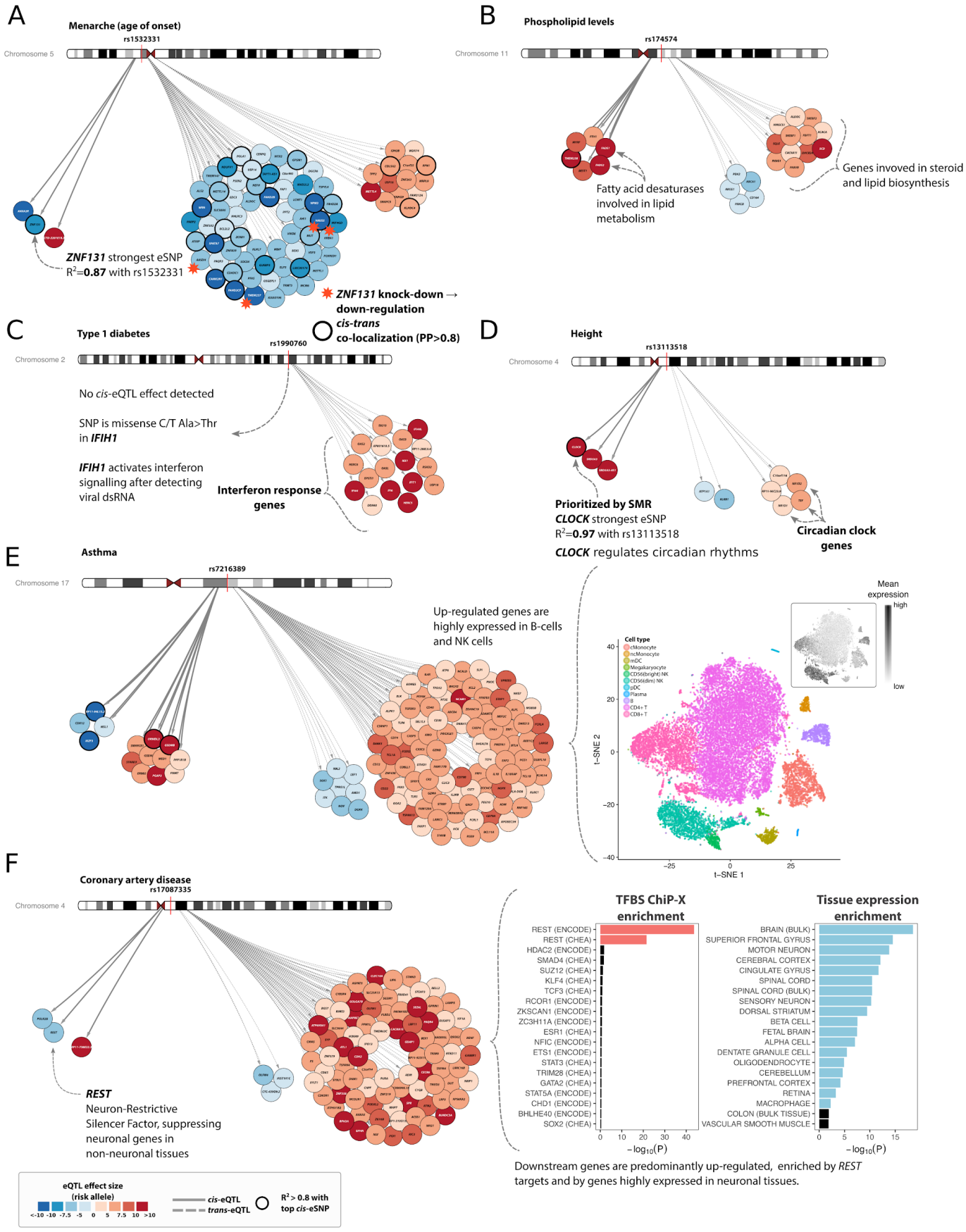


Figure 4. Examples of *cis*- and *trans*-eQTLs. (**A**) *Cis*-eQTL on *ZNF131* is prioritized because several *trans*-eQTL genes are down-regulated by *ZNF131* in functional study. (**B**) Phospholipid-associated SNP shows *cis*-and *trans*-eQTLs on lipid metabolism genes. (**C**) Type I diabetes associated SNP has no *cis*-eQTLs, but *trans*-eQTL genes point to interferon signaling pathway. (**D**) Circadian rhythm genes *CLOCK* (in *cis*) and *NR1D1*, *NR1D2*, *TEF* (in *trans*) identified for height associated SNP. (**E**) eQTLs for asthma SNP tag cell type abundance of B and NK cells. (**F**) *Trans*-eQTL genes for *REST* locus are highly enriched for REST transcription factor targets and for neuronal expression.

***Trans*-eQTLs extend insight for loci with multiple *cis*-eQTL effects.** In the *FADS1*/*FADS2* locus, rs174574 is associated with lipid levels[24] and affects 17 genes in *trans* (**Figure 4B**). The strongest *cis*-eQTLs modulate the expression of *FADS1*, *FADS2* and *TMEM258*, with latter being in high LD with GWAS SNP (R^2>0.8, 1kG p1v3 EUR). *FADS1* and *FADS2* have been implicated[24] since they regulate fatty acid synthesis, and consistent with their function, *trans*-eQTL genes from this locus are highly enriched for triglyceride metabolism ($P < 4.1\times10^{-9}$, GeneNetwork[25] REACTOME pathway enrichment). Since this locus has extensive LD, variant and gene prioritization is difficult: conditional analyses in 4,339 sample subset showed that each of *cis*-eQTL gene is influenced by more than one SNP, but none of these are in high LD with rs174574 (R^2 < 0.8, 1kG p1v3, EUR). As such, our *trans*-eQTL analysis results are informative for implicating *FADS1* and *FADS2*, whereas *cis*-eQTLs are not.

***Trans*-eQTLs can shed light on loci with no detectable *cis*-eQTLs.** rs1990760 is associated with multiple immune-related traits (Type 1 Diabetes (T1D), Inflammatory bowel disease (IBD), Systemic Lupus Erythematosis (SLE) and psoriasis[26–29]). For this SNP we identified 17 *trans*-eQTL effects, but no detectable gene-level *cis*-eQTLs in blood (**Figure 4C**) and GTEx. However, the risk

allele for this SNP causes an Ala946Thr amino acid change in the RIG-1 regulatory domain of MDA5 (encoded by *IFIH1* - Interferon Induced With Helicase C Domain 1), outlining one possible mechanism leading to the observed *trans*-eQTLs. MDA5 acts as a sensor for viral double-stranded RNA, activating interferon I signalling among other antiviral responses. All the *trans*-eQTL genes were up-regulated relative to risk allele to T1D, and 9 (52%) are known to be involved in interferon signaling (**Supplementary Table 12**).

***Trans*-eQTLs can reveal cell type composition effects of the trait-associated SNP.** *Trans*-eQTL effects can also show up as a consequence of a SNP that alters cell-type composition. For example, the asthma-associated SNP rs7216389[30] has 14 *cis*-eQTL effects, most notably on *IKZF3*, *GSDMB*, and *ORMDL3* (**Figure 4E**). SMR prioritized all three *cis*-genes equally (**Extended Data Figure 13**), making it difficult to draw biological conclusions (similar as we observed for the *FADS* locus). However, 94 out of the 104 *trans*-eQTL genes were up-regulated by the risk allele for rs7216389 and were mostly expressed in B cells and natural killer cells[31] (**Figure 4E**). *IKZF3* is part of the Ikaros transcription factor family that regulates B-cell proliferation[31,32], suggesting that a decrease of *IKZF3* leads to an increased number of B cells and concurrent *trans*-eQTL effects caused by cell-type composition differences.

Some *trans*-eQTLs influence genes strongly expressed in tissues other than blood. We observed *trans*-eQTL effects on genes that are hardly expressed in blood, indicating that our *trans*-eQTL effects are informative for non-blood related traits as well: rs17087335, which is associated with coronary artery disease[33], affects the expression of 88 genes in *trans* (**Figure 4F**), that are highly expressed in brain (hypergeometric test, ARCHS4 database, q-value = 2.58×10^{-17}; **Figure 4F, Supplementary Table 13**), but show very low expression in blood. SNPs linked with rs17087335 (R^2 > 0.8, 1kG p1v3 EUR) are associated with height (rs2227901, rs3733309 and

rs17081935)[34,35], and platelet count (rs7665147)[20]. The minor alleles of these SNPs downregulate the nearby gene *REST* (RE-1 silencing transcription factor), although none of these variants is in LD (R^2<0.2, 1kG p1v3 EUR) with the lead *cis*-eQTL SNP for *REST*. *REST* is a TF that downregulates the expression of neuronal genes in non-neuronal tissues[36,37]. It also regulates the differentiation of vascular smooth muscles, and is thereby associated with coronary phenotypes[38]. 85 out of 88 (96.6%) of the *trans*-eQTL genes were upregulated relative to the minor allele and were strongly enriched by transcription factor targets of REST (hypergeometric test for ENCODE REST ChIP-seq, q-value = 1.36×10^{-42} , **Figure 4F**). As such, *trans*-eQTL effects on neuronal genes implicate *REST* as the causal gene in this locus.

***Trans*-eQTLs identify pathways not previously associated with a phenotype.** Some *trans*-eQTLs suggest the involvement of pathways which are not previously thought to play a role for certain complex traits: SMR analysis prioritized *CLOCK* as a potential causal gene in the height-associated locus on chr 4q12 ($P_{SMR}=3\times10^{-25}$; $P_{HEIDI}=0.02$; **Figure 4D**). In line with that, height-associated SNP rs13113518[34] is also in high LD (R^2>0.8, 1kG p1v3 EUR) with the top *cis*-eQTL SNP for *CLOCK*. The upregulated TF CLOCK forms a heterodimer with TF BMAL1, and the resulting protein complex regulates circadian rhythm[39]. Three known circadian rhythm *trans*-eQTL genes (*TEF, NR1D1* and *NR1D2*) showed increased expression for the trait-increasing allele, suggesting a possible mechanism for the observed *trans*-eQTLs through binding of CLOCK:BMAL1. *TEF* is a D-box binding TF whose gene expression in liver and kidney is dependent on the core circadian oscillator and it regulates amino acid metabolism, fatty acid metabolism and xenobiotic detoxification (Gachon et al., 2006)**.** *NR1D1* and *NR1D2* encode the transcriptional repressors Rev-ErbA alpha and beta, respectively, and form a negative feedback loop to suppress *BMAL1* expression[40]. *NR1D1* and *NR1D2* have been reported to be associated

with osteoblast and osteoclast functions[41], revealing a possible link between circadian clock genes and height.

Unlinked trait-associated SNPs converge on the same downstream genes in *trans*. We subsequently ascertained, per trait, whether unlinked trait-associated variants showed *trans*-eQTL effects on the same downstream gene. Here we observed 47 different traits where at least four independent variants affected the same gene in *trans*, 3.4× higher than expected by chance (P = 0.001; two-tailed two-sample test of equal proportions; **Supplementary Table 8**). For SLE, for example, we observed that the gene expression levels of *IFI44L*, *HERC5*, *IFI6*, *IFI44*, *RSAD2*, *MX1*, *ISG15*, *ANKRD55*, *OAS3*, *OAS2*, *OASL* and *EPSTI1* (nearly all interferon genes) were affected by at least three SLE-associated genetic variants, clearly showing the involvement of interferon signaling in SLE (**Figure 5**).

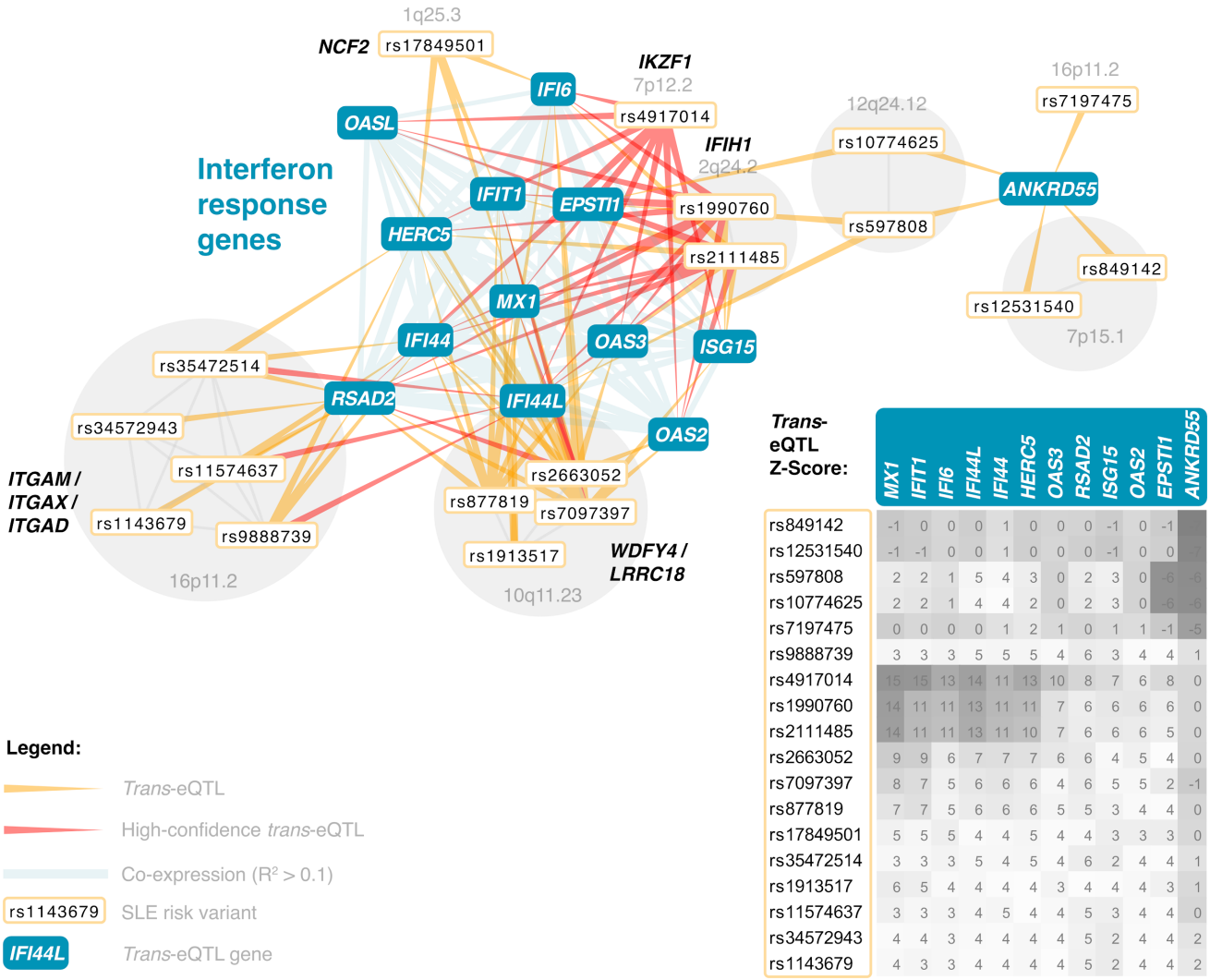


Figure 5. SNPs associated with SLE converge to the shared cluster of interferon response genes. Shown are genes which are affected by at least three independent GWAS SNPs. SNPs in partial LD are grouped together. Heat map indicates the direction and strength of individual *trans*-eQTL effects (Z-scores).

This convergence of multiple SNPs on the same genes lends credence to recent hypotheses with regards to the 'omnigenic' architecture of complex traits[8]: indeed multiple unlinked variants do affect the same 'core' genes. The recent omnigenic model[42] proposes a strategy to partition between core genes, which have direct effects on a disease, and peripheral genes, which can only affect disease risk indirectly through regulation of core genes. In **Supplementary Equations**, we

show that this model also implies a correlation between polygenic risk scores and expression of core genes. We therefore studied this systematically by aggregating multiple associated variants into polygenic scores and ascertaining how they correlate with gene expression levels.

eQTSs identify key driver genes for polygenic traits

To ascertain the coordinated effects of trait-associated variants on gene expression, we used available GWAS summary statistics to calculate PGSs for 1,263 traits in 28,158 samples (**Online Methods, Supplementary Table 14**). We reasoned that when a gene shows expression levels that significantly correlate with the PGS for a specific trait (an expression quantitative trait score; eQTS), the downstream *trans*-eQTL effects of the individual risk variants converge on that gene, and hence, that the gene may be a driver of the disease.

Our meta-analysis identified 18,210 eQTS effects (FDR < 0.05), representing 689 unique traits (54%) and 2,568 unique genes (13%; **Supplementary Table 15**, **Figure 1A**). As expected, most eQTS associations represent blood cell traits (**Extended Data Figure 14, Supplementary Table 16**): for instance the PGS for mean corpuscular volume correlated positively with the expression levels of genes specifically expressed in erythrocytes, such as genes coding for hemoglobin subunits. However, we also identified eQTS associations for genes that are known drivers of other traits.

For example, 11 out of 26 genes associating with the PGS for high density lipoprotein levels (HDL[43,44]; FDR<0.05; **Figure 6A**) have previously been associated with lipid or cholesterol metabolism (**Supplementary Table 18**). *ABCA1* and *ABCG1*, which positively correlated with the PGS for high HDL, mediate the efflux of cholesterol from macrophage foam cells and participate in

HDL formation. In macrophages, the downregulation of both *ABCA1* and *ABCG1* reduced reverse cholesterol transport into the liver by HDL[45] (**Figure 6B**). The genetic risk for high HDL was also negatively correlated with the expression of the low density lipoprotein receptor *LDLR* (strongest eQTS $P=3.35\times10^{-20}$) known to cause hypercholesterolemia[46]. Similarly, the gene encoding the TF SREBP-2, which is known to increase the expression of *LDLR*, was downregulated (strongest eQTS $P=3.08\times10^{-7}$). The negative correlation between *SREBF2* expression and measured HDL levels has been described before[47], indicating that the eQTS reflects an association with the actual phenotype. Zhernakova et al. proposed a model where down-regulation of *SREBF2* results in the effect on its target gene *FADS2*. We did not observe a significant HDL eQTS effect on *FADS2* (all eQTS $P>0.07$), possibly because the indirect effect is too small to detect. We hypothesize that HDL levels in blood can result in a stronger reverse cholesterol transport into the liver, which may result in downregulation of *LDLR*[48]

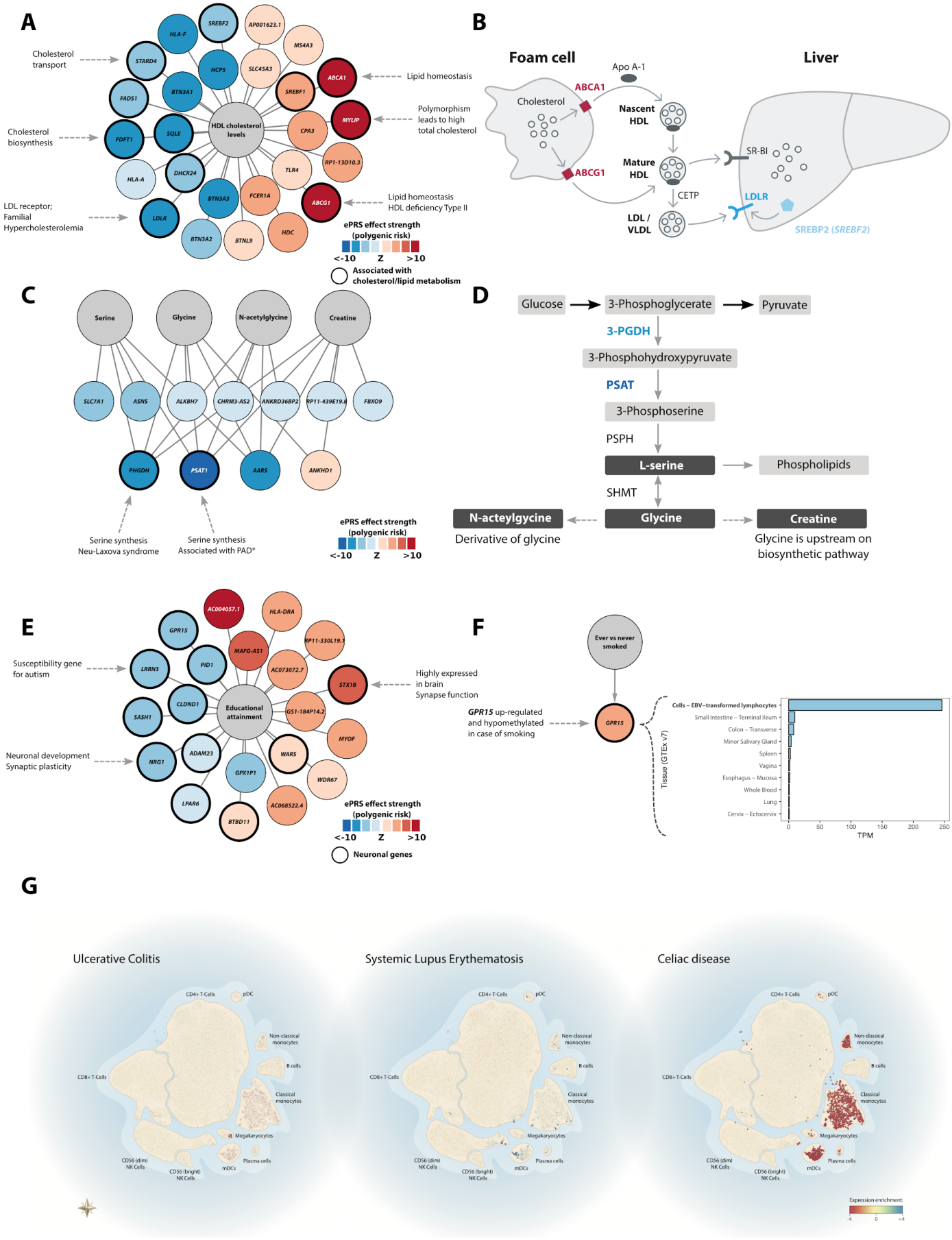


Figure 6. Examples of eQTS. (**A**) Polygenic risk score (PRS) for high density lipoprotein associates to lipid metabolism genes. (**B**) The role of *ABCA1*, *ABCG1*, and *SREBF2* in cholesterol transport. (**C**) Polygenic scores for serine, glycine, n-acetylglycine and creatine levels negatively associate with gene expression of *PHGDH*, *PSAT1*, and *AARS*. (**D**) Serine biosynthesis pathway. (**E**) PRS for educational attainment identifies genes with neuronal functions. (**F**) Polygenic score for smoking status upregulates *GPR15*, which plays a role in lymphocyte differentiation. (**G**) eQTS genes for immune-related diseases are enriched for genes specifically expressed in certain blood cell types.

eQTS analysis also identified genes relevant for non-blood traits, such as the association of *GPR15* ($P=3.7\times10^{-8}$, FDR<0.05; **Figure 6F**) with the trait 'ever versus never smoking'[49]. *GPR15* is a biomarker for smoking[50] that is overexpressed and hypomethylated in smokers[51]. We observe strong *GPR15* expression in lymphocytes (**Figure 6F**), suggesting that the association with smoking could originate from a change in the proportion of T cells in blood[52]. As *GPR15* is involved in T cell homing and has been linked to colitis and inflammatory phenotypes, it is hypothesized to play a key role in smoking-related health risks[53].

The PGS for another non-blood trait, educational attainment[54], correlated significantly with the expression of 21 genes (FDR<0.05; **Figure 6E**, **Supplementary Table 15**). Several of the strongly associated genes are known to be involved in neuronal processes (**Supplementary Table 19**) and show expression in neuronal tissues (GTEx v7, **Extended Data Figure 15**). *STX1B* (strongest eQTS $P=1.3\times10^{-20}$) is specifically expressed in brain, and its encoded protein, syntaxin 1B, participates in the exocytosis of synaptic vesicles and synaptic transmission[55]. Another gene highly expressed in brain, *LRRN3* (Leucine-rich repeat neuronal protein 3; strongest eQTS $P=1.7\times10^{-11}$) was negatively associated with the PGS for educational attainment, and has been associated with autism susceptibility[56]. The downregulated *NRG1* (neuregulin 1; strongest eQTS $P=4.5\times10^{-7}$),

encodes a well-established growth factor involved in neuronal development and has been associated to synaptic plasticity[57]. *NRG1* was also positively associated with the PGS for monocyte levels[20] (strongest eQTS $P=1.5\times10^{-7}$), several LDL cholesterol traits (e.g. medium LDL particles[44]; strongest eQTS $P=6.2\times10^{-8}$), coronary artery disease[33] (strongest eQTS $P=1.5\times10^{-6}$) and body mass index in females[58] (strongest eQTS $P=9.2\times10^{-12}$).

eQTS can also identify pathways known to be associated with monogenic diseases**.** For example, the PGSs for serine, glycine, the glycine derivative n-acetylglycine and creatine[59,60] (**Figure 6C**) were all negatively associated with the gene expression levels of *PHGDH*, *PSAT1* and *AARS* ($P < 5.3\times10^{-7}$). *PHGDH* and *PSAT1* encode crucial enzymes that regulate the synthesis of serine and, in turn, glycine[61] (**Figure 6D**), while n-acetylglycine and creatine form downstream of glycine[62]. Mutations in *PSAT1* and *PHGDH* can result in serine biosynthesis defects including phosphoserine aminotransferase deficiency[63], phosphoglycerate dehydrogenase deficiency[64], and Neu-Laxova syndrome[65], all diseases characterized by low concentrations of serine and glycine in blood and severe neuronal manifestations. *AARS* encodes alanyl-tRNA synthetase, which links alanine to tRNA molecules. A mutation in *AARS* has been linked to Charcot Marie Tooth disease[66], while the phenotypically similar hereditary sensory neuropathy type 1 (HSN1[67]) can be caused by a mutation in the gene encoding serine palmitoyltransferase. The gene facilitates serine's role in sphingolipid metabolism[68]. Disturbances in this pathway are hypothesized to be central in the development the neuronal symptoms[69], suggesting a link between *AARS* expression and the serine pathway. Unexpectedly, the genetic risk for higher levels of these amino acids was associated with lower expression of *PHGDH*, *PSAT1*, and *AARS*, implying the presence of a negative feedback loop that controls serine synthesis.

We next evaluated 6 immune diseases for which sharing of loci has been reported previously, and also observed sharing of downstream eQTS effects for these diseases (**Supplementary Table 20**). For example, the interferon gene *STAT1* was significantly associated with T1D, celiac disease (CeD), IBD and primary biliary cirrhosis (PBC). However, some of these genes are also marker genes for specific blood-cell types, such as *CD79A*, which showed a significant correlation with T1D and PBC. To test whether disease-specific eQTS gene signatures are reflected by blood cell proportions, we investigated single-cell RNA-seq data[31] (**Online Methods**; **Figure 6G**). For ulcerative colitis (a subtype of IBD), we observed significant depletion of expression in megakaryocytes. SLE eQTS genes were enriched for antigen presentation (GeneNetwork $P=1.3\times10^{-5}$) and interferon signaling (GeneNetwork $P=1.4\times10^{-4}$), consistent with the well-described interferon signature in SLE patients[70,71]. Moreover, the SLE genes were significantly enriched for expression in mature dendritic cells, whose maturation depends on interferon signaling[72]. For CeD, we observed strong depletion of eQTS genes in monocytes and dendritic cells, and a slight enrichment in CD4+ and CD8+ T cells. The enrichment of cytokine (GeneNetwork $P=1.6\times10^{-15}$) and interferon (GeneNetwork $P=7.8\times10^{-13}$) signaling among the CeD eQTS genes is expected as a result of increased T cell populations.

Cell-type-specificity of eQTS associations

We next ascertained to what extent these eQTS associations can be replicated in non-blood tissues. We therefore aimed to replicate the significant eQTS effects in 1,460 LCL samples and 762 iPSC samples. Due to the fact these cohorts have a comparatively low sample sizes and study different cell types, we observed limited replication: 10 eQTS showed significant replication effect (FDR<0.05) in the LCL dataset, with 9 out of those (90%) showing the same effect direction as in

the discovery set (**Extended Data Figure 16A, Supplementary Table 17**). For iPSCs, only 5 eQTS showed a significant effect (**Extended Data Figure 16B, Supplementary Table 17**). Since only a few eQTS associations are significant in non-blood tissues and the majority of identified eQTS associations are for blood-related traits, we speculate these effects are likely to be highly cell-type specific. This indicates that large-scale eQTL meta-analyses in other tissues could uncover more genes on which trait-associated SNPs converge.

Discussion

We here performed *cis*-eQTL, *trans*-eQTL and eQTS analyses in 31,684 blood samples, reflecting a six-fold increase over earlier large-scale studies[1,5]. We identified *cis*-eQTL effects for 88.3% and *trans*-eQTL effects for 32% of all genes that are expressed in blood.

We observed that *cis*-eQTL SNPs map close to the TSS or TES of the *cis*-gene: for the top 20% strongest *cis*-eQTL genes, 84.1% of the lead eQTL SNPs map within 20kb of the gene, indicating that these are variants immediately adjacent to the start or end of transcripts that primarily drive *cis*-eQTL effects. The trait-associated variants that we studied showed a different pattern: 77.4% map within 20kb of the closest protein-coding gene, suggesting that the genetic architecture of *cis*-eQTLs is different from disease-associated variants. This is supported by the epigenetic differences that we observed between these two groups and can also partly explain, why we did not observe significantly increased overlap between genes prioritized using pathway enrichment analysis[15] and genes prioritized using our *cis*-eQTLs.

In contrast, for numerous traits we observed that multiple unlinked *trans*-eQTL variants often converge on genes with a known role in the biology for these traits (e.g. the involvement of interferon genes in SLE).

We therefore focused on *trans*-eQTL and eQTS results to gain insight into trait-relevant genes and pathways (**Figures 4, 6**). We estimate that 17.4% of our *trans*-eQTLs are driven by transcriptional regulation, whereas the remaining fraction is driven by not-yet-identified mechanisms. Our results support a model which postulates that, compared to *cis*-eQTLs, weaker distal and polygenic effects converge on core (key driver) genes that are more relevant to the traits and more specific for trait-relevant cell types (**Figure 1B**). The examples we have highlighted demonstrate how insights can be gained from our resource, and we envision similar interpretation strategies can be applied to the other identified *trans*-eQTL and eQTS effects. The catalog of genetic effects on gene expression we present here (available at www.eqtlgen.org) is a unique compendium for the development and application of novel methods that prioritize causal genes for complex traits[14,73], as well as for interpreting the results of genome-wide association studies.

Methods

Cohorts

eQTLGen Consortium data consists of 31,684 blood and PBMC samples from 37 datasets, pre-processed in a standardized way and analyzed by each cohort analyst using the same settings (**Online Methods**). 26,886 (85%) of the samples added to discovery analysis were whole blood samples and 4,798 (15%) were PBMCs, and the majority of samples were of European ancestry (**Supplementary Table 1**). The gene expression levels of the samples were profiled by Illumina (N=17,421; 55%), Affymetrix U291 (N=2,767; 8.7%), Affymetrix HuEx v1.0 ST (N=5,075; 16%) expression arrays and by RNA-seq (N=6,422; 20.3%). A summary of each dataset is outlined in **Supplementary Table 1.** Detailed cohort descriptions can be found in the **Supplementary Note**. Each of the cohorts completed genotype and expression data pre-processing, PGS calculation, *cis*-eQTL-, *trans*-eQTL- and eQTS-mapping, following the steps outlined in the online analysis plans, specific for each platform (see **URL**s) or with slight alterations as described in **Supplementary Table 1** and the **Supplementary Note**. All but one cohort (Framingham Heart Study), included non-related individuals into the analysis.

Genotype data preprocessing

The primary pre-processing and quality control of genotype data was conducted by each cohort, as specified in the original publications and in the **Supplementary Note**. The majority of cohorts used genotypes imputed to 1kG p1v3 or a newer reference panel. GenotypeHarmonizer[74] was used to harmonize all genotype datasets to match the GIANT 1kG p1v3 ALL reference panel and

to fix potential strand issues for A/T and C/G SNPs. Each cohort tested SNPs with minor allele frequency (MAF) > 0.01, Hardy-Weinberg P-value > 0.0001, call rate > 0.95, and MACH r^2 > 0.5.

Expression data preprocessing

Illumina arrays

Illumina array datasets expression were profiled by HT-12v3, HT-12v4 and HT-12v4 WGDASL arrays. Before analysis, all the probe sequences from the manifest files of those platforms were re-mapped to GRCh37.p10 human genome and transcriptome, using SHRiMP v2.2.3 aligner[75] and allowing 2 mismatches. Probes mapping to multiple locations in the genome were removed from further analyses.

For Illumina arrays, the raw unprocessed expression matrix was exported from GenomeStudio. Before any pre-processing, the first two principal components (PCs) were calculated on the expression data and plotted to identify and exclude outlier samples. The data was normalized in several steps: quantile normalization, $\log_2$ normalization, probe centering and scaling by the equation $Expression_{Probe,Sample}$ = ($Expression_{Probe,Sample}$ - MeanProbe) / $Std.Dev._{Probe}$. Genes showing no variance were removed. Next, the first four multidimensional scaling (MDS) components, calculated based on non-imputed and pruned genotypes using plink v1.07[76], were regressed out of the expression matrix to account for population stratification. We further removed up to 20 first expression-based PCs that were not associated to any SNPs, as these capture non-genetic variation in expression. Each cohort also ran MixupMapper[77] software to identify incorrectly labeled genotype-expression combinations, and to remove identified sample mix-ups.

Affymetrix arrays

Affymetrix array-based datasets used the expression data previously pre-processed and quality controlled as indicated in the **Supplementary Note**.

RNA-seq

Alignment, initial quality control and quantification differed slightly across datasets, as described in the **Supplementary Note**. Each cohort removed outliers as described above, and then used Trimmed Mean of M-values (TMM) normalization and a counts per million (CPM) filter to include genes with >0.5 CPM in at least 1% of the samples. Other steps were identical to Illumina processing, with some exceptions for the BIOS Consortium datasets (**Supplementary Note**).

Cis-eQTL mapping

Cis-eQTL mapping was performed in each cohort using a pipeline described previously[1]. In brief, the pipeline takes a window of 1Mb upstream and 1Mb downstream around each SNP to select genes or expression probes to test, based on the center position of the gene or probe. The associations between these SNP-gene combinations was calculated using a Spearman correlation. Next, 10 permutation rounds were performed by shuffling the links between genotype and expression identifiers and re-calculating associations. The false discovery rate (FDR) was determined using 10 meta-analyzed permutations: for each gene in the real analysis, the most significant association was recorded, and the same was done for each of the permutations,

resulting in a gene-level FDR. *Cis*-eQTLs with a gene-level FDR < 0.05 (corresponding to P < 1.829×10^{-5}) and tested in at least two cohorts were deemed significant.

Trans-eQTL mapping

Trans-eQTL mapping was performed using a previously described pipeline[1] while testing a subset of 10,317 SNPs previously associated with complex traits. We required the distance between the SNP and the center of the gene or probe to be >5Mb. To maximize the power to identify *trans*-eQTL effects, the results of the summary statistics based or iterative conditional *cis*-eQTL mapping analyses (**Supplementary Note**) were used to correct the expression matrices before *trans*-eQTL mapping. For that, top SNPs for significant conditional *cis*-eQTLs were regressed out from the expression matrix. Finally, we removed potential false positive *trans*-eQTLs caused by reads cross-mapping with *cis* regions (**Supplementary Note**).

Genetic risk factor selection

Genetic risk factors were downloaded from three public repositories: the EBI GWAS Catalogue[78] (downloaded 21.11.2016), the NIH GWAS Catalogue and Immunobase (www.immunobase.org; accessed 26.04.2016), applying a significance threshold of $P \leq 5\times10^{-8}$. Additionally, we added 2,706 genome-wide significant GWAS SNPs from a recent blood trait GWAS[20]. SNP coordinates were lifted to hg19 using the *liftOver* command from R package rtracklayer v1.34.1[79] and subsequently standardized to match the GIANT 1kG p1v3 ALL reference panel. This yielded 10,562 SNPs (**Supplementary Table 2**). We tested associations between all risk factors and

genes that were at least 5Mb away to ensure that that they did not tag a *cis*-eQTL effect. All together, 10,317 trait-associated SNPs were tested in *trans*-eQTL analyses.

eQTS mapping

PGS trait inclusion

Full association summary statistics were downloaded from several publicly available resources (**Supplementary Table 13**). GWAS performed exclusively in non-European cohorts were omitted. Filters applied to the separate data sources are indicated in the **Supplementary Note**. All the dbSNP rs numbers were standardized to match GIANT 1kG p1v3, and the directions of effects were standardized to correspond to the GIANT 1kG p1v3 minor allele. SNPs with opposite alleles compared to GIANT alleles were flipped. SNPs with A/T and C/G alleles, tri-allelic SNPs, indels, SNPs with different alleles in GIANT 1kG p1v3 and SNPs with unknown alleles were removed from the analysis. Genomic control was applied to all the P-values for the datasets not genotyped by Immunochip or Metabochip. Additionally, genomic control was skipped for one dataset that did not have full associations available[80] and for all the datasets from the GIANT consortium, as for these genomic control had already been applied. All together, 1,263 summary statistic files were added to the analysis. Information about the summary statistics files can be found in the **Supplementary Note** and **Supplementary Table 14**.

PGS calculation

A custom Java program, GeneticRiskScoreCalculator-v0.1.0c, was used for calculating several PGS in parallel. Independent effect SNPs for each summary statistics file were identified by double-

clumping by first using a 250kb window and subsequently a 10Mb window with LD threshold R^2=0.1. Subsequently, weighted PGS were calculated by summing the risk alleles for each independent SNP, weighted by its GWAS effect size (beta or log(OR) from the GWAS study). Four GWAS P-value thresholds ($P<5\times10^{-8}$, 1×10^{-5}, 1×10^{-4} and 1×10^{-3}) were used for constructing PGS for each summary statistics file.

Pruning the SNPs and PGS

To identify a set of independent genetic risk factors, we conducted LD-based pruning as implemented in PLINK 1.9[81] with the setting --indep-pairwise 50 5 0.1. This yielded in 4,586 uncorrelated SNPs (R^2<0.1, GIANT 1kG p1v3 ALL).

To identify the set of uncorrelated PGS, ten permuted *trans*-eQTL Z-score matrices from the combined *trans*-eQTL analysis were first confined to the pruned set of SNPs. Those matrices were then used to identify 3,042 uncorrelated genes, based on Z-score correlations (absolute Pearson R < 0.05). Next, permuted eQTS Z-score matrices were confined to uncorrelated genes and used to calculate pairwise correlations between all genetic risk scores to define a set of 1,873 uncorrelated genetic risk scores (Pearson R^2 < 0.1).

Empirical probe matching

To integrate different expression platforms (four different Illumina array models, RNA-seq, Affymetrix U291 and Affymetrix Hu-Ex v1.0 ST) for the purpose of meta-analysis, we developed an empirical probe-matching approach. We used the pruned set of SNPs to conduct per-platform meta-analyses for all Illumina arrays, for all RNA-seq datasets, and for each Affymetrix dataset separately, using summary statistics from analyses without any gene expression correction for

principal components. For each platform, this yielded an empirical *trans*-eQTL Z-score matrix, as well as ten permuted Z-score matrices, where links between genotype and expression files were permuted. Those permuted Z-score matrices reflect the gene-gene or probe-probe correlation structure.

We used RNA-seq permuted Z-score matrices as a gold standard reference and calculated for each gene the Pearson correlation coefficients with all the other genes, yielding a correlation profile for each gene. We then repeated the same analysis for the Illumina meta-analysis, and the two different Affymetrix platforms. Finally, we correlated the correlation profiles from each array platform with the correlation profiles from RNA-seq. For each array platform, we selected the probe showing the highest Pearson correlation with the corresponding gene in the RNA-seq data and treated those as matching expression features in the combined meta-analyses. This yielded 19,960 genes that were detected in RNA-seq datasets and tested in the combined meta-analyses. Genes and probes were matched to Ensembl v71[82] (see **URLs**) stable gene IDs and HGNC symbols in all the analyses.

Cross-platform replications

To test the performance of the empirical probe-matching approach, we conducted discovery *cis*-, *trans*- and eQTS meta-analyses for each expression platform (RNA-seq, Illumina, Affymetrix U291 and Affymetrix Hu-Ex v1.0 ST arrays; array probes matched to 19,960 genes by empirical probe matching). For each discovery analysis, we conducted replication analyses in the three remaining platforms, observing strong replication of both *cis*-eQTLs, *trans*-eQTLs and eQTS in different platforms, with very good concordance in allelic direction.

Meta-analyses

We meta-analyzed the results using a weighted Z-score method[1], where the Z-scores are weighted by the square root of the sample size of the cohort. For *cis*-eQTL and *trans*-eQTL meta-analyses, this resulted in a final sample size of N=31,684. The combined eQTS meta-analysis included the subset of unrelated individuals from the Framingham Heart Study, resulting in a combined sample size of 28,158.

Quality control of the meta-analyses

For quality control of the overall meta-analysis results, MAFs for all tested SNPs were compared between eQTLGen and 1kG p1v3 EUR (**Extended Data Figure 3**), and the effect direction of each dataset was compared against the meta-analyzed effect (**Extended Data Figure 2A-C**).

FDR calculation for *trans*-eQTL and eQTS mapping

To determine nominal P-value thresholds corresponding to FDR=0.05, we used the pruned set of SNPs for *trans*-eQTL mapping and permutation-based FDR calculation, as described previously[1]. We leveraged those results to determine the P-value threshold corresponding to FDR=0.05 and used this as a significance level in *trans*-eQTL mapping in which all 10,317 genetic trait-associated SNPs were tested. In the eQTS analysis, an analogous FDR calculation was performed using a pruned set of PGSs. We analyzed only SNP/PGS-gene pairs tested in at least two cohorts.

Positive and negative set of *trans*-eQTLs

Based on the results of integrative *trans*-eQTL mapping, we defined true positive (TP) and true negative (TN) sets of *trans*-eQTLs. TP set was considered as all significant (FDR<0.05) *trans*-eQTLs. TN set of *trans*-eQTLs was selected as non-significant (max absolute meta-analysis Z-score 3; all FDR>0.05) SNP-gene combinations, adhering to following conditions:

1. The size of TN set was set equal to the size of TP set (59,786 *trans*-eQTLs).
2. Each SNP giving *trans*-eQTL effects on X genes in the TP set, is also giving *trans*-eQTL effects on X genes in the TN set.
3. Each gene that is affected in *trans* by Y SNPs in the TP set, is also affected in *trans* by Y SNPs in the TN set.
4. Adhere to the correlation structure of the SNPs: if two SNPs are in perfect LD, they affect the same set of genes, both in the TP set and in the TN set.
5. Adhere to the correlation structure of the genes: if two genes are perfectly co-expressed, they are affected by the same SNPs, both in the TP set and in the TN set.

This set of TN *trans*-eQTLs was used in subsequent enrichment analyses as the matching set for comparison.

Conditional *trans*-eQTL analyses

We aimed to estimate how many *trans*-eQTL SNPs were likely to drive both the *trans*-eQTL effect and the GWAS phenotype. The workflow of this analysis is shown in **Extended Data Figure 6**. We

used the integrative *trans*-eQTL analysis results as an input, confined ourselves to those effects which were present in the datasets we had direct access to (BBMRI-BIOS+EGCUT; N=4,339), and showed nominal $P < 8.3115\times 10^{-06}$ in the meta-analysis of those datasets. This P-value threshold was the same as in the full combined trans-eQTL meta-analysis and was based on the FDR=0.05 significance threshold identified from the analysis run on the pruned set of GWAS SNPs after removal of cross-mapping effects. We used the same methods and SNP filters as in the full combined *trans*-eQTL meta-analysis, aside from the FDR calculation, which was based on the full set of SNPs, instead of the pruned set of SNPs.

For each significant *trans*-eQTL SNP, we defined the locus by adding a ±1Mb window around it. Next, for each *trans*-eQTL gene we ran iterative conditional *trans*-eQTL analysis using all loci for given *trans*-eQTL gene. We then evaluated the LD between all conditional top *trans*-eQTL SNPs and GWAS SNPs using a 1 Mb window and $R^2>0.8$ (1kG p1v3 EUR) as a threshold for LD overlap.

Trans-eQTL mediation analysis

To identify potential mediators of *trans*-eQTL effects we used a G x E interaction model:

$$t = \beta_0 + \beta_1 \times s + \beta_2 \times m + \beta_3 \times s \times m$$

Where t is the expression of the *trans*-eQTL gene, s is the *trans*-eQTL SNP, and m is the expression of a potential mediator gene within 100kb of the *trans*-eQTL SNP. On top of the gene expression normalization that we used for the rest of our analysis, we used a rank-based inverse normal transformation to enforce a normal distribution before fitting the linear model, identical to the normalization used by Zhernakova et al.[47] in their G × E interaction eQTL analyses. We fitted this model separately on each of the cohorts that are part of the BIOS consortium. We transformed the interaction P-values to Z-scores and used the weighted Z-score method[83] to perform a meta-

analysis on the in total 3,831 samples. The Benjamini & Hochberg procedure[84] was used to limit the FDR to 0.05. The plots in **Extended Data Figure 9** are created with the default normalization, the regression lines are the best-fitting lines between the mediator gene and the *trans* eQTL gene, stratified by genotype. We used a Fisher's exact test to calculate the enrichment of significant (FDR ≤ 0.05) interactions between our TP *trans*-eQTLs and the interactions identified in the TN *trans*-eQTL set.

TF and tissue enrichment analyses

We downloaded the curated sets of known TF targets and tissue-expressed genes from the Enrichr web site[85,86]. TF target gene sets included TF targets as assayed by ChIP-X experiments from ChEA[87] and ENCODE[88,89] projects, and tissue-expressed genes were based on the ARCHS4 database[90]. Those gene sets were used to conduct hypergeometric over-representation analyses as implemented into the R package ClusterProfiler[91].

SMR analyses

To gain further insight into genes that are important in the biology of the trait, we used the combined *cis*-eQTL results to perform SMR[14] for 16 large GWAS studies (**Supplementary Table 20**). We derived *cis*-eQTL beta and standard error of the beta (SE(beta)) from the Z-score and the MAF reported in 1kG v1p3 ALL, using the following formulae[14]

$$beta = z / (\sqrt{(2p(1-p)(n+z^2))}$$

$SE(beta) = 1 / (\sqrt{(2p(1-p)(n+z^2))})$

Where p is the MAF and n is the sample size.

The *cis*-eQTLs were converted to the dense BESD format. The 1kG p1v3 ALL reference panel was also used to calculate LD, and SMR analysis was run using the SMR software v0.706 without any P-value cut-offs on either GWAS or eQTL input.

DEPICT

We applied DEPICT v194[15] to the same 16 recent GWAS traits as above (**Supplementary Table 20**), using all variants that attain a genome-wide significant P-value threshold. Specifically, we looked at the gene prioritization and gene set enrichment analyses to compare the results with the output of other prioritization methods (SMR [14]).

Comparison of gene prioritization with DEPICT and SMR

To investigate the consistency between results from two gene prioritization methods, we compared the enrichment of overlapping genes for 16 GWAS traits (**Supplementary Table 20**). We confined ourselves to genes that were tested in SMR and that fell within the DEPICT loci, and tested whether genes significant in SMR (P-value < 0.05 / number of tested genes) and DEPICT (FDR < 0.05) were enriched (one-sided Fisher's Exact Test).

Epigenetic marks enrichment

We ascertained epigenetic properties of the lead *cis*-eQTL SNPs, and contrasted these to a set of 3,688 trait-associated SNPs that were associated with either blood-related traits (such as mean corpuscular volume or platelet counts) or immune-mediated diseases. The SNPs were annotated with histone and chromatin marks information from the Epigenomics Roadmap Project. We summarized the information by calculating the overlap ratio across 127 human cell types between the epigenetic marks and the SNP within a window size of +/- 25bp: if a SNP co-localizes with a mark for all 127 cell-types, the score for that SNP will be 1; if a SNP co-localizes with a mark for none of the cell-types, the score will be 0.

The reason we chose only SNPs associated to blood-related traits and immune-mediated diseases was to minimize potential confounding due to a subtle bias in the Epigenomics Roadmap Project towards blood cell-types: 29 of the 127 cell-types that we studied were blood cell types. However, when redoing the epigenetic enrichment analysis, while excluding these blood cell types, we did not see substantial differences in the enriched and depleted histone marks.

Chromosomal contact analyses

Capture Hi-C overlap for *cis*-eQTLs

To assess whether *cis*-eQTL lead SNPs overlapped with chromosomal contact as measured using Hi-C data, we used promoter capture Hi-C data[92], downloaded from CHiCP[93] (see **URLs**). We took the lead eQTL SNPs and overlapped these with the capture Hi-C data and studied the 10,428 cis-eQTL genes for which this data is available. We then checked whether the Capture Hi-C target

maps within 5kb of the lead SNP. Of 508 *cis*-eQTL genes that mapped over 100 kb from the TSS or TES, 223 overlapped capture Hi-C data (27.8%). Of 7,984 *cis*-eQTL genes that mapped within 100kb from the TSS or TES, 1,641 overlapped capture Hi-C data (17.0%, Chi^2 test $P = 10^{-14}$). To ensure this was not an artefact, we performed the same analysis, while flipping the location of the capture Hi-C target with respect to the location of the bait, and did not observe any significant difference (Chi^2 test $P = 0.59$).

Hi-C overlap enrichment analysis for *trans*-eQTLs

To assess whether *trans*-eQTLs were enriched for chromosomal contacts as measured using Hi-C data, we downloaded the contact matrices for the human lymphoblastoid GM12878 cell line[19] (GEO accession GSE63525). We used the intrachromosomal data at a resolution of 10kb with mapping quality of 30 or more (MAPQGE30), and normalized using the KRnorm vectors. For each of the 59,786 *trans*-eQTLs, we evaluated whether any contact was reported in this dataset. We divided each *trans*-eQTL SNP and any of their proxies (R^2>0.8, 1kG p1v3, EUR, acquired from SNiPA[94]; **URLs**) in 10kb blocks. The *trans*-eQTL genes were also assigned to 10kb blocks, and to multiple blocks if the gene was more than 10kb in length (length between TSS and TES, Ensembl v71). For each individual *trans*-eQTL SNP-gene pair, we then determined if there was any overlap with the Hi-C contact matrices. We repeated this analysis using the true negative set of *trans*-eQTLs described before to generate a background distribution of expected contact.

Data availability

Full summary statistics from eQTLGen meta-analyses are available on the eQTLGen website: www.eqtlgen.org which was built using the MOLGENIS framework[95].

Code availability

Individual cohorts participating in the study followed the analysis plans as specified in the **URL**s or with slight alterations as described in the **Methods** and **Supplementary Note**. All tools and source code, used for genotype harmonization, identification of sample mixups, eQTL mapping, meta-analyses and for calculating polygenic scores are freely available at https://github.com/molgenis/systemsgenetics/.

Acknowledgments

The cohorts participating in this study list the acknowledgments in the cohort-specific supplemental information in Supplementary Note.

We thank i2QTL CONSORTIUM for providing the iPSC replication results.

We thank Kate McIntyre for editing the final text.

This work is supported by a grant from the European Research Council (ERC Starting Grant agreement number 637640 ImmRisk) to LF and a VIDI grant (917.14.374) from the Netherlands Organisation for Scientific Research (NWO) to LF. We thank the UMCG Genomics Coordination Center, MOLGENIS team, the UG Center for Information Technology, and the UMCG research IT program and their sponsors in particular BBMRI-NL for data storage, high performance compute and web hosting infrastructure. BBMRI-NL is a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO) [grant number 184.033.111].

URLs

Full summary statistics from this study, www.eqtlgen.org

ExAC pLI scores, http://exac.broadinstitute.org/downloads;

Ensembl v71 annotation file,

ftp://ftp.ensembl.org/pub/release-71/gtf/homo_sapiens;

Reference for genotype harmonizing,

ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/GIANT.phase1_release_v3.2010

1123.snps_indels_svs.genotypes.refpanel.ALL.vcf.gz.tgz

eQTLGen analysis plan for Illumina array datasets,

https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook;

eQTLGen analysis plan for RNA-seq datasets,

https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook-for-RNA-seq-data;

eQTLGen analysis plan for Affymetrix array datasets,

https://github.com/molgenis/systemsgenetics/wiki/QTL-mapping-analysis-cookbook-for-Affymetrix-expression-arrays;

GenotypeHarmonizer, https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer;

Protocol to resolve sample mixups, https://github.com/molgenis/systemsgenetics/wiki/Resolving-mixups;

Enrichr gene set enrichment libraries,

http://amp.pharm.mssm.edu/Enrichr/;

GeneOverlap package for enrichment analyses,

https://www.bioconductor.org/packages/release/bioc/html/GeneOverlap.html;

SHRiMP aligner used for re-mapping Illumina probes,

http://compbio.cs.toronto.edu/shrimp/;

EBI GWAS Catalogue,

https://www.ebi.ac.uk/gwas/;

Immunobase,

http://www.immunobase.org/;

ClusterProfiler package used for tissue enrichment analyses,

http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html;

Capture Hi-C data,

https://www.chicp.org/

SNiPA, used to acquire proxy SNPs,

http://snipa.helmholtz-muenchen.de/snipa3/

Regulatory Circuits, used to acquire TF data,

www.RegulatoryCircuits.org

References:

1. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
2. Kirsten, H. *et al.* Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci†. *Hum. Mol. Genet.* **24,** 4746–4763 (2015).
3. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* **100,** 228–237 (2017).
4. Jansen, R. *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum. Mol. Genet.* **26,** 1444–1451 (2017).
5. Joehanes, R. *et al.* Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18,** 16 (2017).
6. Brynedal, B. *et al.* Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am. J. Hum. Genet.* **100,** 581–591 (2017).
7. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *Am. J. Hum. Genet.* **100,** 571–580 (2017).
8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
9. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9,** 96 (2017).
10. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention

Setting. *Circulation* **135,** 2091–2101 (2017).

11. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
12. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18,** 86 (2017).
13. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
14. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48,** 481–487 (2016).
15. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).
16. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10,** e1004383 (2014).
17. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. methods* **13,** 366–370 (2016).
18. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. methods* **14,** 61–64 (2017).
19. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
20. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).
21. Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514,** 92–97 (2014).

22. Ding, Y. *et al.* ZNF131 suppresses centrosome fragmentation in glioblastoma stem-like cells through regulation of HAUS5. *Oncotarget* **8,** 48545–48562 (2017).

23. Oh, Y. & Chung, K. C. Small ubiquitin-like modifier (SUMO) modification of zinc finger protein 131 potentiates its negative effect on estrogen signaling. *J. Biol. Chem.* **287,** 17517–17529 (2012).

24. Lemaitre, R. N. *et al.* Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet.* **7,** e1002193 (2011).

25. Deelen, P. *et al*. Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis. bioRxiv preprint (2018).

26. Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7,** e1002216 (2011).

27. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).

28. Gateva, V. *et al.* A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1228–1233 (2009).

29. Yin, X. *et al.* Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. *Nat. Commun.* **6,** 6916 (2015).

30. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *New Engl. J. Med.* **363,** 1211–1221 (2010).

31. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-

eQTLs and co-expression QTLs. *Nat. Genet.* **50,** 493–497 (2018).

32. Wang, J. H. *et al.* Aiolos regulates B cell activation and maturation to effector state. *Immunity* **9,** 543–553 (1998).
33. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).
34. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
35. He, M. *et al.* Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* **24,** 1791–1800 (2015).
36. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267,** 1360 LP-1363 (1995).
37. Chong, J. A. *et al.* REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80,** 949–957 (1995).
38. Cheong, A. *et al.* Downregulated REST transcription factor is a switch enabling critical potassium channel expression and cell proliferation. *Mol. cell* **20,** 45–52 (2005).
39. Dibner, C., Schibler, U. & Albrecht, U. The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu. Rev. Physiol.* **72,** 517–549 (2010).
40. Bass, J. & Lazar, M. A. Circadian time signatures of fitness and disease. *Sci.* **354,** 994–999 (2016).
41. Song, C. *et al.* REV-ERB agonism suppresses osteoclastogenesis and prevents ovariectomy-induced bone loss partially via FABP4 upregulation. *FASEB J. : Off. Publ. Fed. Am. Soc. Exp. Biol.* **32,** 3215–3228 (2018).
42. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic

inheritance. *bioRxiv* (2018).

43. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
44. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).
45. Wang, X. *et al.* Macrophage ABCA1 and ABCG1, but not SR-BI, promote macrophage reverse cholesterol transport in vivo. *J. Clin. Investig.* **117,** 2216–2224 (2007).
46. Goldstein, J. L. & Brown, M. S. Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. *J. Biol. Chem.* **249,** 5153–5162 (1974).
47. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49,** 139–145 (2017).
48. Singh, A. B., Kan, C. F. K., Shende, V., Dong, B. & Liu, J. A novel posttranscriptional mechanism for dietary cholesterol-mediated suppression of liver LDL receptor expression. *J. Lipid Res.* **55,** 1397–1407 (2014).
49. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42,** 441–447 (2010).
50. Kõks, S. & Kõks, G. Activation of GPR15 and its involvement in the biological effects of smoking. *Exp. Biol. Med.* **242,** 1207–1212 (2017).
51. van Iterson, M., van Zwet, E. W., BIOS Consortium & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18,** 19 (2017).
52. Bauer, M., Fink, B., Seyfarth, H.-J., Wirtz, H. & Frille, A. Tobacco-smoking induced GPR15-

expressing T cells in blood do not indicate pulmonary damage. *BMC Pulm. Med.* **17,** 159 (2017).

53. Kõks, G. *et al.* Smoking-Induced Expression of the GPR15 Gene Indicates Its Potential Role in Chronic Inflammatory Pathologies. *Am. J. Pathol.* **185,** 2898–2906 (2015).
54. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
55. Smirnova, T., Miniou, P., Viegas-Pequignot, E. & Mallet, J. Assignment of the human syntaxin 1B gene (STX) to chromosome 16p11.2 by fluorescence in situ hybridization. *Genomics* **36,** 551–553 (1996).
56. Sousa, I. *et al.* Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Mol. Autism* **1,** 7 (2010).
57. Agarwal, A. *et al.* Dysregulated expression of neuregulin-1 by cortical pyramidal neurons disrupts synaptic plasticity. *Cell reports* **8,** 1130–1145 (2014).
58. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).
59. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7,** 11122 (2016).
60. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46,** 543–550 (2014).
61. El-Hattab, A. W. Serine biosynthesis and transport defects. *Mol. Genet. Metab.* **118,** 153–159 (2016).
62. Leuzzi, V., Alessandrì, M. G., Casarano, M., Battini, R. & Cioni, G. Arginine and glycine

stimulate creatine synthesis in creatine transporter 1-deficient lymphoblasts. *Anal. Biochem.* **375,** 153–155 (2008).

63. Hart, C. E. *et al.* Phosphoserine aminotransferase deficiency: a novel disorder of the serine biosynthesis pathway. *Am. J. Hum. Genet.* **80,** 931–937 (2007).
64. Klomp, L. W. *et al.* Molecular characterization of 3-phosphoglycerate dehydrogenase deficiency--a neurometabolic disorder associated with reduced L-serine biosynthesis. *Am. J. Hum. Genet.* **67,** 1389–1399 (2000).
65. Shaheen, R. *et al.* Neu-Laxova syndrome, an inborn error of serine metabolism, is caused by mutations in PHGDH. *Am. J. Hum. Genet.* **94,** 898–904 (2014).
66. McLaughlin, H. M. *et al.* A recurrent loss-of-function alanyl-tRNA synthetase (AARS) mutation in patients with Charcot-Marie-Tooth disease type 2N (CMT2N). *Hum. Mutat.* **33,** 244–253 (2012).
67. Auer-Grumbach, M. Hereditary sensory neuropathy type I. *Orphanet J. rare Dis.* **3,** 7 (2008).
68. Hanada, K. Serine palmitoyltransferase, a key enzyme of sphingolipid metabolism. *Biochim. et Biophys. Acta* **1632,** 16–30 (2003).
69. Glinton, K. E. *et al.* Disturbed phospholipid metabolism in serine biosynthesis defects revealed by metabolomic profiling. *Mol. Genet. Metab.* **123,** 309–316 (2018).
70. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. United States Am.* **100,** 2610–2615 (2003).
71. Bennett, L. *et al.* Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* **197,** 711–723 (2003).
72. Pantel, A. *et al.* Direct type I IFN but not MDA5/TLR3 activation of dendritic cells is required

for maturation and metabolic shift to glycolysis after poly IC stimulation. *PLoS Biol.* **12,** e1001759 (2014).

73. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99,** 1245–1260 (2016).
74. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. notes* **7,** 901 (2014).
75. Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5,** e1000386 (2009).
76. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
77. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinforma.* **27,** 2104–2111 (2011).
78. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids Res.* **45,** D896–D901 (2017).
79. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25,** (2009).
80. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48,** 1031–1036 (2016).
81. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).
82. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids Res.* **46,** D754–D761 (2018).
83. Zaykin, D. V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J. Evol. Biol.* **24,** 1836–1841 (2011).

84. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

85. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14,** 128 (2013).

86. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids Res.* **44,** W90–W97 (2016).

87. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinforma.* **26,** 2438–2444 (2010).

88. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Sci.* **306,** 636–640 (2004).

89. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).

90. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9,** 1366 (2018).

91. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : J. Integr. Biol.* **16,** 284–287 (2012).

92. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

93. Schofield, E. C. *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinforma.* **32,** 2511–2513 (2016).

94. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmuller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31,** (2015).

95. Swertz, M. *et al.* The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinforma.* **11,** 12 (2010).

Untargeted metabolome- and transcriptome-wide association study identifies causal genes modulating metabolite concentrations in urine

Reyhan Sönmez Flitman[1,3*], Bita Khalili[1,3] , Zoltan Kutalik[2,3], Rico Rueedi[1,3], Sven Bergmann[1,3,4*]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

[2]University Center for Primary Care and Public Health, University of Lausanne, Switzerland, Lausanne, Switzerland.

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

[4]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa.

*Correspondence: reyhan.sonmez@unil.ch (R.S.) and sven.bergmann@unil.ch (S.B.)

Summary

In this study we investigate the results of a metabolome- and transcriptome-wide association study to identify genes influencing the human metabolome. We used RNAseq data from lymphoblastoid cell lines (LCLs) derived from 555 Caucasian individuals to characterize their transcriptome. As for the metabolome we took an untargeted approach using binned features from ^{1}H nuclear magnetic resonance spectroscopy (NMR) of urine samples from the same subjects allowing for data-driven discovery of associated compounds (rather than working with a limited set of quantified metabolites).

Using pairwise linear regression we identified 21 study-wide significant associations between metabolome features and gene expression levels. We observed the most significant association between the gene *ALMS1* and two adjacent metabolome features at 2.0325 and 2.0375 ppm. By using our previously developed metabomatching methodology, we found N-Acetylaspartate (NAA) as the potential underlying metabolite whose urine concentration is correlated with *ALMS1* expression. Indeed, a number of metabolome- and genome-wide association studies (mGWAS) had already suggested the locus of this gene to be involved in regulation of N-acetylated compounds, yet were not able to identify unambiguously the exact metabolite, nor to disambiguate between *ALMS1* and *NAT8*, another gene found in the same locus as the mediator gene. The second highest significant association was observed between *HPS1* and two metabolome features at 2.8575 and 2.8725 ppm. Metabomatching of the association profile of *HPS1* with all metabolite features pointed at trimethylamine (TMA) as the most likely underlying metabolite. mGWAS had previously implicated a locus containing *HPS1* to be associated with TMA concentrations in urine but could not disambiguate this association signal from *PYROXD2*, a gene in the same locus. We used Mendelian randomization to show for both *ALMS1* and *HPS1* that their expression is causally linked to the respective metabolite concentrations.

Our study provides evidence that the integration of metabolomics with gene expression data can support mQTL analysis, helping to identify the most likely gene involved in the modulation of the metabolite concentration.

Key words: transcriptomics, untargeted metabolomics, genome wide association studies, *ALMS1*, *NAT8*, *HPS1*, *PYROXD2*, N-acetylated compounds, N-Acetylaspartate, trimethylamine

Introduction

Genome-wide association studies (GWAS) have identified thousands of common variants that are associated with complex traits [1], but the regulatory mechanisms behind these associations mostly remain poorly understood. Pinpointing causal variants is difficult, since the lead variants associated with a trait are often in high linkage disequilibrium (LD) with other variants in the same region with only a slightly lower association signal. Such associated LD blocks typically contain several genes or functional elements, preventing the accurate identification of causal genes. Furthermore, some trait associated variants fall into intergenic regions of the genome with no obvious functional role at all.

A number of studies reported that trait associated genetic variants are significantly enriched in expression quantitative trait loci (eQTLs), suggesting that many trait associated variants affect the phenotype by altering gene expression [2-5]. There is also a growing body of literature highlighting the more pronounced effects of genetic variants on molecular traits compared to phenotypic traits [6-9]. This is not surprising, since molecular traits representing fundamental biological processes such as gene expression and metabolism are intermediates in the genotype to trait causality chain.

With high-throughput measurements becoming more accessible and widespread, integration of molecular traits into association studies has become a central challenge in the field. Such synthesis allows investigating the interplay between different organisational layers of a biological system. Despite metabolism and gene expression regulation both being fundamental biological processes that are commonly studied as molecular phenotypes, there are very few studies in humans that focus on the interplay between them. Several studies investigated the relationship between untargeted serum metabolites and whole blood gene expression in humans [10-12], but, to the best of our knowledge no transcriptome- and metabolome-wide association study has been performed using urine metabolome data of healthy human subjects.

Most metabolome- and genome-wide association studies (mGWAS) reporting metabolite quantitative trait loci (mQTL) use targeted approaches where the concentrations of a limited number of metabolites are estimated from the metabolome data generated by mass spectrometry or NMR spectroscopy. This targeted approach is limited to the number of known quantifiable metabolites in the biofluid under study. In the current study we adopted an untargeted approach, making use of the entire metabolomic data captured by binned ^{1}H NMR spectra as our molecular traits. So here we present an untargeted metabolome- and transcriptome-wide association study using the entire NMR spectral information to characterize the urine metabolomes of 555 subjects and RNAseq data of lymphoblastoid cell lines (LCLs) derived from the same set of individuals. LCL have been widely used in genomic studies and proven their worth as faithful surrogates of primary tissues for studying both gene expression variation among individuals and the genetic architecture underlying regulatory variation of gene expression [13-16]. LCLs thus present an interesting system whose genetic variance in expression resembles that of the cell types affecting the urine metabolome, with the added advantage of not being influenced by immediate environmental factors such as recent changes in the diet or exposure to a drug. Despite having limited statistical power and using surrogate tissue, we identified two strong associations between gene expression levels and urine metabolome features, which allowed us to refine previous links between the corresponding genes and metabolites.

Materials and Methods

Study samples

Our 555 transcriptomics and metabolome profiles were measured in a randomly selected subset of individuals from CoLaus (Cohorte Lausannoise), a population-based cross-sectional study of 6,188 participants residing in Lausanne, Switzerland [17]. Recruitment to the cohort was done on the basis of a simple, non-stratified random selection of the entire Lausanne population aged 35 to 75 in 2003. The 555 samples selected for this study had a mean age of 55 (min=35, max=75) and 53% of them were women.

Metabolomics data

We used two metabolomics data sets; the first dataset was acquired at baseline for 555 subjects and the second dataset was acquired five years later for a subset of 301 subjects. Baseline urinary metabolic profiles were generated using one-dimensional proton nuclear magnetic resonance (NMR) spectroscopy. NMR spectra were acquired at 300 K on a Bruker 16.4 T Avance II 700 MHz NMR spectrometer (Bruker Biospin, Rheinstetten, Germany) using a standard ^{1}H detection pulse sequence with water suppression. The spectra were referenced to the TSP signal and phase and baseline corrected. We binned the spectra into chemical shift increments of 0.005 ppm, obtaining metabolome profiles of 2,200 metabolome features, of which 1,276 remain after filtering for missing values [18]. Lastly, the dataset was log10-transformed and standardised first across features and then samples, to make samples and feature intensities comparable.

The follow-up data was acquired with an Avance III HD 600 NMR spectrometer. These spectra were referenced to the TSP signal and phase and baseline corrected. We binned the chemical shifts into 0.005 ppm bins. After removing water and urea spectral regions (4.55-5.00 ppm and 5.5-6.1 ppm), the dataset was log10-transformed and standardised first across features then samples, to make samples and feature intensities comparable. Lastly, we performed principal component analysis (PCA) to detect outliers and 33 spectra with components scores below/above 3.5 standard deviations from the average of all components scores were removed. Our final metabolic dataset includes 1,289 features.

Gene expression data

Total RNA was extracted from Epstein–Barr-virus-transformed lymphoblastoid cell lines (LCLs) by following the Illumina TruSeq v2 RNA Sample Preparation protocol (Illumina, Inc., San Diego, CA) by the Department of Genetic Medicine and Development at the University of Geneva. Next mRNA sequencing was performed on the Illumina HiSeq2000 platform producing 49bp paired-end reads. Paired-end reads were mapped to human genome assembly GRCh37 (hg19) with GEMTools using GENCODE v15 as gene annotation [19]. The reads were then filtered for correct orientation of the two ends and a minimum quality score of 150 while allowing 5 mismatches at both ends. Gene level read counts were quantified with an in-house script. This resulted in expression profiles of 45,470 genes for 555 individuals, which were quantified as RPKM values. Later, we transformed RPKM values by applying log-transformation [$\log_2$(1+RPKM)] and then standardisation across samples to make genes comparable. For our

analysis we removed all genes on the sex chromosomes, as well as mitochondrial DNA genes from the gene expression data, resulting in 43,614 genes to use in the association analysis.

Genotypic data

Genotyping was performed by using the Affymetrix GeneChip Human Mapping 500 K array set and the imputation was carried out for HapMap II SNPs. Further details of genotype calling and the imputation can be found in [18].

Association analysis

All statistical analyses were performed using Matlab [20]. Urine metabolome features were rank normalized in order to have comparable intensities before they were used as response variables in regression.

We used a linear regression model for each pair of metabolome feature (as the response variable) and gene expression level (as the explanatory variable). The model also included the following common confounding factors: age, sex, the first four principal components of the genotypic data (correcting for population stratification) and the first 10 principal components of the gene expression data (correcting for potential batch effects). We tested 1,276 metabolome features for association with the expression of 19,123 protein coding and 24,491 non-coding genes. For the completeness of the analysis we did not apply any a-priori exclusion criteria to remove genes from the analysis. As a consequence, the distribution of genes RPKM values with significant associations were evaluated to ensure close to normality distribution for accurate regression estimations. We applied a nominal Bonferroni threshold for multiple testing p_{max} = 0.05/(125×1,109) = 3.6×10^{-7} by taking into account the effective number of tests which we estimated to be 125 for metabolome features and 1,109 for genes (i.e. the number of principal components explaining more than 95% of the data [21]). Only associations with p-value below p_{max} were considered significant.

Metabomatching

Metabomatching is a method to identify metabolites underlying associations of SNPs with metabolome features [18, 22]. It compares the association profile of a given variable with all metabolome features across the full ppm range, so-called pseudospectrum, with NMR spectra of pure metabolites available in public databases such as HMDB [23]. For each metabolite m, metabomatching defines a set of features $F_\delta(m)$ that contains all the features f that fall within a δ ppm vicinity of any NMR spectrum peak of m listed in the database. Metabomatching then computes the sum

$$(1)\quad s(F_\delta(m)) = \sum_{f\in F_\delta(m)} \left(\frac{\hat{\beta}_f}{\widehat{SE}_f}\right)^2,$$

where $\hat{\beta}_f$ and $\widehat{SE}_f$ are the point estimates of feature f effect size and its standard error. Assuming a χ^2-distribution for the sum with $|F_\delta(m)|$ degrees of freedom, metabomatching defines a score for each m as the negative logarithm of the nominal p-value corresponding to the observed sum. These scores are calculated for all the metabolites with ^{1}H NMR spectrum in the

database, allowing to rank them based on their likelihood to underlie the association of the variable with the metabolome features.

Although metabomatching was originally developed to use SNP-metabolome associations, recently it has been shown that it can also use co-varying features of metabolome data itself to identify metabolites [24]. In the present study we use metabomatching to identify metabolites that are associated with gene expression.

Mendelian randomization

We performed Mendelian randomization (MR) analysis [25, 26] to assess the causal relationship between gene expression and metabolite concentration. While we used SNPs as instrumental variables (IVs), gene expression and metabolome features were interchangeably used as exposure and outcome to determine the direction of causality. For the MR analysis, we used summary statistics from mQTL/eQTL studies with higher statistical power [27, 28]. Causal effects were estimated by using the Wald method where the effect of a genetic variant on the outcome is divided by the effect of the same genetic variant on the exposure [29]. Next, ratio estimates from different instruments (SNPs) were combined by the inverse variance weighted method (IVW) to calculate the causal estimate [30].

We selected significant SNPs from relevant eQTL/mQTL studies as our IVs. To detect the independent SNPs, we used a stepwise pruning approach where we first selected the strongest lead eQTL/mQTL and then pruned the rest of the SNPs in a stepwise manner if they were correlated with the lead SNP ($r^2 > 0.2$). We repeated the pruning process with the next available SNP until there were no SNPs left to prune. We used Cochran's Q test to determine heterogeneity among the candidate instruments [31]. The SNPs were pruned in a stepwise manner from the model until the model did not show any more signs of heterogeneity (Cochran's Q statistic p-value > 0.05/#of original instruments). We also applied more robust MR analysis methods than IVW, such as the median estimator and MR-Egger regression to evaluate the significance of the causal estimates [32]. These methods are known to have more relaxed MR assumptions and they can tolerate the violation of the exclusion-restriction assumption for some instruments. For all MR analysis we used the Mendelian Randomization package implemented in R [33].

Analysis & Results

Association analysis

We performed an untargeted metabolome- and transcriptome-wide association study by pairwise linear regression of log-transformed expression levels of each of the 43,614 genes (as response variable) onto each of the 1,276 metabolome features (as explanatory variable). The metabolome features resulted from binning the raw urinary NMR spectra with a bin-size of 0.005 ppm, and rank normalizing each bin passing QC (see Methods). The gene expression levels, quantified as RPKM, were measured using RNAseq on lymphoblastoid cell lines derived from the same set of 555 subjects.

Figure 1 shows the qq-plot of all pairwise associations. It is well calibrated, and only two association p-values (both involving the *ALMS1* gene, see below) are highly significant (FDR<0.05). Yet, applying an adjusted Bonferroni threshold of 3.6×10^{-7} to account for the effective number of independent variables (see Methods) we identified 25 additional marginally significant feature-gene associations. The 27 association pairs involved 22 unique genes and 25 unique features. We did not apply any a-priori exclusion criteria to remove genes from the analysis. Instead, we inspected the expression value distributions of these 22 significant genes in order to identify cases in which the small p-value may be due to a problematic distribution of the expression values. Indeed, we observed that some of the genes had zero expression values for a sizable fraction of the samples and very low expression values otherwise. Based on the distributions we filtered out all genes that had more than 95% RPKM values equal to 0 and a maximum RPKM value over all samples lower than 1. Applying this rather mild filtering removed 11,547 out of the 43,614 autosomal genes (26%) and 1,994 out of 19,123 protein-coding genes (10%). Amongst the 22 marginally significant associations five (23%) were removed. Expression distributions of the discarded as well as remaining genes are presented in Supplementary Figure 1 and 2, respectively. We report the remaining 21 significant associations corresponding to 17 unique genes and 19 unique features in Table 1.

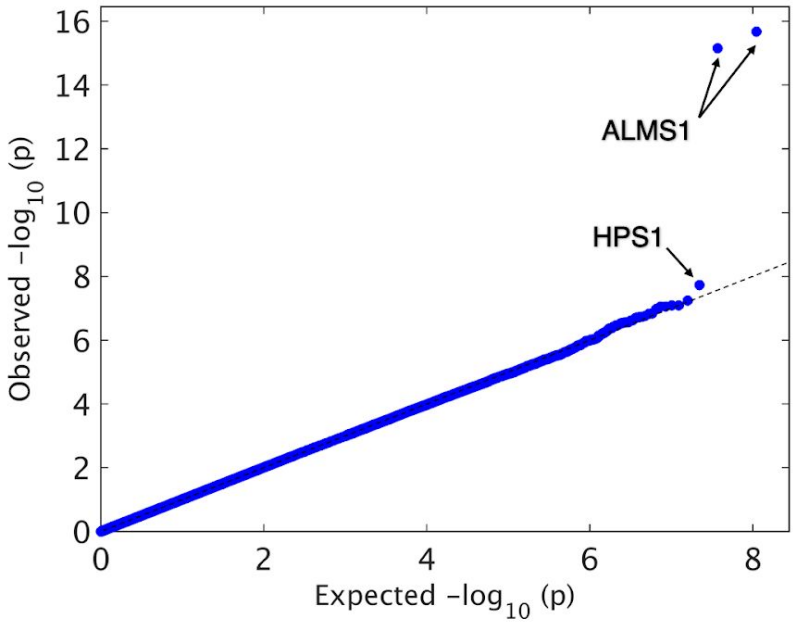


Figure1: QQ-plot showing -log10(p)-values of metabolome- and transcriptome-wide association analysis. The features that significantly associate with *ALMS1* expression are ranking as 1st, 2nd and 8th; the features associated with *HPS1* expression are ranking as 3rd and 5th and the features associated with *ALMS1P* expression are ranking as 6th and 7th.

Genes			Metabolite	Association		Published as mGWAS
Gene ID	Chr	Gene symbol	Feature(s)	*X*	*P*	Body fluid
ENSG00000116127	2	ALMS1	2.0375, 2.0325, 2.0275	0.51, 0.50, 0.32	2.1×10^{-16}, 7.0×10^{-16}, 2.7×10^{-7}	Serum [34,35] Urine[27,36]
ENSG00000107521	10	HPS1	2.8725, 2.8575	-0.31, -0.29	1.9×10^{-8}, 8.0×10^{-8}	Serum [34,35] Urine[27]
ENSG00000256029	1	RP11-190A12.7	3.0925	0.24	5.6×10^{-8}	Serum [35]
ENSG00000163016	2	ALMS1P	2.0325, 2.0375	0.24, 0.23	8.1×10^{-8}, 1.5×10^{-7}	Serum [34,35] Urine[27]
ENSG00000219355	12	RPL31P52	2.8675	-0.23	1.0×10^{-7}	
ENSG00000149089	11	APIP	2.7925	-0.27	1.5×10^{-7}	Serum [34,35] Urine[27]
ENSG00000102786	13	INTS6	4.3775	-0.34	1.9×10^{-7}	Serum [34,35]
ENSG00000115295	2	CLIP4	3.9525	-0.31	1.9×10^{-7}	Serum [34,35]
ENSG00000173702	3	MUC13	1.6425	0.24	2.1×10^{-7}	Serum [35]
ENSG00000228360	7	RP11-365F18.1	1.3675	0.22	2.4×10^{-7}	
ENSG00000267273	19	CTC-543D15.3	1.2125	0.22	2.4×10^{-7}	
ENSG00000184276	11	DEFB108B	8.1525	0.23	2.7×10^{-7}	
ENSG00000175895	8	PLEKHF2	2.2975	0.31	2.8×10^{-7}	Serum [34]
ENSG00000267267	17	CTD-3199J23.4	2.6475	-0.22	2.9×10^{-7}	
ENSG00000213650	7	RP11-760D2.7	6.7025	0.23	3.0×10^{-7}	
ENSG00000140297	15	GCNT3	3.1075	0.27	3.3×10^{-7}	Serum [34,35]
ENSG00000211611	2	IGKV6-21	3.2175	-0.22	3.5×10^{-7}	

Table 1: 21 study-wide significant associations from metabolome- and transcriptome-wide association analysis, corresponding to 17 unique genes and 19 unique features. Abbreviations: GeneID - Ensembl Gene ID (NCBI build 37), Chr - chromosome, *X* - effect size, *P* - P-value.

Metabolite discovery

To find the metabolites underlying these significant associations between gene expression levels and metabolome features we used metabomatching. Metabomatching has been previously established as an effective tool for prioritizing candidate metabolites underlying SNP-metabolome features association profiles, so-called pseudospectra [18, 27]. In this study we used association profiles of genes which had at least one significantly associated metabolite feature as input to metabomatching and found that the pseudospectra of *ALMS1* and *ALMS1P* matched well with the N-Acetylaspartate (NAA) NMR spectrum and that the pseudospectrum of *HPS1* matched well with the trimethylamine (TMA) NMR spectrum (Figure 2).

As shown in Table 1, the expression of *ALMS1* significantly associates with three neighboring features at 2.0375 ppm (p-value= 2×10^{-16}), 2.0325 ppm (p-value=7×10^{-16}) and 2.0275 ppm (p-value=3×10^{-7}). There are few metabolites with resonances in this region and usually a singlet signal in this area is interpreted as the N-acetylated resonance detected in the ^{1}H NMR spectrum of N-acetylated compounds [37]. As illustrated in Figure 2A, among the top three metabolites suggested by metabomatching that have a peak at 2.03 ppm, the only one with the highest intensity peak at this position is NAA. Also the presence of a secondary peak in the pseudospectrum at 7.9225 ppm matches well with one of the lower intensity peaks of the NMR spectrum of NAA reported at 7.92 ppm in HMDB, even though the association p-value of this feature is below the Bonferroni threshold (p-value=2×10^{-4}). Similarly, metabomatching the pseudospectrum of *ALMS1P* (*ALMS1* pseudogene) points to NAA as the most likely matching N-acetylated compound (Supplementary Figure 3). The metabolome features pointing to NAA are the same features as in *ALMS1* but with lower association p-values (2.0375 ppm with p-value=1×10^{-7}, 2.0325 ppm with p-value= 8×10^{-8}, 7.9225 ppm with p-value= 2×10^{-4}).

The reference spectrum of NAA in the Urinary Metabolome Database (UMDB) that we used for metabomatching was recorded in water. In order to verify that the peaks of this spectrum are comparable to those of NAA in urine, we spiked NAA into pooled urine samples from our collection at a concentration of 10 mM and recorded its ^{1}H NMR spectrum. Inspecting the 5 multiplet regions of NAA, we concluded that the NAA peak positions are very similar in both solvents (Supplementary Figure 4). To further investigate if a better match exists among all the N-acetylated family of compounds, we built a library consisting of all N-acetylated compounds proton NMR spectra available in HMDB and the Biological Magnetic Resonance Data Bank (BMRB). NAA remained the best metabomatching hit for the *ALMS1* pseudospectrum (Supplementary Figure 5). Figure 3 illustrates the relationship between *ALMS1* gene expression level and the NAA metabolite concentration where every point in the plot represents a study sample and each of the samples are color coded according to the genotype at rs7566315 SNP, that is an eQTL of *ALMS1* and mQTL of NAA.

The third and fifth strongest associations in Table 1 are between *HPS1* gene expression and two neighboring metabolome features at 2.8725 ppm (p-value=2×10^{-8}) and 2.8575 ppm (p-value= 8×10^{-8}), respectively. Figure 2B shows the metabomatching result of the *HPS1* pseudospectrum. Among the top three metabolites suggested by metabomatching, trimethylamine (TMA) is the most plausible metabolite driving the association pattern, as it is the only metabolite with its highest intensity NMR peak at 2.86 ppm region and no missing peaks. Schematic representation of the match between pseudospectra and the NMR spectra for both *ALMS1* and *HPS1* can be seen in Supplementary Figure 6.

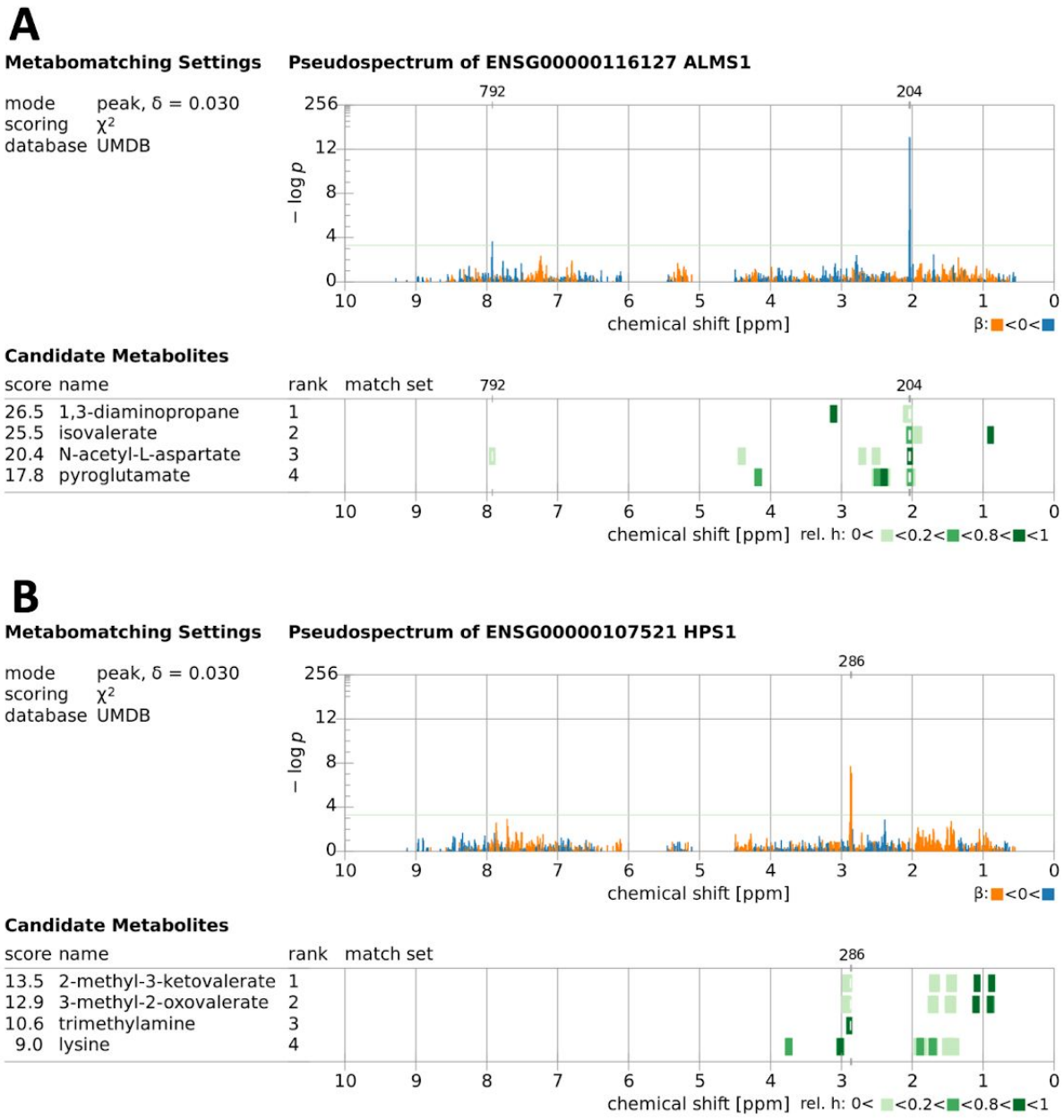


Figure 2: Metabomatching figures showing the pseudospectra derived from gene expression - metabolome features associations [22]. The features in each pseudospectrum are color-coded by the sign of the effect size and the four highest ranking candidate metabolites are listed on the lower left with their reference NMR spectra shown on the right (color coding indicating their relative peak intensities). A) CoLaus urine metabolome-*ALMS1* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.03 ppm and 7.92 ppm regions which match well with the highest intensity peak of NAA and one of the lower intensity peaks of the NAA NMR spectrum, respectively. B) CoLaus urine metabolome - *HPS1* gene expression association profile metabomatching figure. Leading features allowing metabolite identification are at 2.87 and 2.86 ppm which match well with TMA singlet.

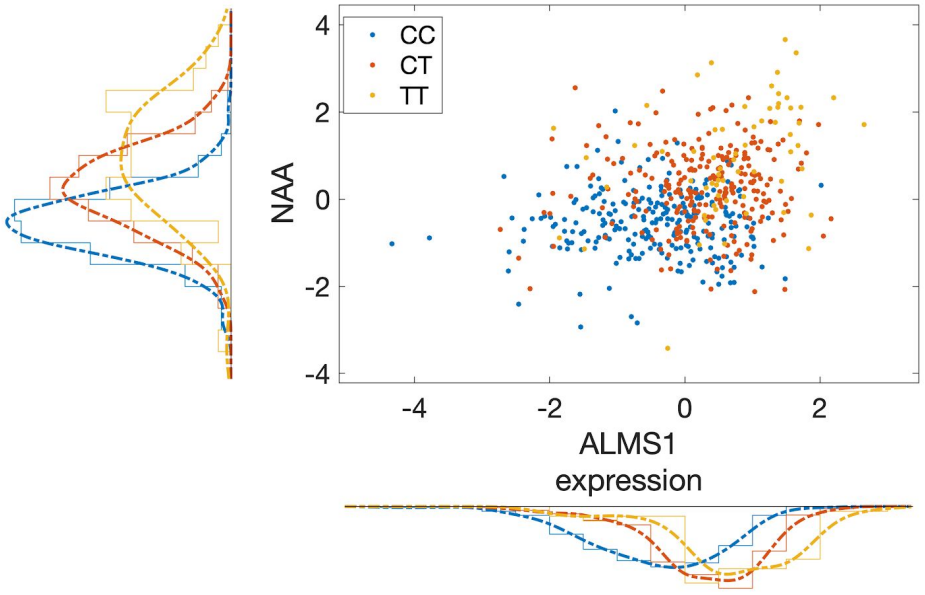


Figure 3: SNP rs7566315, showing a mQTL effect on NAA and an eQTL effect on *ALMS1* gene expression. Each point represents a study sample. NAA concentration is approximated by the feature at 2.0375 ppm that is log10 transformed after feature- and sample-wise z-scoring (y-axis). *ALMS1* expression is quantified as log2 transformed RPKM+1 values (x-axis). Color code represents the genotype of rs7566315 (legend).

Validation of *ALMS1* and *HPS1* associations

To the best of our knowledge, there is no other study with urine NMR spectra and expression data of LCLs derived from the same subjects that is of comparable or larger sample size, precluding proper out-of-sample replication of our results. We have, however, access to additional urine NMR spectra from samples collected for a subset of 301 CoLaus subjects in a follow-up study conducted five years after the baseline data collection. We note that the follow-up NMR data are not independent from the baseline data, yet they were obtained from physically different samples collected at a significantly later time and processed in a different NMR spectrometer and facility. As for the expression data, we only have those from LCLs derived from blood taken at baseline, so we could only test whether the associations we observed between baseline metabolomics and baseline transcriptomics measurements would persist as associations between follow-up metabolomics and baseline transcriptomics data.

We thus asked whether our significant and marginally significant results can be confirmed also using the follow up metabolomics data. We focused on the *ALMS1* and *ALMS1P* gene expression association with NAA and the *HPS1* gene expression association with TMA. As baseline and follow-up urine NMR data were each processed and binned individually, the features did not correspond one-to-one between the studies. To test the association of these three genes with relevant features, we selected all features within +/- 0.03 ppm neighborhood of top features associated with these genes from baseline dataset; i.e. 2.0375 ppm for *ALMS1* and *ALMS1P*, and 2.8575 ppm for *HPS1*. This resulted in 12 features to test for each of the genes. We used a Bonferroni multiple testing corrected p-value threshold of 0.05/(12 features x 3 genes) = 1.4 $\times 10^{-3}$.

In the follow-up, *ALMS1* gene expression level significantly associated with three neighboring features at 2.042 ppm (p-value=5.1×10^{-7}), 2.037 ppm (p-value=3.7×10^{-6}) and 2.032 ppm (p-value=3.9×10^{-4}), likely corresponding to the features at 2.0375 and 2.0325 ppm in the baseline association study. *HPS1* gene expression level significantly associated with 2 features at 2.869 ppm (p-value=2.2×10^{-5}) and 2.859 ppm (p-value=1.3×10^{-3}) that likely correspond to the features at 2.8725 and 2.8575 ppm in the baseline dataset. *ALMS1P* however did not show any significant association with candidate features in the follow-up study. Supplementary Table 1 summarises our validation results.

Comparison with mGWAS results

We performed an association study with metabolome features in the NAA and TMA NMR peak regions using data from 826 individuals of the CoLaus cohort for whom the urinary NMR spectra are available (similar to [18]). Figure 4A shows the locuszoom figure of SNPs in loci surrounding *ALMS1*/*NAT8* locus with significant association p-values with metabolome feature at 2.0375 ppm. The SNPs most strongly associated with this metabolome feature are correlated with each other and lie within a locus containing *ALMS1*, *ALMS1-IT1*, *NAT8* and *ALMS1P* genes (r^2>0.8). In Figure 4B, we show the p-values for association of expression values from nine genes with five different metabolome features that represent all multiplet regions of NAA (see Supplementary Figure 4 for a wider range of genes in the locus). *ALMS1* and *ALMS1P* have the most significant association results with the 2.0375 ppm feature, compared to the rest of the genes. Concordantly, *ALMS1* and *ALMS1P* gene expression levels are associated more significantly to the feature at 7.9225 ppm, the secondary feature in our NAA identification, compared to the other genes at the locus. Figure 5A shows the significant association pattern of SNPs in the loci surrounding *HPS1*/*PYROXD2* locus with metabolome feature at 2.8725 ppm and Figure 5B shows the significance level for association of expression values from seven genes with the same metabolome feature. Even though the SNPs with the most significant association with feature 2.8725 are physically located closer to *PYROXD2* gene rather than *HPS1* gene, the expression level of *PYROXD2* does not show significant association with this feature. Inspecting the list of published mGWAS in humans [38], we found that the SNPs in both *ALMS1* and *HPS1* loci have been previously reported to associate with a number of metabolic traits (Tables 2). The *ALMS1* locus has previously been associated with a number of N-acetylated compounds, while *HPS1* locus has been associated with various metabolites including trimethylamine and dimethylamine [18, 27, 36]. In mGWAS studies determining the mediator genes is not a straightforward procedure, as mQTL SNPs are indistinguishable from neighboring SNPs in LD, and mediator genes of the mQTLs are often inferred based on their physical proximity to the SNPs or functional relevance. Consequently, published mGWAS studies were not able to distinguish between *NAT8* and *ALMS1* or *HPS1* and *PYROXD2* as mediator genes of NAA and TMA, respectively. In contrast, in the current association study we use gene expression data allowing us to pinpoint *ALMS1* and *HPS1* as mediator genes.

Reference	Platform	Biofluid	Locus	Reported mGWAS results
Nicholson *et al.* 2011	MS + NMR	Urine + Plasma	ALMS1, NAT8	N-acetylated compounds
Montoliu et al. 2013	NMR	Urine	ALMS1	N-acetylated compounds
Rueedi *et al.* 2014	NMR	Urine	ALMS1	2.0375 (suggested as N-acetylated compounds)
Raffler et al. 2015	NMR	Urine	NAT8	2.031 (metabomatching: N-acetyl L-aspartate)
Suhre et al.2011	MS	Serum	NAT8	N-acetylornithine
Yu *et al.* 2014	MS	Serum	NAT8	N-acetylornithine
Shin *et al.* 2014	MS	Serum	NAT8	N-acetyllysine, Unknown compounds
Nicholson *et al.* 2011	MS + NMR	Urine + Plasma	HPS1, PYROXD2	Trimethylamine (urine), Dimethylamine (plasma)
Rueedi *et al.* 2014	NMR	Urine	PYROXD2	Trimethylamine, unknown compound, 1.8025
Raffler et al. 2015	NMR	Urine	PYROXD2	2.854 (metabomatching: trimethylamine)
Raffler et al. 2013	NMR	Plasma	PYROXD2	2.757
Rhee *et al.* 2013	MS	Plasma	HPS1	Asymmetric dimethylarginine
Krumsiek *et al.* 2012	MS	Serum	HPS1, PYROXD2	Multiple compunds, Unknown compounds
Hong *et al.* 2013	MS	Serum	HPS1	Caprolactam
Shin *et al.* 2014	MS	Serum	PYROXD2	Unknown compounds

Table 2: List of published mGWAS results in humans concerning *ALMS1*/*NAT8* and *HPS1*/*PYROXD2* loci. MS:Mass Spectrometry, numbers in reported mGWAS results section refer to NMR spectral shift positions in ppm.

To further evaluate the possible regulation of NAA and TMA by other genes suggested by published mGWAS studies, we investigated the metabomatching plots of these genes in order to see if they pointed to any N-acetylated compounds/TMA. The investigated genes either (a) were the target of an eQTL SNP that is mQTL of NAA/TMA, or (b) were within 500kb of *ALMS1*/*HPS1*. However none of these candidate genes produced a pseudospectrum containing even a single nominally significant signal pointing to NAA/TMA.

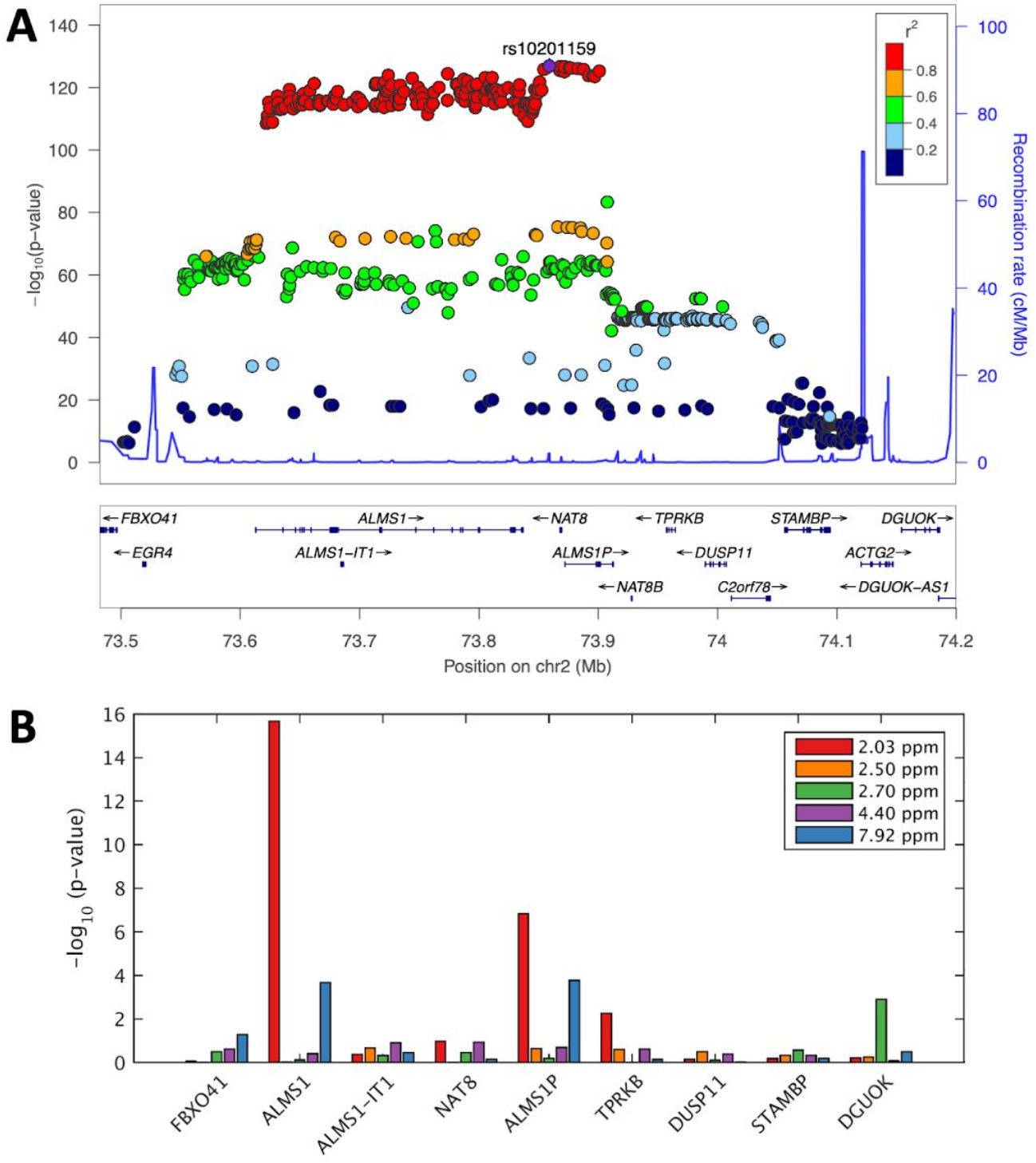


Figure 4: A) LocusZoom plot for *ALMS1/NAT8* locus, where the SNPs are associated with metabolome feature at 2.0375 ppm, LD colored with respect to lead mQTL. B) Bar plot shows -$\log_{10}$ transformed p-values from associating expression values of nine genes in the locus with the five NAA features.

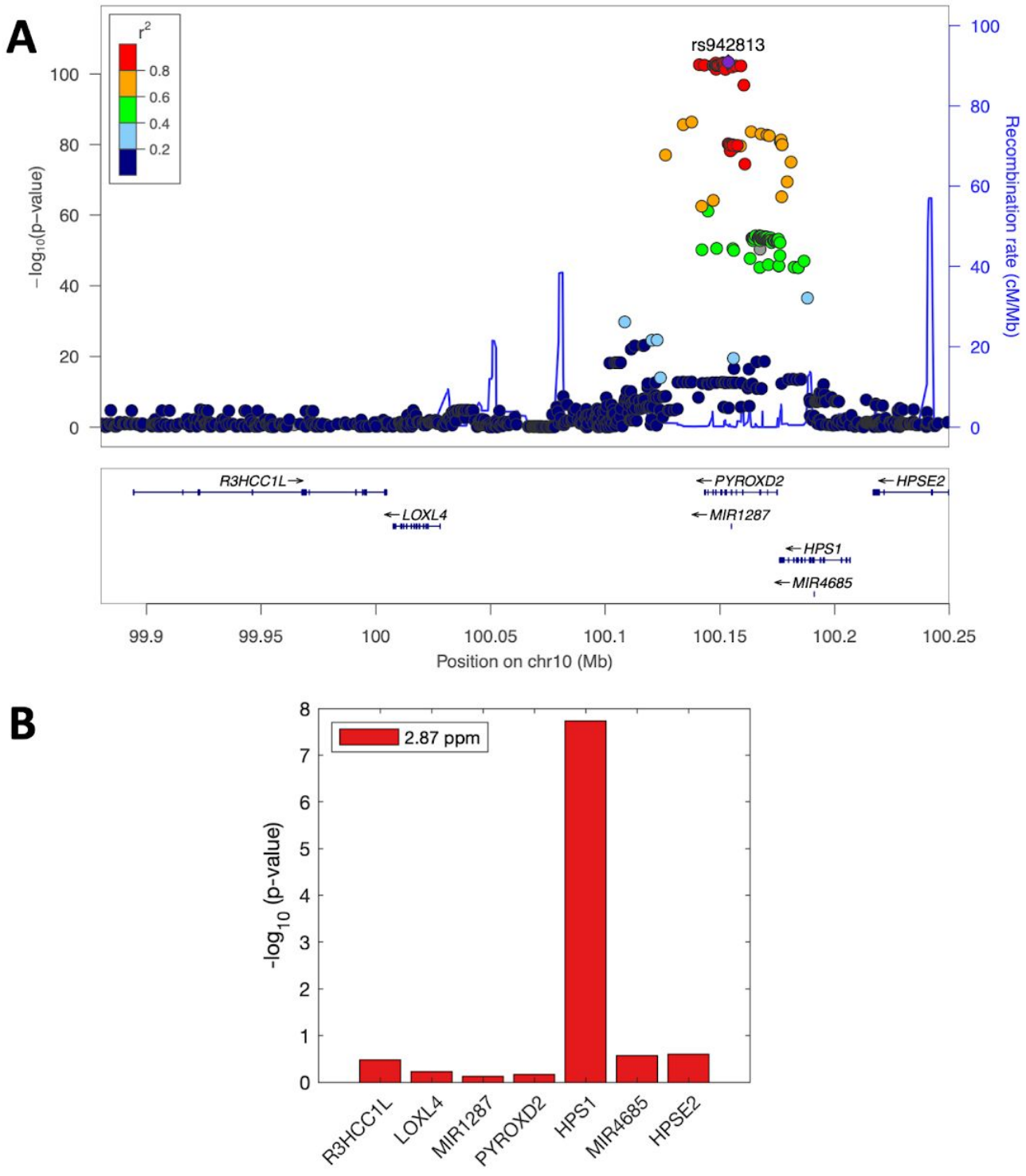


Figure 5: A) LocusZoom plot for *HPS1/PYROXD2* locus, where the SNPs are associated with metabolome feature at 2.8725 ppm, LD colored with respect to lead metaboliteQTL. B) Bar plot shows -$\log_{10}$ transformed p-values from associating expression values of seven genes in the locus with the TMA feature at 2.87 ppm.

Causality analysis

We performed MR analysis using summary statistics from the eQTLGen Consortium [28] and Raffler et al. [27] for eQTL and untargeted mQTL results, respectively. We investigated both the causal effect of the gene expression on the metabolite concentration and vice versa for the *ALMS1*-NAA and *HPS1*-TMA gene-metabolite pairs.

In the MR analysis where we investigated the causal effect of *ALMS1* gene on NAA concentration, instrumental variables (IVs) were selected among the SNPs that were reported as significant eQTLs (FDR<0.05) in eQTLGen and that were also measured in Raffler et al., resulting in 86 SNPs. By applying the stepwise pruning approach (see Methods) we found 14 independent SNPs as candidate IVs. Next, we performed Cochran's Q test to detect heterogeneity among these 14 SNPs and removed a further three of those, resulting in 11 SNPs as potentially valid IVs to use in the MR analysis (see Methods). As for the outcome, we used NMR peak intensities as proxies for the concentration of NAA as there were no targeted studies reporting summary statistics explicitly for NAA concentration. To this end we used the peak at 2.0308 ppm reported in Raffler et al., as this peak is the highest peak in the NAA spectrum and often used to estimate the concentration of N-acetylated compounds (NAC) [36, 39]. NAA has other NMR peaks in its spectrum, yet the observed intensities at these peaks are much lower and therefore difficult to detect robustly by NMR spectroscopy. Indeed these peaks were only weakly correlated amongst themselves and with the main peak at 2.03 ppm region (pearson rho<0.5), so they were too noisy to define a more robust estimate of the NAA concentration than the main peak on its own. For these reasons we decided to perform our MR analysis using only the intensity measure at 2.03 ppm as outcome, which implies therefore that we studied the causality of any NAC rather than NAA specifically. Causal effect estimates given by different meta-analysis methods are reported in Table 3. All methods agreed on *ALMS1* expression level being causal for NAC concentrations.

For the completeness of the analysis, we also tested the causal effect of NAC on *ALMS1* gene expression level. IVs were selected among the SNPs that were reported as significant mQTLs (p-value < 1×10^{-6}) in Raffler et al. [27]. Amongst the cis-eQTLs of *ALMS1* from eQTLGen, most candidate IVs seemed to have direct pleiotropic effect on *ALMS1* expression in cis, reflected by the strong heterogeneity between their expected and observed effects. To overcome this problem we sought to use also trans-eQTLs of *ALMS1*, however none of the candidate IVs were measured in the trans-eQTL study of eQTLGen. As an alternative, we performed an association study between the candidate IVs and *ALMS1* gene expression level as measured in CoLaus and used these eQTL results in the MR analysis. Overall, we identified 26 significant mQTLs for the 2.03 ppm feature in Raffler et al. (p-value < 1×10^{-6}) which corresponded to six independent SNPs. Two of the six candidate IVs exhibited pleiotropic effects and they were removed from the analysis. Finally, we had four SNPs as potentially valid IVs to use in the MR analysis (see Methods). Causal effect estimates given by different meta-analysis methods are reported in Table 3. None of the methods found NAC concentration to be causal for *ALMS1* gene expression level. However, it should be noted that due to low sample size of trans-eQTL study, this particular MR analysis was underpowered.

	Method	Causal Effect Size Estimate	Std. Error	95% CI	P-value	Cochran's Q-statistic p-value
ALMS1 -> NAC	Inverse variance weighted	0.967	0.061	0.847 - 1.087	$< 2\times10^{-16}$	0.2323
	Weighted median	1.111	0.075	0.965 - 1.257	$< 2\times10^{-16}$	NA
	MR - Egger	0.994	0.092	0.812 - 1.175	$< 2\times10^{-16}$	0.1776
	Maximum-likelihood	0.999	0.065	0.872 - 1.126	$< 2\times10^{-16}$	0.249
NAC -> ALMS1	Inverse variance weighted	-0.015	0.264	-0.532 - 0.502	0.955	0.7443
	Weighted median	0.122	0.321	-0.507 - 0.751	0.704	NA
	MR - Egger	1.495	1.976	-2.377 - 5.368	0.449	0.7256
	Maximum-likelihood	-0.015	0.266	-0.535 - 0.505	0.955	0.7443

Table 3: MR results for testing causal effect of *ALMS1* gene expression levels on N-acetylated compounds (*ALMS1* -> NAC) and MR results for testing causal effect of N-acetylated compounds on *ALMS1* gene expression levels (NAC -> *ALMS1*) using summary statistics data.

For the MR analysis of the *HPS1* gene, IVs were selected among the SNPs that were reported as significant eQTLs (FDR<0.05) in eQTLGen and that were also measured in Raffler et al. [27]. As for the outcome, similarly to NAA, there were no studies reporting targeted summary statistics for TMA concentration, therefore we used the NMR peak intensities to estimate the concentration of TMA. According to HMDB, TMA has one singlet at 2.89 ppm where the peak position ranges from 2.79 to 2.99 ppm. In the Raffler et al. dataset we used the intensity of feature at 2.8541 ppm as a proxy of TMA concentration. For the MR analysis we had 77 candidate SNPs six of which were selected as valid IVs as they were independent and did not exhibit heterogeneity (see Methods). Causal effects estimated by using different meta-analysis methods are reported in Table 4. All of the methods agreed on *HPS1* gene expression having a causal effect on TMA concentration.

We also explored the causal effect in the other direction, testing the causal effect of TMA concentration on *HPS1* gene expression. There were 87 significant mQTLs in Raffler et al. [27] that were also measured in eQTLGen. By applying the stepwise pruning approach and removing the SNPs showing heterogeneity (see Methods) we had 18 SNPs to use as IVs in the MR analysis. Causal effects estimated by using different meta-analysis methods are reported in Table 4. All of the methods agreed on TMA concentration being causal on *HPS1* expression. To sum up, the estimated causal effect size of *HPS1* on TMA ranged from 0.27 to 0.37 depending on the method, while the causal effect size of TMA on *HPS1* was around -0.09, pointing to the existence of a negative feedback loop.

	Method	Causal Effect Size Estimate	Std. Error	95% CI	P-value	Cochran's Q-statistic p-value
HPS1 -> TMA	Inverse variance weighted	0.266	0.094	0.082 - 0.450	0.005	0.0803
	Weighted median	0.311	0.072	0.170 - 0.453	$< 2\times10^{-16}$	NA
	MR - Egger	0.37	0.126	0.123 - 0.617	0.003	0.0852
	Maximum-likelihood	0.267	0.094	0.083 - 0.452	0.004	0.0829
TMA -> HPS1	Inverse variance weighted	-0.089	0.012	-0.113 - -0.065	< 2×10-16	0.0958
	Weighted median	-0.09	0.011	-0.111 - -0.068	$< 2\times10^{-16}$	NA
	MR - Egger	-0.086	0.013	-0.111 - -0.061	$< 2\times10^{-16}$	0.0758
	Maximum-likelihood	-0.09	0.012	-0.114 - -0.066	$< 2\times10^{-16}$	0.1258

Table 4: MR results for testing causal effect of *HPS1* gene expression level on TMA (*HPS1* -> TMA) and MR results for testing causal effect of TMA on *HPS1* gene expression level (TMA -> *HPS1*) using summary statistics data.

Discussions & Conclusion

In this study, we present a metabolome- and transcriptome-wide association study using matching RNA-seq and NMR urine profiles from 555 subjects of the CoLaus cohort. This is the first time such a study is performed on untargeted urine metabolome of healthy individuals. In contrast to targeted approaches that are restricted to a limited set of urine metabolites, our association study uses the binned features of the entire ^{1}H NMR spectra as metabolic traits. We identified one gene (*ALMS1*) whose association with two adjacent NMR features around 2.03 ppm is highly significant, surviving even the most conservative correction for multiple hypotheses testing. 16 additional genes are associated with metabolic features with marginal significance with p-values below an adjusted threshold accounting for the estimated number of independent variables (see Table 1). Among the top 17 genes, 10 are in loci with SNPs that have been previously reported as mQTLs. This shows the sensitivity of our study to extract likely candidates of metabolically relevant genes, despite its small sample size and low power.

We used metabomatching to search for promising metabolite candidates underlying gene expression-metabolome features associations. This approach was particularly insightful for our top hit *ALMS1*, as well as the strongest marginally significant association involving *HPS1*: Both genes had previously been implicated by mGWAS linking their loci to compound families. However, in both cases the reported mQTL also harbored other genes, leaving the exact gene-metabolite association ambiguous.

Specifically, the locus associated through mGWAS with N-acetylated compounds includes both *ALMS1* and the *NAT8* gene [18, 27, 36, 39], and the latter seemed to be the more likely candidate due to its known N-acetyltransferase activity. Yet, our association study using transcriptomics data only implicates *ALMS1* and not *NAT8*. Thus, while we cannot rule out a functional role of *NAT8*, the mQTLs of this locus likely act, at least predominantly, as eQTLs through *ALMS1*, pointing to its regulatory role in modulating the compound concentration. This metabolic role of *ALMS1* is also supported through its known role in Alström syndrome characterised by metabolic deficits (PMC6327082) and kidney health disorder phenotypes [40]. Interestingly, in the mGWAS reported by Montoliu et al. using data from a Brazilian cohort, the authors observed the association between N-acetylated compounds and the SNPs located in *ALMS1*/*NAT8* locus with stronger SNP associations in the *ALMS1* gene rather than *NAT8* [39] . They argued that the high ethnic diversity of their study population might have been responsible for breaking down the linkage disequilibrium in the *ALMS1*/*NAT8* region of the genome, resulting in a stronger association for SNPs close or in the *ALMS1* gene compared to other studies.

Our study also sheds more light onto the involved compound: Applying metabomatching on the pseudospectrum from association of all NMR features with the *ALMS1* expression level using a database composed of all N-acetylated compounds NMR spectra, suggested NAA as the best matching metabolite due to the presence of a secondary peak at 7.92 ppm and not missing any high intensity peaks unlike other N-acetylated compounds (Supplementary Figure 5). Interestingly, NAA is the second most abundant metabolite in the brain and involved in neural signalling by serving as a source of acetate for lipid and myelin synthesis in oligodendrocytes [41]. NAA can be detected in urine of both healthy and unhealthy individuals in low concentrations [42] and it has a long history of being a surrogate marker of neural health and a broad measure of cognitive performance [43, 44]. Recently it has been shown that NAA correlates with time measures of neuropsychological performance [45]. The signals of SNPs in

ALMS1 by GWAS with intellectual phenotypes such as self-reported ability in mathematics [46, 47] might therefore be due to its role in modulating NAA. This conjecture of course assumes that NAA levels in relevant brain tissues reflect those in urine and that the *ALMS1* expression variation, and in particular its genetic component, in LCLs or blood, can serve as a proxy for brain tissue. As for *HPS1*, our second strongest association of a gene expression level with urine NMR features, we note that mGWAS previously associated its locus with TMA levels [18, 27, 36]. Yet, most of these studies, including the aforementioned GWAS using a Brazilian cohort [39] considered the *PYROXD2* gene, which is in the same locus, as the most likely modulator of TMA concentrations due to its known function as pyridine nucleotide-disulfide oxidoreductase. While we cannot rule out that this gene is indeed involved in TMA metabolism, in contrast to *HPS1* we have no evidence for association of *PYROXD2* expression levels with TMA. Thus, our data indicates that the mQTLs of this locus act predominantly as eQTLs through *HPS1*, pointing to its regulatory role in modulating TMA.

Our work illustrates the potential of metabolome- and transcriptome-wide association studies for deciphering gene-metabolite relationships. In particular, even with our modest sample size of 555 matched profiles we already had enough power to detect one significant and several marginally significant associations. Moreover, our two strongest associations pinpointed genes in loci implicated by mGWAS as the most likely candidates for transcriptional metabolite regulation. We also showed the possibility of extending correlative work and studying the causal relationship between gene expression levels and metabolite concentrations. Our Mendelian randomization study supported the causal role of *ALMS1* gene expression levels on N-acetylated compound concentration, whereas for *HPS1* we observed a negative feedback loop between its expression levels and TMA metabolite concentrations. Furthermore, this work demonstrated that our metabomatching tool, whose usefulness for elucidating candidate metabolites from mGWAS association profiles [18, 22] as well as auto-correlation signals in NMR data [24] was demonstrated previously, performs equally well on pseudospectra generated by association with gene expression levels.

Our study has many limitations: First, we only had access to gene expression levels of LCLs. While blood and such blood-derived cells are the easiest samples one can obtain from healthy subjects, their expression levels in many cases may only reflect poorly those of the relevant cells and tissues. Furthermore, metabolic reactions are of course driven by enzymes whose protein concentration determines the metabolic rate, and variation in gene expression levels is only one source of variation in active enzyme concentration (next to post-transcriptional and post-translational modifications, as well as their decay rate). Second, metabolite concentrations in urine correspond to excess that is cleared from the body, which depends on food intake and provide a poor proxy for many metabolite concentrations in their relevant location. Nevertheless, our study shows the promise of co-analyzing two or more distinct molecular traits observed in the same cohort.

Acknowledgements

This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1) and the NIH (grant R03 CA211815).

Author Contributions

RS, RR and SB designed the project. RS carried out the computational analysis and prepared the results. BK and RR provided data and feedback on the metabolomics analysis. ZK guided the MR analysis. All authors discussed the results and provided feedback on the manuscript that was written by RS, BK and SB.

Declaration of Interests

The authors declare no competing interests.

References

1. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic acids research, 2018. **47**(D1): p. D1005-D1012.
2. Consortium, G., *The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-660.
3. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-1195.
4. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS genetics, 2010. **6**(4): p. e1000888.
5. Ward, L.D. and M. Kellis, *Interpreting non-coding variation in complex disease genetics.* Nature biotechnology, 2012. **30**(11): p. 1095.
6. Lloyd-Jones, L.R., et al., *The genetic architecture of gene expression in peripheral blood.* The American Journal of Human Genetics, 2017. **100**(2): p. 228-237.
7. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773.
8. Wright, F.A., et al., *Heritability and genomics of gene expression in peripheral blood.* Nature genetics, 2014. **46**(5): p. 430.
9. Suhre, K., et al., *A genome-wide association study of metabolic traits in human urine.* Nature Genetics, 2011. **43**(6): p. 565-569.
10. Bartel, J., et al., *The Human Blood Metabolome-Transcriptome Interface.* PLoS Genet, 2015. **11**(6): p. e1005274.
11. Burkhardt, R., et al., *Integration of Genome-Wide SNP Data and Gene-Expression Profiles Reveals Six Novel Loci and Regulatory Mechanisms for Amino Acids and Acylcarnitines in Whole Blood.* PLoS Genet, 2015. **11**(9): p. e1005510.
12. Inouye, M., et al., *Metabonomic, transcriptomic, and genomic variation of a population cohort.* Molecular systems biology, 2010. **6**(1).
13. Bullaughey, K., et al., *Expression quantitative trait loci detected in cell lines are often present in primary tissues.* Human molecular genetics, 2009. **18**(22): p. 4296-4303.
14. Çalışkan, M., et al., *The effects of EBV transformation on gene expression levels and methylation profiles.* Human molecular genetics, 2011. **20**(8): p. 1643-1652.

15. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type–dependent manner.* Science, 2009. **325**(5945): p. 1246-1250.

16. Ding, J., et al., *Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals.* The American Journal of Human Genetics, 2010. **87**(6): p. 779-789.

17. Firmann, M., et al., *The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome.* BMC cardiovascular disorders, 2008. **8**(1): p. 6.

18. Rueedi, R., et al., *Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links.* PLoS genetics, 2014. **10**(2): p. e1004132.

19. GEM-Tools, *(v1.7.1)*.

20. MATLAB, *8.5.0.197613 (R2015a)*. 2015, The MathWorks Inc.: Natick, Massachusetts.

21. Gao, X., J. Starmer, and E.R. Martin, *A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2008. **32**(4): p. 361-369.

22. Rueedi, R., et al., *Metabomatching: Using genetic association to identify metabolites in proton NMR spectroscopy.* PLoS computational biology, 2017. **13**(12): p. e1005839.

23. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018.* Nucleic acids research, 2018. **46**(D1): p. D608-D617.

24. Khalili, B., et al., *Automated analysis of large-scale NMR data generates metabolomic signatures and links them to candidate metabolites.* bioRxiv, 2019: p. 613935.

25. Burgess, S., D.S. Small, and S.G. Thompson, *A review of instrumental variable estimators for Mendelian randomization.* Statistical methods in medical research, 2017. **26**(5): p. 2333-2355.

26. Davey Smith, G. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International journal of epidemiology, 2003. **32**(1): p. 1-22.

27. Raffler, J., et al., *Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality.* PLoS genetics, 2015. **11**(9): p. e1005487.

28. Võsa, U., et al., *Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis.* BioRxiv, 2018: p. 447367.

29. Wald, A., *The fitting of straight lines if both variables are subject to error.* The Annals of Mathematical Statistics, 1940. **11**(3): p. 284-300.

30. Hartung, J., G. Knapp, and B.K. Sinha, *Statistical meta-analysis with applications*. Vol. 738. 2011: John Wiley & Sons.

31. Greco M, F.D., et al., *Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome.* Statistics in medicine, 2015. **34**(21): p. 2926-2940.

32. Staley, O.Y.J., *MendelianRandomization: Mendelian Randomization Package. R package version 0.4.1.* 2019.

33. Staley, O.Y.J., *MendelianRandomization: Mendelian Randomization Package*. 2019, <james.staley@bristol.ac.uk>.

34. Engelke, U.F., et al., *N-acetylated metabolites in urine: proton nuclear magnetic resonance spectroscopic study on patients with inborn errors of metabolism.* Clinical chemistry, 2004. **50**(1): p. 58-66.

35. Kastenmüller, G., et al., *Genetics of human metabolism: an update.* Human molecular genetics, 2015. **24**(R1): p. R93-R101.

36. Nicholson, G., et al., *A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection.* PLoS genetics, 2011. **7**(9): p. e1002270.

37. Montoliu, I., et al., *Current status on genome–metabolome-wide associations: an opportunity in nutrition research.* Genes & nutrition, 2013. **8**(1): p. 19.

38. Chambers, J.C., et al., *Genetic loci influencing kidney function and chronic kidney disease.* Nature genetics, 2010. **42**(5): p. 373-375.

39. Simmons, M., C. Frondoza, and J. Coyle, *Immunocytochemical localization of N-acetyl-aspartate with monoclonal antibodies.* Neuroscience, 1991. **45**(1): p. 37-45.

40. Masaharu, M., et al., *N-acetyl-l-aspartic acid, N-acetyl-α-l-aspartyl-l-glutamic acid and β-citryl-l-glutamic acid in human urine.* Clinica Chimica Acta, 1982. **120**(1): p. 119-126.

41. Barker, P.B., *N-acetyl aspartate—a neuronal marker?* Annals of neurology, 2001. **49**(4): p. 423-424.

42. Jung, R.E., et al., *Biochemical markers of intelligence: a proton MR spectroscopy study of normal human brain.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 1999. **266**(1426): p. 1375-1379.

43. Patel, T. and J.B. Talcott, *Moderate relationships between NAA and cognitive ability in healthy adults: implications for cognitive spectroscopy.* Frontiers in human neuroscience, 2014. **8**: p. 39.

44. Davies, G., et al., *Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function.* Nature communications, 2018. **9**(1): p. 1-16.

45. Lee, J.J., et al., *Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals.* Nature genetics, 2018. **50**(8): p. 1112-1121.

Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3360−3368

pubs.acs.org/jpr

Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites

Bita Khalili,[†,‡,∥] Mattia Tomasoni,[†,‡,∥] Mirjam Mattei,[†,‡,∥] Roger Mallol Parera,[†,‡] Reyhan Sonmez,[†,‡] Daniel Krefl,[†,‡] Rico Rueedi,[†,‡,∥] and Sven Bergmann*[,†,‡,§,∥]

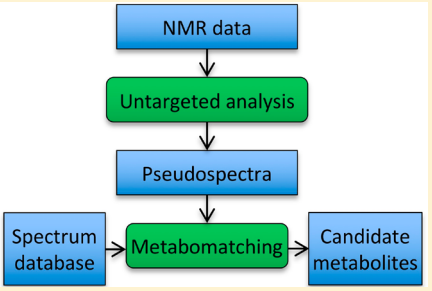
[†]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland

[‡]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[§]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town 7700, South Africa

Supporting Information

ABSTRACT: Identification of metabolites in large-scale ^{1}H NMR data from human biofluids remains challenging due to the complexity of the spectra and their sensitivity to pH and ionic concentrations. In this work, we tested the capacity of three analysis tools to extract metabolite signatures from 968 NMR profiles of human urine samples. Specifically, we studied sets of covarying features derived from principal component analysis (PCA), the iterative signature algorithm (ISA), and averaged correlation profiles (ACP), a new method we devised inspired by the STOCSY approach. We used our previously developed metabomatching method to match the sets generated by these algorithms to NMR spectra of individual metabolites available in public databases. On the basis of the number and quality of the matches, we concluded that ISA and ACP can robustly identify ten and nine metabolites, respectively, half of which were shared, while PCA did not produce any signatures with robust matches.



KEYWORDS: *1D NMR automated analysis, metabolite identification, modular analysis, STOCSY, ISA, pseudoquantification, NMR spectroscopy, untargeted metabolomics*

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for metabolomic profiling. NMR spectroscopy does not consume the sample and has high accuracy and reproducibility. Single proton NMR spectroscopy (^{1}H NMR) can be used to generate one-dimensional spectra of biofluids at high throughput and low cost, facilitating the generation of large sets of spectral data.

A first step in NMR spectral analysis is usually to identify the main metabolites giving rise to a given spectrum or set of spectra. This is nontrivial, since human biofluids typically contain a large number of individual metabolites and their corresponding peak positions may overlap and are often affected by the pH, ionic strength, and overall protein content of the fluid.

For small sets of samples, expert analysis is therefore still the most accurate means for metabolite identification, yet for large sets this approach is costly, time-consuming, and potentially less reproducible. As a result, various methods have been suggested to assist or fully automate metabolite identification.

In their landmark paper on statistical total correlation spectroscopy (STOCSY), Cloarec et al. showed that analyzing the correlation patterns between features across a sizable collection of ^{1}H NMR spectra has great potential for metabolite identification.[1] This is because features corresponding to the same molecule (or molecules whose concentrations covary) tend to be significantly correlated in large data sets. Analyzing data from 612 mouse urine samples, they observed that the correlation matrix exhibited correlated peaks of features characteristic of valeramide, glucose, hippurate, 2-oxoglutarate, 3-hydroxyphenylpropionate, citrate, as well as methylamine, dimethylamine, and trimethylamine.

Subsequent variations of STOCSY attempted to make clusters of NMR peaks to simplify the interpretation of the information stored in the correlation matrix from STOCSY analysis. Recoupled-STOCSY (R-STOCSY) employs a variable size bucketing method to reduce the dimensionality of NMR data and statistical recoupling of variables (SRV) to identify correlations between distant clusters.[2] Iterative-STOCSY (I-STOCSY) aims at separating the intermetabolite connections from intrametabolite connection by recursively applying STOCSY analysis first from a selected *driver peak* and then for all peaks correlating with the driver peak above a specific threshold.[3] Subset optimization by reference matching (STORM) selects subsets of ^{1}H NMR spectra that contain specific spectroscopic signatures of biomarkers differentiating between different human populations.[4]

Once metabolite identification has been achieved, the next challenge is to quantify metabolite concentrations. This process works robustly only for a relatively small set of metabolites, and requires expert refinement when using publicly available quantification tools such as BATMAN,[5] FOCUS,[6] BAYESIL,[7]

Received: May 6, 2019
Published: July 18, 2019

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

ASICS,[8] AQuA,[9] or rDolphin.[10] This is unsatisfactory in light of the fact that a sizable number of metabolites have been identified in human biofluids. For example, the latest version of the Human Metabolome Database (HMDB 4.0)[11] includes more than 1500 metabolites with ^{1}H NMR spectra, 179 of which have been identified in urine[12] and 67 in serum.[13,14]

There are several reasons why it remains difficult to perform fully automated quantification from large-scale NMR data for the vast majority of metabolites. First, human biofluids contain a large number of individual metabolites whose concentrations vary across several orders of magnitude. This makes it difficult to disentangle the contributions of metabolites with low concentrations, in particular when their NMR features are not unique. Second, the exact feature positions depend on the biofluid and may have been different when acquiring reference spectra. Third, while the number of reference spectra continues to grow, reference databases are certainly not yet exhaustive.

In two recent studies, we demonstrated that the limitations of targeted NMR metabolomics can be addressed by linking metabolites to external variables.[15,16] Specifically, in the context of genome-wide association studies (GWAS) applied to metabolomics (known as mGWAS) the aim is to associate metabolites with genotypic variants. We observed that the effect of a genetic variant on the concentration of a metabolite often translates into associations with all or many features of the metabolite NMR spectrum. The set of association scores with all measured features provides a *pseudospectrum* across the full range of ppm covered by the ^{1}H NMR spectra. The challenge is then to identify the metabolite underlying the most significant associations. To this end we developed the analysis tool *metabomatching*, which takes as input a pseudospectrum and a collection of reference spectra for individual metabolites found, for example, in HMDB.[11] Our previous work showed that metabomatching works well to prioritize the most likely metabolite candidates for pseudospectra derived from metabolome feature association with genotypes.[15−17]

In the present work, we tested the metabomatching methodology for identifying metabolites that vary across large collection of samples without the need for any external variables associated with this variation. We investigated three methods to identify covarying spectral features within large-scale NMR data: principal component analysis (PCA), the iterative signature algorithm (ISA), and averaged correlation profiles (ACP) inspired by the STOCSY approach. For each method, we devised a principled way for processing their output into pseudospectra. In addition, we extended metabomatching to process the respective outputs of the methods, and implemented a permutation-based robustness test to assess the quality of the matches. This allowed us to compare the matches across different methods, and assess the consistency or complementarity of the methods. We incorporated our analysis for unsupervised generation of metabolomic signatures from large-scale NMR data and integration with metabomatching (including further documentation) into the *metabomodules* software tool, which is publicly available at https://github.com/BergmannLab/metabomodules-docker.

METHODS

Preprocessing

For this study, we used 968 ^{1}H NMR spectra acquired from urine samples from the CoLaus cohort.[15] The samples are referenced to the TSP signal, phase-corrected, and baseline-corrected.

We used the FOCUS[6] tool to align and bin the spectra to a resolution of about 0.02 ppm, and a correspondingly large number of 687 peaks. To obtain this resolution, we set the downsampling frequency parameter *window.fs* to 1 (no downsampling), the sliding window length for spectral segmentation *window.length* to 0.03 ppm, the minimum peak width parameter *peak.DFL* to 0.02 ppm, and the peak sample frequency parameter *peak.pS* to 0.2, keeping the rest of the parameters at their default values.

To normalize the data, we log-transformed, standardized across features (thereby normalizing the concentration of each sample), then standardized across samples (thereby making intensities comparable).

Confounding

In order to allow for the identification of metabolites that may be hidden by confounding, we additionally generated a data set of residuals, created by regressing out the confounders from the feature metabolome. The main confounding factors of the NMR data that we investigated here are age, sex, serum creatinine, and urine creatinine.[15,16]

Metabomatching

Our original metabomatching method was designed to match the NMR spectra of individual metabolites recorded in a database with pseudospectra from the association between metabolome features and an external variable, typically a SNP genotype. For a metabolite m, metabomatching computes the sum

$$s(F_m) = \sum_{f \in F_m} z_f^2 \tag{1}$$

where F_m is the set of N_f features that fall within a neighborhood of any peak of m according to the database, and z_f denotes the significance value for feature f, and is given by $z_f = \hat{\beta}_f / \widehat{SE}_f$, where $\hat{\beta}_f$ and $\widehat{SE}_f$ refer to the point estimates of the effect size and its standard error, respectively. Under the null hypothesis of normally and independently distributed z_f, the sum s follows a χ^2-distribution with N_f degrees of freedom, and metabomatching defines the score for m as the negative logarithm of the nominal p-value for the sum. This score is then used to rank all tested metabolites as metabolites with more similar NMR spectra to a given pseudospectrum achieve higher scores.

In addition to pseudospectra from regression analysis, provided as columns headed by *beta*, *se*, and *p*, we extended metabomatching to accept pseudospectra produced by PCA, ACP, and ISA as columns headed by *pca*, *cr*, and *isa* respectively. For ACP pseudospectra, metabomatching translates a correlation c to a z-score with the Fisher transformation $z = \lambda \operatorname{arctanh}(c)$. For independent features, $\lambda = \sqrt{N-3}$ produces z-scores with unit standard deviation, where N is the number of samples across which the correlations are computed. However, since proximal features are usually not independent, metabomatching allows for a user-provided estimate for λ (obtained from the pairwise feature−feature correlation matrix), or re-estimates λ from the given correlations. For ISA and PCA pseudospectra, metabomatching standardizes the loadings or module scores.

We used the plus/minus mode of metabomatching since features are z-scored and have positive and negative signs. This allows for detecting metabolites corresponding either to the

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

negative or positive features (see metabomatching documentation at https://github.com/rrueedi/metabomatching for more details).

We also introduced a measure of the quality of a match, which allows to compare matches between different pseudospectra. This *adjusted score* is obtained by reshuffling the pseudospectrum, and defining a heuristic p-value by the number N_p of all N_r reshuffled pseudospectra that produce a metabomatching score (for any reference spectrum) higher than the metabomatching score of the input pseudospectrum with the highest ranked reference spectrum. This p-value is defined as $(N_p + 1)/(N_r + 1)$, and the adjusted score as $-\log(p)$. We used $N_r = 9999$, which sets the upper limit for the adjusted score to 4.

For the reshuffling to be consistent with the structure of NMR spectra, metabomatching identifies *cut points* that separate the pseudospectrum into peak-preserving clusters of features and only reshuffles these clusters. The cut points are obtained as follows. Let f_i be the positions on the chemical shift axis of the metabolome features, sorted such that $f_i < f_{i+1}$, and C the set of cut points. First, metabomatching populates C with features bordering a *gap*, that is a region absent from the spectrum and larger than δ_{gap} (i.e., $f_i > f_{i-1} + \delta_{gap}$), with $\delta_{gap} = 0.3$ ppm as default value. Next it sorts the remaining features by their corresponding absolute-valued z-scores. Starting from the feature with the lowest absolute z-score it adds features to C provided they have a distance greater than δ_{min} to any features already assigned to C and an absolute z-score below a threshold z_{min}. Default values are 0.04 ppm for δ_{min} and the standard deviation of all absolute z-scores across the features of a given pseudospectrum for z_{min}.

ACP: Averaged Correlation Profile

ACP is a greedy approach to generate a list L of feature pairs and their corresponding correlation profiles c as input for metabomatching: (1) We compute and sort all pairwise correlations C_{ij} between features f_i and f_j separated by at least 0.1 ppm. (2) Starting with the feature pair $P = (i,j)$ with the highest correlation, we successively add feature pairs to L unless there is already a feature pair in L whose features are within 0.1 ppm of f_i and f_j, respectively. (3) For each feature pair in L, we define an *averaged correlation profile* as the average of the correlation profiles of f_i and f_j: $c_k = (C_{ik} + C_{jk})/2$. The correlation profiles of strongly correlated features are similar, consequently their average is similar to both of them. Crucially, the average does not contain an element equal to 1, as $c_i = c_j = (1 + C_{ij})/2 < 1$ given that $C_{ij} < 1$ in real data. For our analysis, we limit L to 179 averaged correlation profiles, 179 being the number of spectra in UMDB, the reference database on which metabomatching will run.

As an alternative approach, we tried agglomerative clustering of features, iteratively joining features (or sets of features) whose correlation was above a threshold C_{min}. At each step, we averaged joined (sets of) features into a metafeature and recomputed its correlation to all remaining (meta)features. We then built a correlation profile for each feature cluster by averaging the correlations profiles of the component features.

ISA: Iterative Signature Algorithm

ISA is a biclustering method first developed for modular analysis of gene expression data.[18,19] ISA uses a heuristic iterative procedure starting with random features to refine *modules*, consisting of self-consistent subsets of features *and* samples. Each module is defined for a set of two thresholds, determining how extreme the features and samples are allowed to be. Importantly, scanning through an array of thresholds usually identifies a set of modules (or module families) that is smaller than the number of samples or features.

We first ran the ISA algorithm to generate modules from the NMR data using the default values for the parameters except the following: (1) we changed both row and column thresholds from the default values {1, 2, 3} to {1, 2, 3, 4, 5, 6} to produce modules containing fewer rows or columns that are more likely to represent single metabolites, (2) we increased the number of seeds from 100 to 250, and (3) we lowered the correlation threshold below which ISA considers two modules equal to one another from 0.95 to 0.50 to favor diversity in modules.

We allowed feature scores to be either positive or negative, while sample scores were always positive. This means that modules can include features which have on average higher or lower intensities in the selected samples than for the remaining samples. Modules which include such a mixture are likely to represent (at least) two metabolites whose concentrations are inversely related to each other (like a substrate and its product).

This procedure generated 216 modules. To select the 179 modules as an input for metabomatching, we sorted them by the size of their basin of attraction. To measure the basin size, we ran ISA a second time with the same parameters, but on 10 000 seeds, turning off the *sweeping* option, and keeping all converged modules (by setting the *purge* option to false). We then assigned a run 2 module to the basin of the run 1 module with which it had the highest correlation (provided that correlation be greater than 0.5). We assumed that the run 1 modules attractor basin size is approximately equal to the count of modules from run 2 which were assigned to them in the previous step. Finally, we passed to metabomatching the 179 modules from run 1 with largest basins.

PCA: Principal Component Analysis

We used the sklearn library for Python (2.7) to compute all principal components of the preprocessed data. Specifically, we used the *decomposition*.PCA object, with *n_components* set to 687 and *svd_solver* set to *full*.

Identification of Metabolites

After running metabomatching on the pseudospectra generated by ISA, ACP and PCA, we applied a filtering algorithm to select only the most robust matches among all pseudospectra. The filtering passes only those pseudospectra which achieve an adjusted score above 2 with their top metabolite match (ensuring the pseudospectra finds a reasonable match by metabomatching) and have at least one peak with z-score above 4 (ensuring there is a strong signal). Note that multiple pseudospectra can match with the same metabolite NMR spectrum.

Out of the 179 pseudospectra from ACP, 10 pseudospectra matching different metabolites passed the filtering. Among these, only for one pseudospectrum, i.e., 3.87 and 3.75, the top match to hydroxypropionate (Figure S24) did not look convincing because of slightly worse matches to mannitol and arabitol (Figure S24).

Out of the 179 pseudospectra from ISA, 19 pseudospectra matching different metabolites passed the filtering. Among these, we discarded 9 for one or more of the following three reasons: (1) There are several metabolites which all achieve high adjusted scores (Figures S25 and S26). (2) There is at least one strong peak in the pseudospectrum that does not match with any of the spectra of the best matching metabolites (see Figures S27−S32). This may happen for pseudospectra with a large

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

number of peaks. These pseudospectra are not necessarily biologically irrelevant and might carry a signature for two or more related metabolites. (3) The pseudospectrum passed the filtering, yet did not appear to have sufficiently strong signals at the peak positions of its putative matching metabolites (Figures S33 and S34).

We analyzed all 687 pseudospectra from PCA and observed an elevation in metabomatching scores for the last principal components (all between components #505−687, which jointly explain only 1% of variation; Figures S1A, S35−S39). However, since these components explain almost no variation in the metabolome and the last 9 principal components matched the same metabolite, hippurate (Figure S1D), we investigated whether these matches rely on numerical instability. Indeed, when we removed one feature from all feature pairs that correlated above 0.95 (i.e., 8 features from 34 feature pairs all belonging to hippurate multiplets regions 7.54−7.56, 7.63−7.65, 7.83−7.84, and 3.96 ppm), and ran metabomatching on all pseudospectra generated from the principal components of the remaining features, only five passed the filtering (Figure S1F). However, none of these seemed convincing when applying the same curation as for the ISA pseudospectra.

Pseudoquantification of Metabolite Concentrations

We use the term pseudoquantification as this approach should not be confused with traditional quantification approach which sometimes rely on experiments that target a specific metabolite and often require the use of proprietary software operated by an expert.

We perform our pseudoquantification by using discrete integration to estimate relative metabolite concentrations, according to

$$c_i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H_k}\sum_{l_k \le s_j \le r_k} I_{ij}\Delta_j \qquad (2)$$

where K is the number of multiplets, H_k the number of protons in multiplet k, $[l_k, r_k]$ the range of multiplet k, s_j the chemical shift of feature j, I_{ij} the intensity of this feature in individual i, and Δ_j the width of the bin at s_j. For example, for hippurate $K = 4$, $H = [2, 2, 1, 2]$, $l = m - 0.025$, $r = m + 0.025$, where $m = [3.98, 7.54, 7.65, 7.84]$. We then evaluated this relative concentration first using the peak positions from the reference spectrum as listed in UMDB, and second using the peak positions as suggested by the pseudospectrum found by the modular approach.

To perform the pseudoquantification based on the pseudospectra of the modules, we defined a set of multiplet positions to use in eq 2 for each module that robustly identified a metabolite. This set is composed of all the chemical shifts from the module of interest with z-scores above 3 and within 0.025 ppm of the multiplet positions of the matching metabolite in reference database. It includes all relevant peaks from the metabolite detected by the modular analysis.

■ ANALYSIS AND RESULTS

In this work, we show that metabomatching can be used with pseudospectra capturing the internal structure of large-scale NMR data rather than their correlation with external variables. Specifically, our premise is that in sizable sample collections there is sufficient power for methods identifying coherent features that may point to the same metabolite.

We used three methods for identifying weighted sets of covarying spectral features from large-scale NMR data that can be used as input for metabomatching (see Methods for more details and Figure 1 for an illustration of the workflow).

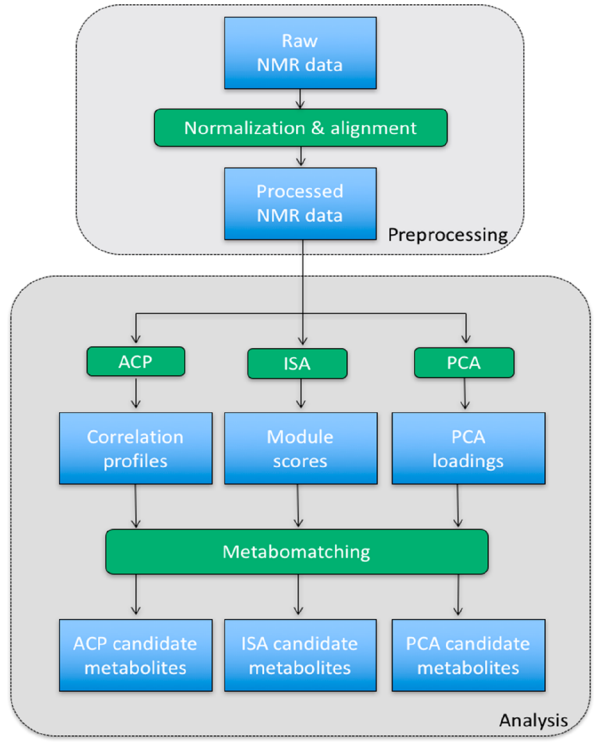


Figure 1. Workflow for unsupervised analysis of large-scale NMR data. Raw ^{1}H NMR data are normalized then aligned. These processed profiles are used as input for the averaged correlation profile (ACP), iterative signature algorithm (ISA), and principal component analysis (PCA) methods, which output correlation profiles, module scores, and PCA loadings, respectively. These outputs constitute possible pseudospectra for metabomatching, which identifies the most plausible candidate metabolites underlying the coherent feature variations.

Correlation-Based Pseudospectra

Our first approach to select covarying features was to use the correlations between features across all samples. We faced two challenges: First, using the correlations of a given feature with all features as a pseudospectrum would break the scoring algorithm in metabomatching, due to self-correlation of features that result in infinite z-scores. Second, as generating only a limited number of pseudospectra was desired ranking the input sets was necessary to select only the most relevant ones.

To address these challenges, we devised an algorithm that ranks all pairs of sufficiently distant features and computes averaged correlation profiles (ACP) for pairs of highly correlated features. Strictly speaking, ACPs are not the correlations but they are the average correlation profiles, with the premise that for highly correlated feature pairs the correlations to other features tend to be similar, while none of the averages equals to one. Within metabomatching these ACPs are then translated into z-scores using the Fisher z-transformation (see Methods for more details). We also tried hierarchical clustering to define pseudospectra from multiple highly similar features (see Methods), but this approach did not work as well.

Iterative Signature Algorithm

ISA has been designed for the unsupervised identification of coherent subsets in large-scale data.[18,19] Specifically, coherence

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

between features is *not* defined by total correlation across all samples, but rather by a subset of samples for which a set of features takes more extreme values than for the rest of the samples. Such a joint set of features and samples is called a *module*. In order to obtain a pseudospectrum for each module we averaged each feature's score (whether part of the module or not) across the samples assigned to the module. These averages were then transformed into *z*-scores. By definition the features of the module have the most extreme *z*-scores, yet other features that were just below the threshold may also have a sizable contribution. We then used these *z*-scores as input for metabomatching.

Principal Component Analysis

We also used PCA to compute the loadings of all features onto the eigenvectors of the sample−sample correlation matrix across all features. These eigenvectors (or eigensamples) characterize independent axes of variation in sample space. The corresponding eigenvalues reflect the fraction of total variation explained by each eigenvector. It was not clear a priori which principal components might characterize variation due to single metabolites. We therefore applied metabomatching to all of them. Specifically, we generated pseudospectra by standardizing the loadings corresponding to each eigensample (see Methods for more details).

Many Pseudospectra Defined by the ACP Method and ISA Match to Urine Metabolites

We observed a trend of elevated metabomatching scores for pseudospectra corresponding to principal components with small eigenvalues (starting from component #505), jointly explaining only 1% of variation (Figure S1A). The last nine principal components matched to hippurate, but disappeared when running PCA on the metabolome stripped of features that are highly correlated to other features (see Methods; Figures S1D and S1F). Additionally, the adjusted scores of all potential hits decreased significantly for the stripped metabolome (Figure S1E). We therefore concluded that PCA is not well-suited for generating robust metabolite signatures.

In contrast, our ACP method and ISA resulted in a sizable number of pseudospectra for which metabomatching produced robust matches to urine metabolites (see Figure 2 and Methods for details). Specifically, both ACP and ISA identified feature sets pointing to glucose, citrate, ethanol, hippurate, and P-hydroxyphenylacetate (Figures S2−S11). Glucose and hippurate were among the metabolites identified by Cloarec et al. with the correlation matrix of mouse urine NMR data.[1]

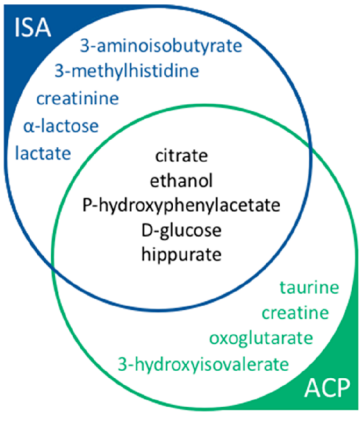


Figure 2. Urine metabolites that were robustly matched by metabomatching to pseudospectra derived from average correlation profiles (ACP, green), the iterative signature algorithm (ISA, blue), or both methods (black).

P-hydroxyphenylacetate shares an aromatic ring with 3-hydroxyphenylpropionate, another metabolite highlighted in the original STOCSY paper.[1] Both compounds are part of phenylalanine metabolism and occur as products of bacterial degradation of aromatic compounds. In human urine high concentrations of these compounds may reflect an overgrown *Clostridium* species in gut microbiota, which has been associated with autism spectrum disorders.[20]

In healthy humans, urine glucose should be low, but concentrations may be elevated due to diabetes or chronic kidney disease (CKD), conditions which are prevalent in the CoLaus population.

Citrate is an additive commonly used by the food industry and it is also synthesized as an intermediate product in the tricarboxylic acid cycle, a central pathway that releases stored energy from fat, proteins, and carbohydrates. Low urinary citrate is associated with CKD and kidney stone formation.

There are a number of metabolites that matched to pseudospectra generated only by one of the two methods. With ISA, we found modules matching 3-aminoisobutyrate, an end product of nucleic acid metabolism that has been considered a potential biochemical marker for cancer[21] (Figure S12); creatinine, a breakdown product of creatine, whose high and stable concentration in urine is often used for normalization (Figure S13); lactose (Figure S14); and lactate, the bacterial breakdown product of lactose (Figure S15). Conversely, ACP produced correlation profiles matching to taurine (Figure S16), an organic compound widely distributed in animal tissues and a major constituent of bile; creatine (Figure S14); oxoglutarate (α-ketoglutarate), an important biological compound produced by deamination of glutamate, and an intermediate in the Krebs cycle (Figure S18); and 3-hydroxyisovalerate, a byproduct of valine, leucine, and isoleucine degradation and a marker for biotin deficiency[22] (Figure S19). These compounds are all common urine metabolites that can exist in high concentrations.

Correcting features for significant covariates, both methods also found set of features matching carnitine, which owes its name to its high concentration in meat (see Methods for more details). While it is produced in both animal and plant cells, this may explain why we could detect its signature in human urine samples.

Metabolite Concentration Pseudoquantification with NMR Features of Matched Pseudospectra

We next investigated whether the sets of weighted NMR features generated by ISA or the ACP method can not only be used to identify metabolites, but also facilitate their pseudoquantification. This pseudoquantification approach aims to estimate the relative concentration of the metabolites in untargeted ^{1}H NMR of urine samples. We performed our pseudoquantification by computing the area under the peak of each multiplet in the metabolite spectrum using discrete integration, dividing it by the number of protons associated with the multiplet and then averaging scaled areas over all multiplets in the metabolite spectrum (see Methods for more details). We hypothesized that the leading features selected by our algorithms for a certain metabolite may be better suited for pseudoquantification than the full reference spectra from public databases such as UMDB. There are two possible reasons for this. First, the exact feature positions extracted from the data by ISA or the ACP method

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

Table 1. Correlation between Pseudoquantification and Measured Biomarkers of Glucose and Ethanol

urine metabolite	feature source	multiplet positions (ppm)	related biomarker	correlation [95% CI]
glucose	UMDB	3.23, 3.40, 3.46	serum	0.46
		3.52, 3.73, 3.82	glucose	[0.41, 0.52]
		3.88, 4.63, 5.22		
glucose	ACP: 3.48 and 5.24	3.40, 3.48, 4.65	serum	0.48
		5.24	glucose	[0.43, 0.54]
glucose	ACP: 3.89 and 5.24	3.82, 3.89, 4.65	serum	0.44
		5.24	glucose	[0.38, 0.49]
glucose	ISA: module #16	3.25, 3.41, 3.48	serum	0.50
		3.50, 3.89, 4.65	glucose	[0.44, 0.55]
		5.24		
ethanol	UMDB	1.17, 3.65	serum	0.29
			CDT	[0.23, 0.35]
ethanol	ACP: 1.18 and 3.67	1.18, 3.67	serum	0.16
			CDT	[0.10, 0.22]
ethanol	ISA: module #57	1.18, 3.67	serum	0.16
			CDT	[0.10, 0.22]
EtG	Nicholas et al.[23]	1.24, 3.30, 3.52	serum	0.36
		3.71, 3.99, 4.48	CDT	[0.30, 0.42]
EtG	ISA: module #240	1.24, 3.52, 4.47	serum	0.46
			CDT	[0.40, 051]

may be more accurate, even if, by design, they fall within the margin of the matching window of reference spectrum features. Second, for metabolites with several peaks, only a subset might have been picked up by these algorithms. Indeed, both ISA and the ACP method may leave out peaks that did not contribute coherently to the peak set since their signal was too noisy (e.g., due to overlap with those from other metabolites).

We performed our pseudoquantification method (using eq 2 in Methods) to estimate concentrations of glucose and ethanol, for which relevant phenotypes were available in the cohort. For urine glucose, the phenotype was fasting blood glucose. For urine ethanol, relevant biomarkers included serum asialotransferrin and disialotransferrin, which combined are known as carbohydrate-deficient transferrin (CDT), a biomarker for heavy alcohol use. Furthermore, self-reported alcohol consumption was available. These biomarkers were measured in a different biofluid (i.e., blood), which was collected on the same day as the urine sample. We argue that detecting significant correlations between our estimated metabolite concentrations and these biomarkers provides a proof of concept that our pseudoquantification is reliable, and comparing correlations between different pseudoquantification approaches provides a means to evaluate them.

The ^{1}H NMR spectra of glucose has nine multiplets. Including all these multiplets chemical shifts from the UMDB database (Table 1) to perform pseudoquantification, we obtain a correlation of 0.46 (with a 95% confidence interval (CI) of [0.41, 0.52]) between the estimated concentration of urine glucose and fasting blood glucose. This correlation increases to 0.50 [0.44, 0.55] if the subset of seven multiplets from ISA module #16 (Figure 3A) is used for pseudoquantification (see Methods for details). From the 179 ACP pseudospectra, two robustly matched glucose, one from averaging the correlation profiles from feature pair 3.48 and 5.24 and another from averaging correlation profiles from 3.89 and 5.24 (Figure 3B,C), each capturing four out of nine multiplets of glucose (Table 1). Using the subset of peaks from 3.48 and 5.24 ACP pseudospectra we obtain a correlation of 0.48 [0.43, 0.54] between glucose pseudoquantification and fasting blood glucose while using the subset of peaks from 3.89 and 5.24 ACP pseudosspectra a correlation of 0.44 [0.38, 0.49] is obtained. Combining the peak subsets from both pseudospectra to a subset of 6 peaks did not improve the correlation beyond 0.48 [0.43, 0.54] (see Table 1 and Methods for more details on peak sets used for pseudoquantification).

The ^{1}H NMR spectra of ethanol has only two multiplets (as well as one singlet from the hydroxyl group, which can not be discerned in water solutions like urine). Using the reference spectrum from UMDB to perform pseudoquantification, we obtained a relatively low correlation of 0.29 [0.23, 0.35] between the estimated concentration of urine ethanol and CDT levels in serum (Table 1, see Table S1 for other alcohol markers). The ACP method produced one pseudospectra, 1.18 and 3.67, and ISA produced two modules (#57 and #240) that metabomatching matched to ethanol (Figures S6, S7, and S22). The positions of the ACP peak set (i.e., 1.18 and 3.67) were identical to those of ISA module #57 and were more similar to the ethanol spectrum (achieving a higher adjusted score in metabomatching) than those of ISA module #240 (Figure 3D−F). Nevertheless, pseudoquantification for the ACP pseudospectrum and ISA module #57 yielded a lower correlation of 0.16 [0.10, 0.22] with CDT levels than the UMDB reference peaks (0.29 [0.23, 0.35]) (Table 1). This is due to a high correlation of CDT with the features at 1.145−1.155 ppm, which are within a 0.025 ppm neighborhood of the UMDB ethanol peak at 1.17 ppm but not within the same neighborhood of the 1.18 ethanol peak from ACP and ISA module #57 (Figure S20). Yet, these peaks at 1.145−1.155 ppm are unlikely to correspond to ethanol, since their correlation with the other ethanol peak at 3.67 ppm is much weaker than the correlation between the 1.18 and 3.67 ppm peaks. Instead, they may belong to a different metabolite whose concentration is correlated with CDT (Figure S21). In contrast, summing up the intensities over all the features of ISA module #240 with a z-score above 3 as a pseudoquantification measure (in the absence of any multiplet information), we obtained a correlation of 0.51 [0.46, 0.57] with the CDT measurements.

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

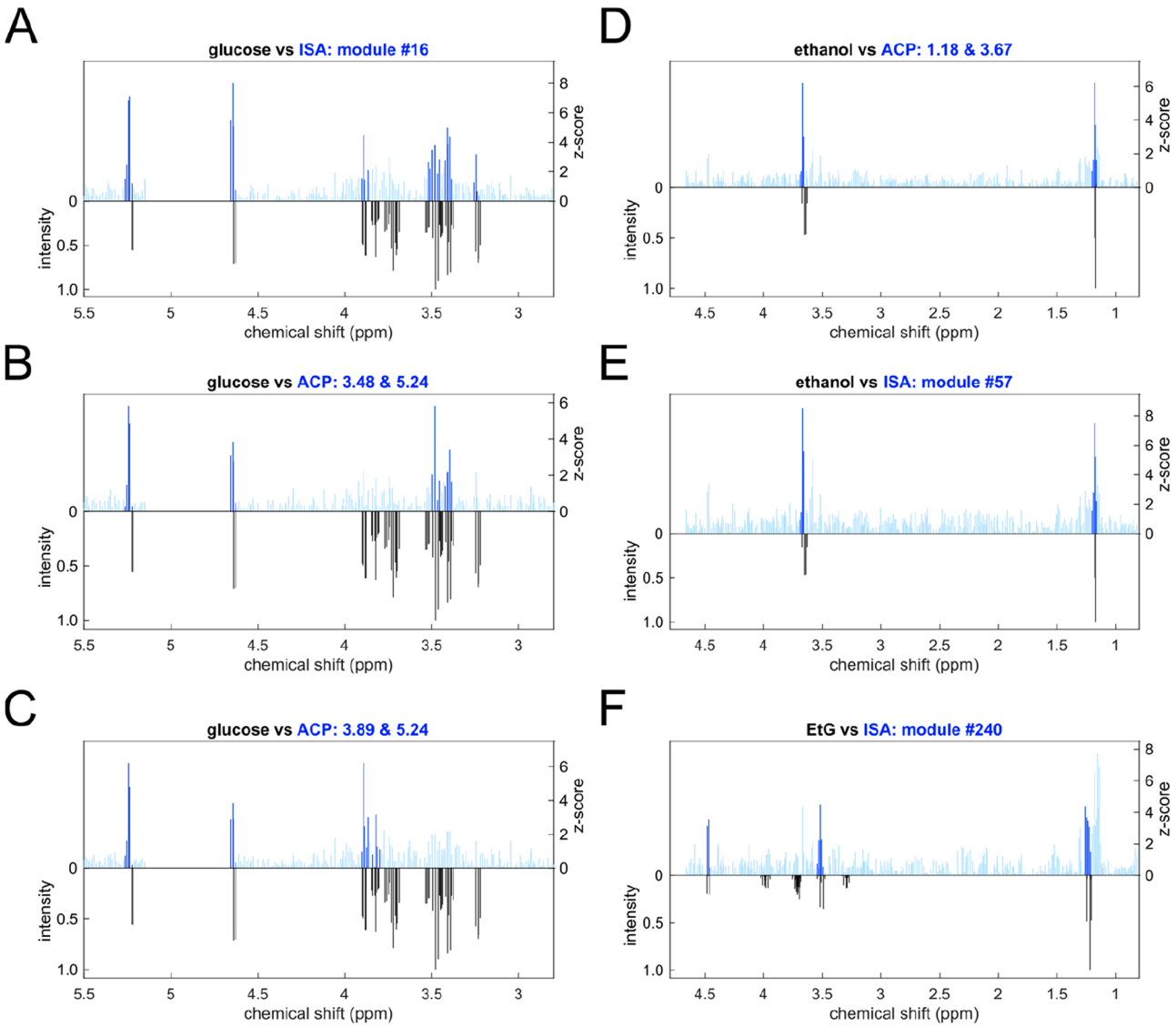


Figure 3. Pseudospectra from ACP and ISA algorithms matching glucose, ethanol, and EtG. Each plot shows the pseudospectrum in blue in the upper half and the reference spectrum from UMDB in black and in the lower half. Dark blue indicates chemical shifts and their ±0.025 ppm vicinity that were used for pseudoquantification.

To better understand why module #240 correlates more strongly to the alcohol consumption biomarker while being a worse match to ethanol than module #57 (Figure 3), we studied whether any of its features point to other compounds related to ethanol metabolism. Indeed, we found that this module contains three features, at 1.26, 3.52, and 4.47 ppm, that individually correlate more strongly to CDT (0.40 [0.34, 0.45], 0.29 [0.23, 0.35], 0.33 [0.27, 0.39], respectively) than the features mapping to ethanol. Interestingly, these features appeared to be close to those of ethyl glucuronide (EtG), a direct product of ethanol nonoxidative metabolism by conjugation with uridine diphosphate (UDP)-glucuronic acid, which had previously been detected in ^{1}H NMR spectra of liver extracts[23] and more recently in human urine of alcohol drinkers.[24] To confirm EtG as a possible match for ISA module #240, we added its features as extracted from Nicholas et al.[23] to the metabomatching library manually, since EtG had no entry in UMDB. We observed that ethanol and EtG spectra together provided a better match to ISA module #240 than ethanol alone (Figure S22). Interestingly, the distance between the two peaks corresponding to the doublet at 4.48 ppm is about 0.0126 ppm, corresponding to a coupling of 8.8 Hz (for a 700 MHz spectrometer) consistent with the coupling of 8 Hz reported in Nicholas et al.[23] (see Figure S23 and Supporting Informationfor more details).

Performing the pseudoquantification of EtG using the peak set of 6 reference positions extracted from Nicholas et al.[23], we obtained a correlation of 0.36 [0.30, 0.42] with CDT levels; performing the pseudoquantification with the 3 feature subset from module #240, we obtained a correlation of 0.46 [0.40, 0.51] (Table 1). This indicates that EtG pseudoquantification correlates better with CDT than ethanol, which is in agreement with the fact that EtG is detectable in urine for a longer time window (2−5 days) than ethanol (12−24 h), and CDT is a marker for heavy alcohol use (at least five drinks a day over a period of 2 weeks before giving the sample[25]). Remarkably, the pseudoquantification facilitated by the three features of module #240 correlates even more strongly with CDT than the full set of EtG reference features, presumably because these features have the best signal-to-noise ratio and optimal position for our data. They may therefore constitute a promising urine biomarker for

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

heavy alcohol consumption. Indeed, while the correlation between EtG pseudoquantification and CDT measure increases to 0.59 [0.46, 0.72] when focusing on subjects who have self-reported heavy drinking, pseudoquantification of module #240 gives rise to a slightly higher correlation of 0.61 [0.48, 0.74].

CONCLUSIONS AND DISCUSSION

In this work, we implemented and tested new methodologies for analyzing large-scale ^{1}H NMR spectroscopy data. Building on previous ideas to use the correlation structure of such data to generate metabolomic signatures, we investigated three complementary methods for generating such signatures and benchmarked the methods in terms of how many of their signatures matched with reference spectra in public databases. By design, these approaches will only identify metabolites with at least two distinct peaks, and therefore complement peak-picking identification approaches, which tend to focus on single peak metabolites.

We found that average correlation profiles (ACP) of highly correlated feature pairs, a method inspired by STOCSY, as well as the iterative signature algorithm (ISA) identified ten and nine metabolites, respectively, five of which overlapped. In contrast, principal component analysis (PCA) did not generate any pseudospectra with robust metabomatching, likely because leading components explain variation driven by many metabolites.

While ACP is designed to pick up individual metabolites with at least two (nonproximal) features in their spectrum (or those of metabolite pairs whose concentrations are coupled), ISA is able to generate modules where many features exhibit coherent variation, yet potentially only over a subset of samples. We believe that this may be particularly useful when integrating data from a heterogeneous set of samples (e.g., including those from diseased or medicated subpopulations).

One interesting property of our modular approach is that the feature sets identified by ACP or ISA do not need to match perfectly with those of the reference spectrum of the corresponding compound. Indeed, the two ACP signatures matching glucose each only cover four and jointly six of the nine glucose peaks, while the ISA module with the best match to glucose includes seven of its peaks. Adding the "missing" peaks in our pseudoquantification slightly reduced the correlation with serum glucose, indicating there is a marginal improvement in the pseudoquantification using ppm positions only from the multiplets found by our algorithms rather than the database. Further work will be needed to substantiate this observation.

Another interesting aspect of our approach is that modular feature sets may match multiple compounds. Our current implementation of metabomatching allows simultaneous identification of up to two compounds. Indeed, our finding that ISA picked up a module whose signature mapped well to ethanol and its specific metabolic product ethyl glucuronide demonstrated the potential power of ISA to identify metabolite pairs within the same pathway. Moreover, the strong correlation of this module with the alcohol abuse marker CDT was likely driven by the fact that ISA can extract context specific covariance, which in this case is strongest in samples with particularly high alcohol consumption. This module also highlighted that using the relevant chemical shifts found by the module rather than all shifts from the reference database can lead to more accurate pseudoquantification of the underlying metabolites, due to different contribution of shifts specific to the experimental conditions in the complex urine spectra.

Extending metabomatching beyond compound pairs is challenging due to the large number of possible trios and higher order combinations, but could be feasible in future work, for example by using metabolic pathway information to limit the number of relevant metabolite combinations to test.

A critical element of our analysis was to transform the signatures generated by the different methods into a universal format (i.e., *z*-scores) as input for our metabomatching tool that we previously developed for the analysis of feature signatures generated by regression on external variables. Indeed, being able to query both internal and external signatures of large-scale NMR data against a reference data set of known spectra from individual metabolites is pivotal for exploring new methods dissecting the auto- and cross-correlation structure for integrative analyses.

In conclusion, we believe that our study using fewer than 1000 samples gives ample evidence for the potential of automated analysis of large-scale NMR data, and that increased sample sizes are likely to result in further identifications and more accurate pseudoquantifications of individual metabolites. To this end, our analysis software, metabomodules, is made publicly available on GitHub https://github.com/BergmannLab/metabomodules-docker.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00295.

Supporting figures and table (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: sven.bergmann@unil.ch.

ORCID

Bita Khalili: 0000-0001-5630-1812

Author Contributions

‖B.K., M.T., and M.M. are co-first authors. R.R. and S.B. are co-last authors. S.B. and R.R. designed the study. The manuscript was written by B.K. and S.B. The correlation-based methods was implemented and applied by B.K. The modular analysis with ISA was performed by M.M. and R.R. PCA was run by M.T., D.K., and B.K. All authors discussed the results and implications, and contributed to the manuscript.

Notes

The authors declare no competing financial interest.

Our analysis software, metabomodules, is made publicly available on GitHub: https://github.com/BergmannLab/metabomodules-docker.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1) and the NIH (grant R03 CA211815).

REFERENCES

(1) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; et al. Statistical Total Correlation Spectroscopy: An Exploratory Approach

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. *Anal. Chem.* **2005**, *77* (5), 1282−1289.
(2) Blaise, B. J.; Navratil, V.; Domange, C.; Shintu, L.; Dumas, M.-E.; Elena-Herrmann, B.; Emsley, L.; Toulhoat, P. Two-Dimensional Statistical Recoupling for the Identification of Perturbed Metabolic Networks from NMR Spectroscopy. *J. Proteome Res.* **2010**, *9* (9), 4513−4520.
(3) Sands, C. J.; Coen, M.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Data-Driven Approach for Metabolite Relationship Recovery in Biological 1H NMR Data Sets Using Iterative Statistical Total Correlation Spectroscopy. *Anal. Chem.* **2011**, *83* (6), 2075−2082.
(4) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M. D.; Nicholson, J. K. Subset Optimization by Reference Matching (STORM): An Optimized Statistical Approach for Recovery of Metabolic Biomarker Structural Information from 1H NMR Spectra of Biofluids. *Anal. Chem.* **2012**, *84* (24), 10694−10701.
(5) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. BATMAN—an R Package for the Automated Quantification of Metabolites from Nuclear Magnetic Resonance Spectra Using a Bayesian Model. *Bioinformatics* **2012**, *28* (15), 2088−2090.
(6) Alonso, A.; Rodríguez, M. A.; Vinaixa, M.; Tortosa, R.; Correig, X.; Julià, A.; Marsal, S. Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis. *Anal. Chem.* **2014**, *86* (2), 1160−1169.
(7) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; et al. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* **2015**, *10* (5), No. e0124219.
(8) Tardivel, P. J. C.; Canlet, C.; Lefort, G.; Tremblay-Franco, M.; Debrauwer, L.; Concordet, D.; Servien, R. ASICS: An Automatic Method for Identification and Quantification of Metabolites in Complex 1D 1H NMR Spectra. *Metabolomics* **2017**, *13* (10), 109.
(9) Röhnisch, H. E.; Eriksson, J.; Müllner, E.; Agback, P.; Sandström, C.; Moazzami, A. A. AQuA: An Automated Quantification Algorithm for High-Throughput NMR-Based Metabolomics and Its Application in Human Plasma. *Anal. Chem.* **2018**, *90* (3), 2095−2102.
(10) Cañueto, D.; Gómez, J.; Salek, R. M.; Correig, X.; Cañellas, N. rDolphin: A GUI R Package for Proficient Automatic Profiling of 1D 1H-NMR Spectra of Study Datasets. *Metabolomics* **2018**, *14* (3), 24.
(11) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608−D617.
(12) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; et al. The Human Urine Metabolome. *PLoS One* **2013**, *8* (9), No. e73076.
(13) Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; et al. The Human Serum Metabolome. *PLoS One* **2011**, *6* (2), No. e16957.
(14) Nagana Gowda, G. A.; Gowda, Y. N.; Raftery, D. Expanding the Limits of Human Blood Metabolite Quantitation Using NMR Spectroscopy. *Anal. Chem.* **2015**, *87* (1), 706−715.
(15) Rueedi, R.; Ledda, M.; Nicholls, A. W.; Salek, R. M.; Marques-Vidal, P.; Morya, E.; Sameshima, K.; Montoliu, I.; Da Silva, L.; Collino, S.; et al. Genome-Wide Association Study of Metabolic Traits Reveals Novel Gene-Metabolite-Disease Links. *PLoS Genet.* **2014**, *10* (2), No. e1004132.
(16) Rueedi, R.; Mallol, R.; Raffler, J.; Lamparter, D.; Friedrich, N.; Vollenweider, P.; Waeber, G.; Kastenmüller, G.; Kutalik, Z.; Bergmann, S. Metabomatching: Using Genetic Association to Identify Metabolites in Proton NMR Spectroscopy. *PLoS Comput. Biol.* **2017**, *13* (12), No. e1005839.
(17) Raffler, J.; Friedrich, N.; Arnold, M.; Kacprowski, T.; Rueedi, R.; Altmaier, E.; Bergmann, S.; Budde, K.; Gieger, C.; Homuth, G.; et al. Genome-Wide Association Study with Targeted and Non-Targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet.* **2015**, *11* (9), No. e1005487.
(18) Ihmels, J.; Bergmann, S.; Barkai, N. Defining Transcription Modules Using Large-Scale Gene Expression Data. *Bioinformatics* **2004**, *20* (13), 1993−2003.
(19) Bergmann, S.; Ihmels, J.; Barkai, N. Iterative Signature Algorithm for the Analysis of Large-Scale Gene Expression Data. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2003**, *67* (3), No. 031902.
(20) Xiong, X.; Liu, D.; Wang, Y.; Zeng, T.; Peng, Y. Urinary 3-(3-Hydroxyphenyl)-3-Hydroxypropionic Acid, 3-Hydroxyphenylacetic Acid, and 3-Hydroxyhippuric Acid Are Elevated in Children with Autism Spectrum Disorders. *BioMed Res. Int.* **2016**, *2016*, 9485412.
(21) Nielsen, H. R.; Killmann, S. A. Urinary Excretion of Beta-Aminoisobutyrate and Pseudouridine in Acute and Chronic Myeloid Leukemia. *J. Natl. Cancer Inst.* **1983**, *71* (5), 887−891.
(22) Ziegler, E. E., Ed.; *Present Knowledge in Nutrition*; Filer, L. J. J., Engl. Ed.; International Life Sciences Inst.: Washington, D.C., 1996.
(23) Nicholas, P. C.; Kim, D.; Crews, F. T.; Macdonald, J. M. Proton Nuclear Magnetic Resonance Spectroscopic Determination of Ethanol-Induced Formation of Ethyl Glucuronide in Liver. *Anal. Biochem.* **2006**, *358* (2), 185−191.
(24) Kim, S.; Lee, M.; Yoon, D.; Lee, D.-K.; Choi, H.-J.; Kim, S. 1D Proton NMR Spectroscopic Determination of Ethanol and Ethyl Glucuronide in Human Urine. *Bull. Korean Chem. Soc.* **2013**, *34* (8), 2413−2418.
(25) Solomons, H. D. Carbohydrate Deficient Transferrin and Alcoholism. *GERMS* **2012**, *2* (2), 75−78.

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368