Network reconstruction for trans acting genetic loci using multi-omics data and prior information

Johann S. Hawe[1,2], Ashis Saha[3], Melanie Waldenberger[4], Sonja Kunze[4], Simone Wahl[4], Martina Müller-Nurasyid[5,6,7,8], Holger Prokisch[9], Harald Grallert[4,10,11], Christian Herder[12,11,13], Annette Peters[10], Konstantin Strauch[5,7,14], Fabian J. Theis[1,15], Christian Gieger[4,10,11], John Chambers[16,17], Alexis Battle[3,18], and Matthias Heinig[1,2,*]

[1] *Institute of Computational Biology, German Research Center for Environmental Health, HelmholtzZentrum München, Neuherberg, Germany*

[2] *Department of Informatics, Technical University of Munich, Garching, Germany*

[3] *Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA*

[4] *Research Unit of Molecular Epidemiology, German Research Center for Environmental Health, HelmholtzZentrum München, Neuherberg, Germany*

[5] *Institute of Genetic Epidemiology, German Research Center for Environmental Health, HelmholtzZentrum München, Neuherberg, Germany*

[6] *IBE, Faculty of Medicine, LMU Munich, 81377 Munich, Germany*

[7] *Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Germany*

[8] *Department of Internal Medicine I (Cardiology), Hospital of the Ludwig-Maximilians-University (LMU) Munich, Munich, Germany*

[9] *Institute of Human Genetics, School of Medicine, Technische Universität München, Munich, Germany*

[10] *Institute of Epidemiology, German Research Center for Environmental Health, HelmholtzZentrum München, Neuherberg, Germany*

[11] *German Center for Diabetes Research (DZD), Neuherberg, Germany*

[12] *Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University, Düsseldorf, Germany*

[13] *Division of Endocrinology and Diabetology, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany*

[14] *Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Munich, Germany*

[15] *Department of Mathematics, Technical University of Munich, Garching, Germany*

[16] *Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK*

[17] *Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore*

[18] *Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA*

[*] *To whom correspondence should be addressed*

May 19, 2020

Abstract

Background: Molecular multi-omics data provide an in-depth view on biological systems, and their integration is crucial to gain insights in complex regulatory processes. These data can be used to explain disease related genetic variants by linking them to intermediate molecular traits (quantitative trait loci, QTL). Molecular networks regulating cellular processes leave footprints in QTL results as so-called *trans* -QTL hotspots. Reconstructing these networks is a complex endeavor and use of biological prior information has been proposed to alleviate network inference. However, previous efforts were limited in the types of priors used or have only been applied to model systems. In this study, we reconstruct the regulatory networks underlying *trans* -QTL hotspots using human cohort data and data-driven prior information.

Results: We devised a strategy to integrate QTL with human population scale multi-omics data and comprehensively curated prior information from large-scale biological databases. State-of-the art network inference methods applied to these data and priors were used to recover the regulatory networks underlying *trans* -QTL hotspots. We benchmarked inference methods and showed, that Bayesian strategies using biologically-informed priors outperform methods without prior data in simulated data and show better replication across datasets. Application of our approach to human cohort data highlighted two novel regulatory networks related to schizophrenia and lean body mass for which we generated novel functional hypotheses.

Conclusion: We demonstrate, that existing biological knowledge can be leveraged for the integrative analysis of networks underlying *trans* associations to deduce novel hypotheses on cell regulatory mechanisms.

Keywords: systems biology, omics, data integration, network inference, prior information, simulation, machine learning, personalized medicine

Background

Genome-wide associations studies (GWAS) have been tremendously successful in discovering disease associated genetic loci. However, establishing causality or obtaining functional explanations for GWAS SNPs is still challenging. In recent years, the focus has shifted from discovery of disease loci to mechanism and explanation, and large efforts have been put into unravelling the functional consequences of GWAS SNPs [1, 2]. These have been made possible through technological advances in measuring genome-wide molecular data in large population cohorts, which further led to a steady increase in biological resources providing simultaneous measurements of different molecular layers (often termed *multi-omics* data). To elucidate disease mechanisms, systems genetics approaches seek to link GWAS SNPs to intermediate molecular traits by identifying quantitative trait loci (QTL) [3, 4], for example for gene expression levels (eQTL) [5–7] or DNA methylation at CpG dinucleotides (meQTL) [8–10].

Genetic variants that are QTL for quantitative molecular phenotypes that reside on a different chromosome are called *trans* -QTL. Previously, *trans* -QTL studies were successful in model systems [11, 12]. Recently, large-scale meta analyses of molecular QTL in very large sample sizes have now been applied to successfully map large numbers of *trans* -QTL in humans [7]. These are particularly interesting, as they have been found to be enriched for disease associations [7, 8, 13]. Yet, the underlying mechanisms leading to such associations can usually not be explained in a straightforward way [6], and in fact, 83% of discovered *trans* -eQTL in human are estimated to still be unexplained [7].

Trans-QTL hotspots [14], where a single genetic locus influences numerous quantitative traits on different chromosomes, can be seen as footprints of regulatory molecular networks and likely encode master regulators. One way of mechanistically explaining the effects of these master regulators is by reverse engineering the regulatory networks, and hence de-

termining the intermediate molecular processes giving rise to the observed *trans* effects, ultimately yielding novel insights into disease pathophysiology [1, 14–16].

A large body of work has focused on inferring regulatory interactions from high-throughput data by individually combining distinct genomic layers like gene expression levels and genotype [6, 17–19] or chromosomal aberration [20] information. Generally, network inference to uncover regulatory mechanisms in biological systems has gotten much interest [15, 21–24]. The emergence of multi-omics data now also allows for establishing networks across more than two omics layers in a holistic approach to obtain more insight into the function of regulatory elements [16]. Major efforts have been made to recover functional interactions from such data, but methods to successfully reverse engineer regulatory networks across multiple omics layers are still lacking [1, 4, 25, 26].

Furthermore, utilizing the wealth of data available from genomic databases as biological prior information can guide the inference of complex multi-omics networks [26–28]. For instance, using known relationships discovered in previous studies as prior knowledge, such as protein-protein interactions (PPIs) or eQTL, can facilitate network reconstruction on novel datasets. Application of priors has been investigated in numerous works [e.g. 15, 27, 29–34], and while several studies show the advantage of using priors in synthetic datasets [22, 31, 33, 34] or model systems [15, 32, 34, 35], relatively few studies apply their inference methodologies to functional genomics data in humans [29, 33, 36, 37]. In case human data is considered, either cell line data are used [36], the inference is restricted to a single pathway [37] or no informative priors are used for this specific context [29]. Zuo *et al.* apply prior based inference to human cancer gene expression data, however, they only use priors based on PPIs extracted from the STRING database and focus on differential expression analysis [33]. What is still missing, is, to comprehensively integrate the vast amount of functional data from large-scale databases [38–41] as prior information in human multi-omic *trans* -QTL studies and to determine the appropriate inference methods.

Here, we developed a novel approach for understanding the molecular mechanisms underlying the statistical associations of *trans* -QTL hotspots by integrating existing biological knowledge and available multi-omics data to infer regulatory networks. We derived a comprehensive set of continuous priors from public datasets such as GTEx, the BioGrid and Roadmap Epigenomics and applied state-of-the-art network inference methods including graphical lasso [42], BDgraph [29] and iRafnet [32], and showed, that methods using data-driven priors outperform non-prior approaches for network reconstruction on simulated data. Moreover, we showed that networks inferred on real-world data using priors can be replicated more faithfully across independent datasets than networks inferred without priors. Finally, we demonstrated, that incorporating existing knowledge with multi-omics data yields novel insights into disease related cellular mechanisms when applied to real-world population cohort data of different omics types and tissues.

Results

Trans-QTL hotspots define regulatory network candidates

In this study, we aimed to reconstruct regulatory networks to explain *trans* quantitative trait locus (*trans* -QTL) hotspots on a molecular level through simultaneous integration of multi-omics data [4]. *Trans*-QTL hotspots have previously been associated with disease [8, 13], and understanding their mechanisms of action can deepen our insights into regulatory pathways and, ultimately, into the disease process.

Our general analysis strategy is depicted in Figure 1A and consists of the following steps: 1) curate QTL hotspots, 2) gather functional data and prior information, 3a+b) benchmark network inference methods in simulation and replication study to select best suited method and 4) infer and interpret networks identified in the cohort data.

We obtained *trans* hotspots from the methylation QTL (meQTL) discovered in whole-

blood in the KORA [43] and LOLIPOP [44] cohorts reported by Hawe and colleagues [10] and the expression QTL (eQTL) published by the eQTLGen consortium [7], yielding a total of 107 and 444 *trans* -loci per QTL type, respectively (Figure 1B, see Methods for details). In addition to the whole-blood derived hotspots, we curated a single *trans* -eQTL hotspot in Skeletal Muscle tissue from GTEx v8 [38, 39], which we analyzed separately.

For each hotspot, we aimed to identify the causal gene at the genetic locus affected by the SNP and the intermediate genes which mediate the observed *trans* associations. To this end, we collected sets of candidate genes with different roles for each locus, which we term 'locus sets' (see Methods). A locus set contains the SNP defining the hotspot, the respective *trans* associated traits (CpGs for meQTL and genes for eQTL, 'eGenes'), *cis* genes encoded near the SNP as candidate causal genes, *trans* genes (for meQTLs, genes in vicinity of the CpGs), as well as transcription factors (TFs) binding near the *trans* associated entities and PPI genes residing on the shortest path between *trans* traits and *cis* genes in a protein-protein interaction (PPI) network, as potential intermediate genes. *Cis* genes form potential candidate regulator genes of the locus, and the inclusion of the PPI and TF binding information allows us to bridge the inter-chromosomal gap between the SNP and the *trans* CpG sites/*trans* eGenes. An overview of entities collected over all loci for both QTL types is given in Figure 1C.

One main aspect of this work is the use of any form of biological prior information, including continuous scores, to guide network inference. We hence collect prior information for all possible edges between entities contained in locus sets in addition to the functional data (Figure 1). In total, four distinct types of edges are annotated with prior information: *SNP-Gene*, *Gene-Gene*, *TF-CpG*/*TF-Gene* and *CpG-Gene* edges. All prior information is generated from matched, public data independent of the data used during network inference (see Methods for details).

Figure 1D indicates the total number of edges annotated with prior information over

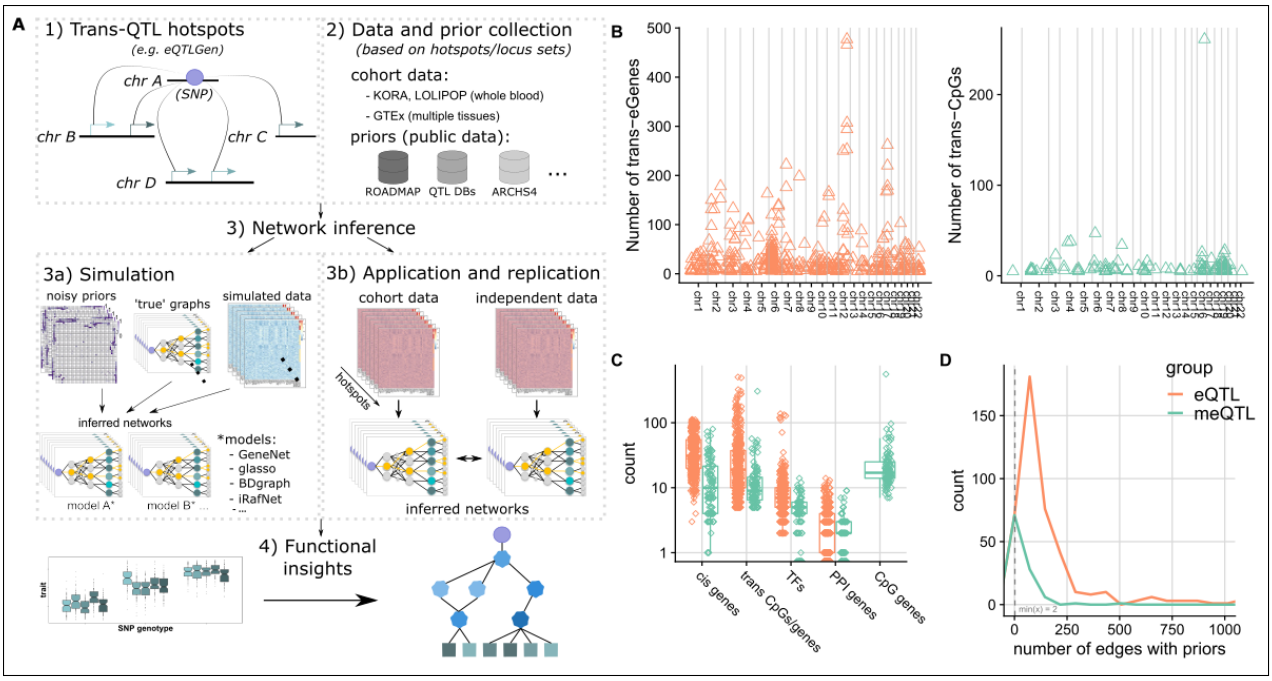

Figure 1: Project overview. Panel **A**) shows a graphical abstract of the analyses performed in this project. Panel **B**) provides a global view on the collected eQTL (orange) and meQTL (green) hotspots. The x-axis indicates ordered chromosomal positions for *trans* eGenes and CpG sites, respectively. Panel **C**) shows the total number of different genomic entities gathered over all hotspots during locus set creation (log scale). Panel **D**) depicts density plots of the number of possible network edges with available prior information (x-axis) over all hotspots, zoomed in to area between 0 and 1000. Same color coding is used in panels **B**-**D**.

all hotspots. For meQTL and eQTL, a minimum of 2 and 3 edges per hotspot show prior evidence, respectively, and most hotspots get only relatively few priors compared to the total number of possible edges (median 26 and 94, respectively). However, in both cases several networks collect priors for over 100 edges (8 and 209 loci with >= 100 priors for meQTL and eQTL). As expected, the total number of edges with prior information per locus correlates with the total number of possible edges in the respective loci, however, the fraction of all possible edges annotated with prior information decreases (Additional File 1, Figure S2).

Benchmark of network inference methods

Simulation study shows benefit of data-driven priors

Numerous methods for regulatory network inference have been proposed (e.g. [42, 45, 46], see also [4]), and, therefore, before investigating individual hotspots in detail we sought to select the method best suited for this study (see Figure 1A step 3). To this end, we performed an extensive simulation study (Figure 1A step 3a) to evaluate the performance of five distinct methodologies (see Table 1 for a method overview) in reconstructing ground truth graphs from simulated data and prior information. Simulated data were matched with the observed QTL-hotspots by preserving the sample size and the total number of input nodes and 100 simulations were performed for each hotspot. We evaluated the impact of priors for different sample sizes by sub-sampling the simulated data and using the full prior matrix. To assess the impact of noise in priors, we inferred networks separately from prior information with varying degrees of noise (up to 100%, see Methods for details) for the complete data.

name	**version**	**repository**	**attribute**	**reference**
BDgraph	2.61	CRAN	MCMC	Mohammadi and Wit (2015) [29]
gLASSO	1.11	CRAN	Graphical lasso	Friedman *et al.* (2008) [42]
GENIE3	1.2.1	bioconductor	Random forests	Huynh-Thu *et al.* (2010) [46]
GeneNet	1.2.13	CRAN	Shrinkage/ FDR	Opgen-Rhein *et al.* (2007) [45]
iRafNet *	1.1-2	CRAN	Random forests	Petralia *et al.* (2015) [32]

Table 1: Overview of the network inference packages used in the simulation study.
** adjusted to make use of parallel processing, see Methods*

We gauge the relative gain in performance attributable to prior information for both *gLASSO* and *BDgraph* by always training two distinct models, one utilizing the provided priors ($gLASSO_P$, $BDgraph_P$) and one without priors (*gLASSO* , *BDgraph*). The implementation of *iRafNet* always requires a prior matrix, whereas both *GeneNet* and

186 $GENIE3$ cannot utilize prior information and hence were trained only with the simulated
187 data. We utilize Matthews Correlation Coefficient (MCC) [47] as a balanced performance
188 measure to compare inferred networks to the respective ground truth (see also [29]). Fig-
189 ures 2A and 2B show the results for the simulation study for all methods (see also Additional
190 File 1, Tables S2, S3, S4 and S5). Overall, both $gLASSO_P$ and $BDgraph_P$ exhibit improved
191 performance with relatively low standard deviation in terms of MCC as compared to their
192 non-prior counterparts, both for low and high sample size settings. The performance of all
193 other methods is affected by low sample sizes, with $BDgraph$ showing slightly better perfor-
194 mance than all other methods. Moreover, both $gLASSO_P$ and $BDgraph_P$ outperform all
195 other methods as long as the prior noise does not exceed 10% ($gLASSO_P$) and 30% of incor-
196 rect edges in the prior graph, in which case $BDgraph$ achieves the highest median MCC over
197 all methods. $GeneNet$ performs well in all simulations, whereas $GENIE3$, $gLASSO$ and
198 $iRafNet$ show about average performance with $iRafNet$ achieving worst results overall.
199 In addition to the curated prior matrices, we also generated a prior matrix reflecting the
200 sparsity of the true graph (column 'rbinom' in Figure 2B and Additional File 1, Tables S2
201 and S3, see also Methods), and our results indicate, that information about sparsity of the
202 underlying network already improves network inference performance. Finally, prior based
203 methods, and specifically $BDgraph_P$, outperform non-prior methods in the task of identify-
204 ing the correct *cis* -gene by recovering associations between the discrete SNP and continuous
205 gene expression data types (Additional File 1, Figure S3), when using independent eQTL
206 data as prior.

207 Inferred networks replicate in independent datasets

208 In addition to the simulation study, we evaluated the methods on real world data from
209 two large population cohorts: the KORA (Cooperative Health Research in the Region of
210 Augsburg) and LOLIPOP (London Life Sciences Population) cohorts (see Figure 1A2 and

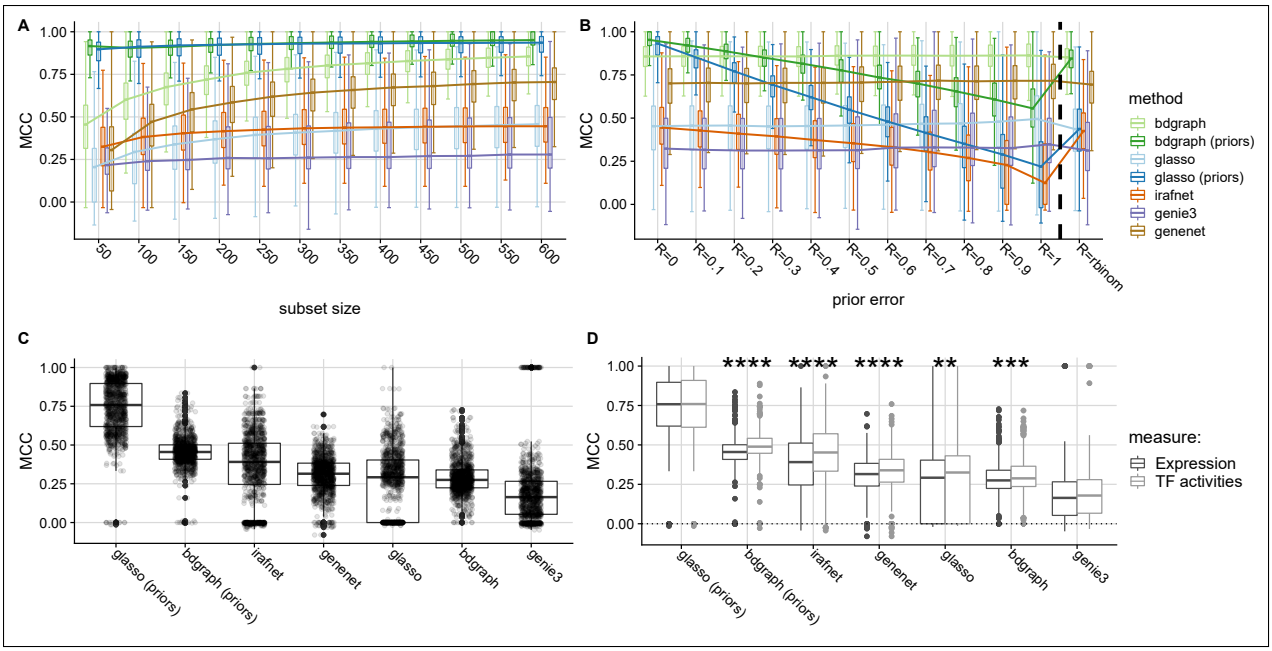

Figure 2: Method comparison results. **(A)** Results of simulation study: y-axis shows the Matthews correlation coefficient (MCC) as compared to the simulated ground truth, x-axis indicates increasing sample size from left to right, colors indicate different inference methods. **(B)** Similar to (A), but x-axis indicates increasing noise in the prior matrix from left to right. Group ('*rbinom*') indicates uniform prior set to reflect degree distribution of true graph. **(C)** shows MCC (y-axis) between networks inferred on KORA and LOLIPOP data for same locus for all methods (x-axis). **(D)** contrasts MCC across cohorts using TF expression (dark gray) versus using substituted TFAs (light gray). Boxplots show medians (horizontal line) and first and third quartiles (lower/upper box borders). Whiskers show $1.5 * IQR$ (inter-quartile range); for (**B**), dots depict individual results and for (**C**), stars indicate significant difference between expression/TFA results for each method (Wilcoxon test, **: $P \leq 0.01$, ***: $P \leq 0.001$, ****: $P \leq 0.0001$)

Methods). Data from both cohorts were generated from whole-blood samples and contain imputed genotypes as well as microarray measurements of gene expression and DNA methylation for a total of 683 (KORA) and 612 (LOLIPOP) samples. Since for these data no ground truth is available, we evaluate robustness of the networks inferred by the individual methods via cross cohort replication. For each hotspot, we collect data for all genes, CpGs and the SNP in the locus set for KORA and LOLIPOP and separately inferred networks in both cohorts for all models. Obtained networks were then compared between cohorts

using MCC to get a quantitative estimate of how robust the network inference is across different datasets for the same hotspot, yielding scores for KORA versus LOLIPOP and vice versa (i.e. one network functioning as the reference). Results of this analysis are shown in Figure 2C. With respect to MCC, models supplied with prior information ($gLASSO_P$, $BDgraph_P$ and $iRafNet$) show the best performance, with $gLASSO_P$ coming up as the most robust method, followed by $BDgraph_P$ and $iRafNet$. Noticeably, of the top methods $BDgraph_P$ shows much less variance compared to $gLASSO_P$ and $iRafNet$. Ignoring prior information lead to a drop in performance for both $gLASSO$ and $BDgraph$, which leads to $GeneNet$ outperforming both methods. Finally, $GENIE3$ shows worst performance in this setting.

Estimated transcription factor activities as a proxy to TF activation

Transcription factor activities (TFAs) estimated from transcription factor binding sites (TFBS) and gene expression data have been suggested as an alternative to using TF gene expression in inference tasks [48], since a transcription factor's expression level alone might not reflect the actual activity of a TF (driven for instance by its phosphorylation state). To evaluate, whether TFAs could improve our inference, we estimated TFAs for all TFs based on their expression and ChIP-seq derived TFBS from ReMap [49] and ENCODE [50, 51] (see Methods for details). We applied the same cross cohort replication strategy as above and compared MCCs from the TFA based analysis to the previous results using a one-sided Wilcoxon test. Figure 2D shows the results of TFA (light gray boxes) versus gene expression (dark gray boxes) based analysis in terms of MCC for all available hotspots. For all models but $gLASSO_P$ and $GENIE3$, TFAs yield a significantly higher MCC (Wilcoxon test $P < 0.01$) as compared to using the pure expression data (see also Additional File 1, Table S6).

According to the results presented above, detailed investigation of real world data was

focused on networks obtained from $gLASSO_P$ and $BDgraph_P$ and TF expression was substituted by TFA estimates for all subsequent analyses.

Replication of previous findings by simultaneous data integration

Before seeking new mechanistic insights and generating novel hypotheses from *trans* -QTL hotspots, we first checked whether our approach can replicate previous findings. Hawe *et al.* [10] inferred gene regulatory networks from *trans* -meQTL hotspots using a two-step approach involving 1) a random walk on a PPI and ChIP-seq based networks and 2) subsequent local correlation analysis. In contrast, our approach simultaneously integrates all functional data, relying on PPI and ChIP-seq information as prior knowledge, thereby avoiding the need for post-hoc correlation testing of e.g. SNP-gene and CpG-gene edges. For the comparison, we extracted three of their hotspot networks and evaluated the overlap with the networks inferred in this study.

locus	num. nodes	num. edges	common edges	MCC
rs9859077	99 (89)	447 (287)	141	0.52
rs730775	58 (49)	98 (67)	48	0.69
rs7783715	25 (17)	24 (23)	5	0.65

Table 2: Comparison of the networks inferred in this study to the networks extracted from [10]. Numbers in bracket indicate statistics for the networks from the original publication.

Table 2 shows the results of this comparison. Overall, the comparisons indicate relatively strong concordance between the two approaches with MCCs of 0.515, 0.689 and 0.65. Moreover, for all three networks, our simultaneous inference approach yielded more edges and nodes than the two-step approach (56%, 46% and 4% novel edges and 11%, 19%, 47% additional nodes for rs9859077, rs730775 and rs7783715, respectively), which might have been missed by the two-step approach, as it relies on known PPI and ChIP-seq information.

Figure 3 contrasts the two networks obtained for the *rs730775* hotspot using 1) the two-step approach by Hawe *et al.* [10] and 2) the network inferred in this study using $gLASSO_P$,

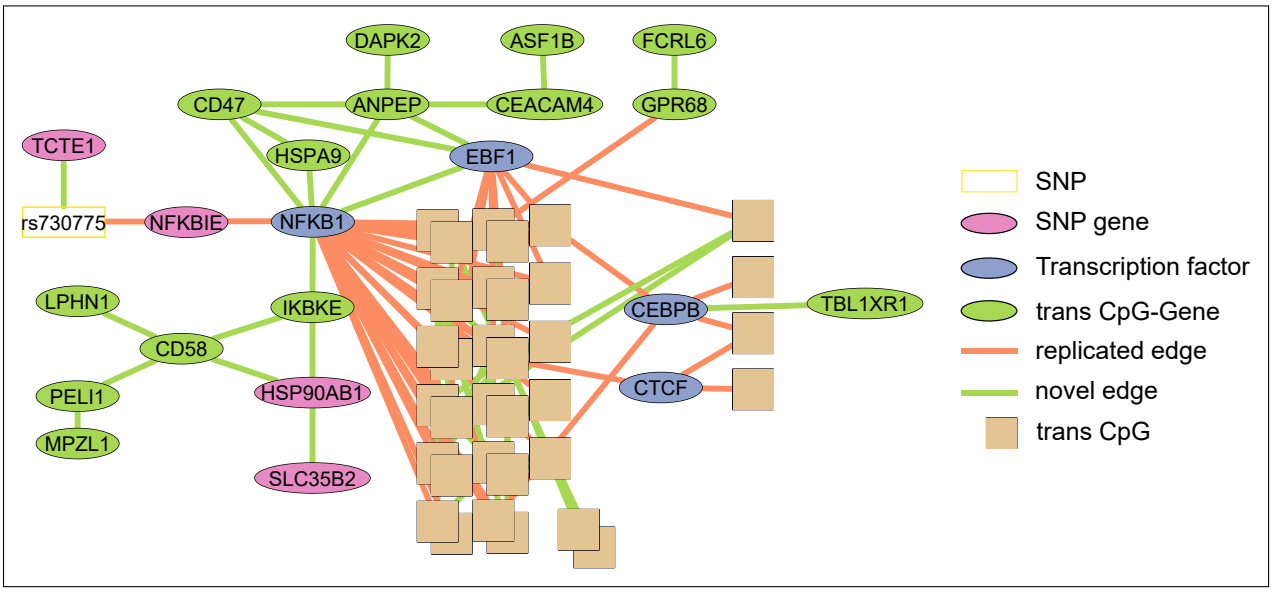

Figure 3: Comparison of the random walk based network reported in [10] and the network inferred from functional omics data in this study for the rs730775 locus. Shown is the complete network constructed from the omics data, edge color indicates replication/novelty. Orange edges: replicated with respect to the random walk network. Green edges: novel in our network. White box: SNP; pink nodes: SNP-genes; blue nodes: TFs; brown boxes: CpGS; green nodes: CpG-genes.

orange edges showing replicated and green edges indicating novel edges. In Hawe *et al.* [10], the authors described a regulatory network involving the *rs730775* SNP connected via *NFKBIE* to *NFKB1* which connects to the trans-CpG sites. This main pathway is also discovered in our approach (i.e. *rs730775* $\leftrightarrow$ *NFKBIE* $\leftrightarrow$ *NFKB1* $\leftrightarrow$ *CpG sites*), in addition to some of the initially reported TFs (blue nodes), of which *NFKB1* is connected to most of the *trans* CpGs (82%, 29 out of 35) as was the case in the original network. However, we also identify patterns of CpG genes (green nodes) connected to the TFs, which were not previously identified. Overall, the integrated approach using prior information leads to high replication of previous networks including novel connections leading to potential new insights in target gene regulation.

A trans regulatory network for a schizophrenia susceptibility locus

In order to demonstrate the effectiveness of our approach in getting mechanistic insights from *trans* -QTL associations, we inferred networks for all meQTL [10] and eQTL [7] hotspots using whole blood data from the KORA and LOLIPOP cohorts using the prior based $gLASSO_P$ and $BDgraph_P$ models (see Methods, all networks are listed in Additional File 2, Table S3). Based on the GWAS catalog (v1.0.2, [52]), graph properties and a custom graph score (see Methods), we prioritized a *trans* acting locus that has previously been associated with schizophrenia (SCZ).

The network involves the *trans* -eQTL locus around the *rs9469210* (alias *rs9274623*[1]) SNP in the Human Leukocyte Antigen (HLA) region on chromosome 6 shown in Figure 4A.

rs9274623 has been associated with SCZ [54] and is a *cis* -eQTL for all three of its directly connected SNP-genes, *PBX2*, *RNF5* and *HLA-DQA1* in the eQTLGen study. *RNF5* showed differential expression for SCZ cases vs controls in addition to its expression being associated with an additional independent SCZ susceptibility SNP (rs3132947, $R^2 = 0.14$

[1]according to SNiPA: https://snipa.helmholtz-muenchen.de/snipa3/, [53]

in 1000 genomes Europeans[2]) located in the HLA locus [55]. Interestingly, *PBX2* has been associated with a SCZ related phenotype in a pharmacogenetics study (clozapine-induced agranulocytosis) [56, 57] and shows direct binding evidence to the *SPI1* promoter region (ReMap TFBS [49]). The transcription factor *SPI1* (*PU.1*) is linked to Alzheimer's Disease likely by impacting neuroinflammatory response [58] and was found to interact with its network neighbor,*RUNX1*, in modulating gene expression [59]. Moreover, *RUNX1* has been implicated in rheumatoid arthritis, a disease negatively associated with SCZ and which hence might share susceptibility genes with SCZ [60]. Interestingly, several genes encoded in the HLA locus, which has been implicated in SCZ and other psychiatric and neurological disorders [61–64], were picked up by our inference downstream of *SPI1* and *RUNX1*. *TCF12* is a paralog of *TCF4* and *TCF3* which are known E-box transcription factors and are expressed in multiple brain regions [65]. *TCF4* loss-of-function mutations are the cause of Pitt-Hopkins syndrome (a syndrome causing mental retardation and behavioral changes amongst other symptoms) [66] and regulatory SNPs relating to *TCF4* have been associated with SCZ [67, 68]. The *NFKB1* pathway has been recognized as an important regulatory and developmental factor of neural processes and was found to be dysregulated in patients with SCZ [69] . Finally, 9 of the 40 discovered *trans* -eGenes of the locus are connected to the SNP via the selected TFs. Of these, *SH3BGRL3* [70] has already been linked to SCZ and *PSEN1* [71], *B9D2* [72], *CXCR5* [73] as well as *DNAJB2* [74] were implicated in other neurological disorders. In addition, the *trans* eGene *RNF114* has previously been shown to play a role in the *NFKB1* pathway [75]. A formal colocalization analysis using fastENLOC [76] showed evidence of a common causal variant underlying the SCZ GWAS signal [77] and each of the eQTLGen *trans* -eQTL of PSEN1, DNAJB2 and CD6 (SNP-level colocalization probability of 0.92, 0.87 and 0.42, respectively; see Methods and Additional File 1, Figure S4).

Our approach highlighted a potential regulatory pathway involving diverse genes related

[2]https://ldlink.nci.nih.gov/?tab=ldmatrix

to SCZ and other neurological disorders. While some of the genes were not previously reported in this specific disease context (e.g. *CD6*, *BRD2*, *DEF8*), their association to this network indicates a potential role in SCZ pathogenesis and additional colocalization analysis hints at a potential causal relationship between these genes and SCZ.

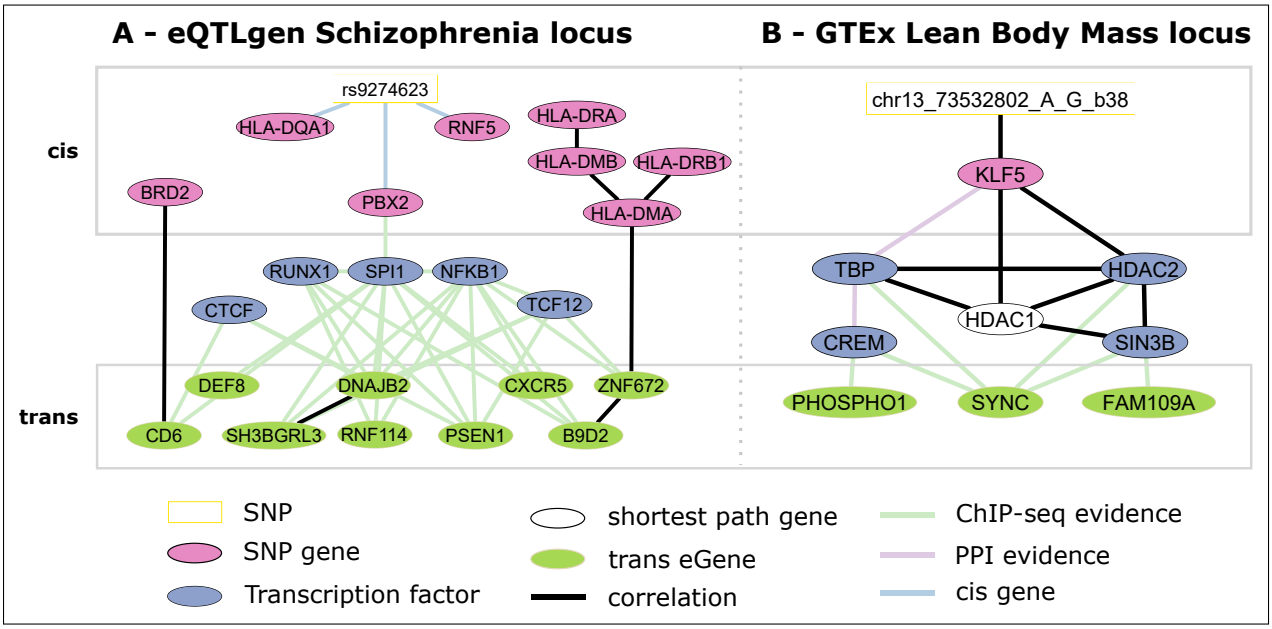

Figure 4: Inferred networks for the schizophrenia susceptibility locus rs9274623 obtained from eQTLgen (A) and the rs9318186 locus obtained from GTEx (B). The white boxes indicate sentinel SNPs, pink ovals indicate SNP-Genes, blue ovals transcription factors and white ones shortest path derived genes. Light green ovals represent genes trans-associated to the SNP. Black edges were inferred during network inference. In addition to being inferred, colored edges indicate ChIP-seq protein-DNA binding evidence (green), protein-protein interaction in the BioGrid (purple) and whether or not a gene is encoded in *cis* of the linked entity (blue).

Application to GTEx Skeletal Muscle tissue

All above analyses were focused on whole-blood data, however, the proposed strategy can be applied to data from any biological context. To demonstrate this, we investigated the recently published *trans* -eQTLs from the GTEx v8 release [38, 78][3]. We identified a single

[3]https://www.gtexportal.org/

LD block in Skeletal Muscle tissue, which is a *trans* -eQTL hotspot (see Methods), and for which we inferred regulatory networks. Since we can't use the same priors, which were initially derived from GTEx, to analyze the same data set, we set out to curate muscle tissue specific priors from independent datasets. We utilized muscle eQTL from Scott *et al.* (2016) [79] and gene expression data curated from the ARCHS[4] [41] database and generated tissue specific TFBS using factorNet [80] on DNAse-seq data obtained from ENCODE [50, 51][4] (see Methods for details). The resulting network for the $gLASSO_P$ model is shown in Figure 4B.

The genetic variant rs9318186 is a *cis* -eQTL of *KLF5* in GTEx v8 Skeletal Muscle ($P = 6.1x10^{-37}$) and a proxy of it ($R^2 = 0.88$) has been associated with *Lean Body Mass* (LBM). *KLF5* itself, too, has been associated with LBM in a transcriptome-wide association study integrating GWAS results with gene expression [81] and with lipid metabolism in *KLF5* knockout mice [82]. In addition, several other genes in the network have been associated with related phenotypes: Both *HDAC1* and *HDAC2* have been found to control skeletal muscle homeostasis in mice [83], work together with *SIN3B* in the SIN3 core complex to regulate gene expression and are involved in muscle development [84]. TATA binding protein (*TBP*) is a well known transcription factor and important for the transcriptional regulation of many eukaryotic genes [85]. The *trans* -eGene *SYNC* was found to interact with dystrobrevin (*DMD* gene) in order to maintain muscle function (during contraction) in mice as well as being associated with neuromuscular disease [86, 87]. In addition, in Seim *et al.* (2018) [88], the authors investigated the relationship between obesity and cancer subtypes and found, that both *PHETA1/FAM109A* expression are associated to Body-Mass-Index (BMI) in esophageal carcinoma in data from The Cancer Genome Atlas (TCGA). *PHOSPHO1* has been found to be involved in metabolism, specifically in energy homeostasis [89], and has also been associated via DNA methylation with BMI [90, 91] and with HDL levels, which have been negatively associated with LBM [92]. Dayeh *et al.* (2016) [93] further showed decreased

[4]https://www.encodeproject.org/

DNA methylation at the *PHOSPHO1* locus in skeletal muscle of diabetic vs. non-diabetic samples. The remaining gene in the network (*CREM*) has not yet been described in the broader context of LBM, but a GWAS meta-analysis executed by Wang *et al.* (2014) [94] hinted at association of a *CREM* SNP (rs1531550, $P = 1.88x10^{-6}$) with elite sprinter status. These results suggest, that *KLF5* may exert its specific functions through transcriptional regulation via the SIN3 core complex including *TBP*, with a potential involvement of *CREM*, of the *trans* -eGenes *PHOSPHO1*, *SYNC* and *PHETA1/FAM109A*.

Discussion

In this study, we introduced a Bayesian framework for the inference of undirected regulatory networks underlying molecular *trans* -QTL hotspots across multi-omics data types using existing prior knowledge. We compiled a comprehensive set of context specific network edge priors from diverse biological databases and applied these together with multi-omics data in different settings. These settings include an extensive simulation study to benchmark state-of-the-art inference methods as well as application to two large population cohorts, which we use for a replication analysis on the one hand and to generate novel hypotheses about molecular disease mechanisms on the other hand. Moreover, by applying our approach a GTEx Skeletal Muscle eQTL hotspot, we showed, that our strategy can be applied to data sets from other tissues, generated with different technologies.

Benchmarking is important for selecting the best possible methods for specific tasks and we hence followed recently published guidelines [95] to perform benchmarking of state-of-the-art network inference methods in 1) a simulation study and 2) a replication analysis. Results from both analyses were then used to select the methods best suited for network inference based on functional multi-omics data from QTL hotspots using prior information.

By inferring networks in over 10,000 simulated data sets, which reflect the distribution of

network parameters obtained from real-world data, we showed, that methods utilizing prior information outperform methods without any prior information in recovering a simulated ground truth, similar to what has been found e.g. in [27, 28, 36]. We further observed that, as expected, too much noise in the prior information significantly reduces method performance. However, only by increasing the noise level, i.e. the percentage of incorrect prior edges, to above 30% decreases the performance for BDgraph below the performance of its non-prior counterpart, indicating that low levels of noise in edge priors still improve network inference, results which are in line with e.g. Wang *et al.* (2013) [30], who used a modified graphical lasso approach, Christley *et al.* (2009) [28], who used an regularized ODE model and Greenfield *et al.* (2013) [27], who used a Bayesian regression framework. We further find, that, both for the prior and non-prior case, the Markov-Chain-Monte-Carlo based $BDgraph_P$ method outperforms respective other methods. However, both the copula approach based BDgraph and the $gLASSO_P$ outperform other methods in recovering mixed edges between discrete SNP allele dosage and continuous gene expression levels, although the tree based methods should be able to incorporate mixed data. While $BDgraph_P$ shows overall better performance than $gLASSO_P$, the graphical lasso exhibits much lower run time which can be an important practical consideration. Our results hence highlight the strong value of using prior information for multi-omics based network reconstruction, and slightly favor BDgraph over the graphical lasso for this kind of inference.

We confirmed the results of the simulation study by extended benchmarking of inference methods in a cross cohort replication analysis on two large multi-omics data sets. Prior based methods showed overall best replication across different cohorts as compared to non-prior methods. In the real-world setting, however, $iRafNet$ performed similarly well as the other two prior methods in contrast to the simulation study and all prior based methods outperform non-prior methods. The good replication of prior based methods across different cohorts shows, that curated priors help to obtain more stable and confident results as com-

pared to using functional data alone. Together with the simulation, these results provide a comprehensive benchmark of established network inference methods and suggest, that priors should be integrated in network inference tasks wherever possible.

Based on the results from the replication and simulation study, we choose the two best (prior based) methods $BDgraph_P$ and $gLASSO_P$ for detailed investigation of networks obtained form real-world cohort data. Using our integrative approach, we were able to reproduce and expand upon previous results from a step-wise network analysis approach presented in [10]. Of three of the locus networks described in their study, we reconstructed most of the edges and found additional edges, allowing more mechanistic interpretations for the function of specific transcription factors in relation to DNA methylation. One reason for finding additional edges is, that these could not be detected by the previous approach, since the authors focused on using established PPI and protein-DNA interactions and did not test all possible edges in the functional data. In contrast, our integrated approach considers all edges regardless of available prior evidence and associations will emerge, if the signal in the functional data alone or in addition to the prior evidence is strong enough.

Next, we utilized the two top performing methods ($BDgraph_P$ and $gLASSO_P$) to infer networks from *trans* -eQTL hotspots and found, that our strategy can be used to recover known biology on the one hand and generate novel hypotheses about the molecular basis of diseases on the other hand. For a schizophrenia (SCZ) susceptibility locus, we identified several known SCZ (e.g. *RNF5, HLA genes* [55, 61]) and related (e.g. *PBX2* [56, 57]) genes in the inferred locus network. Caution is needed for the interpretation of the candidates based on *cis* -eQTL, because of the haplotype structure of the HLA locus. However, our candidate *PBX2* is defined by its connections in the network to the *trans* genes and, therefore, independent of the *cis* eQTL. Expanding upon similar previous observations based on *trans* eQTL [7], the integrated network analysis including associated *trans* genes prioritizes *PBX2*, which was not possible using *cis* -eQTL alone. It was previously hypothesized, that

RUNX1 is involved in SCZ due to a negative association of SCZ with rheumatoid arthritis [60]. Our network corroborates this hypothesis and further allows for generating novel hypotheses about the involvement of other genes (e.g. *BRD2*, *DEF8* and *RNF114*), which could potentially play a role in schizophrenia. Moreover, we further substantiated these results by a formal colocalization analysis of the *trans* -eQTL and schizophrenia GWAS [77] signals of the *trans* genes linked in the network, which revealed strong evidence for colocalization of the underlying genetic variants of the disease and molecular traits. As this locus was derived from whole-blood data, interpretation is not straight forward for SCZ. Ideally, this analysis can be followed up in data derived from brain tissue to corroborate findings.

To show, that our approach can be applied across different omics types and data sets, we analyzed a Skeletal Muscle *trans* -eQTL hotspot from GTEx associated with Lean Body Mass. We recovered known genes involved in lipid metabolism (*KLF5* [81, 82]) as well as muscle development and controlling skeletal muscle homeostasis (e.g. *HDAC1*, *HDAC2*, [83]) and maintaining muscle function (*SYNC* [87]). This shows, that the genes linked in the inferred network are overall coherent with the observed phenotype association at this *trans* -acting locus. Moreover, *HDAC1*, *HDAC2* and *SIN3B* have been described to interact together during muscle development [84], and, although these results were described in mice, our results suggest that these genes could exhibit a similar function in human. In addition, we observed an association between *CREM* and *SYNC* in our network, which led us to hypothesize, that *CREM* might also be involved in maintaining muscle function and Lean Body Mass, although is has not been previously linked to these phenotypes. However, additional experimental validation needs to be performed in order to corroborate findings of these computational analyses.

Several practical considerations arise from our findings: First, by investigating the effect of increasing amounts of noise in the prior information in our simulation study, we showed, that some caution needs to be applied when curating continuous prior information from

public biological data to keep noise levels low. Therefore, although $gLASSO_P$ and especially $BDgraph_P$ seem to be robust to low to moderate levels of noise, one might consider using only experimentally validated protein-protein interactions or high quality gene expression data to generate priors. Next, the definition of hotspot locus sets and priors in this study mitigates the $N << P$ problem. This has been a problem sought to be alleviated using specialized approaches in previous applications [4]. Using our approach, the total number of entities (variables) going into the network inference typically does not exceed the total number of available samples in our data sets, and we showed in a simulation study, that priors improve inference also in low sample size settings. Overall, the benefit of the locus sets comes with the risk of missing certain genes needed to fully describe the *trans* effects. For instance, we reason that most relevant genes lie on the shortest path between *cis* and *trans* entities in the PPI network and hence only included those shortest path genes. However, our strategy of curating a stringent set of relevant transcription factors as well as including genes showing protein-protein interactions and all the genes in the vicinity of the hotspot SNP, should enable most key regulator genes to enter the inference process and yields parsimonious and easily interpretable results. In addition, methods have been developed to handle mixed data types, such as e.g. genotypes and gene expression. $BDgraph$, which uses a copula based approach to transform non-normal data, showed better performance in recovering associations between discrete and continuous data types as compared to $gLASSO$ and the tree based methods, and hence should be preferred for applications on mixed data, especially when prior information is available. Finally, while we could use transcription factor binding sites (TFBS) in blood related cell-lines to analyze whole-blood cohort data, context (e.g. tissue) specific TFBS are not yet available for a large number of transcription factors, which potentially limits this approach to fewer applications. However, novel developments to predict TFBS from context specific open chromatin information (e.g. *factorNet* [80]) can help in carrying this strategy to more contexts. As an example, we utilized TFBS predicted using *factorNet* based

on ENCODE [50, 51] DNAse-seq data for analyzing a GTEx Skeletal Muscle *trans* eQTL locus.

Conclusion

This study describes a novel strategy for using comprehensive edge-wise priors from biological data to improve network inference for *trans* -QTL hotspots from human population scale multi-omics data. This facilitates the investigation of their underlying regulatory networks and enables the generation of novel mechanistic hypotheses for disease associated genetic loci. Moreover, we report a rigorous benchmark of state-of-the-art network inference methods for this task both in simulated and real-world data, and highlight the benefit of including biological prior information to guide network inference.

Methods

Cohort data processing

Methylation data were measured using the Infinium Human Methylation 450K BeadChip in both the KORA and the LOLIPOP cohort and methylation beta values obtained as described previously [43, 44]. Quantile normalized methylation beta values were adjusted for Houseman blood cell-type proportion estimates and the first 20 principal components calculated on the array control probes by using residuals of the following linear model:

$$methylation\ \beta \sim 1 + CD4T + CD8T + NK + BCell + Mono + PC1 + \cdots + PC20$$

For expression data, the Illumina HumanHT-12 v3 and Illumina HumanHT-12 v4 expression BeadChips were used in KORA and LOLIPOP, respectively, and processed as described

previously [10, 96]. Only probes common to both arrays were selected for analysis. Expression data were adjusted for potential confounders by regressing log2 transformed expression values against age, sex, RNA integrity number (RIN) as well as RNA amplification plate (KORA) / RNA conversion batch (LOLIPOP) (batch1) and sample storage time (KORA) / RNA extraction batch (LOLIPOP) (batch2) and obtaining the residuals from the linear model:

$$expression \sim age + sex + RIN + batch1 + batch2$$

Additional details on the cohort data and design are presented in [43, 96, 97] (KORA) and [44, 98] (LOLIPOP).

For the inference of the GTEx Skeletal Muscle related network, we used GTEx v8 Skeletal Muscle data [78]. Potential confounders including first 5 genotype PCs, 60 expression PEER factors and measured covariates 'WGS sequencing platform' (HiSeq 2000 or HiSeq X), 'WGS library construction protocol' (PCR-based or PCR-free) and donor sex, were removed from expression data prior to analysis. Processing has been performed as previously described and details can be found elsewhere [78].

Hotspot extraction and construction of locus sets

We extract sub-sets of genomic entities (SNPs, CpGs and genes) on which we perform network inference based on the *trans* -meQTL reported by [10] (Supplementary Table 9 of their study) and eQTLGen *trans* -eQTL [7][5]. For GTEx, we obtained current (GTEx v8) tissue specific *trans* -eQTL from https://www.gtexportal.org/home/datasets[6].

Hotspot extraction. The list of *trans* -meQTL results obtained from [10] was already

[5]obtained from https://eqtlgen.org/trans-eqtls.html
[6]file GTEx_Analysis_v8_trans_eGenes_fdr05.txt

pruned for independent genetic loci and was used as provided in the paper supplement. To remove redundant highly correlated genetic loci, we pruned the eQTLGen *trans* -eQTL by selecting the eQTLs with 1) the highest minor allele frequency and 2) the largest number of trans genes for each LD cluster (1Mbp window, $R^2 > 0.2$). For GTEx, we merged eQTL by combining SNPs with $R^2 > 0.2$ and distance $<$ 1Mbp to independent genetic loci and kept all *trans* -eGenes (eGenes: genes associated with eQTL genotype) of the individual SNPs for this locus. The SNP with the highest MAF was selected as a representative SNP for the hotspot. We defined hotspots as genetic loci with ≥ 5 *trans* associations, yielding a single hotspot for GTEx, 107 for the meQTL and 444 for the eQTLGen data (Additional File 2, Tables S1 and S2). In [10], the authors provide a total of 114 meQTL hotspots per our definition. We discarded 7 of the 114 meQTL hotspots (SNPs rs10870226, rs1570038, rs17420384, rs2295981, rs2685252, rs57743634, rs7924137, as either no *cis* genes are available or no gene expression data were measured for any of the annotated *cis* genes (mostly lincRNAs, miRNAs and pseudogenes; Additional File 1, Table S1), which are needed for locus set definition (see below).

Locus sets. To mitigate the $N << P$ problem in network inference [4], where the number of features or parameters far exceeds the number of samples, we run the inference on a subset of genomic entities (SNPs, genes and CpGs) induced by *trans* hotspots. We therefore gathered all genes, which could be involved in mediating the observed QTL effects and thus were considered during the network inference, in the form of *locus sets* for each hotspot. We bridge the gap between the involved chromosomes by including transcription factor binding site (TFBS) information collected from *ReMap* [49][7] and *ENCODE* [50, 51][8] as well as human protein-protein interaction (PPI) information available via *theBioGrid* [99][9]

[7] http://tagc.univ-mrs.fr/remap/download/All/filPeaks_public.bed.gz

[8] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz

[9] https://downloads.thebiogrid.org/Download/BioGRID/Release-Archive/BIOGRID-3.5.166/BIOGRID-ORGANISM-3.5.166.tab2.zip

(version 3.5.166). We filtered *ReMap* and *ENCODE* TFBS for blood related cell types by selecting all samples which contain at least one of the following terms: "amlpz12_leukemic", "aplpz74_leukemia", "bcell", "bjab", "bl41", "blood", "lcl", "erythroid", "gm", "hbp", "k562", "kasumi", "lymphoblastoid", "mm1s", "p493", "plasma", "sem", "thp1", "u937". Genes in the PPI network were filtered for genes expressed in whole blood (GTEx v6p $RPKM > 0.1$)[10]. We enumerated all entities to be included in the locus set by performing the following steps:

1. Define set S_L for a locus L and add the QTL entities (QTL SNP $\mathcal{S}$ and *trans* -QTL eGenes/CpGs $\mathcal{T} = \{T_1, \ldots, T_q\}$, where q is the number of associated *trans* entities for L)

2. Add all genes encoded within 500kb (1Mbp window) of $\mathcal{S}$ as **SNP-Genes** to S_L (set $\mathcal{G}_C$)

3. For meQTL hotspots, add genes in the vicinity of each $\mathcal{T}_i \in \mathcal{T}$ (previous, next and overlapping genes with respect to the location of $\mathcal{T}_i$) as **CpG-Genes** to S_L (set $\mathcal{G}_T$)

4. Add all **TFs** with binding sites within 50bp of each CpG or binding in the promoter region of each gene over all $\mathcal{T}_i \in \mathcal{T}$ to S_L (set $\mathcal{G}_{TF}$)

5. Add shortest path genes G_{SP}, i.e. genes which connect $\mathcal{G}_C$ (step 2) with $\mathcal{G}_{TF}$ (step 4) according to BioGrid PPIs to S_L

To define G_{SP}, we added only genes which reside on the shortest path between the *trans* entities $\mathcal{T}$ and the SNP-Genes $\mathcal{G}_C$ in the induced PPI sub-network, i.e. containing all genes and their connections which can be linked to either $\mathcal{G}_C$ or the TFs $\mathcal{G}_{TF}$. Specifically, we added the CpGs to the filtered BioGrid PPI network, connected them to the TFs ($\mathcal{G}_{TF}$)

[10]https://storage.googleapis.com/gtex_analysis_v6p/rna_seq_data/GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz

which show binding sites in their vicinity and calculated node weights based on network propagation as described in [10]. We then extracted nodes on paths with maximal total propagation score based on node-wise propagation scores PS. For this, we weighted node scores proportional to $(-1) \times PS$ and then calculate the minimal node-weight paths between *trans* entities $\mathcal{T}$ and SNP-Genes $\mathcal{G}_C$ using the *sp.between()* method of the *RBGL* R package (version 1.56.0, R interface to the Boost Graph Library [100]) and extracted all genes on the resulting shortest paths. All nodes of the generated locus set were subsequently used as inputs to the network inference.

Prior generation

We utilized several data sources to define priors for possible edges between and within different omics levels. Each possible edge between entities in the locus set can only be assigned a single type of prior. Specifically, the different priors include:

- **SNP-to-Gene** priors, for edges between the SNP $\mathcal{S}$ and SNP-Genes $\mathcal{G}_C$
- **Gene-to-Gene** priors, for edges between all gene-gene combinations except TFs $\mathcal{G}_{TF}$ and their eQTL based targets in $\mathcal{T}$
- **CpG-to-Gene** priors, for edges between CpGs in $\mathcal{T}$ and their neighbouring genes $\mathcal{G}_T$
- **TF-to-target** priors, for edges between TFs $\mathcal{G}_{TF}$ and their targets in the *trans* set $\mathcal{T}$

SNP-to-Gene. To obtain SNP-to-Gene edge priors, we downloaded the full GTEx v6p whole-blood eQTL table [11]) and calculated, for each SNP-Gene pair, the local false discovery rate (lFDR, [101]) using the *fdrtool* R package (version 1.2.15). As described in Efron *et al.*

[11] file Whole_Blood_Analysis.v6p.all_snpgene_pairs.txt.gz from https://www.gtexportal.org/home/datasets

(2008) [101], the lFDR represents the Bayesian posterior probability of having a null case (i.e. that the null hypothesis is true) given a test statistic. We therefore defined the prior for a specific SNP $\mathcal{S}$ and a SNP-Gene $\mathcal{G}_C$ as $p_{\mathcal{S}\mathcal{G}_C} = 1 - lFDR_{\mathcal{S}\mathcal{G}_C}$.

Gene-to-Gene. We formulate *Gene-to-Gene* edge priors by combining public GTEx gene expression data [38] with PPI information from the BioGrid [99] to retrieve co-expression p-values and the respective lFDR for pairs of genes connected by a protein - protein interaction. A special case are priors between TFs and their target genes as identified via ChIP-seq (see above), which are not considered as *Gene-to-Gene* edges but are handled separately as described under 'TF-to-target priors' below. GTEx v6p RNA-seq gene expression data were downloaded from the GTEx data portal [12]. Expression data for GTEx were filtered for high quality samples (RIN $\geq$ 6) and log2 transformed, quantile normalized and transferred to standard normal distribution before removing the first 10 principle components to remove potential confounding effects [102]. Priors were derived for all Gene-Gene pairs with PPIs in the BioGRID network, where a gene $\mathcal{G} \in \mathcal{G}_C \cup \mathcal{G}_{TF}$ (for meQTL) or $\mathcal{G} \in \mathcal{G}_C \cup \mathcal{G}_{TF} \cup \mathcal{T}$ (for eQTL). For each pair, we calculated the Pearson correlation p-values in the GTEx expression data and subsequently determined the lFDR over all p-values. The prior for two genes $\mathcal{G}_A$ and $\mathcal{G}_B$ was then set to $p_{\mathcal{G}_A\mathcal{G}_B} = 1 - lFDR_{\mathcal{G}_A\mathcal{G}_B}$.

CpG-to-Gene. For the *CpG-to-Gene* priors (meQTL context only), we utilized two strategies, distinguishing between TF-CpG priors (i.e. priors between CpGs and TFs showing binding sites near the CpG site, described below under 'TF-to-target priors') and CpG-to-Gene priors (i.e. where the gene itself is encoded near the CpG). For the *CpG-to-Gene* priors, we utilized the genome-wide chromHMM [103] states (15 states model) identified in

[12] https://www.gtexportal.org/home/datasets

the Roadmap Epigenomics project [40][13]. These states reflect functional chromatin states in 200bp windows and were obtained using histone mark combinations as identified via ChIP-sequencing. We quantified a CpGs potential to affect a nearby gene, p_{T_x}, by retrieving the proportion of Roadmap cell-lines in which the CpG resides within a transcription start site (TSS) related state (see Table 3). We further adjusted the p_{T_x} by weighting state information according to the Houseman blood cell type estimates available from our data. To this end, we took the population mean for each of the Houseman cell proportion estimates and multiplied them with the chromHMM state proportions. A specific CpG-to-Gene prior for a CpG $\mathcal{T}_i \in \mathcal{T}$ and a gene $\mathcal{G}_{T_i} \in \mathcal{G}_T$ was then set to $p_{\mathcal{T}_i \mathcal{G}_{T_i}} = p_{T_x}$, if the genomic distance $d(\mathcal{T}_i, \mathcal{G}_T) <= 200bp$.

STATE NO.	MNEMONIC	DESCRIPTION
1	**TssA**	**Active TSS**
2	**TssAFlnk**	**Flanking Active TSS**
3	TxFlnk	Transcr. at gene 5' and 3'
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	EnhG	Genic enhancers
7	Enh	Enhancers
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin
10	**TssBiv**	**Bivalent/Poised TSS**
11	**BivFlnk**	**Flanking Bivalent TSS/Enh**
12	EnhBiv	Bivalent Enhancer
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

Table 3: Description of chromHMM states used in our analyses as given at https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html. Bold faced states were defined as 'active transcription' states and used to set CpG-Gene priors.

TF-to-target priors. We formulate separate priors for all edges between transcription factors $\mathcal{G}_{TF}$ and *trans* CpGs (meQTL) and *trans* genes (eQTL) in $\mathcal{T}$. Priors were only set for TF-to-CpG edges were we observe a TF binding site (from ReMap/ENCODE, see above)

[13]obtained from https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html

within 50bp of the CpG. For TF-to-Gene edges, we only considered pairs were the TF has a binding site 2,000bp upstream and 1,000 downstream of the gene's TSS. In both cases, if the TFBS criteria are met, we set a fixed large prior of 0.99 for all $\mathcal{G}_{TF}$-$\mathcal{T}$ pairs to represent the strong protein-DNA interaction evidence of ChIP-seq data.

Finally, the priors for all remaining possible edges which were not set based on one of the criteria described above, e.g. for SNP-to-Gene edges without eQTL in the GTEx data, were set to a small pseudo-prior $p_{pseudo} = 10e^{-7}$.

Ground truth network generation, data simulation and prior randomization

We performed a simulation experiment for each of the meQTL hotspots. For each SNP $\mathcal{S}$ and its corresponding locus set $\mathcal{S}_L$, we first collect the corresponding prior matrix $\mathcal{P}_S$ with priors defined as described above. We generate 10 noisy ($\mathcal{G}_N$) ground truth graphs $\mathcal{G}_N^{10}, \mathcal{G}_N^{20} \ldots \mathcal{G}_N^{100}$ by switching edges in the graph while keeping the degree distribution of a sampled graph $\mathcal{G}_T$. $\mathcal{G}_T$ is generated using all entities of $\mathcal{S}_L$ by uniformly sampling from $\mathcal{P}_S$, i.e. $\mathcal{G}_T$ contains an edge e_{ij} for each element p_{ij} of $\mathcal{P}_S$, if $p_{ij} > p_{pseudo}$ and $runif(0,1) <= p_{ij}$, where $runif(0,1)$ generates uniformly distributed random numbers between [0,1]. This procedure effectively introduces noise in the study. For instance, by switching 10% of the edges from $\mathcal{G}_T$ to generate $\mathcal{G}_N^{10}$, and making sure, that the new edges are not present as priors in $\mathcal{P}_S$, we introduce a noise level of 10% when comparing $\mathcal{P}_S$ to $\mathcal{G}_N^{10}$. We simulate data for each $\mathcal{G}_S \in \{mathpzcG_T, \mathcal{G}_N^i; i \in \{10, 20, \ldots, 100\}$ using the *bdgraph.sim()* method of the *BDgraph* package with parameters: p=$|\mathcal{S}_L|$ (number of nodes), graph=$\mathcal{G}_S$, N=612 (number of samples in LOLIPOP) and mean = 0. This approach generates normally distributed data with a covariance structure as defined by the ground truth graph. We want to assess the impact of

having discrete (genotype) data present for the network inference. To this end, we converted the SNP variable in the simulated data to genotype dosages (0,1,2) reflecting the allele frequencies of the genetic variant used in this simulation run. Specifically, we transformed the Gaussian data obtained from *bdgraph.sim()* to discrete values using the frequencies of the individual dosages for the SNP in the LOLIPOP data as quantile cut points. For each of these simulated data individually, we infer the network models and compare the inferred networks to the respective ground truth graphs $\mathcal{G}_T, \mathcal{G}_N^{10}, \ldots, \mathcal{G}_N^{100}$. We added one additional comparison, evaluating a prior on the density of the observed graph. For this, we estimated a single prior value reflecting the desired density for all edges based on a binomial model. We use the number of edges $|E_{\mathcal{G}_T}|$ of all sampled graphs $\mathcal{G}_T$ for a single run, the total number of possible edges $|E_T| = (N * (N-1))/2$, with N the total number of available nodes, and set the prior as

$$p_{rbinom} = max(\frac{1}{N_S} * \frac{\sum_{\mathcal{G}_T} |E_{\mathcal{G}_T}|}{|E_T|}, p_{pseudo}),$$

where N_S is the number of sampled graphs (i.e. the number of randomizations). For each hotspot, we repeated the above simulation procedure 100 times to obtain stable results.

Network inference

Based on the data and priors gathered for the individual hotspots, we set out to infer the regulatory networks which are best supported by these data. We evaluated several state-of-the art methods with respect to their applicability to this problem, both in a simulation study (see above) and via replication of inferred networks in real-world data from two large human population based cohorts. We applied *GeneNet* [45, 104], the graphical lasso [*glasso*, 42], *BDgraph* [29], *iRafNet* [32] as well as *GENIE3* [46] on the individual data to reconstruct regulatory networks using the respective *CRAN*[14] and *bioconductor*[15] R packages. An overview

[14]https://cran.r-project.org/
[15]https://www.bioconductor.org/

on the used inference methods and package versions is given in Table 1. Methods were chosen to reflect a range of different approaches (i.e. shrinkage based partial correlation in *GeneNet* , Bayesian MCMC sampling in *BDgraph* , lasso in *gLASSO* and tree based inference in *iRafNet* and *GENIE*3), based on whether or not implementation was readily available and whether prior knowledge could be incorporated. The well known *GeneNet* and *GENIE*3 methods are not capable of utilizing prior information, but were used as a reference for comparison to the other methods.

GeneNet For the application of GeneNet we first filtered any CpG probes from the data containing missing values. We then estimated the regulatory network by calling first the *ggm.estimate.pcor* followed by the *network.test.edges* and *extract.network* methods, all with default parameters.

GENIE3 To infer networks with GENIE3, we again used the NA filtered data (see above) with the *GENIE3* method of the package followed by the *getLinkList* method using default parameters. GENIE3 generates a ranked list of regulatory links which do not relate to any statistical measure and hence a cutoff for the link weights has to be identified manually[16]. To define an optimal cutoff, we first divide the list of weights into 200 quantiles (marking 200 distinct cutoffs) if the number of unique link weights exceeded 200. We then extracted for each cutoff the respective regulatory network and compared it to a scale free topology analogously to the approach used in [105], generating R^2 values indicating the goodness-of-fit to the topology. To choose the final network, we followed the approach suggested by Zhang *et al.* (2005) [105], which suggests to use networks with $R^2 > 0.8$. If none of our networks fit that criteria, we choose the network with the highest R^2.

[16] see also https://bioconductor.org/packages/release/bioc/vignettes/GENIE3/inst/doc/GENIE3.html

BDgraph We used BDgraph to infer networks under consideration of prior information as well as without prior information ($BDgraph$ and $BDgraph_P$) using the *bdgraph* method of the $BDgraph$ CRAN package (version 2.61). The following parameters were set: *method = "gcgm"*, *iter = 10000*, *burnin = 5000*. We further set the *g.prior* parameter to the prior matrix collected for the hotspots and the *g.start* parameter to the incidence matrix obtained from the prior matrix by setting all entries with prior information > 0.5 to 1 and all others to 0. For comparison with the no prior case, we kept all parameters the same but omitted the *g.start* and *g.prior* parameters. The graph was then obtained from the fitted model using the *select* method of the package with parameter *cut = 0.9*, thereby only choosing edges with a posterior probability of at least 0.9.

glasso Similar to BDgraph, we utilized the graphical lasso both with and without prior information. To infer the graphical lasso models, we used the *glasso* method available in the *glasso* CRAN package and set the parameter *penalize.diagonal = FALSE*. The *glasso* takes a regularization parameter λ, which implies either strong penalization of edges (high λ) or weak penalization (low λ) of parameters. This parameter can also be supplied as a matrix Λ of size $n \times n$ (where n is the number of nodes/variables) in order to supply individual parameters for individual edges. We integrated the prior information by first transforming the prior matrix $\mathcal{P}$ such that $\Lambda = 1 - \mathcal{P}$ and then supplying Λ as the regularization matrix containing values for each possible edge. This approach is similar to what has been proposed in [30, 31]. In addition, we screened a selection of penalization factors ω for both the prior and the none prior case to construct the optimal graphical lasso network with respect to the Bayesian Information Criterion (BIC). For the prior case, we included ω in the model by setting $\Lambda = \Lambda \times \omega$). For the non-prior case, we set $\lambda = \omega$. We performed 5-fold cross validation and inferred the model for all $\omega \in \{0.01, 0.015, ..., 1\}$ on the training set (containing 80% of the data) and then selected the ω yielding the minimal mean BIC

value on the test data over all folds to generate the final network.

iRafNet We use *iRafNet* to infer networks using prior information (it is not possible to run it without specifying priors). We called the *iRafNet* method of the package, setting the parameters *ntrees = 1000*, *mtry = round(sqrt(ncol(data)-1))*, and *npermut = 5* using the data filtered for missing values (see above) and then used the *Run_permutation* method with the same parameters. The final network was extracted using the *iRafNet_network* method by supplying the output of the previous method calls and setting the FDR cutoff parameter *TH = 0.05*. We used a custom implementation of *iRafNet* adjusted to make use of multiple CPUs which we made available at https://github.com/jhawe/irafnet_custom.

Method evaluation via simulation study and cross cohort replication

To identify the inference method best suited for our application, we evaluated all described network inference methods independently on the simulated data as to 1) their ability to reconstruct the underlying ground truth network as well as 2) their robustness to noise in the supplied prior information. We further compared networks inferred independently on the different cohort data to assess stability of the network inference across different, yet similar, data. Performance was measured in terms of Matthew's Correlation Coefficient (MCC) [29, 47, 106] between the inferred networks and the respective ground truth (simulation study) and the inferred networks on the different cohorts (cross cohort replication). It is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (1)$$

MCC was calculated using the *compare()* method as implemented in the *BDgraph* package

(version 2.61).

Transcription factor activities

We calculated transcription factor activities for all TFs extracted from the ReMap/ENCODE (see above) using the *plsgenomics* R package's $TFA.estimate()$ method (version 1.5-2) [107]. As input, we used the full expression matrix from KORA and LOLIPOP individually as well as the TFBS information encoded as an incidence matrix indicating for each TF its target genes. Target genes were defined as genes with an TFBS within their promoter region (2,000bp upstream and 1,000bp downstream of the TSS).

Network prioritization and final network creation

Networks were inferred for each of the 107 meQTL and 444 eQTLGen *trans* hotspots with $gLASSO_P$ and $BDgraph_P$, yielding networks with a median number of 67 and 20 edges for $gLASSO_P$ and 72 and 27 for $BDgraph_P$ over all hotspots, respectively. We filtered and ranked the networks based on the following criteria.

GWAS filtering. We filtered genetic loci with hits in genome-wide association studies (GWAS) using the current version (v1.0.2) of the GWAS catalog [52]. We extracted high LD (>0.8) SNPs and SNP aliases using the SNiPA tool [53] for each hotspot SNP. If any of the extracted SNP rsIDs had a match in the GWAS catalog, the hotspot's inferred network was permitted for downstream analysis.

Network ranking. We utilized a self devised graph score for prioritizing final models for further investigation. The graph score reflects desirable biological properties, which can be assumed for the networks underlying the *trans* -QTL hotspots. The score is formulated such that 1) the adjacency of SNP-genes and SNPs is rated positively, 2) the presence of trans entities is rated positively if they are not connected directly to the SNP and 3) high

graph density is rated negatively (i.e. sparser graphs yield higher scores). Specifically, the graph score S_G for an inferred graph G is defined as:

$$S_G = -log10(D_G) * [\frac{1}{|\mathcal{G}_C|}(\sum_{i=1}^{|G_S|} 1 - \sum_{i=1}^{|\overline{G_S}|} 1) + \frac{1}{|\mathcal{T}|}(\sum_{i=1}^{|G_T|} 1 - \sum_{i=1}^{|\overline{G_T}|} 1)]$$

where: D_G is the graph density, $\mathcal{G}_C$ is the set of all SNP-Genes, $\mathcal{T}$ is the set of all *trans* entities, G_S is the set of all SNP-genes adjacent to the SNP in G or directly connected to another SNP-Gene, $\overline{G_S}$ is the set of SNP-Genes in G but not connected directly to the SNP or one of the other SNP-Genes, G_T is the set of trans entities in G which can be reached from any SNP-Gene without traversing the SNP or another *trans* gene first and $\overline{G_T}$ is the set of *trans* genes directly connected to the SNP. Only the cluster containing the SNP, i.e. the SNP itself and any nodes reachable from the SNP via any path in G, is considered for calculating S_G; if the SNP is not present or no SNP gene has been selected in the final graph the score is set to 0.

In addition to the graph score, we ranked networks according to the total number of edges and nodes to prioritize smaller networks for detailed analysis.

Graph merging. Finally, we constructed hotspot networks containing only high confidence edges by merging the individually obtained networks from the two cohorts (KORA and LOLIPOP) and keeping only edges and nodes present in both networks. Nodes without any adjacent edges are not included in the final graph.

Priors for skeletal muscle tissue

We downloaded Muscle tissue eQTL generated by Scott *et al.* (2016) [79] from https://theparkerlab.med.umich.edu/data/papers/doi/10.1038/ncomms11764/ and used local FDRs calculated from the provided p-values to define SNP-Gene priors. Gene expression data for Muscle tissue were obtained from the ARCHS[4] [41] database. We downloaded all relevant

Muscle expression data using the keywords "Skeletal_Muscle" with the ARCHS4 loader[17] (N=194 samples). Expression data were normalized using the *ComBat* method implemented in the *sva* R package, providing dataset series ID as batch parameter.

TFBS prediction for muscle tissue. We used *factorNet* [80] to predict transcription factor binding sites from DNAse-seq chromatin accessibility data obtained from muscle cell lines. First, we trained a *factorNet* model for all TFs available for the K562 cell-line in ReMap [49]. ReMap ChIP-seq peaks functioned as a ground truth during training, DNAse-seq data from ENCODE[18] [50, 51] and DNA sequence information formed the inputs. We downloaded DNAse-seq data for the LHCN-M2 muscle cell-line from ENCODE in bigWig format for hg38[19]. *FactorNet* was then run with default parameters, using as input 1) the DNA sequence and 2) the bigWig DNAse track for each of the trained ChIP-seq tanscription factors (N=179 TFs from ReMap). High confidence TFBS were extracted by setting a *factorNet* score cutoff of 0.999, merging overlapping regions and then retaining only regions with a $width < W_{0.95}$, where $W_{0.95}$ is the 95th percent quantile of the widths of all obtained regions.

Colocalization analysis

GWAS summary statistics for schizophrenia were identified using the GWAS Atlas [108] [20] and downloaded from http://walters.psycm.cf.ac.uk/clozuk_pgc2.meta.sumstats.txt.gz. Whole-blood *trans* -eQTL summary statistics for all SNP-Gene pairs from eQTLgen were downloaded from the eQTLgen website[21]. We used *fastENLOC* [76, 109][22] to calculate colocalization probabilities as described in the *fastENLOC* Github README using default parameters. To generate probabilistic eQTL annotations, we used *DAP-G* [110, 111][23] and

[17]https://github.com/jhawe/archs4_loader
[18]dataset ENCFF971AHO
[19]dataset *ENCFF639MPM*
[20]https://atlas.ctglab.nl/
[21]https://www.eqtlgen.org/trans-eqtls.html, file 'Full trans-eQTL summary statistics'
[22]https://github.com/xqwen/fastenloc
[23]https://github.com/xqwen/dap/

769 created PIP files as needed using *TORUS* [112][24]. For LD block definition, we utilized data
770 available from LDetect [113][25].

771 Software environment

772 In case no other information is given above, all calculations were performed using standard
773 Unix commands and version 3.5.2 of the R statistical computing language [26] on a Cen-
774 tos 7 Unix system. The Docker image used in this project is available from dockerhub at
775 https://hub.docker.com/repository/docker/jhawe/r3.5.2_custom. The workflows for both
776 the cohort and the simulation studies were implemented in Snakemake [114] and can be
777 found on Github at https://github.com/jhawe/bggm. All calculations performed to arrive
778 at the discussed results in this article can be obtained using the code in the pipeline. Data
779 to run the workflow can be made available upon reasonable request by the authors.

780 Declarations

781 Availability of data and material

782 **Data.** All public data information and the respective sources are given in the methods
783 section, including URLs for downloading the data where possible. The meQTL associations
784 from Hawe et al. were directly obtained from the supplementary table 3 of the paper [10] and
785 eQTLGen *trans* -eQTL directly from the eQTLGen browser[27]. The lists of derived hotspots
786 for both data sets are made available in the supplement of this paper. Cohort data can be
787 made available upon reasonable request by the authors.

788 **Code.** The complete code used in this project is provided via Github at https://github.

[24]https://github.com/xqwen/torus
[25]https://bitbucket.org/nygcresearch/ldetect-data/src/master/
[26]https://www.r-project.org/
[27]https://eqtlgen.org/trans-eqtls.html

789 com/jhawe/bggm. The analyses were implemented in the form of a Snakemake pipeline
790 [114]. The software environment used to calculate the results is available as a Docker im-
791 age via docker hub at https://hub.docker.com/repository/docker/jhawe/r3.5.2_custom, the
792 corresponding dockerfile is available at the project's Github repository.

793 Ethics approval and consent to participate

794 All KORA participants have given written informed consent and the study was approved
795 by the Ethics Committee of the Bavarian Medical Association. The LOLIPOP study is
796 approved by the National Research Ethics Service (07/H0712/150) and all participants gave
797 written informed consent.

798 Consent for publication

799 KORA project agreement for this study was granted under K141/15g. The views expressed
800 are those of the author(s) and not necessarily those of the Imperial College Healthcare NHS
801 Trust, the NHS, the NIHR or the Department of Health.

802 Competing interests

803 FJT reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and
804 ownership interest in Cellarity, Inc. and Dermagnostix. The other authors declare that they
805 have no competing interests.

806 Funding

807 MH gratefully acknowledges funding by the Federal Ministry of Education and Research
808 (BMBF, Germany) in the project eMed:confirm (01ZX1708G). JC is supported by the Sin-
809 gapore Ministry of Health's National Medical Research Council under its Singapore Transla-

tional Research Investigator (STaR) Award (NMRC/STaR/0028/2017). AB is supported by the NIH grant 1R01MH109905. The LOLIPOP study is supported by the National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre Imperial College Healthcare NHS Trust, the NIHR Official Development Assistance (ODA, award 16/136/68), the European Union FP7 (EpiMigrant, 279143) and H2020 programs (iHealth-T2D, 643774). The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The German Diabetes Center is funded by the German Federal Ministry of Health (Berlin, Germany), the Ministry of Culture and Science of the state North Rhine-Westphalia (Düsseldorf, Germany), and grants from the German Federal Ministry of Education and Research (Berlin, Germany) to the German Center for Diabetes Research e.V. (DZD).

Authors' contributions

MH conceived the study, JH performed the analyses. AB and AS assisted with use of GTEx v8 data. AB and FT contributed to the design of the data analysis strategy. CG, MW, KS, CH, SK, SW, HP, HG, AP, and MM provided KORA cohort data and JC the LOLIPOP data. JH and MH wrote the manuscript with input from all authors. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank the participants and research staff of LOLIPOP who made the study possible. The KORA-Study Group consists of A. Peters (speaker), J. Heinrich, R. Holle, R. Leidl, C.

Meisinger, K. Strauch and their co-workers, who are responsible for the design and conduct of the KORA studies. We gratefully acknowledge the contribution of all members of field staff conducting the KORA study. Finally, we are grateful to all study participants of KORA for their invaluable contributions to this study.

References

[1] Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome biology **18**(1), 83 (2017). doi:10.1186/s13059-017-1215-1

[2] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J.: 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics **101**(1), 5–22 (2017). doi:10.1016/j.ajhg.2017.06.005

[3] Civelek, M., Lusis, A.J.: Systems genetics approaches to understand complex traits. Nature reviews. Genetics **15**(1), 34–48 (2014). doi:10.1038/nrg3575. NIHMS150003

[4] Hawe, J.S., Theis, F.J., Heinig, M.: Inferring Interaction Networks From Multi-Omics Data. Frontiers in Genetics **10**, 535 (2019). doi:10.3389/fgene.2019.00535

[5] Gilad, Y., Rifkin, S.A., Pritchard, J.K.: Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics **24**(8), 408–415 (2008). doi:10.1016/J.TIG.2008.06.001

[6] Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.-x., Nguyen, Q.T., Demirkale, C.Y., Feolo, M.L., Sharopova, N.R., Sturcke, A., Schäffer, A.A., Heard-Costa, N., Chen, H., Liu, P.-c., Wang, R., Woodhouse, K.A., Tanriverdi, K., Freedman, J.E., Raghavachari, N., Dupuis, J., Johnson, A.D., O'Donnell, C.J., Levy, D., Munson, P.J.: Integrated genome-wide analysis of expression quantitative trait loci aids interpretation

of genomic association studies. Genome Biology **18**(1), 16 (2017). doi:10.1186/s13059-016-1142-6

[7] Võsa, U., Claringbould, A., Westra, H.-j., Bonder, M.J., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Alvaes, I., Fave, M.-j., Agbessi, M., Christiansen, M., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Beutner, F., Consortium, B., Brugge, H., Kähönen, M., Kim, Y., Knight, J.C., Kovacs, P., Krohn, K., Stegle, O., Battle, A., Yang, J., Visscher, P.M., Scholz, M.: Unraveling the polygenic architecture of complex traits using blood eQTL meta- analysis. bioRxiv, 1–57 (2018). doi:10.1101/447367

[8] Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., Slieker, R.C., Jhamai, P.M., Verbiest, M., Suchiman, H.E.D., Verkerk, M., van der Breggen, R., van Rooij, J., Lakenberg, N., Arindrarto, W., Kielbasa, S.M., Jonkers, I., van 't Hof, P., Nooren, I., Beekman, M., Deelen, J., van Heemst, D., Zhernakova, A., Tigchelaar, E.F., Swertz, M.A., Hofman, A., Uitterlinden, A.G., Pool, R., van Dongen, J., Hottenga, J.J., Stehouwer, C.D.A., van der Kallen, C.J.H., Schalkwijk, C.G., van den Berg, L.H., van Zwet, E.W., Mei, H., Li, Y., Lemire, M., Hudson, T.J., Slagboom, P.E., Wijmenga, C., Veldink, J.H., van Greevenbroek, M.M.J., van Duijn, C.M., Boomsma, D.I., Isaacs, A., Jansen, R., van Meurs, J.B.J., 't Hoen, P.A.C., Franke, L., Heijmans, B.T.: Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics **49**(1), 131–138 (2016). doi:10.1038/ng.3721

[9] Husquin, L.T., Rotival, M., Fagny, M., Quach, H., Zidane, N., McEwen, L.M., MacIsaac, J.L., Kobor, M.S., Aschard, H., Patin, E., Quintana-Murci, L.: Exploring the genetic basis of human population differences in DNA methylation and

their causal impact on immune gene regulation. Genome Biology **19**(1), 222 (2018). doi:10.1186/s13059-018-1601-3

[10] Hawe, J.S., Lehne, B.C., Wilson, R., Loh, M., Heinig, M., Gieger, C., Waldenberger, M., Chambers, J.C.: Genetic variation influencing DNA methylation provides new insights into the molecular pathways regulating genomic function. Manuscript in preparation (2020)

[11] West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W., Clair, D.A.S.: Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. Genetics **175**(3), 1441–1450 (2007). doi:10.1534/GENETICS.106.064972

[12] Albert, F.W., Bloom, J.S., Siegel, J., Day, L., Kruglyak, L.: Genetics of trans-regulatory variation in gene expression. eLife **7** (2018). doi:10.7554/ELIFE.35471

[13] Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., 't Hoen, P.A.C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B.J., Franke, L.: Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics **45**(10), 1238–1243 (2013). doi:10.1038/ng.2756

[14] Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L.V., de Haan, G., Su, A.I., Jansen, R.C.: Genetical genomics: spotlight on QTL hotspots. PLoS genetics **4**(10), 1000232 (2008). doi:10.1371/journal.pgen.1000232

[15] Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., Schadt, E.E.: Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics **40**(7), 854–861 (2008). doi:10.1038/ng.167

[16] Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D.: Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics **16**(2), 85–97 (2015). doi:10.1038/nrg3868

[17] Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A., Lusis, A.J.: An integrative genomics approach to infer causal associations between gene expression and disease. Nature genetics **37**(7), 710–7 (2005). doi:10.1038/ng1589

[18] Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., Van Den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M., Jansen, R.C.: Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America **104**(5), 1708–1713 (2007). doi:10.1073/pnas.0610429104

[19] Luijk, R., Dekkers, K.F., van Iterson, M., Arindrarto, W., Claringbould, A., Hop, P., Boomsma, D.I., van Duijn, C.M., van Greevenbroek, M.M.J., Veldink, J.H., Wijmenga, C., Franke, L., 't Hoen, P.A.C., Jansen, R., van Meurs, J., Mei, H., Slagboom, P.E., Heijmans, B.T., van Zwet, E.W.: Genome-wide identification of directed gene networks

928 using large-scale population genomics data. Nature Communications **9**(1), 3097 (2018).
929 doi:10.1038/s41467-018-05452-6

930 [20] Mine, K.L., Shulzhenko, N., Yambartsev, A., Rochman, M., Sanson, G.F.O., Lando,
931 M., Varma, S., Skinner, J., Volfovsky, N., Deng, T., Brenna, S.M.F., Carvalho,
932 C.R.N., Ribalta, J.C.L., Bustin, M., Matzinger, P., Silva, I.D.C.G., Lyng, H., Gerbase-
933 DeLima, M., Morgun, A.: Gene network reconstruction reveals cell cycle and antiviral
934 genes as major drivers of cervical cancer. Nature communications **4**, 1806 (2013).
935 doi:10.1038/NCOMMS2693

936 [21] Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., Thorsson,
937 V.: The inferelator: An algorithn for learning parsimonious regulatory networks from
938 systems-biology data sets de novo. Genome Biology **7**(5) (2006). doi:10.1186/gb-2006-
939 7-5-r36

940 [22] Lam, K.Y., Westrick, Z.M., Müller, C.L., Christiaen, L., Bonneau, R.: Fused Re-
941 gression for Multi-source Gene Regulatory Network Inference. PLoS Computational
942 Biology **12**(12), 1–23 (2016). doi:10.1371/journal.pcbi.1005157

943 [23] Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath,
944 H., Thareja, G., Wahl, A., DeLisle, R.K., Gold, L., Pezer, M., Lauc, G., El-Din Selim,
945 M.A., Mook-Kanamori, D.O., Al-Dous, E.K., Mohamoud, Y.A., Malek, J., Strauch,
946 K., Grallert, H., Peters, A., Kastenmüller, G., Gieger, C., Graumann, J.: Connecting
947 genetic risk to disease end points through the human blood plasma proteome. Nature
948 Communications **8**, 14357 (2017). doi:10.1038/ncomms14357

949 [24] Castro, J.C., Valdés, I., Gonzalez-García, L.N., Danies, G., Cañas, S., Winck, F.V.,
950 Ñústez, C.E., Restrepo, S., Riaño-Pachón, D.M.: Gene regulatory networks on trans-
951 fer entropy (GRNTE): A novel approach to reconstruct gene regulatory interactions

applied to a case study for the plant pathogen Phytophthora infestans. Theoretical Biology and Medical Modelling **16**(1), 1–15 (2019). doi:10.1186/s12976-019-0103-7

[25] Kamoun, A., Idbaih, A., Dehais, C., Elarouci, N., Carpentier, C., Letouzé, E., Colin, C., Mokhtari, K., Jouvet, A., Uro-Coste, E., Martin-Duverneuil, N., Sanson, M., Delattre, J.-Y., Figarella-Branger, D., de Reyniès, A., Ducray, F., Adam, C., Andraud, M., Aubriot-Lorton, M.-H., Bauchet, L., Beauchesne, P., Bielle, F., Blechet, C., Campone, M., Carpentier, A.F., Carpiuc, I., Cazals-Hatem, D., Chenard, M.-P., Chiforeanu, D., Chinot, O., Cohen-Moyal, E., Colin, P., Dam-Hieu, P., Desenclos, C., Desse, N., Dhermain, F., Diebold, M.-D., Eimer, S., Faillot, T., Fesneau, M., Fontaine, D., Gaillard, S., Gauchotte, G., Gaultier, C., Ghiringhelli, F., Godard, J., Gueye, E.M., Guillamo, J.S., Hamdi-Elouadhani, S., Honnorat, J., Kemeny, J.L., Khallil, T., Labrousse, F., Langlois, O., Laquerriere, A., Larrieu-Ciron, D., Lechapt-Zalcman, E., Guérinel, C.L., Levillain, P.-M., Loiseau, H., Loussouarn, D., Maurage, C.-A., Menei, P., Motsuo Fotso, M.J., Noel, G., Parker, F., Peoc'h, M., Polivka, M., Quintin-Roué, I., Ramirez, C., Ricard, D., Richard, P., Rigau, V., Rousseau, A., Runavot, G., Sevestre, H., Tortel, M.C., Vandenbos, F., Vauleon, E., Viennet, G., Villa, C., Villa, C.: Integrated multi-omics analysis of oligodendroglial tumours identifies three subgroups of 1p/19q co-deleted gliomas. Nature Communications **7**, 11263 (2016). doi:10.1038/ncomms11263

[26] Huang, S., Chaudhary, K., Garmire, L.X.: More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in genetics **8**, 84 (2017). doi:10.3389/fgene.2017.00084

[27] Greenfield, A., Hafemeister, C., Bonneau, R.: Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics **29**(8), 1060–1067 (2013). doi:10.1093/bioinformatics/btt099

[28] Christley, S., Nie, Q., Xie, X.: Incorporating Existing Network Information into Gene Network Inference. PLoS ONE **4**(8), 6799 (2009). doi:10.1371/journal.pone.0006799

[29] Mohammadi, A., Wit, E.C.: Bayesian Structure Learning in Sparse Gaussian Graphical Models. Bayesian Analysis **10**(1), 109–138 (2015). doi:10.1214/14-BA889

[30] Wang, Z., Xu, W., Lucas, F.A.S., Liu, Y.: Incorporating prior knowledge into Gene Network Study. Bioinformatics **29**(20), 2633–2640 (2013). doi:10.1093/bioinformatics/btt443

[31] Li, Y., Jackson, S.A.: Gene Network Reconstruction by Integration of Prior Biological Knowledge. G3 (Bethesda, Md.) **5**(6), 1075–9 (2015). doi:10.1534/g3.115.018127

[32] Petralia, F., Wang, P., Yang, J., Tu, Z.: Integrative random forest for gene regulatory network inference. Bioinformatics **31**(12), 197–205 (2015). doi:10.1093/bioinformatics/btv268

[33] Zuo, Y., Cui, Y., Yu, G., Li, R., Ressom, H.W.: Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. BMC Bioinformatics **18**(1), 99 (2017). doi:10.1186/s12859-017-1515-1

[34] Studham, M.E., Tjärnberg, A., Nordling, T.E.M., Nelander, S., Sonnhammer, E.L.L.: Functional association networks as priors for gene regulatory network inference. Bioinformatics **30**(12), 130–138 (2014). doi:10.1093/bioinformatics/btu285

[35] Gustafsson, M., Hörnquist, M.: Gene expression prediction by soft integration and the elastic net - Best performance of the DREAM3 gene expression challenge. PLoS ONE **5**(2) (2010). doi:10.1371/journal.pone.0009134

[36] Siahpirani, A.F., Roy, S.: A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Research **45**(4), 1–22 (2017). doi:10.1093/nar/gkw963

[37] Pei, B., Shin, D.G.: Reconstruction of biological networks by incorporating prior knowledge into bayesian network models. Journal of Computational Biology **19**(12), 1324–1334 (2012). doi:10.1089/cmb.2011.0194

[38] The GTEx Consortium: The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science **348**, 648–660 (2015). doi:10.1126/science.1262110

[39] Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y.S., Parsana, P., Segrè, A.V., Strober, B.J., Zappala, Z., Cummings, B.B., Gelfand, E.T., Hadley, K., Huang, K.H., Lek, M., Li, X., Nedzel, J.L., Nguyen, D.Y., Noble, M.S., Sullivan, T.J., Tukiainen, T., MacArthur, D.G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A.R., Lockhart, N.C., Moore, H.M., Rao, A., Struewing, J.P., Volpi, S., Brigham, L.E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W.F., Lonsdale, J.T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J.A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B.A., Gillard, B.M., Karasik, E., Kumar, R., Miklos, M., Moser, M.T., Jewell, S.D., Montroy, R.G., Rohrer, D.C., Valley, D., Mash, D.C., Davis, D.A., Sobin, L., Barcus, M.E., Branton, P.A., Abell, N.S., Balliu, B., Delaneau, O., Frésard, L., Gamazon, E.R., Garrido-Martín, D., Gewirtz, A.D.H., Gliner, G., Gloudemans, M.J., Han, B., He, A.Z., Hormozdiari, F., Li, X., Liu, B., Kang, E.Y., McDowell, I.C., Ongen, H., Palowitch, J.J., Peterson, C.B., Quon, G., Ripke, S., Saha, A., Shabalin, A.A., Shimko, T.C., Sul, J.H., Teran, N.A., Tsang, E.K., Zhang, H., Zhou, Y.H., Bustamante, C.D., Cox, N.J., Guigó, R., Kellis, M., McCarthy, M.I., Conrad, D.F.,

Eskin, E., Li, G., Nobel, A.B., Sabatti, C., Stranger, B.E., Wen, X., Wright, F.A., Ardlie, K.G., Dermitzakis, E.T., Lappalainen, T., Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B., Handsaker, R.E., Kashin, S., Karczewski, K.J., Nguyen, D.T., Trowbridge, C.A., Barshir, R., Basha, O., Bogu, G.K., Chen, L.S., Chiang, C., Damani, F.N., Ferreira, P.G., Hall, I.M., Howald, C., Im, H.K., Kim, Y., Kim-Hellmuth, S., Mangul, S., Monlong, J., Muñoz-Aguirre, M., Ndungu, A.W., Nicolae, D.L., Oliva, M., Panousis, N., Papasaikas, P., Payne, A.J., Quan, J., Reverter, F., Sammeth, M., Scott, A.J., Sodaei, R., Stephens, M., Urbut, S., Van De Bunt, M., Wang, G., Xi, H.S., Yeger-Lotem, E., Zaugg, J.B., Akey, J.M., Bates, D., Chan, J., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J.A., Feinberg, A.P., Fernando, M.S., Halow, J., Hansen, K.D., Haugen, E., Hickey, P.F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kibriya, M.G., Lee, K., Li, J.B., Li, Q., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M.T., Molinie, B., Nelson, J., Neri, F.J., Park, Y., Pierce, B.L., Rinaldi, N.J., Rizzardi, L.F., Sandstrom, R., Skol, A., Smith, K.S., Snyder, M.P., Stamatoyannopoulos, J., Tang, H., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C.R., Carithers, L.J., Vaught, J.B., Gould, S.E., Lockart, N.C., Martin, C., Addington, A.M., Koester, S.E., Undale, A.H., Smith, A.M., Tabor, D.E., Roche, N.V., McLean, J.A., Vatanian, N., Robinson, K.L., Valentino, K.M., Qi, L., Hunter, S., Hariharan, P., Singh, S., Um, K.S., Matose, T., Tomaszewski, M.M., Barker, L.K., Mosavel, M., Siminoff, L.A., Traino, H.M., Flicek, P., Juettemann, T., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S.J., Zerbino, D.R., Craft, B., Goldman, M., Haeussler, M., Kent, W.J., Lee, C.M., Paten, B., Rosenbloom, K.R., Vivian, J., Zhu, J.: Genetic effects on gene expression across human tissues. Nature **550**(7675), 204–213 (2017). doi:10.1038/nature24277

[40] The Roadmap Epigenomics Consortium: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015). doi:10.1038/nature14248

[41] Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., Ma'ayan, A.: Massive mining of publicly available RNA-seq data from human and mouse. Nature Communications **9**(1), 1366 (2018). doi:10.1038/s41467-018-03751-6

[42] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008). doi:10.1093/biostatistics/kxm045

[43] Pfeiffer, L., Wahl, S., Pilling, L.C., Reischl, E., Sandling, J.K., Kunze, S., Holdt, L.M., Kretschmer, A., Schramm, K., Adamski, J., Klopp, N., Illig, T., Hedman, Å.K., Roden, M., Hernandez, D.G., Singleton, A.B., Thasler, W.E., Grallert, H., Gieger, C., Herder, C., Teupser, D., Meisinger, C., Spector, T.D., Kronenberg, F., Prokisch, H., Melzer, D., Peters, A., Deloukas, P., Ferrucci, L., Waldenberger, M.: DNA methylation of lipid-related genes affects blood lipid levels. Circulation. Cardiovascular genetics **8**(2), 334–42 (2015). doi:10.1161/CIRCGENETICS.114.000804

[44] Chambers, J.C., Loh, M., Lehne, B., Drong, A., Kriebel, J., Motta, V., Wahl, S., Elliott, H.R., Rota, F., Scott, W.R., Zhang, W., Tan, S.T., Campanella, G., Chadeau-Hyam, M., Yengo, L., Richmond, R.C., Adamowicz-Brice, M., Afzal, U., Bozaoglu, K., Mok, Z.Y., Ng, H.K., Pattou, F., Prokisch, H., Rozario, M.A., Tarantini, L., Abbott, J., Ala-Korpela, M., Albetti, B., Ammerpohl, O., Bertazzi, P.A., Blancher, C., Caiazzo, R., Danesh, J., Gaunt, T.R., de Lusignan, S., Gieger, C., Illig, T., Jha, S., Jones, S., Jowett, J., Kangas, A.J., Kasturiratne, A., Kato, N., Kotea, N., Kowlessur, S., Pitkäniemi, J., Punjabi, P., Saleheen, D., Schafmayer, C., Soininen, P., Tai, E.S., Thorand, B., Tuomilehto, J., Wickremasinghe, A.R., Kyrtopoulos, S.A., Aitman, T.J., Herder, C., Hampe, J., Cauchi, S., Relton, C.L., Froguel, P., Soong, R., Vineis, P., Jarvelin, M.R., Scott, J., Grallert, H., Bollati, V., Elliott, P., McCarthy, M.I., Kooner, J.S.: Epigenome-wide association of DNA methylation markers

in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: A nested case-control study. The Lancet Diabetes and Endocrinology **3**(7), 526–534 (2015). doi:10.1016/S2213-8587(15)00127-8

[45] Opgen-Rhein, R., Strimmer, K.: From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology **1**(1), 37 (2007). doi:10.1186/1752-0509-1-37

[46] Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE **5**(9), 12776 (2010). doi:10.1371/journal.pone.0012776

[47] Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Structure (1975). doi:10.1016/0005-2795(75)90109-9

[48] Arrieta-Ortiz, M.L., Hafemeister, C., Bate, A.R., Chu, T., Greenfield, A., Shuster, B., Barry, S.N., Gallitto, M., Liu, B., Kacmarczyk, T., Santoriello, F., Chen, J., Rodrigues, C.D.A., Sato, T., Rudner, D.Z., Driks, A., Bonneau, R., Eichenberger, P.: An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Molecular systems biology **11**(11), 839 (2015). doi:10.15252/msb.20156236

[49] Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., Ballester, B.: ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Research **46**(D1), 267–275 (2018). doi:10.1093/nar/gkx1092

[50] ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012). doi:10.1038/nature11247

[51] Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., Onate, K.C., Graham,

1098 K., Miyasato, S.R., Dreszer, T.R., Strattan, J.S., Jolanki, O., Tanaka, F.Y., Cherry,
1099 J.M.: The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic
1100 acids research **46**(D1), 794–801 (2018). doi:10.1093/nar/gkx1081

1101 [52] Buniello, A., MacArthur, J.A., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone,
1102 C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O.,
1103 Whetzel, P.L., Amode, R., Guillen, J.A., Riat, H.S., Trevanion, S.J., Hall, P., Junk-
1104 ins, H., Flicek, P., Burdett, T., Hindorff, L.A., Cunningham, F., Parkinson, H.: The
1105 NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted
1106 arrays and summary statistics 2019. Nucleic Acids Research **47**(D1), 1005–1012 (2019).
1107 doi:10.1093/nar/gky1120

1108 [53] Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., Kastenmüller, G.: SNiPA: An inter-
1109 active, genetic variant-centered annotation browser. Bioinformatics **31**(8), 1334–1336
1110 (2015). doi:10.1093/bioinformatics/btu779

1111 [54] Goes, F.S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski,
1112 I., Nestadt, G., Kenny, E.E., Vacic, V., Peters, I., Lencz, T., Darvasi, A., Mulle,
1113 J.G., Warren, S.T., Pulver, A.E.: Genome-wide association study of schizophrenia
1114 in Ashkenazi Jews. American Journal of Medical Genetics Part B: Neuropsychiatric
1115 Genetics **168**(8), 649–659 (2015). doi:10.1002/ajmg.b.32349

1116 [55] de Jong, S., van Eijk, K.R., Zeegers, D.W.L.H., Strengman, E., Janson, E.,
1117 Veldink, J.H., van den Berg, L.H., Cahn, W., Kahn, R.S., Boks, M.P.M., Ophoff,
1118 R.A., PGC Schizophrenia (GWAS) Consortium, T.P.S.G.: Expression QTL analy-
1119 sis of top loci from GWAS meta-analysis highlights additional schizophrenia can-
1120 didate genes. European journal of human genetics : EJHG **20**(9), 1004–8 (2012).
1121 doi:10.1038/ejhg.2012.38

[56] Kanazawa, T., Bousman, C.A., Liu, C., Everall, I.P.: Schizophrenia genetics in the genome-wide era: A review of Japanese studies. npj Schizophrenia **3**(1), 2–7 (2017). doi:10.1038/s41537-017-0028-2

[57] Saito, T., Ikeda, M., Mushiroda, T., Ozeki, T., Kondo, K., Shimasaki, A., Kawase, K., Hashimoto, S., Yamamori, H., Yasuda, Y., Fujimoto, M., Ohi, K., Takeda, M., Kamatani, Y., Numata, S., Ohmori, T., ichi Ueno, S., Makinodan, M., Nishihata, Y., Kubota, M., Kimura, T., Kanahara, N., Hashimoto, N., Fujita, K., Nemoto, K., Fukao, T., Suwa, T., Noda, T., Yada, Y., Takaki, M., Kida, N., Otsuru, T., Murakami, M., Takahashi, A., Kubo, M., Hashimoto, R., Iwata, N.: Pharmacogenomic Study of Clozapine-Induced Agranulocytosis/Granulocytopenia in a Japanese Population. Biological Psychiatry **80**(8), 636–642 (2016). doi:10.1016/j.biopsych.2015.12.006

[58] Rustenhoven, J., Smith, A.M., Smyth, L.C., Jansson, D., Scotter, E.L., Swanson, M.E.V., Aalderink, M., Coppieters, N., Narayan, P., Handley, R., Overall, C., Park, T.I.H., Schweder, P., Heppner, P., Curtis, M.A., Faull, R.L.M., Dragunow, M.: PU.1 regulates Alzheimer's disease-associated genes in primary human microglia. Molecular neurodegeneration **13**(1), 44 (2018). doi:10.1186/s13024-018-0277-1

[59] Hu, Z., Gu, X., Baraoidan, K., Ibanez, V., Sharma, A., Kadkol, S., Munker, R., Ackerman, S., Nucifora, G., Saunthararajah, Y.: RUNX1 regulates corepressor interactions of PU.1. Blood **117**(24), 6498–508 (2011). doi:10.1182/blood-2010-10-312512

[60] Watanabe, Y., Nunokawa, A., Kaneko, N., Muratake, T., Arinami, T., Ujike, H., Inada, T., Iwata, N., Kunugi, H., Itokawa, M., Otowa, T., Ozaki, N., Someya, T.: Two-stage case–control association study of polymorphisms in rheumatoid arthritis susceptibility genes with schizophrenia. Journal of Human Genetics **54**(1), 62–65 (2009). doi:10.1038/jhg.2008.4

[61] Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, T.S.P.G.-W.A.S.G.: Genome-wide association study identifies five new schizophrenia loci. Nature genetics **43**(10), 969–76 (2011). doi:10.1038/ng.940

[62] Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., Olincy, A., Amin, F., Cloninger, C.R., Silverman, J.M., Buccola, N.G., Byerley, W.F., Black, D.W., Crowe, R.R., Oksenberg, J.R., Mirel, D.B., Kendler, K.S., Freedman, R., Gejman, P.V.: Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature **460**(7256), 753–7 (2009). doi:10.1038/nature08192

[63] International Schizophrenia Consortium, I.S., Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature **460**(7256), 748–52 (2009). doi:10.1038/nature08185

[64] Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O.P.H., Mors, O., Mortensen, P.B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A.D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.-J., Giegling, I., Rasmussen, H.B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J.M., Magnusdottir, B.B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J.R., Haraldsson, M., Fossdal, R., Thorgeirsson, T.E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemeney, L.A., Genetic Risk and Outcome in Psychosis (GROUP), Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A.C., Ge, D., Yoon, J.L., Shianna, K.V., Freimer, N.B.,

Cantor, R.M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E.G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M.M., Rietschel, M., Matthews, P.M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D.B., Stefansson, K., Collier, D.A.: Common variants conferring risk of schizophrenia. Nature **460**(7256), 744–7 (2009). doi:10.1038/nature08186

[65] Quednow, B.B., Brinkmeyer, J., Mobascher, A., Nothnagel, M., Musso, F., Gründer, G., Savary, N., Petrovsky, N., Frommann, I., Lennertz, L., Spreckelmeyer, K.N., Wienker, T.F., Dahmen, N., Thuerauf, N., Clepce, M., Kiefer, F., Majic, T., Mössner, R., Maier, W., Gallinat, J., Diaz-Lacava, A., Toliat, M.R., Thiele, H., Nürnberg, P., Wagner, M., Winterer, G.: Schizophrenia risk polymorphisms in the TCF4 gene interact with smoking in the modulation of auditory sensory gating. Proceedings of the National Academy of Sciences of the United States of America **109**(16), 6271–6 (2012). doi:10.1073/pnas.1118051109

[66] Zweier, C., Peippo, M.M., Hoyer, J., Sousa, S., Bottani, A., Clayton-Smith, J., Reardon, W., Saraiva, J., Cabral, A., Göhring, I., Devriendt, K., de Ravel, T., Bijlsma, E.K., Hennekam, R.C.M., Orrico, A., Cohen, M., Dreweke, A., Reis, A., Nürnberg, P., Rauch, A.: Haploinsufficiency of TCF4 Causes Syndromal Mental Retardation with Intermittent Hyperventilation (Pitt-Hopkins Syndrome). The American Journal of Human Genetics **80**(5), 994–1001 (2007). doi:10.1086/515583

[67] Jung, M., Häberle, B.M., Tschaikowsky, T., Wittmann, M.-T., Balta, E.-A., Stadler, V.-C., Zweier, C., Dörfler, A., Gloeckner, C.J., Lie, D.C.: Analysis of the expression pattern of the schizophrenia-risk and intellectual disability gene TCF4 in the developing and adult brain suggests a role in development and plasticity of cortical and hippocampal neurons. Molecular autism **9**, 20 (2018). doi:10.1186/s13229-018-0200-1

[68] Huo, Y., Li, S., Liu, J., Li, X., Luo, X.-J.: Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nature Communications **10**(1), 670 (2019). doi:10.1038/s41467-019-08666-4

[69] Roussos, P., Katsel, P., Davis, K.L., Giakoumaki, S.G., Lencz, T., Malhotra, A.K., Siever, L.J., Bitsios, P., Haroutunian, V.: Convergent findings for abnormalities of the NF-κB signaling pathway in schizophrenia. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology **38**(3), 533–9 (2013). doi:10.1038/npp.2012.215

[70] Saia-Cereda, V.M., Cassoli, J.S., Schmitt, A., Falkai, P., Nascimento, J.M., Martins-de-Souza, D.: Proteomics of the corpus callosum unravel pivotal players in the dysfunction of cell signaling, structure, and myelination in schizophrenia brains. European Archives of Psychiatry and Clinical Neuroscience **265**(7), 601–612 (2015). doi:10.1007/s00406-015-0621-1

[71] Bagyinszky, E., Youn, Y.C., An, S.S.A., Kim, S.: The genetics of Alzheimer's disease. Clinical interventions in aging **9**, 535–51 (2014). doi:10.2147/CIA.S51571

[72] Dowdle, W.E., Robinson, J.F., Kneist, A., Sirerol-Piquer, M.S., Frints, S.G.M., Corbit, K.C., Zaghloul, N.A., van Lijnschoten, G., Mulders, L., Verver, D.E., Zerres, K., Reed, R.R., Attié-Bitach, T., Johnson, C.A., García-Verdugo, J.M., Katsanis, N., Bergmann, C., Reiter, J.F., Reiter, J.F.: Disruption of a Ciliary B9 Protein Complex Causes Meckel Syndrome. The American Journal of Human Genetics **89**(1), 94–110 (2011). doi:10.1016/j.ajhg.2011.06.003

[73] Stuart, M.J., Singhal, G., Baune, B.T.: Systematic review of the neurobiological relevance of chemokines to psychiatric disorders. Frontiers in Cellular Neuroscience **9**(September), 1–15 (2015). doi:10.3389/fncel.2015.00357

[74] Sanchez, E., Darvish, H., Mesias, R., Taghavi, S., Firouzabadi, S.G., Walker, R.H., Tafakhori, A., Paisán-Ruiz, C.: Identification of a Large DNAJB2 Deletion in a Family with Spinal Muscular Atrophy and Parkinsonism. Human mutation **37**(11), 1180–1189 (2016). doi:10.1002/humu.23055

[75] Rodriguez, M.S., Egaña, I., Lopitz-Otsoa, F., Aillet, F., Lopez-Mato, M.P., Dorronroso, A., Lobato-Gil, S., Sutherland, J.D., Barrio, R., Trigueros, C., Lang, V.: The RING ubiquitin E3 RNF114 interacts with A20 and modulates NF-κB activity and T-cell activation. Cell Death and Disease **5**(8) (2014). doi:10.1038/cddis.2014.366

[76] Wen, X., Pique-Regi, R., Luca, F.: Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLOS Genetics **13**(3), 1006646 (2017). doi:10.1371/journal.pgen.1006646

[77] Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J.H., McCarroll, S.A., Baune, B.T., Breen, G., Byrne, E.M., Dannlowski, U., Eley, T.C., Hayward, C., Martin, N.G., McIntosh, A.M., Plomin, R., Porteous, D.J., Wray, N.R., Caballero, A., Geschwind, D.H., Huckins, L.M., Ruderfer, D.M., Santiago, E., Sklar, P., Stahl, E.A., Won, H., Agerbo, E., Als, T.D., Andreassen, O.A., Bækvad-Hansen, M., Mortensen, P.B., Pedersen, C.B., Børglum, A.D., Bybjerg-Grauholm, J., Djurovic, S., Durmishi, N., Pedersen, M.G., Golimbet, V., Grove, J., Hougaard, D.M., Mattheisen, M., Molden, E., Mors, O., Nordentoft, M., Pejovic-Milovancevic, M., Sigurdsson, E., Silagadze, T., Hansen, C.S., Stefansson, K., Stefansson, H., Steinberg, S., Tosato, S., Werge, T., Harold, D., Sims, R., Gerrish, A., Chapman, J., Abraham, R., Hollingworth, P., Pahwa, J., Denning, N., Thomas, C., Taylor, S., Powell, J., Proitsi, P., Lupton, M., Lovestone, S., Passmore, P., Craig, D., McGuinness, B., Johnston, J., Todd, S., Maier, W., Jessen, F., Heun, R.,

Schurmann, B., Ramirez, A., Becker, T., Herold, C., Lacour, A., Drichel, D., Nothen, M., Goate, A., Cruchaga, C., Nowotny, P., Morris, J.C., Mayo, K., O'Donovan, M., Owen, M., Williams, J., Achilla, E., Barr, C.L., Böttger, T.W., Cohen, D., Curran, S., Dempster, E., Dima, D., Sabes-Figuera, R., Flanagan, R.J., Frangou, S., Frank, J., Gasse, C., Gaughran, F., Giegling, I., Hannon, E., Hartmann, A.M., Heißerer, B., Helthuis, M., Horsdal, H.T., Ingimarsson, O., Jollie, K., Kennedy, J.L., Köhler, O., Konte, B., Lang, M., Lewis, C., MacCaba, J., Malhotra, A.K., McCrone, P., Meier, S.M., Mill, J., Nöthen, M.M., Pedersen, C.B., Rietschel, M., Rujescu, D., Schwalber, A., Sørensen, H.J., Spencer, B., Støvring, H., Strohmaier, J., Sullivan, P., Vassos, E., Verbelen, M., Collier, D.A., Kirov, G., Owen, M.J., O'Donovan, M.C., Walters, J.T.R.: Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics **50**(3), 381–389 (2018). doi:10.1038/s41588-018-0059-2

[78] The Genotype Tissue Expression Consortium: The GTEx Consortium atlas of genetic regulatory effects across human tissues The Genotype Tissue Expression Consortium. bioRxiv (2019). doi:10.1101/787903

[79] Scott, L.J., Erdos, M.R., Huyghe, J.R., Welch, R.P., Beck, A.T., Wolford, B.N., Chines, P.S., Didion, J.P., Narisu, N., Stringham, H.M., Taylor, D.L., Jackson, A.U., Vadlamudi, S., Bonnycastle, L.L., Kinnunen, L., Saramies, J., Sundvall, J., Albanus, R.D., Kiseleva, A., Hensley, J., Crawford, G.E., Jiang, H., Wen, X., Watanabe, R.M., Lakka, T.A., Mohlke, K.L., Laakso, M., Tuomilehto, J., Koistinen, H.A., Boehnke, M., Collins, F.S., Parker, S.C.J.: The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nature Communications **7**(1), 11764 (2016). doi:10.1038/ncomms11764

[80] Quang, D., Xie, X.: FactorNet: A deep learning framework for predicting cell type spe-

cific transcription factor binding from nucleotide-resolution sequential data. Methods (November 2018), 1–8 (2019). doi:10.1016/j.ymeth.2019.03.020

[81] Singh, A.N., Gasman, B.: Disentangling the genetics of sarcopenia: prioritization of NUDT3 and KLF5 as genes for lean mass & HLA-DQB1-AS1 for hand grip strength with the associated enhancing SNPs & a scoring system. BMC Medical Genetics **21**(1), 40 (2020). doi:10.1186/s12881-020-0977-6

[82] Oishi, Y., Manabe, I., Tobe, K., Ohsugi, M., Kubota, T., Fujiu, K., Maemura, K., Kubota, N., Kadowaki, T., Nagai, R.: SUMOylation of Krüppel-like transcription factor 5 acts as a molecular switch in transcriptional programs of lipid metabolism involving PPAR-δ. Nature Medicine **14**(6), 656–666 (2008). doi:10.1038/nm1756

[83] Moresi, V., Carrer, M., Grueter, C.E., Rifki, O.F., Shelton, J.M., Richardson, J.A., Bassel-Duby, R., Olson, E.N.: Histone deacetylases 1 and 2 regulate autophagy flux and skeletal muscle homeostasis in mice. Proceedings of the National Academy of Sciences of the United States of America **109**(5), 1649–54 (2012). doi:10.1073/pnas.1121159109

[84] Silverstein, R.A., Ekwall, K.: Sin3: a flexible regulator of global gene expression and genome stability. Current Genetics **47**(1), 1–17 (2005). doi:10.1007/s00294-004-0541-5

[85] Lee, T.I., Young, R.A.: Transcription of Eukaryotic Protein-Coding Genes. Annual Review of Genetics **34**(1), 77–137 (2000). doi:10.1146/annurev.genet.34.1.77

[86] Zhang, J., Bang, M.-L., Gokhin, D.S., Lu, Y., Cui, L., Li, X., Gu, Y., Dalton, N.D., Scimia, M.C., Peterson, K.L., Lieber, R.L., Chen, J.: Syncoilin is required for generating maximum isometric stress in skeletal muscle but dispensable for muscle cytoarchitecture. American journal of physiology. Cell physiology **294**(5), 1175–82 (2008). doi:10.1152/ajpcell.00049.2008

[87] Brown, S.C., Torelli, S., Ugo, I., De Biasia, F., Howman, E.V., Poon, E., Britton, J., Davies, K.E., Muntoni, F.: Syncoilin upregulation in muscle of patients with neuromuscular disease. Muscle & Nerve **32**(6), 715–725 (2005). doi:10.1002/mus.20431

[88] Seim, I., Jeffery, P.L., Chopin, L.K.: Gene expression profiling of The Cancer Genome Atlas supports an inverse association between body mass index (BMI) and major oesophageal tumour subtypes. bioRxiv, 378778 (2018). doi:10.1101/378778

[89] Oldknow, K., Morton, N.M., Yadav, M., Rajoanah, S., Huesa, C., Bunger, L., Ferron, M., Karsenty, G., MacRae, V., Milan, J.L., Farquharson, C.: An emerging role of phospho1 in the regulation of energy metabolism. Bone Abstracts (2013). doi:10.1530/boneabs.01.OC6.6

[90] Mittelstraß, K., Waldenberger, M.: DNA methylation in human lipid metabolism and related diseases. Current Opinion in Lipidology **29**(2), 116–124 (2018). doi:10.1097/MOL.0000000000000491

[91] Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W.R., Kunze, S., Tsai, P.-C., Ried, J.S., Zhang, W., Yang, Y., Tan, S., Fiorito, G., Franke, L., Guarrera, S., Kasela, S., Kriebel, J., Richmond, R.C., Adamo, M., Afzal, U., Ala-Korpela, M., Albetti, B., Ammerpohl, O., Apperley, J.F., Beekman, M., Bertazzi, P.A., Black, S.L., Blancher, C., Bonder, M.-J., Brosch, M., Carstensen-Kirberg, M., de Craen, A.J.M., de Lusignan, S., Dehghan, A., Elkalaawy, M., Fischer, K., Franco, O.H., Gaunt, T.R., Hampe, J., Hashemi, M., Isaacs, A., Jenkinson, A., Jha, S., Kato, N., Krogh, V., Laffan, M., Meisinger, C., Meitinger, T., Mok, Z.Y., Motta, V., Ng, H.K., Nikolakopoulou, Z., Nteliopoulos, G., Panico, S., Pervjakova, N., Prokisch, H., Rathmann, W., Roden, M., Rota, F., Rozario, M.A., Sandling, J.K., Schafmayer, C., Schramm, K., Siebert, R., Slagboom, P.E., Soininen, P., Stolk, L., Strauch, K., Tai, E.-S., Tarantini, L.,

Thorand, B., Tigchelaar, E.F., Tumino, R., Uitterlinden, A.G., van Duijn, C., van Meurs, J.B.J., Vineis, P., Wickremasinghe, A.R., Wijmenga, C., Yang, T.-P., Yuan, W., Zhernakova, A., Batterham, R.L., Smith, G.D., Deloukas, P., Heijmans, B.T., Herder, C., Hofman, A., Lindgren, C.M., Milani, L., van der Harst, P., Peters, A., Illig, T., Relton, C.L., Waldenberger, M., Jarvelin, M.-R., Bollati, V., Soong, R., Spector, T.D., Scott, J., McCarthy, M.I., Elliott, P., Bell, J.T., Matullo, G., Gieger, C., Kooner, J.S., Grallert, H., Chambers, J.C.: Epigenome-wide association study of body mass index , and the adverse outcomes of adiposity. Nature **541**(7635), 81–86 (2017). doi:10.1038/nature20784.Epigenome-wide

[92] Pietrobelli, A., Lee, R.C., Capristo, E., Deckelbaum, R.J., Heymsfield, S.B.: An independent, inverse association of high-density-lipoprotein-cholesterol concentration with nonadipose body mass. The American Journal of Clinical Nutrition **69**(4), 614–620 (1999). doi:10.1093/ajcn/69.4.614

[93] Dayeh, T., Tuomi, T., Almgren, P., Perfilyev, A., Jansson, P.-A., de Mello, V.D., Pihlajamäki, J., Vaag, A., Groop, L., Nilsson, E., Ling, C.: DNA methylation of loci within ABCG1 and PHOSPHO1 in blood DNA is associated with future type 2 diabetes risk. Epigenetics **11**(7), 482–488 (2016). doi:10.1080/15592294.2016.1178418

[94] Wang, G., Padmanabhan, S., Miyamoto-Mikami, E., Fuku, N., Tanaka, M., Miyachi, M., Murakami, H., Cheng, Y.-C., Mitchell, B.D., Austin, K.G., Pitsiladis, Y.P.: Gwas of elite jamaican, african american and japanese sprint athletes: 2254 may 30, 945 am - 1000 am. Medicine & Science in Sports & Exercise **46**(5S) (2014)

[95] Weber, L.M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P.P., Boulesteix, A.-L., Saeys, Y., Robinson, M.D.: Essential guidelines for computational

method benchmarking. Genome Biology **20**(1), 125 (2019). doi:10.1186/s13059-019-1738-8

[96] Schramm, K., Marzi, C., Schurmann, C., Carstensen, M., Reinmaa, E., Biffar, R., Eckstein, G., Gieger, C., Grabe, H.-J., Homuth, G., Kastenmüller, G., Mägi, R., Metspalu, A., Mihailov, E., Peters, A., Petersmann, A., Roden, M., Strauch, K., Suhre, K., Teumer, A., Völker, U., Völzke, H., Wang-Sattler, R., Waldenberger, M., Meitinger, T., Illig, T., Herder, C., Grallert, H., Prokisch, H.: Mapping the genetic architecture of gene regulation in whole blood. PloS one **9**(4), 93844 (2014). doi:10.1371/journal.pone.0093844

[97] Holle, R., Happich, M., Löwel, H., Wichmann, H.E.: KORA - A research platform for population based health research. Gesundheitswesen **67**(SUPPL. 1) (2005). doi:10.1055/s-2005-858235

[98] Lehne, B., Drong, A.W., Loh, M., Zhang, W., Scott, W.R., Tan, S.-T., Afzal, U., Scott, J., Jarvelin, M.-R., Elliott, P., McCarthy, M.I., Kooner, J.S., Chambers, J.C.: A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biology **16**(1), 37 (2015). doi:10.1186/s13059-015-0600-x

[99] Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2019 update. Nucleic acids research **47**(D1), 529–541 (2019). doi:10.1093/nar/gky1079

[100] Siek, J., Lee, L.-Q., Lumsdaine, A.: The Boost Graph Library - User Guide and Reference Manual. Addison-Wesley, Amsterdam (2002)

[101] Efron, B., *et al.*: Microarrays, empirical bayes and the two-groups model. Statistical science **23**(1), 1–22 (2008)

[102] Parsana, P., Ruberman, C., Jaffe, A.E., Schatz, M.C., Battle, A., Leek, J.T.: Addressing confounding artifacts in reconstruction of gene co-expression networks. Genome Biology **20**(1), 94 (2019). doi:10.1186/s13059-019-1700-9

[103] Ernst, J., Kellis, M.: ChromHMM: automating chromatin-state discovery and characterization. Nature Methods **9**(3), 215–216 (2012). doi:10.1038/nmeth.1906

[104] Schäfer, J., Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics **21**(6), 754–764 (2004). doi:10.1093/bioinformatics/bti062. https://academic.oup.com/bioinformatics/article-pdf/21/6/754/506488/bti062.pdf

[105] Zhang, B., Horvath, S.: A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology **4**(1) (2005). doi:10.2202/1544-6115.1128. arXiv:1403.6652v2

[106] Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 1–13 (2020). doi:10.1186/s12864-019-6413-7

[107] Boulesteix, A.-L., Strimmer, K.: Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. Theoretical biology & medical modelling **2**, 23 (2005). doi:10.1186/1742-4682-2-23

[108] Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., Posthuma, D.: A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics **51**(9), 1339–1348 (2019). doi:10.1038/s41588-019-0481-0

[109] Pividori, M., Rajagopal, P.S., Barbeira, A.N., Liang, Y., Melia, O., Bastarache, L., Park, Y., Consortium, T.G., Wen, X., Im, H.K.: PhenomeXcan: Mapping the genome to the phenome through the transcriptome. bioRxiv, 833210 (2019). doi:10.1101/833210

[110] Wen, X., Lee, Y., Luca, F., Pique-Regi, R.: Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. American Journal of Human Genetics **98**(6), 1114–1129 (2016). doi:10.1016/j.ajhg.2016.03.029

[111] Lee, Y., Luca, F., Pique-Regi, R., Wen, X.: Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. bioRxiv, 1–46 (2018). doi:10.1101/316471

[112] Wen, X.: Effective QTL Discovery Incorporating Genomic Annotations. bioRxiv, 032003 (2015). doi:10.1101/032003

[113] Berisa, T., Pickrell, J.K.: Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics **32**(2), 283–285 (2016). doi:10.1093/bioinformatics/btv546

[114] Köster, J., Rahmann, S.: Snakemake–a scalable bioinformatics workflow engine. Bioinformatics **28**(19), 2520–2522 (2012). doi:10.1093/bioinformatics/bts480