nature genetics

Tensor decomposition for multiple-tissue gene expression experiments

Victoria Hore[1], Ana Viñuela[2], Alfonso Buil[3], Julian Knight[4], Mark I McCarthy[4,5], Kerrin Small[2] & Jonathan Marchini[1,4]

Genome-wide association studies of gene expression traits and other cellular phenotypes have successfully identified links between genetic variation and biological processes. The majority of discoveries have uncovered *cis*–expression quantitative trait locus (eQTL) effects via mass univariate testing of SNPs against gene expression in single tissues. Here we present a Bayesian method for multiple-tissue experiments focusing on uncovering gene networks linked to genetic variation. Our method decomposes the 3D array (or tensor) of gene expression measurements into a set of latent components. We identify sparse gene networks that can then be tested for association against genetic variation across the genome. We apply our method to a data set of 845 individuals from the TwinsUK cohort with gene expression measured via RNA-seq analysis in adipose, lymphoblastoid cell lines (LCLs) and skin. We uncover several gene networks with a genetic basis and clear biological and statistical significance. Extensions of this approach will allow integration of different omics, environmental and phenotypic data sets.

Studies of cellular phenotypes are transforming understanding of the genetic influences on complex traits. Genomic screens of gene expression levels[1], chromatin accessibility[2], chromatin state[3] and protein levels[4] are all helping to elucidate how genetics is related to disease mechanisms. Over the last few years, eQTL mapping has emerged as a key component in this research and has led to the identification of many genetic variants affecting gene expression. Typically, these studies involve assaying gene expression in a single tissue or cell type, although multiple-tissue experiments are beginning to emerge as a way to uncover the principles of gene regulation.

The standard paradigm for single-tissue eQTL studies involves testing the expression of each gene or transcript against SNP genotypes in a local region to identify *cis*-eQTLs. This approach has been successful, with recent large eQTL studies suggesting that there will be at least one *cis*-eQTL for almost all expressed genes[5]. Multiple-tissue approaches can increase the power to find *cis*-eQTLs[6]; however, as *cis*-eQTLs are estimated to account for only 30–40% of the heritability of expression levels[7,8], there is a need to identify *trans*-eQTLs to account for the remaining heritability.

The detection of *trans*-eQTLs and networks of genes with related expression patterns is hard both computationally and statistically. Testing all genes against all SNPs via tens of thousands of genome-wide scans incurs a substantial penalty for multiple testing. In addition, *trans*-eQTL effect sizes tend to be smaller than those for *cis*-eQTLs, making the detection of *trans*-eQTLs harder[9]. For these reasons, scans for *trans*-eQTLs usually work with a reduced set of genetic variants, such as those associated with disease traits[9,10]. In general, the approach of carrying out very large numbers of marginal statistical tests (of one SNP versus one gene at a time) ignores the complex structure of these data sets. The expression levels of each gene will likely be due to a mixture of several different sources related to underlying biology as well as confounding factors.

In this paper, we present a new method for the analysis of multiple-tissue gene expression experiments with a specific focus on identifying *trans*-eQTLs and gene networks. The data from such experiments can be viewed as a '3D' array, or tensor, with dimensions representing individuals, genes and tissues (**Fig. 1**). Our method decomposes this tensor into a number of latent components (or factors) that represent major modes of variation in the data set. Each of these components consists of three vectors of scores (or loadings) that indicate the relative contribution of each individual, gene and tissue. For example, if a consistent pattern of gene expression across a network of genes occurs in a subset of tissues, with a different magnitude in each individual, then our model aims to represent this in a single component. Such signals might naturally arise because of transcription factor genes that have multiple targets throughout the genome. If the expression level or function of a gene is altered by *cis*-acting genetic variants, then we would likely observe different magnitudes of effect across individuals.

One useful way to think about the approach is as analogous to the use of principal-components analysis (PCA) applied to '2D' (individual-by-SNP) genetic data sets. PCA is used routinely to decompose genome-wide SNP data sets into components of variation that are then used to understand population structure (for example, see ref. 11). Here we aim to decompose

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [3]Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland. [4]Wellcome Trust Centre for Human Genetics, Oxford, UK. [5]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. Correspondence should be addressed to J.M. (marchini@stats.ox.ac.uk).

Received 11 September 2015; accepted 22 June 2016; published online 1 August 2016; doi:10.1038/ng.3624

higher-dimensional data sets into components that uncover real biology. Our method has several notable properties:

- Our approach is developed in a Bayesian framework and we use a sparse 'spike-and-slab' prior[12], allowing the gene loadings of each component to have a unique level of sparsity. This allows us to shrink gene effects to zero so that we can more clearly infer which genes are involved in gene networks.
- Individual scores represent the magnitude of the effect of each component across individuals, analogous to the individual scores that are usually plotted in PCA for genetic data sets. We use these scores as phenotypes in genome-wide SNP scans to identify genetic variants that drive each component. The number of components we test (a few hundred) is typically much smaller than the number of genes (tens of thousands), which substantially reduces the multiple-testing burden as compared to approaches that test all genes against all genetic variants in all tissues.
- We do not claim that all genes identified in a network will reach genome-wide significance thresholds with the driving SNPs. However, when the method is applied to real data sets, we find that the majority of genes are nominally significant.
- The tissue scores vector indicates the 'activity' of the component for each tissue. By examining the entries of the tissue scores matrix across components, we can make inferences about how many components are shared across tissues.
- Our model also allows for non-sparse components that might be expected to arise from confounding effects, such as batch effects or sequencing properties.
- In addition, the model can naturally accommodate missing data, such as samples without gene expression data for subsets of tissues, which is a real and prevalent feature of multiple-tissue experiments.

Our motivation for this work stemmed from similar approaches that have emerged in the field of neuroscience to uncover shared signals across different high-dimensional imaging modalities[13,14]. Most tensor decomposition methods[15–17] are not able to handle missing data or invoke sparsity on the components, although there are some exceptions[18]. Our model is the first tensor decomposition method to our knowledge to use spike-and-slab priors with model fitting carried out using Variational Bayes (VB; Online Methods). Via extensive simulations (**Supplementary Note**), we show that our method has the best performance in terms of estimation of component individual scores and recovery of sparsity patterns in gene loadings, when compared to other matrix and tensor decomposition methods, and is well powered to detect *trans*-eQTL signals and gene networks. Our method is implemented in a software package called SDA (Sparse Decomposition of Arrays) (see URLs).

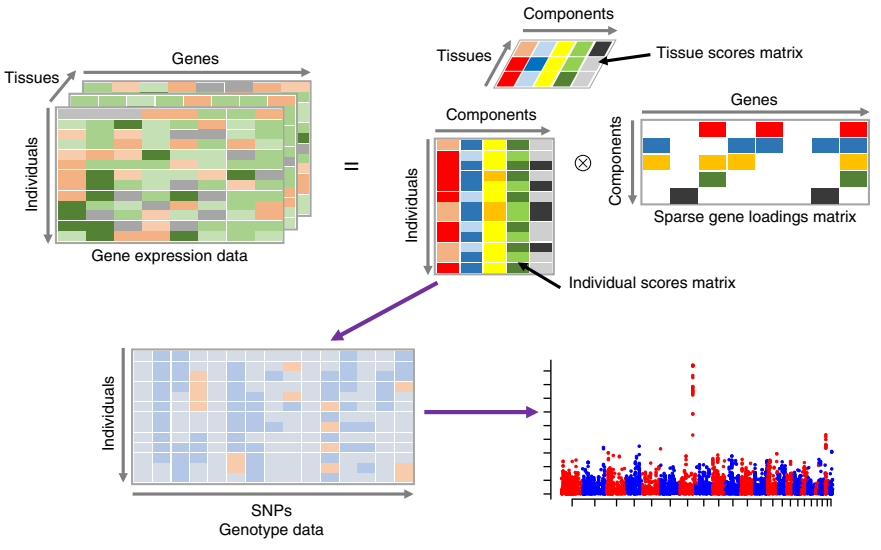

Figure 1 Graphical representation of the method. The gene expression data tensor (top left) is decomposed into the product of an individual scores matrix, a tissue scores matrix and a gene loadings matrix (top right). Columns of the individual scores matrix are then used as phenotypes in a GWAS using SNP genotypes (bottom left) to uncover genetic variation correlated with the latent components (bottom right).

RESULTS

We have validated our approach by applying it to RNA-seq data from the TwinsUK cohort, which consist of gene expression measured in 845 related individuals in adipose, LCLs and skin[19,20]. To focus on robustly identified components, we applied our method ten times to the TwinsUK RNA-seq data set and combined the results across runs via clustering (Online Methods). After clustering, we identified 236 robust components for further investigation. Examination of the tissue scores matrix is informative in determining in which tissues each component is active (**Supplementary Fig. 1**). We found that the majority of the 236 components were active in a single tissue (57 in adipose, 74 in LCLs and 70 in skin). There were 20 components that were active in all three tissues, 14 components that were active in just adipose and skin, and 1 component that was active in adipose and LCLs. Full details on these 236 components are given in the **Supplementary Data**.

The individual scores vectors of these components were then used as phenotypes in genome-wide scans using SNP genotype data imputed with the 1000 Genomes Project Phase 1 reference panel. We used a *P*-value threshold of 1×10^{-10} to assign significance (Online Methods). There were 26 components that reached this level of significance: 5 components were active in just one tissue (1 in adipose and 4 in LCLs), 20 components were active in all three tissues, and 1 component was active in just adipose and skin. The majority of these components were clearly uncovering *cis*-eQTLs. In all but two of these components, we identified pairs of SNPs (significantly associated with our component scores) and genes (with a nonzero loading) that had previously been identified as *cis*-eQTLs in the MuTHER (Multiple-Tissue Human Expression Resource) and GTEx (Genotype–Tissue Expression) studies[7,21]. These components exhibited very sparse gene loadings, with a single localized cluster of high gene loadings and highly significant SNP associations in the flanking region (**Supplementary Figs. 2–21**). Methodology for the detection of *cis*-eQTLs is well established and is best carried out using focused analysis that looks for such effects at SNPs flanking each gene. Our main focus was on uncovering *trans*-eQTLs and gene networks, so we did not pursue the *cis*-eQTLs that our method found any further.

The remaining six components were less sparse in their gene loadings and exhibited patterns of gene loadings and SNP associations that highlighted gene expression networks with substantial biological relevance. For these networks, the majority of gene loadings tended to be unidirectional, suggesting that the components are identifying a directional

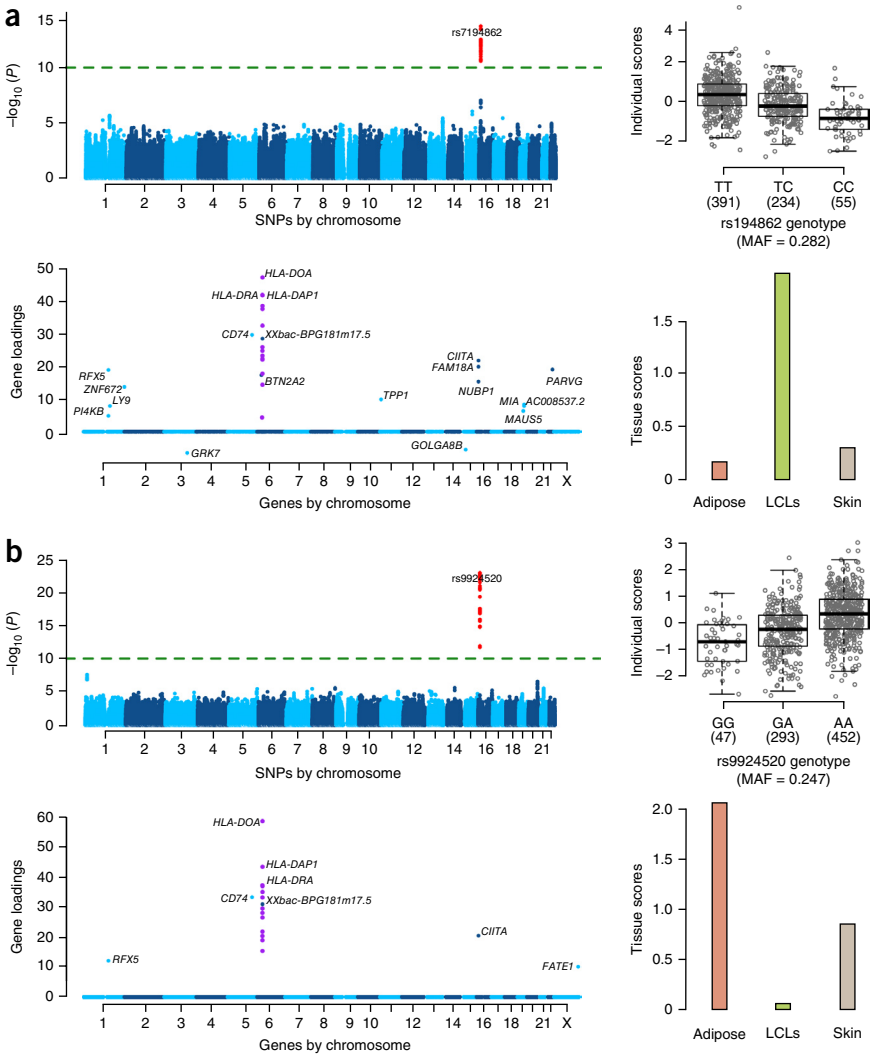

Figure 2 MHC class II regulation. (**a**,**b**) The plots show two components identifying a similar network in different tissues. Top left, GWAS with the component's individual scores vector as a phenotype. The dashed line corresponds to a *P*-value thresholds of 1×10^{-10}. Top right, box plots of individual scores stratified by genotype at the lead GWAS SNP. Box plots show the median and upper and lower quartiles, with whiskers extending to either 1.5 times the interquartile range (IQR) or to the most extreme data point if this was within 1.5 times the IQR. Bottom left, gene loadings for the component. Only gene loadings with PIP >0.5 are shown. Bottom right, tissue scores vector for the component shown as a bar plot.

effect on expression. These components are summarized in **Figures 2–6**, which show the gene loadings, SNP genome-wide association studies (GWAS) and tissue activation patterns. The majority of genes identified in each of these networks had nominally significant *P* values in the relevant tissues (**Supplementary Table 1**). We also applied PCA and independent-component analysis (ICA) to the TwinsUK data set and found that neither of these approaches uncovered the gene networks reported here; more details are given in the **Supplementary Note**.

We found two clustered components (**Fig. 2**) with individual scores that exhibited significant SNP associations in the gene *CIITA* on chromosome 16p13 (**Supplementary Fig. 22**). The first component was active mostly in adipose and skin and had a lead SNP, rs9924520 ($P = 1.33 \times 10^{-23}$, minor allele frequency (MAF) = 0.247), that is an intronic variant of *CIITA*. The second component was active mostly in LCLs and had a lead SNP, rs7194862 ($P = 1.74 \times 10^{-14}$, MAF = 0.282), that is 5′ to *CIITA*. The rs9924520 and rs7194862 SNPs are in strong linkage disequilibrium (LD; $r^2 = 0.82$). Both components showed a cluster of major histocompatibility complex (MHC) class II genes on chromosome 6 with nonzero gene loadings. In addition, two other genes had significant gene loadings in both components (*RFX5* on chromosome 1 and *CD74* on chromosome 5). *CIITA* is known to be a master controller in the regulation of MHC class II gene expression[22]. Its gene product is recruited to the proximal promoter regions of the classical MHC class II genes (HLA-DP, HLA-DR and HLA-DQ genes) and to the HLA-DM and HLA-DO genes and *CD74* (encoding the molecular chaperone invariant chain, which associates with the MHC class II complex) through protein–protein interactions with other components of the MHC class II enhanceosome, which includes *RFX5*. The direct associations of SNPs rs9924520 and rs7194862 with expression levels of all the genes identified in our components (in all three tissues) after correction for covariates and 15 factors derived using the method PEER[23] are detailed in **Supplementary Table 2** (Online Methods). Both SNPs were strongly associated with *HLA-DOA* and *HLA-DMA* in adipose and skin (*P* values in the range 2.89×10^{-8} to 5.56×10^{-19}) and with *CIITA* in adipose ($P = 2.08 \times 10^{-11}$ and 1.44×10^{-12} for rs9924520 and rs7194862, respectively). However, neither SNP reached a strict Bonferroni-corrected *P*-value threshold for *trans* analysis of $9.05 \times 10^{-13} = 5 \times 10^{-8}/(3 \times 18{,}409)$ (obtained by accounting for genome-wide testing across all genes in all tissues) with any of the other genes in the three tissues. These results suggest that, although a *trans*-eQTL association would have been found between SNPs in the *CIITA* region and expression at two MHC class II genes via a marginal *trans* analysis, the more extensive network of genes recovered by our components would not have been uncovered.

A component with significant associations in the gene *NLRC5* (also known as *CITA*) on chromosome 16q13 is shown in **Figure 3** (**Supplementary Fig. 23**). The lead SNP rs289749 ($P = 1.34 \times 10^{-11}$, MAF = 0.3) is an intronic variant of *NLRC5*. The component showed a cluster of genes on chromosome 6 with nonzero gene loadings that included MHC class I genes (*HLA-C*, *HLA-B*, *HLA-F*, *HLA-A* and *HLA-E*), BTN genes (*BTN3A2*, *BTN3A1*, *BTN3A3*, *BTN2A2* and *BTN2A1*), *TAP1*, *TAP2*, *PSMB8* and *PSMB9*. Overexpression of *NLRC5* is known to increase the mRNA levels of genes encoding human MHC class I molecules and proteins functioning in the MHC class I–mediated antigen presentation pathway, including *B2M* (β2 microglobulin), *TAP1* (transporter associated with antigen processing 1) and *PSMB9* (proteasome subunit β type 9)[24]. *B2M*, *TAP1* and *PSMB9* all had significant gene loadings in the component. The direct associations of SNP rs289749 with the expression

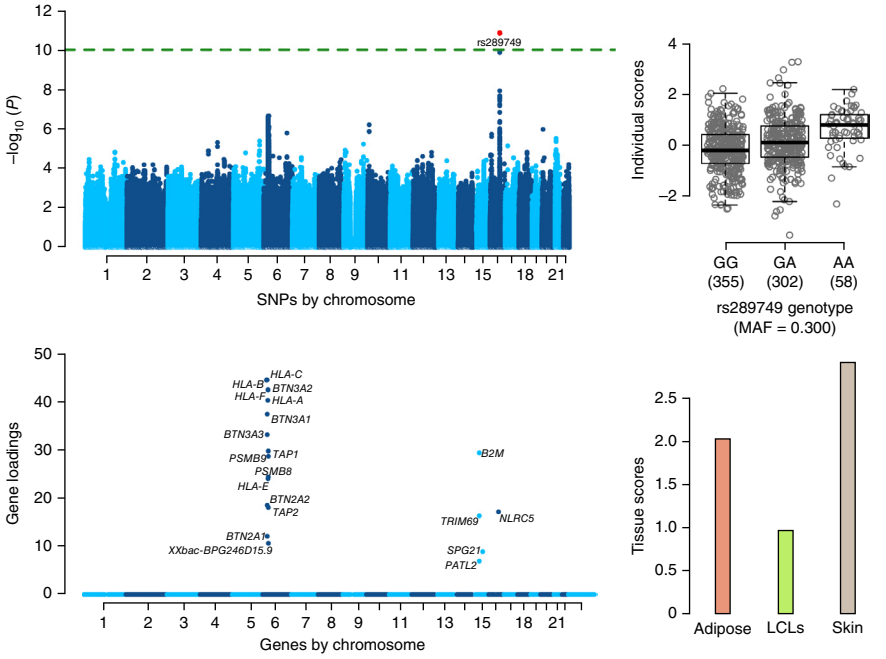

Figure 3 MHC class I regulation. Top left, GWAS with the component's individual scores vector as a phenotype. Top right, box plot of individual scores stratified by genotype at the lead GWAS SNP, rs289749. Bottom left, gene loadings for the component. Only gene loadings with PIP >0.5 are shown. Bottom right, tissue scores vector for the component shown as a bar plot.

levels of all the genes in the component in all three tissues are detailed in **Supplementary Table 3**. In skin, rs289749 was strongly associated with *NLRC5* ($P = 1.37 \times 10^{-28}$) and moderately associated with several MHC class I genes (*HLA-F* ($P = 3.02 \times 10^{-12}$), *HLA-A* ($P = 1.22 \times 10^{-9}$) and *HLA-B* ($P = 1.35 \times 10^{-10}$)), although none of these associations would pass a Bonferroni-corrected significance level in a *trans* analysis (9.05×10^{-13}). *P* values for association between rs289749 and other genes in this component suggest that the link between *NLRC5* and BTN, TAP and PSMB genes or the *B2M* gene would not have been recovered using a traditional *trans* analysis. In addition, these direct associations failed to provide evidence for the signal in either adipose or LCLs.

A component significantly associated with a cluster of SNPs near *LSM11* on chromosome 5q33.3 that is known to be involved in histone RNA processing[25] is shown in **Figure 4** (**Supplementary Fig. 24**). The gene loadings of our component showed a striking cluster of 23 histone genes in the chromosome 6p21 cluster as well as the gene *HIST2H2BE* in the 1q21 cluster (**Fig. 4**, purple points). There were also additional signals at other histone genes on chromosomes 1q42 (*HIST3H2A*), 11q23 (*H2AFX*) and 12p12 (*HIST4H4*). SNP rs6882516 ($P = 2.39 \times 10^{-15}$, MAF = 0.206) is in the 3′ UTR of *LSM11* and is predicted to be a microRNA-binding site using mirSNP[26]. Key histone gene regulatory factors are organized in a limited number of subnuclear foci. It is known that cell cycle–dependent phosphorylation of p220^{NPAT} by cyclin E–CDK2, which induces histone gene transcription, occurs at these intranuclear sites. p220^{NPAT} colocalizes with both (i) the histone gene clusters on chromosomes 1q21 and 6p21 and (ii) the protein subunit gene *LSM11* (ref. 13). A set of 31 significant genes (loadings with posterior inclusion probability (PIP) >0.5; Online Methods) showed Gene Ontology *P* values of 1.91×10^{-25} and 1.40×10^{-24} for the terms 'nucleosome organization' and 'chromatin assembly or disassembly', respectively. The tissue scores indicate that this component was only active in LCLs. The direct associations of SNP rs6882516 with expression levels of *LSM11* and the other genes in this component in all three tissues are detailed in **Supplementary Table 4**. The SNP was significantly associated with *LSM11* in LCLs ($P = 5.57 \times 10^{-33}$) and had *P* values in the range 2.65×10^{-12} to 1.17×10^{-12} with three histone genes in our component with extreme gene loadings (*HIST1H1C*, *HIST1H2BJ* and *HIST1H2BK*). Although these associations are encouraging, they did not pass a strict *trans*-analysis significance level, and additionally these direct associations did not uncover the link between *LSM11* and the histone gene cluster at 1q21 (the *P* value for rs6882516 and *HIST2H2BE* in LCLs was 5.40×10^{-9}).

A component significantly associated with a SNP near *USP18* is shown in **Figure 5** (**Supplementary Fig. 25**). The lead SNP rs2401506 ($P = 9.82 \times 10^{-16}$, MAF = 0.358) is 5 kb upstream of *USP18*. The set of 160 genes in the loadings with PIP >0.5 showed

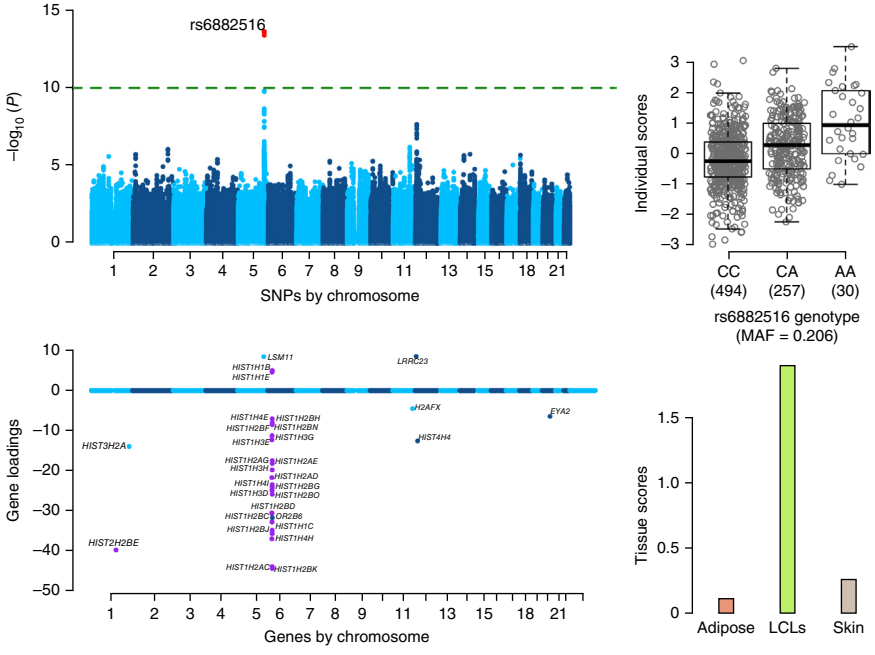

Figure 4 Histone RNA processing. Top left, GWAS with the component's individual scores vector as a phenotype. Top right, box plot of individual scores stratified by genotype at the lead GWAS SNP, rs6882516. Bottom left, gene loadings for the component. Only gene loadings with PIP >0.5 are shown. Bottom right, tissue scores vector for the component shown as a bar plot.

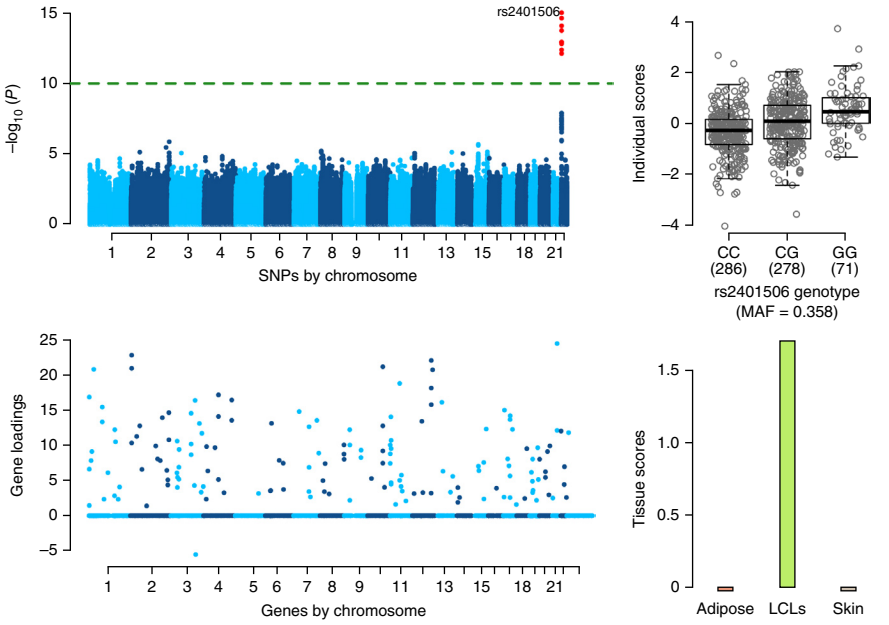

Figure 5 Type I interferon response. Top left, GWAS with the component's individual scores vector as a phenotype. Top right, box plot of individual scores stratified by genotype at the lead GWAS SNP, rs2401506. Bottom left, gene loadings for the component. Only gene loadings with PIP >0.5 are shown. Bottom right, tissue scores vector for the component shown as a bar plot.

Gene Ontology *P* values of 1.73×10^{-42} and 1.23×10^{-38} for the terms 'defense response to virus' and 'response to type I interferon', respectively. Of the 70 genes annotated by 'response to type I interferon', we found 28 with nonzero gene loadings (**Supplementary Fig. 26**). These included all four members of the 2′–5′ oligoadenylate synthetase (OAS) gene family (*OAS1*, *OAS2*, *OAS3* and *OASL*) known to be actively induced by interferons[27], the genes *STAT1* and *STAT2*, which are key mediators of type I and type III interferon signaling, several interferon γ–inducible protein (IFI) genes (*IFI6*, *IFI44L*, *IFI16*, *IFIH1*, *IFIT1*, *IFIT3*, *IFIT5*, *IFIT2*, *IFITM1*, *IFITM2* and *IFI35*) and the genes *MX1* and *MX2* also related to interferon signaling. USP18, a type I interferon–induced protein that deconjugates the ubiquitin-like modifier ISG15 (which is also in our component) from target proteins[28], has an important function in downregulation of interferon responses[29,30] and significantly inhibits tumor growth[31]. The tissue scores indicated that this component was only active in LCLs. The direct associations of SNP rs2401506 with the 160 genes identified in this component across all three tissues are detailed in **Supplementary Table 5**. There was only evidence of association in LCLs, with several genes obtaining *P* values smaller than 1×10^{-8} (*IFIT1*, *PLSCR1*, *STAT1*, *CMPK2*, *RSAD2* and *EIF2AK2*), but none were significant when accounting for genome-wide testing across all genes, suggesting that this network of genes would not have been uncovered by a scan of all SNPs versus all genes.

Two significant associations on separate chromosomes for a component with a striking cluster of nonzero gene loadings for zinc-finger genes on chromosome 19 are shown in **Figure 6**. SNP rs17611866 ($P = 5.40 \times 10^{-21}$, MAF = 0.251) on chromosome 16 is a missense variant in *ZNF75A*, which is one of six zinc-finger genes in a local cluster. The flanking genes *ZNF263* and *TIGD7* had nonzero gene loadings (**Supplementary Fig. 27**). SNP rs12630796 ($P = 5.10 \times 10^{-17}$, MAF = 0.487) on chromosome 3 is an intronic SNP in *SENP7*. A SNP in high LD with this SNP (rs13320918, $P = 7.34 \times 10^{-15}$, MAF = 0.377) has been shown to be a microRNA QTL for miR-1270 ($P = 1.71 \times 10^{-10}$), which is located in a zinc-finger gene cluster on chromosome 19p12 (ref. 32). In a separate study, four other intronic SNPs in *SENP7* (rs2553419, rs2682386, rs9859077 and rs2141180), all in high LD with each other and with rs13320918, were shown to correlate with *cis*-acting regulation of *SENP7* expression in $CD4^+$ and $CD8^+$ lymphocytes and *trans*-acting regulation of *ZNF154*, *ZNF274* and *ZNF814* (ref. 33), which all reside within a ~250-kb region on chromosome 19q13.43 (**Supplementary Fig. 28**).

The direct associations of SNPs rs12630796 and rs17611866 with *SENP7* on chromosome 3 and genes with nonzero gene loadings in the component in all three tissues are detailed in **Supplementary Table 6**. This analysis partially recovered the signal that we found using our method (see the **Supplementary Note** for more details).

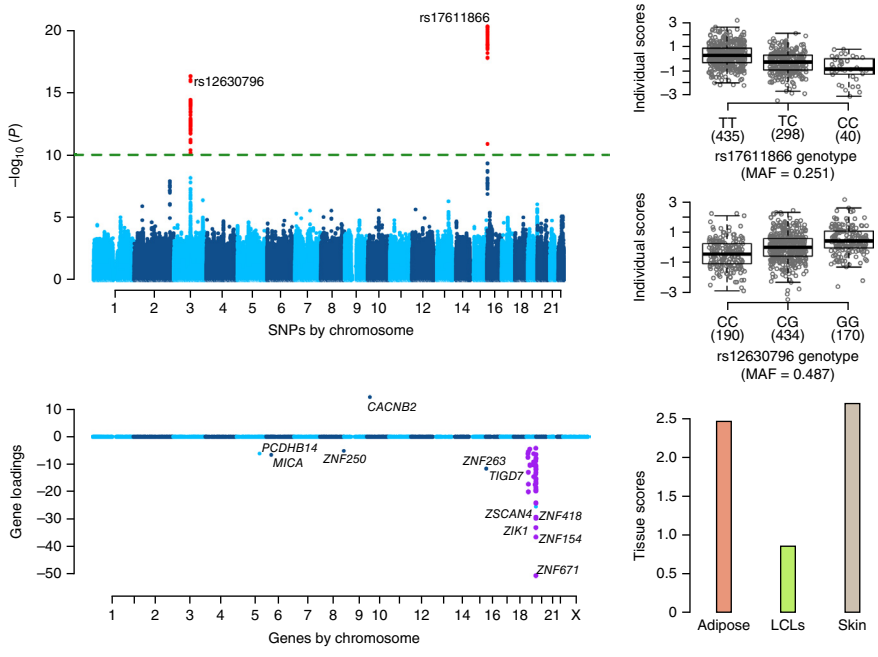

Figure 6 Zinc-finger gene network. Top left, GWAS with the component's individual scores vector as a phenotype. Top right, box plots of individual scores stratified by genotype at the lead GWAS SNPs, rs17611866 and rs12630796. Bottom left, gene loadings for the component, with zinc-finger genes on chromosome 19 highlighted in purple. Only gene loadings with PIP >0.5 are shown. Bottom right, tissue scores vector for the component shown as a bar plot.

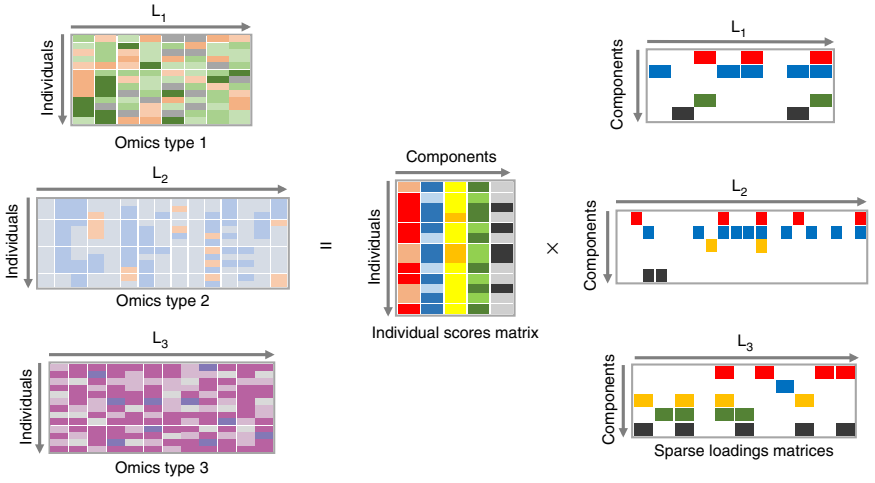

Figure 7 Multiple-omics data integration. Graphical representation of a linked decomposition for several genomic assays. A matrix decomposition is applied to each data type. The matrix decompositions identify a different loadings matrix for each data type and a shared individual scores matrix.

It can be challenging to interpret the large number of components that are produced by sparse matrix and tensor decomposition methods. By clustering components across independent runs of the method and then selecting components with genetic associations, we have shown that it is possible to identify gene networks with clear biological relevance. However, we have found evidence that the components without genetic associations also capture important variance in the data. Many components had individual scores vectors that were significantly associated with variables measuring properties of the sequencing; these components were mostly dense, with several thousand nonzero gene loadings (**Supplementary Figs. 29–31** and **Supplementary Table 7**). Similarly, we identified several components that were significantly associated with measured phenotypes, including age, body mass index (BMI) and cholesterol levels (**Supplementary Fig. 32**). We found two components that showed association with age. These components are shown in **Supplementary Figures 33** and **34**. The most significant molecular function ontology term for both components was 'oxidoreductase activity', with P values of 1.9×10^{-24} and 2.1×10^{-22}.

In addition, we have found that it can be useful to examine the components from a single run of the method. Specifically, we focused on the best run of ten that produced the highest value of the model negative free energy (Online Methods). We identified all components highlighted in **Figures 2–6** with significant or very close to significant GWAS P values. In addition, we found several components that identified *KLF14* as a master *trans* regulator[34] (for example, see **Supplementary Fig. 35**). More details are given in the **Supplementary Note** and the **Supplementary Data**.

A previous analysis of a similar set of samples in the MuTHER study[7] using a microarray-based gene expression experiment called 518, 491 and 493 *trans*-eQTL SNPs at a normal GWAS P-value threshold of 5×10^{-8}. The study reported an FDR of <10% at this threshold; however, only ~5% of these signals replicated at a nominal significance threshold of 0.05 in at least one of five other studies. The overlap with our results included (i) a SNP, rs7714390, on chromosome 5 (near our lead SNP rs6882516) associated with two histone genes (*HIST1H2BK* on chromosome 6, $P = 8 \times 10^{-9}$ in LCLs; *HIST2H2BE* on chromosome 1, $P = 3.2 \times 10^{-8}$ in LCLs); (ii) a SNP, rs220377, on chromosome 16 (near our lead SNP rs17611866) associated with a zinc-finger gene (*ZNF667* on chromosome 19, $P = 2.9 \times 10^{-9}$ in LCLs); and (iii) several associated SNPs near rs4731702 that overlapped with the *KLF14* network (P values between 4.4×10^{-8} and 2.2×10^{-15}). This analysis did not identify the type I interferon network or the MHC networks that we found in our analysis.

DISCUSSION

We have described a new algorithm for efficient tensor decomposition for multiple-tissue gene expression data sets and have demonstrated its usefulness on a real three-tissue data set to uncover sparse gene networks with clear biological and statistical significance. A marginal analysis of all SNPs versus all genes would not have uncovered these networks in the same way or with as much power. For example, no aspect of the type I interferon component would have been identified. We have further shown in simulations that our method has good power to detect sparse gene networks correlated with genetic variants and dense confounding factors.

This approach complements current eQTL analysis pipelines that tend to focus mainly on identifying *cis*-eQTLs in one tissue at a time. Analysis of cross-tissue effects usually proceeds in a subsequent step by comparing effect sizes across tissues. Our method focuses on decomposing the complete multiple-tissue data set into components of variance with varying levels of sparsity. We then test each component against genetic variation across the genome to uncover underlying eQTL effects, ensuring robustness by only considering components that are consistently found across multiple runs. We view our approach as complementary to an association analysis of all SNPs versus all genes, as it requires two orders of magnitude fewer tests and has more power to detect SNP associations with gene networks.

In general, we find that dense components uncovered by our method show high levels of significance with confounding variables, and the method additionally uncovers many very sparse components that represent *cis*-eQTLs. More interestingly, we find six components with intermediate levels of sparsity with gene loadings spread across multiple chromosomes that represent gene networks showing a highly significant association with genetic variants. In all six of these components, we are able to link the gene networks they describe to known biology. In the future, it will be natural to apply this method to gene expression data sets with even more tissues, such as that being collected by the GTEx Project[21] or the Allen Institute for Brain Science (AIBS) human microarray data set[35].

There are several interesting ways in which this model can be extended or changed. The method can be naturally extended to higher-dimensional data sets. For example, 4D multiple-tissue gene expression experiments through time and/or under different experimental conditions (**Supplementary Fig. 36**).

One assumption of our model is that the gene loadings pattern of a component is constant across active tissues, which may or may not be true depending upon the data set being analyzed. One way to overcome this would be to develop a model that applies a matrix decomposition to the gene expression matrix for each tissue but with a linked individual scores matrix (**Supplementary Fig. 37**). A downside of such an approach is that it would substantially increase the number of unknown parameters in the factorization. However, this model would allow variation in the gene loadings between tissues if

there were indeed clear differences and might be a way of combining together components found by our tensor method (like those describing MHC class II regulation pathways) with clearly similar gene loadings. However, it may also be necessary to model the similarity between gene loadings to aid estimation, given the larger parameter space. This approach has strong connections to sparse canonical correlations analysis (CCA)[36] and unsupervised multiview learning[37].

Such a linked matrix decomposition method could also be used to integrate different genomic data sets. The model has no constraint that the set of matrices being jointly decomposed have the same dimensions. So, for example, matrices of gene expression and epigenetic measurements could be jointly decomposed to uncover relevant shared biology (**Fig. 7**). Example applications might include joint decomposition of different omics data sets collected on cancer samples from the International Cancer Genome Consortium (ICGC; see URLs). This model can be extended further to tensors of different data types (**Supplementary Fig. 38**).

URLs. Sparse Decomposition of Arrays (SDA) software, https://jmarchini.org/sda/; International Cancer Genome Consortium (ICGC), https://dcc.icgc.org/projects/details; fastICA, https://cran.r-project.org/web/packages/fastICA/index.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to A. Dahl, W. Kretzschmar, K. Sharp, L. Elliot and S. Myers for helpful discussions about the method and interpretation of the results. The TwinsUK cohort was funded by the Wellcome Trust and the European Community's Seventh Framework Programme (FP7/2007-2013). The study also receives support from the NIHR Clinical Research Facility at Guy's and St Thomas' NHS Foundation Trust and the NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. SNP genotyping was performed by the Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR. A.V. and A.B. were supported by European Union Framework Programme 7 grant EuroBATS (259749). V.H. acknowledges the EPSRC for funding through a studentship at the Life Sciences Interface program of the University of Oxford's Doctoral Training Center. J.M. acknowledges support from the ERC (grant 617306).

AUTHOR CONTRIBUTIONS

V.H. and J.M. developed the method. V.H. carried out all analysis. J.M. and V.H. wrote the manuscript. A.V., A.B. and K.S. provided the TwinsUK data set. A.V., A.B., J.K., M.I.M. and K.S. advised on interpretation of the results.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
2. Degner, J.F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
3. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
4. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
5. Pai, A.A., Pritchard, J.K. & Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.* **11**, e1004857 (2015).
6. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
7. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
8. Price, A.L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
9. Westra, H.-J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
10. Yao, C. *et al.* Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* **131**, 536–549 (2015).
11. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
12. Mitchell, T.J. & Beauchamp, J.J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **83**, 1023–1032 (1988).
13. Groves, A.R., Beckmann, C.F., Smith, S.M. & Woolrich, M.W. Linked independent component analysis for multimodal data fusion. *Neuroimage* **54**, 2198–2217 (2011).
14. Groves, A.R. *et al.* Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *Neuroimage* **63**, 365–380 (2012).
15. Kolda, T.G. & Bader, B.W. Tensor decompositions and applications. *SIAM Rev.* **51**, 455–500 (2009).
16. Yener, B. *et al.* Multiway modeling and analysis in stem cell systems biology. *BMC Syst. Biol.* **2**, 63 (2008).
17. Hoff, P.D. Hierarchical multilinear models for multiway data. *Comput. Stat. Data Anal.* **55**, 530–543 (2011).
18. Khan, S.A., Leppaaho, E. & Kaski, S. Bayesian multi-tensor factorization. Preprint at https://arxiv.org/abs/1412.4679 (2014).
19. Buil, A. *et al.* Gene–gene and gene–environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
20. Brown, A.A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3**, e01381 (2014).
21. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
22. Reith, W., LeibundGut-Landmann, S. & Waldburger, J.M. Regulation of MHC class II gene expression by the class II transactivator. *Nat. Rev. Immunol.* **5**, 793–806 (2005).
23. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
24. Kobayashi, K.S. & van den Elsen, P.J. NLRC5: a key regulator of MHC class I–dependent immune responses. *Nat. Rev. Immunol.* **12**, 813–820 (2012).
25. Pillai, R.S. *et al.* Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes Dev.* **17**, 2321–2333 (2003).
26. Liu, C. *et al.* MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* **13**, 661 (2012).
27. Melchjorsen, J. *et al.* Differential regulation of the *OASL* and *OAS1* genes in response to viral infections. *J. Interferon Cytokine Res.* **29**, 199–207 (2009).
28. Potu, H., Sgorbissa, A. & Brancolini, C. Identification of USP18 as an important regulator of the susceptibility to IFN-α and drug-induced apoptosis. *Cancer Res.* **70**, 655–665 (2010).
29. Malakhova, O.A. *et al.* UBP43 is a novel regulator of interferon signaling independent of its ISG15 isopeptidase activity. *EMBO J.* **25**, 2358–2367 (2006).
30. François-Newton, V. *et al.* USP18-based negative feedback control is induced by type I and type III interferons and specifically inactivates interferon α response. *PLoS One* **6**, e22200 (2011).
31. Burkart, C. *et al.* Usp18 deficient mammary epithelial cells create an antitumour environment driven by hypersensitivity to IFN-λ and elevated secretion of Cxcl10. *EMBO Mol. Med.* **5**, 967–982 (2013).
32. Huan, T. *et al.* Genome-wide identification of microRNA expression quantitative trait loci. *Nat. Commun.* **6**, 6601 (2015).
33. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* **6**, 6326 (2015).
34. Small, K.S. *et al.* Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
35. Hawrylycz, M.J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
36. Witten, D.M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
37. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **23**, 2031–2038 (2013).

ONLINE METHODS

Bayesian sparse tensor decomposition model. We use Y to denote the 3D array, or tensor, containing preprocessed gene expression measurements. Y has dimensions $N \times L \times T$, where N is the number of individuals, L is the number of genes and T is the number of tissues. We model Y as follows

$$Y_{nlt} = \sum_{c=1}^{C} A_{nc} B_{tc} X_{cl} + \varepsilon_{nlt}$$

where C is the number of components (also called factors). A is an $N \times C$ matrix with the cth column containing the individual scores of the cth component. B is a $T \times C$ matrix with the cth column containing the tissue scores of the cth component. X is a $C \times L$ matrix with the cth row containing the gene loadings of the cth component.

The error term is modeled as

$$\varepsilon_{nlt} \sim N(0, \lambda_{lt}^{-1})$$

where λ_{lt} is the precision of the error term at the lth gene in the tth tissue.

We deal with missing samples for a given tissue by not including them in the model likelihood. We introduce an indicator variable I_{nt} that equals 1 when gene expression has been measured in tissue t for sample n and zero otherwise. The likelihood is then given by

$$P(Y|\Theta) = \prod_{n,l,t} P(Y_{nlt}|\Theta)^{I_{nt}}$$

where Θ is the vector of model parameters.

We fit this model in a Bayesian framework and place priors on the entries of the matrices A, B and X and also on the precisions λ_{lt}. A key prior is the one we place on the elements of the gene loadings matrix X. We wish to encourage sparsity in the rows of this matrix, so we use a hierarchical spike-and-slab prior[38] of the form

$$\begin{aligned} X_{cl} &\sim p_{cl} N(0, \beta_c^{-1}) + (1 - p_{cl})\delta_0 \\ \beta_c &\sim Gamma(e, f) \\ p_{cl} &\sim \rho_c Beta(q, r) + (1 - \rho_c)\delta_0 \\ \rho_c &\sim Beta(s, z) \end{aligned}$$

For the purposes of making inference easier (**Supplementary Note**), we use the equivalent factorization of the spike-and-slab distribution as $X_{cl} = W_{cl} S_{cl}$ where

$$\begin{aligned} W_{cl} &\sim N(0, \beta_c^{-1}) \\ S_{cl} &\sim \text{Bernoulli}(p_{cl}) \end{aligned}$$

For the elements of A and B, we use standard normal priors $A_{nc} \sim N(0,1)$ and $B_{tc} \sim N(0,1)$.

Model fitting. We fit this model using Variational Bayes (VB)[39], which approximates the posterior distribution $P(\Theta|Y) \approx Q(\Theta)$. The approach iteratively refines the estimate $Q(\Theta)$, by minimizing the Kullback–Leibler (KL) divergence between $Q(\theta)$ and $P(Y|\theta)$ or equivalently maximizes the negative free energy. Once converged, $Q(\theta)$ can be used to approximate properties of the posterior distribution. The full details of the parameter factorization we use, the resulting VB update equations and details of parameter initialization are given in the **Supplementary Note.** The resulting algorithm has complexity $O(NLTC^2)$ and can be run on a multicore server. For the TwinsUK data analyzed in this paper, the method took 20 h for each of the ten runs using eight threads.

Our model has the ability to shrink an entire component to zero ($\rho_c = 0$) and effectively remove that component from the model. In this way, our model can adaptively choose the number of components it needs. Just a small amount of experimentation is needed to find a large enough value of C so that components start being shrunk to zero. For the TwinsUK data, we fit the model with 1,000 components and found that in all ten runs of the method around 50 components would always be estimated as zero.

Summarizing the Variational Bayesian posterior approximation. The VB posterior for every entry of the gene loadings matrix X_{cl} has the same spike-and-slab form as the prior. We use this distribution to calculate the expected value, denoted $E_Q(X_{cl})$. We also calculate a PIP that X_{cl} is not equal to zero, which is equal to $E_Q(S_{cl})$. We use the PIPs to infer a network of genes for each component consisting of the genes with PIP >0.5. We summarize the individual and tissue scores vectors in a similar way by using the expected values of the VB posterior, $E_Q(A_{cl})$ and $E_Q(B_{cl})$, respectively.

Identifying robust components. The model is complex and has a large number of parameters, and there is no guarantee that the VB algorithm will find a global solution when optimizing the bound on the marginal likelihood. Running the method multiple times highlights this issue. Some components are found consistently across multiple runs, whereas other components only occur in a small number of runs. For example, our method often uncovers components that show strong *cis*-eQTL signals when using the associated component scores as phenotypes. To identify robust components, we implemented a method that clusters similar components across different runs. We then focus on large clusters containing components from multiple different runs and use these as the basis for our search for novel signals.

More specifically, we run our method ten times and store the individual and tissue scores, gene loadings and PIPs. We calculate the absolute correlation between the individual scores for all pairs of components across the ten runs. Hierarchical clustering is then used to group components into clusters, using one minus the absolute correlation as a dissimilarity measure. The clustering is terminated when no correlations between clusters are above 0.6.

The components within each cluster are then combined. We take the mean of the individual scores, tissue scores and gene loadings and the median PIPs. The individual scores for each component cluster are then used as a phenotype against a genome-wide data set of SNPs on the same individuals to identify which components have a genetic basis. We apply quantile normalization to the individual scores before testing for association with SNPs. Tissue scores are thresholded to obtain tissue activity patterns. The distribution of tissue scores tends to be trimodal with one well-defined mode centered on zero, so a threshold can easily be picked to set small score values to zero. We only test averaged components calculated from clusters with a minimum (user-defined) membership size, to focus on components that are robustly and consistently identified across runs.

Analysis of the TwinsUK data set. Gene expression levels were measured for 845 female twins from the TwinsUK cohort using whole-transcriptome sequencing (RNA-seq), with data in three tissues (adipose, LCLs and skin) for the majority of the individuals[19,20]. Experiments were performed using the Illumina TruSeq sample preparation kit with sequencing on a HiSeq 2000 machine. Reads were mapped to the GRCh37 reference genome using BWA (v0.5.9)[40]. Only reads that map uniquely were used. We ran the method using RPKM values after performing the following pre-processing and normalization steps: (i) removal of genes with >20% zeros in all three tissues resulting in 18,409 genes, (ii) quantile normalization of expression data within each tissue, and (iii) rank-based transformation of each gene onto a standard normal.

Samples were genotyped on a combination of the HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo Illumina arrays. Samples were imputed using the 1000 Genomes Project Phase 1 reference panel (data freeze 10 November 2010) using IMPUTE2 (ref. 41) and filtered (MAF < 0.01 and IMPUTE info value < 0.8). Imputed genotypes were available on 795 of the 845 individuals.

We also used 11 concurrently measured phenotypes that were available on the samples (age, BMI, weight, height, total cholesterol, HDL cholesterol, LDL cholesterol (calc), total triglycerides, adiponectin, insulin and glucose) and variables derived from the sequencing. Specifically, we used (i) the mode of the insert size calculated for each sample, which can vary between sequencing library preps, (ii) the GC content of the reads from a sample, which can vary because of biochemical differences in library preparation and lane effects, (iii) the date of sequencing, and (iv) the primer index.

We ran our method ten times on the data set and combined components across runs via clustering (see above). **Supplementary Figure 39** shows the resulting distribution of cluster size. Only clusters with more than or equal to five components were then retained for GWAS.

We used a linear mixed model[42] to test an individual scores vector as a phenotype against the SNP genotypes. The scores vector was subset down to the 795 individuals for whom imputed genotype data were available. We used a Bonferroni-corrected significance threshold of 1×10^{-10}, calculated by scaling a genome-wide significance threshold of 5×10^{-8} by 500 to account for the multiple GWAS we perform.

Testing associations between individual scores vectors and phenotypes and batch variables was also performed using a linear mixed model[42], again only using 795 individuals. Only one member of each twin pair was used in the associations with age. The categorical batch variables, date and primer index, were dealt with by creating binary vectors (one for each category) and individually using these as a fixed effect in the linear mixed model.

Gene ontology analysis was carried out using the TopGO R package. Gene ontology analysis evaluates whether a particular set of genes is enriched for a GO term in comparison to a background gene set. TopGO uses Fisher's exact test to obtain a *P* value for enrichment based on the expected and observed number of genes with a GO term. Of the 18,409 genes used in this analysis, 13,965 have GO annotations. To determine a significance level for this analysis, we randomly sampled 10,000 sets of genes of a random size and performed an enrichment analysis on each set. We took the smallest *P* value from each gene set to create a null distribution and used this distribution to estimate a significance level of 1%.

We used a linear mixed model[42] to perform direct associations between the SNPs and the (normalized) expression levels of genes involved in our components. To account for unmeasured confounding factors, we fit the PEER model[23] to each tissue's expression data with 15 factors and used these as covariates in the mixed model. In addition to the PEER factors, we also included two phenotypes (age and BMI) and two tissue-specific batch variables (GC content mean and insert size mode) as covariates.

Application of fastICA. We used the R package fastICA (see URLs) to apply ICA to the TwinsUK data set. We concatenated the normalized expression data from the three tissues into a single matrix. Only 618 of 845 individuals had expression data on all three tissues, so this matrix had 618 rows and $3 \times 18{,}409$ columns. We fit the maximum number of components possible (618). We selected the 200 components where the measure of kurtosis of the gene loadings was >3.5 and ran a GWAS against all SNPs. We also tested the components' individual scores against the known confounding variables from the sequencing. More details are given in the **Supplementary Note**.

38. Lucas, J. *et al.* in *Bayesian Inference for Gene Expression and Proteomics* (eds. Do, K.-A., Muller, P. & Vannucci, M.) 1–25 (2006).
39. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. & Saul, L.K. An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233 (1999).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
42. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

doi:10.1038/ng.3624