ARTICLES

# Genetics of gene expression and its effect on disease

Valur Emilsson[1,2], Gudmar Thorleifsson[1], Bin Zhang[2], Amy S. Leonardson[2], Florian Zink[1], Jun Zhu[2], Sonia Carlson[2], Agnar Helgason[1], G. Bragi Walters[1], Steinunn Gunnarsdottir[1], Magali Mouy[1], Valgerdur Steinthorsdottir[1], Gudrun H. Eiriksdottir[1], Gyda Bjornsdottir[1], Inga Reynisdottir[1], Daniel Gudbjartsson[1], Anna Helgadottir[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Unnur Styrkarsdottir[1], Solveig Gretarsdottir[1], Kristinn P. Magnusson[1], Hreinn Stefansson[1], Ragnheidur Fossdal[1], Kristleifur Kristjansson[1], Hjortur G. Gislason[3], Tryggvi Stefansson[3], Bjorn G. Leifsson[3], Unnur Thorsteinsdottir[1], John R. Lamb[2], Jeffrey R. Gulcher[1], Marc L. Reitman[4], Augustine Kong[1], Eric E. Schadt[2]* & Kari Stefansson[1]*

**Common human diseases result from the interplay of many genes and environmental factors. Therefore, a more integrative biology approach is needed to unravel the complexity and causes of such diseases. To elucidate the complexity of common human diseases such as obesity, we have analysed the expression of 23,720 transcripts in large population-based blood and adipose tissue cohorts comprehensively assessed for various phenotypes, including traits related to clinical obesity. In contrast to the blood expression profiles, we observed a marked correlation between gene expression in adipose tissue and obesity-related traits. Genome-wide linkage and association mapping revealed a highly significant genetic component to gene expression traits, including a strong genetic effect of proximal (*cis*) signals, with 50% of the *cis* signals overlapping between the two tissues profiled. Here we demonstrate an extensive transcriptional network constructed from the human adipose data that exhibits significant overlap with similar network modules constructed from mouse adipose data. A core network module in humans and mice was identified that is enriched for genes involved in the inflammatory and immune response and has been found to be causally associated to obesity-related traits.**

The comprehensive assessment of molecular quantities in biological samples using high-throughput technologies has already led to the identification of disease subtypes[1,2], novel genes and gene structures[3,4], and biomarkers for disease[5], as well as the elucidation of transcriptional networks associated with disease traits[6–8]. The analysis of genotypes and gene expression data in animal models and human cell lines has proven useful for identifying genetic determinants of expression traits[1,9–13] and for mapping genes in regions linked to complex traits[6,10,11,14]. In general, such studies provide the means to examine the overall genetic complexity of gene expression traits, including a characterization of the relative effect of *cis* versus *trans* control[15,16].

Associating patterns of gene expression with DNA and complex trait variation is necessarily limited to those changes that are reflected in the transcriptional network. Although a number of studies have highlighted the importance of post-transcriptional alterations in gene activity that induce changes in biological processes[17], variation in protein structure and state may be reflected in the transcriptional network because such variation often induces a change in transcript stability, rates of transcription, transport of RNA from the nucleus, alternative splicing events, and other processes that affect expression levels[1]. Importantly, given the context specificity of many critical biological processes[18] and the fact that most common diseases are thought to be the outcome of a complex interaction between many genetic loci and the environment, it follows that there are obvious advantages to studying the genetics of gene expression in cells that represent the *in vivo* state.

Towards this end, we collected blood and subcutaneous adipose tissues in a population-based sampling of hundreds of Icelandic subjects ranging in age from 18 to 85 years old. These cohorts are referred to as the Icelandic Family Blood (IFB) cohort ($N = 1{,}002$) and the Icelandic Family Adipose (IFA) cohort ($N = 673$) (see Supplementary Table 1 for cohort description). A number of clinical traits including differential blood cell count as well as biometric traits such as body mass index (BMI), percentage body fat (PBF, measured by bioimpedance) and waist-to-hip ratio (WHR) were collected for all subjects of the IFB and the IFA cohorts (Supplementary Table 1). The relatively large sample size used in this study design provided the means to assess the relationship between sequence variants and gene expression with more statistical power than previous studies[12,13,16].

## Gene–clinical trait correlations

Expression profiles produced for this study contained measurements of relative abundances of 23,720 transcripts, representing 84% of the 24,060 protein-coding genes annotated in the Ensembl database (v.33)[19]. Given that probes overlapping single nucleotide polymorphisms (SNPs) may give rise to artificial signals, we sequenced a number of probes implicated as strong expression quantitative trait loci (eQTL) in 470 subjects from the IFB (see Supplementary Results and Supplementary Table 2). In short, we found that probes overlapping SNPs is not a concern in the present study.

The distribution of biometric traits such as BMI in our cohorts is not unlike the distribution that one would encounter in the general Western population, with BMI ranging from 16 to 70 and a median of

[1]deCODE genetics, 101 Reykjavik, Iceland. [2]Rosetta Inpharmatics, LLC, 401 Terry Ave N, Seattle, Washington 98109, USA. [3]Department of Surgery, National University Hospital, 101 Reykjavik, Iceland. [4]Merck Research Laboratories, Rahway, New Jersey 07065, USA.
*These authors contributed equally to this work.

28.8 (Supplementary Fig. 2a). Given the known associations of biometric traits with age and sex, and the fact that gene expression traits in blood have been found to be correlated with these covariates as well as with white blood cell counts[20], we adjusted for these covariates using multiple linear regression (Methods) in all analyses of correlation between gene expression and clinical traits, as well as in the analyses of the genetic component of gene expression (see below). In blood, fixing the false discovery rate (FDR)[21] at 5%, we found 2,172 (9.2%) gene expression traits to be correlated with BMI, 1,098 (4.6%) with PBF, and 711 (3.0%) with WHR in the IFB cohort (Supplementary Table 3). In adipose tissue, at a 5% FDR, the expression levels of 17,080 (72.0%) genes were correlated with BMI, 16,977 (71.6%) with PBF, and 14,901 (62.8%) with WHR (Supplementary Table 3). Thus, there is at least an order of magnitude more expression traits that are significantly correlated with these biometric traits in adipose tissue than in blood. Furthermore, 2,784 of the gene expression traits in adipose tissue explained more than 10% of the BMI variation in the IFA ($R^2 \geq 0.1$, $P \leq 10^{-15}$; see Supplementary Fig. 2b), whereas none of the expression traits in blood achieved this level of correlation. To ensure equivalent statistical power for making these detections between the tissues, we compared these associations in the 553 subjects represented in both the IFB and IFA cohorts. Using these paired samples, we found an even more marked difference between the two tissues (Supplementary Table 3). For example, there was a notable 34.6-fold enrichment of expression traits correlated with BMI in adipose tissue compared with blood using the 553 subjects (FDR $\leq$ 0.01), whereas this enrichment was 13.9-fold in the full data sets.

Overall, our results suggest that a substantial fraction of the transcriptional network in adipose tissue, together with infiltrated macrophages[22–24], is associated with the obesity of subjects. There are several reasons why this strong relationship between gene expression levels in adipose tissue and obesity should not come as a surprise. First, obesity is a disorder of excessive body fat. Second, the physiology and morphology of the adipocyte is known to be drastically altered in obese subjects[25]. Third, the number of macrophages is markedly increased in the adipose tissue of obese subjects, and they have been shown to have an important role in obesity and related metabolic disorders[22–24].

## Heritability of gene expression traits

The subjects in the IFB and IFA cohorts were clustered into multigenerational families (for details, see Methods). In the case of the IFB cohort, it was possible to cluster 938 out of the 1,002 subjects into 209 families, whereas for the IFA cohort, 570 out of the 673 subjects clustered into 124 families. Using this family structure, we estimated the heritability of each of the 23,720 gene expression traits, both with and without adjusting for sex, age, cell count (IFB only) and BMI (IFA only). The number of traits with statistically significant heritability is summarized in Table 1. With no adjustment, the number of significantly heritable traits at a 5% FDR was 13,910 in IFB and 16,825 in IFA, or 58.6% and 70.9% of all assessed transcripts, respectively. For those significantly heritable expression traits in the IFA and IFB cohorts, the genetic variance component on average explained nearly 30% of the variation observed (Supplementary Fig. 2c). After adjustment, the number of heritable traits fell by as much as 26% (Table 1). When combined with the high heritability estimated for the expression traits, these results indicate that a significant proportion of the heritability mediated by BMI or differential cell count is also reflected by a large number of gene expression traits. The heritability values (percentage) of all expression traits for the different types of adjustments and in both cohorts are listed in Supplementary Tables 4 and 5.

## Detection of *cis* and *trans* eQTL

All subjects in the two tissue cohorts were genotyped using a framework set of 1,732 microsatellites and were used for genome-wide linkage analysis. Because one of the main aims of this analysis was to detect eQTL signals that are proximal to the physical locations of genes corresponding to the expression traits (referred to here as *cis*-acting eQTL signals), this analysis was restricted to the 20,877 expression traits that had well-defined map positions (NCBI Build 34). For comparison, the eQTL analysis was performed both with and without adjusting the trait values for sex, age, differential cell-count (IFB only) and BMI (IFA only).

We defined a *cis*-acting eQTL signal for a given expression trait as the logarithm of the odds (eLOD) score at the nearest microsatellite to the location of the corresponding probe. The number of traits with significant *cis* eQTL is summarized in Table 1. For instance, at a 5% FDR and without any adjustment, we observed significant *cis* eQTL for 1,970 (9.4%) traits in blood and 1,215 (5.8%) traits in the adipose tissues. After adjusting for sex, age and blood cell counts in IFB, the number of *cis* eQTL signals increased to 2,529. In adipose tissue, this number was 1,307 after adjusting for age and sex and was 1,489 after also adjusting for BMI (Table 1). Out of the 1,489 significant *cis*-acting eQTL in adipose tissue, 762 (51.2%) were also observed in blood. Furthermore, expression traits with high heritability in both blood and adipose tissue showed greater reproducibility between the tissues (Fig. 1a). Here, 70% of all expression traits within the upper 25th percentile for heritability in blood that had a significant *cis*-acting eQTL in adipose tissue, also had a significant *cis* eQTL in blood (Fig. 1a). In fact, the proportion of significant *cis* eQTL signals in both tissues was notably higher for traits with greater levels of differential expression or heritability (Fig. 1b). The *cis*-acting eQTL LOD scores for each of the expression traits in the different cohorts are listed in Supplementary Tables 4 and 5.

Our finding of a strong genetic effect associated with *cis* signals in these tissues is consistent with results from previous studies[1,11–13]. The results on the detection of eQTL signals that were distal to the physical locations of the genes corresponding to the expression traits (referred to here as *trans*-acting linkage signals) are shown in the Supplementary Results and in Supplementary Table 6. We note that

**Table 1 | Heritability, *cis* eQTL and *cis* eSNP detection**

| | | IFB* | | IFA† | | |
|---|---|---|---|---|---|---|
| Variable | FDR or $\eta$‡ | No adjustment | Age, sex and cell-count adjusted | No adjustment | Age and sex adjusted | Age, sex and BMI adjusted |
| Heritability | 0.05 | 13,910 | 10,364 | 16,825 | 16,714 | 15,727 |
| | 0.01 | 10,829 | 8,047 | 12,309 | 12,392 | 11,251 |
| | $\eta$ | 0.68 | 0.55 | 0.78 | 0.77 | 0.75 |
| *cis* eQTL | 0.05 | 1,970 | 2,529 | 1,215 | 1,307 | 1,489 |
| | 0.01 | 1,256 | 1,567 | 737 | 773 | 820 |
| | $\eta$ | 0.40 | 0.44 | 0.33 | 0.32 | 0.37 |
| *cis* eSNPs | 0.05 | 2,417 | 2,714 | 3,048 | 3,149 | 3,364 |
| | 0.01 | 1,827 | 2,026 | 2,271 | 2,323 | 2,506 |
| | $\eta$ | 0.33 | 0.32 | 0.37 | 0.35 | 0.36 |

The number of *cis* eQTL and *cis* eSNPs were as determined for a unique set of gene expression traits, for example the single most significant *cis* eSNP for any given trait.
* Multiple regression analysis in blood, adjusting for sex and age as (age × sex) or for age, sex and differential cell-count as (age + neutrophil + monocyte + lymphocyte) × sex.
† Multiple regression analysis in adipose, adjusting for sex and age as (age × sex) or for age, sex and BMI as (age + log(BMI)) × sex.
‡ The proportion of significant tests, $\eta$, was estimated as $\eta = 1 - \pi_0$ (see Methods for details).

the number of traits with significant *trans* eQTL in blood and adipose tissue are 50 times fewer than the number of expression traits with significant *cis* eQTL, consistent with what has been found in other studies[1,9,12,14]. Finally, although others have reported hotspots of localized linkage activity in a number of species[1,6,9–11,13,14], we failed to detect such activity beyond what was expected by chance (Supplementary Results).

### Identification of *cis* and *trans* eSNPs

For the identification of sequence variants that have *cis* and *trans* regulatory effects on expression traits, we selected a subset of 150 unrelated (excluding all first-degree relatives) subjects who donated both blood and adipose tissue, and performed a whole-genome genotyping of these samples employing 317,503 SNPs using the Illumina platform[26]. The strongest *cis* effect for a given expression trait was then mapped by testing all SNPs located within a 2 megabase (Mb) window centred at the location of the probe corresponding to the expression trait, again restricting the analysis to the 20,877 genes with well defined positions in the genome. For each expression trait, because multiple correlated SNPs were tested for *cis* association, simulation was used to adjust the *P* value of the most significant expression (e)SNPs (see Methods). The effect of testing multiple expression traits was, as before, taken into account by means of the FDR approach[21]. The number of significant *cis*-acting eSNPs is summarized in Table 1. Assuming an FDR of 5%, we detected *cis* eSNPs for 2,417 (11.5%) expression traits in blood and 3,048 (14.6%) traits in adipose without any adjustment (Table 1). After adjusting for sex, age and cell count in blood, the number of *cis* eSNPs increased to 2,714 (Table 1). After adjusting for sex, age and BMI in the adipose tissue, the number of *cis* eSNPs increased to 3,364 (Table 1). Thus, we detected 650 more gene expression traits with significant *cis* eSNPs in the adipose tissue than in blood. This difference may reflect a more homogenous cell population in adipose tissue compared to blood, granting greater power to detect the *cis* effect in adipose. Furthermore, the number of significant *cis* eSNPs observed in both blood and adipose tissue increased as the heritability increased (Fig. 2a). For example, at an FDR of 1%, at least 50% of all SNPs that were *cis*-acting in blood and within the upper 25th percentile for heritability or differential expression were also *cis*-acting in adipose tissue (Fig. 2a).

Figure 2b summarizes the number of significant *cis* associations plotted as a function of heritability and differential expression. As observed in our analysis of *cis*-acting eQTL signals, the number of significant *cis* eSNPs increases with greater heritability scores or greater differential expression (Fig. 2b). A direct comparison of the results obtained from the genome-wide linkage and association analyses of *cis*-acting signals revealed a marked agreement between these two approaches (Supplementary Results and Supplementary Fig. 3). The significance of the *trans* association effect was assessed using the FDR approach[21], and the number of significant *trans* eSNPs summarized in Supplementary Table 6, again showing significantly fewer effects in *trans* than in *cis*, as described and discussed in Supplementary Results.

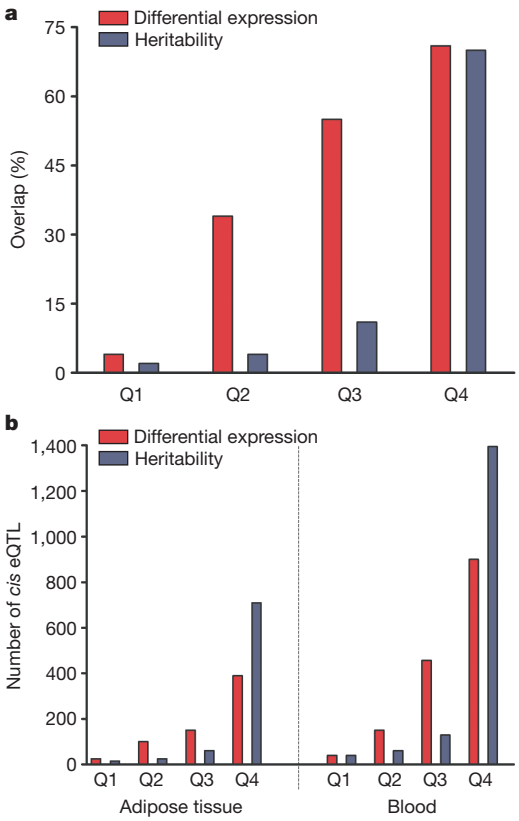


**Figure 1 | eQTL mapping in human blood and adipose tissue.** Individuals from large multi-generational families were genotyped for 1,732 microsatellites, and linkage analysis was performed on 20,877 standardized gene expression traits (see Methods for detail). Expression traits, ranked according to their differential expression or heritability strength, were binned into quartiles (Q1 → Q4), each comprised of 5,939 genes. **a**, Shown is the fraction of traits that have varying levels of differential expression and heritability with significant *cis*-acting eQTL in adipose that reproduced in blood at 1% FDR. **b**, Shown is the number of significant *cis* eQTL in both tissues, as a function of differential expression and heritability at 1% FDR.

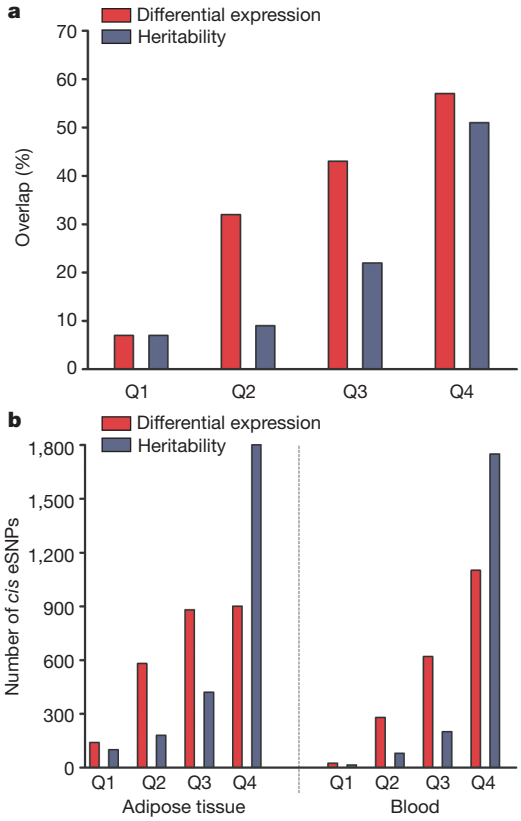


**Figure 2 | Genome-wide association screens for eSNPs.** A subset of 150 unrelated subjects who donated both blood and adipose tissue were genotyped at 317,503 tag SNPs (ILMN). The *cis* eSNP effects were assessed using linear regression on 20,877 standardized gene expression traits (see Methods for details). As described in Fig. 1, all traits were binned into quartiles at varying strengths of differential expression or heritability. **a**, Shown is the fraction of traits at varying degrees of differential expression or heritability with significant *cis* eSNPs in adipose tissue that reproduced in blood at 1% FDR. **b**, Shown is the number of significant *cis* associations in both tissues plotted as a function of heritability and differential expression.

### Characterizing the transcriptional network

The analysis of gene expression traits in a large sample of individuals allows for a direct and unbiased assessment of the connectivity structure of transcriptional networks[27]. This further provides a basis for the identification of key functional modules within such networks that contribute to disease risk[28]. We have previously described the characterization of transcriptional networks based on brain, adipose and liver tissues in a cross between two inbred strains of mice (referred to here as the B × H cross)[29–31]. Building on this approach, we constructed extensive, sex-specific, gene co-expression networks based on the human adipose tissue data to identify modules strongly associated with obesity and, more generally, comparing the structure of this human network to that constructed in the mouse B × H cross using similar tissues. The adipose co-expression network was constructed by considering all pair-wise correlations among the most differentially expressed genes detected in this tissue (Methods). The resulting gene–gene correlation matrix was then transformed into an adjacency matrix in which the connectivity of a given gene was defined as the sum of its connection strengths with all other genes in the network[27]. The gene–gene interconnectivity represented in this matrix (referred to here as the connectivity map) was then characterized using a topological overlap measure[28]. The identification of functional modules of highly co-regulated genes in the resulting network was carried out using a dynamic programming procedure to search the network for sets of maximally interconnected genes[29].

Figure 3a depicts the connectivity map for the male human adipose tissue as a heat map of the topological overlap matrix. In this type of map, the rows and the columns represent genes in a symmetric fashion, and the colour intensity represents the interaction strength between genes. This connectivity map highlights that genes in the adipose transcriptional network fall into distinct network modules, where genes within a given module are more highly interconnected with each other (blocks along the diagonal of the matrix) than with genes in other modules, as has been described previously for mice[30]. A comparison of the connectivity structure between the male and female human adipose networks reveals a number of network modules that are well conserved between gender, both in terms of gene identities and the connectivity strength (hub status or centrality; see Supplementary Figs 6 and 7). However, there are also network modules that are strictly gender specific (Supplementary Fig. 6).

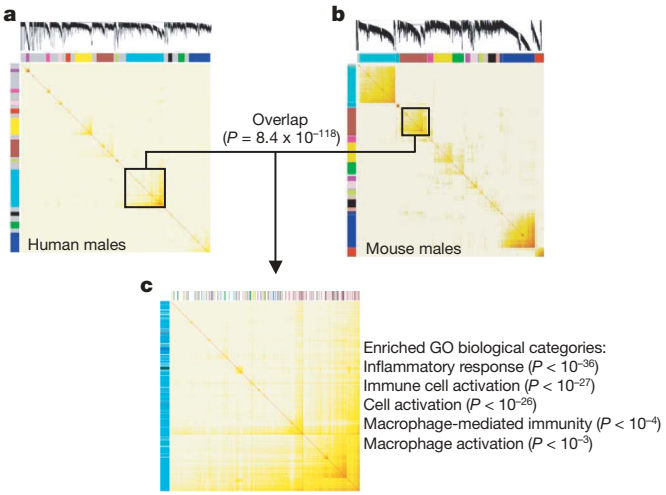


**Figure 3 | The human and mouse gene transcriptional networks.** **a**, Clustering of the connectivity matrix for the top 25% most differentially expressed genes in the male human adipose data. In the heat map, rows and columns represent genes in a symmetric fashion. The colour intensity signifies the connection strength between two genes, with red colour representing the strongest connection and white representing no connection. The colour bars along the *x*- and *y*-axes delineate the highly interconnected gene modules. **b**, Same as **a**, but for the male mouse B × H adipose data. **c**, The turquoise module in the male human network (**a**) is significantly overlapping the male mouse brown module (**b**), as well as the turquoise module in human females. The GO list in **c** shows the enrichment of inflammatory pathways in the conserved module.

An explicit comparison of the human and mouse adipose gene co-expression networks revealed a single core module in humans that was highly conserved in mice (Fig. 3a–c). The mouse module corresponding to this human module (Fig. 3b) is very significantly enriched for genes with eQTL that co-localize with obesity-associated-trait QTLs as well as for genes shown to be in a causal relationship with obesity-associated traits[31]. This mouse module significantly overlapped the human network module (Fig. 3a), with 196 out of the 673 (~29%) genes in the mouse module overlapping the set of 886 genes in the corresponding human module (only 8 were expected to overlap by chance; Fisher's Exact Test, $P = 8.4 \times 10^{-118}$). In addition, the Gene Ontology (GO) Biological Process categories that were enriched in this conserved network module were virtually identical in mouse and human (Supplementary Table 7). This conserved module was also strongly indicative of macrophage function for a number of reasons. First, GO Biological Process categories enriched in this module relate to inflammatory response and macrophage activation pathways. Second, well known macrophage-specific cell-surface markers such as *EMR1* and *CD68* are represented in the mouse and human modules. Third, using a recently constructed mouse body gene expression atlas comprised of more than 60 tissues and cell lines[31], this conserved module had an over-representation of genes enriched for expression in bone-marrow-derived macrophages (Fisher's Exact Test, $P < 1 \times 10^{-21}$), spleen, thymus and lymphoid tissue (Fisher's Exact Test, $P < 1 \times 10^{-20}$). These findings are consistent with results from recent studies showing that the adipose tissue secretes factors that regulate a wide variety of physiological states, including energy homeostasis and the immune response[25]. Given all of these significant enrichments and the association of this module to macrophage function and metabolic traits, we refer to it as the macrophage-enriched metabolic network (MEMN).

Because the mouse MEMN described above had previously been shown to be significantly enriched for genes associated with obesity[31], we investigated whether a similar association to obesity could be detected for the corresponding human module. Our results show that the expression of 868 (or 98%) of the 886 genes in the human MEMN module were significantly correlated with BMI in adipose tissue at an FDR of 1%, indicating that the human MEMN module may have a key role in obesity. Although the connection between inflammation and metabolic disorders such as obesity and diabetes has been reported previously[25], these data suggest that there may be many immune pathways or entire networks functioning in the adipose tissue. In fact, a number of genes previously identified and validated as being in a causal relationship with obesity-associated phenotypes are represented in this module, and perturbing many of these genes perturbs the entire module (see Supplementary Results for additional information).

If the MEMN module has a role in human obesity, then variations in DNA that result in expression changes in genes in the MEMN module should, in the obese, be enriched for variations that are associated with obesity. Therefore, we combined genotype and gene expression data to identify the SNP in the vicinity of each gene in the human MEMN module (Fig. 3a) that was most strongly associated with the corresponding gene expression trait. We then tested these variants jointly for association to BMI and PBF—the biometric traits most widely used to assess human obesity. Of the 886 expression traits represented in this module, 785 had a well defined genomic position and were used in this analysis. A selection of 768 *cis* eSNPs for the blood and adipose tissue data were successfully genotyped in a cohort of 8,685 individuals measured for BMI and 1,939 for PBF (Table 2). We used multiple linear regression analysis to test the association of the sex- and age-adjusted trait values to genotype counts for all the *cis* eSNPs jointly (see Methods for details).

**Table 2 | Association of eSNPs to obesity traits**

| Cohort | Trait | All *P* (*N*) | All *P* adjusted* | Male *P* (*N*) | Male *P* adjusted* | Female *P* (*N*) | Female *P* adjusted* |
|---|---|---|---|---|---|---|---|
| Human MEMN | | | | | | | |
| IFA | BMI | $3.8 \times 10^{-6}$ (8,685) | 0.0051 | 0.033 (3,606) | 0.24 | 0.0022 (5,079) | 0.049 |
| | PBF | 0.0011 (1,939) | 0.047 | 0.20 (906) | 0.47 | 0.12 (1,035) | 0.27 |
| IFB | BMI | $5.3 \times 10^{-7}$ (8,685) | 0.0022 | 0.00041 (3,606) | 0.015 | 0.016 (5,079) | 0.16 |
| | PBF | 0.063 (1,939) | 0.46 | 0.64 (906) | 0.87 | 0.39 (1,035) | 0.61 |
| Combined MEMN | | | | | | | |
| IFA | BMI | 0.14 (8,685) | 0.41 | 0.22 (3,606) | 0.39 | 0.27 (5,079) | 0.47 |
| | PBF | 0.010 (1,939) | 0.055 | 0.23 (904) | 0.49 | 0.011 (1,035) | 0.021 |
| IFB | BMI | 0.0014 (8,685) | 0.018 | 0.028 (3,606) | 0.075 | 0.028 (5,079) | 0.081 |
| | PBF | 0.23 (1,939) | 0.51 | 0.46 (904) | 0.73 | 0.45 (1,035) | 0.56 |

785 out of the 886 expression traits in the human turquoise module (see Fig. 3) mapped to a unique position and had a corresponding *cis* eSNP; this corresponds to 768 unique *cis* eSNPs that were used in the analysis. Missing genotypes were substituted with the mean genotype frequency. 128 out of the 146 expression traits in the combined human and mouse MEMN module had a unique map position and a *cis* eSNP; this corresponds to 123 unique *cis* eSNPs that were used in the analysis.
* The adjusted *P* value was based on up to 20,000 sets of simulated genotypes (see Methods).

Furthermore, we constructed 20,000 sets of simulated genotypes for all the variants conditioned on the familial relatedness of the individuals from the Icelandic genealogy database to compare the observed association with that expected to occur by chance, and used these to generate the adjusted *P* values represented in Table 2. In the larger data set with BMI measurements, we find that the *cis* eSNPs selected for genes in the human MEMN module showed some evidence for association to BMI, with *P* values of $3.8 \times 10^{-6}$ (adjusted $P = 0.005$) and $5.3 \times 10^{-7}$ (adjusted $P = 0.002$) for the *cis* eSNPs in adipose tissue and blood, respectively (see Table 2). Although these analyses were crude for individual *cis* eSNPs and the corresponding genes, these results suggest that the human MEMN module is enriched for sequence variants that confer risk of obesity in humans, and that genetic perturbations affecting gene expression traits may more generally perturb networks that in turn lead to increased susceptibility to disease. These data combined offer a glimpse of the complicated network of interactions that could drive at least a portion of obesity in humans, and demonstrate that at least a part of obesity is a property of the macrophage gene network.

## Discussion

Previous studies of the genetics of gene expression in humans have been restricted to lymphoblastoid cell lines with no clinical phenotypes[12,13,16]. Before our study, the validation of this type of data in primary human tissues from subjects scored for clinical traits was lacking. Our analysis of genetic variation and gene expression in population-based sampling of blood and subcutaneous adipose tissue from a large number of extended families begins to fill this gap. We showed that more than 50% of all gene expression traits in adipose tissue are strongly correlated with clinical traits related to obesity, compared to less than 10% in blood. Furthermore, through segregation analysis and genome-wide linkage and association studies, we demonstrated an extensive genetic component underlying gene expression traits in blood and adipose tissue. This was evidenced by detection of heritability as a highly significant contributor to variation in gene expression and by the identification of a large number of significant linkage and association signals for the expression traits in the two tissues, with approximately 50% overlap of genetic signals between the two tissues. Consistent with previous reports, the signals detected using both linkage and association analysis was strongly biased towards *cis*- rather than *trans*-acting genetic signals.

We also constructed an extensive co-expression network on the basis of the human adipose tissue data with the aim of identifying key functional modules within this network that associated with obesity. A core gene expression module, the MEMN module, was identified in humans that has significant overlap with a previously described mouse network module. The gene sets in the core human and mouse modules were highly enriched for genes involved in inflammatory response and macrophage activation pathways. Furthermore, the mouse MEMN module has previously been shown to be enriched for genes that contribute to the risk of obesity, diabetes and atherosclerosis-associated traits. By using the strongest *cis*-acting SNPs for each of the gene expression traits from the human MEMN module and testing them jointly as a group, we observed a significant enrichment of genetic associations to clinical traits related to human obesity in this module. The identification of SNPs that are associated with variation in gene expression provides a level of functional support for such SNPs that makes them ideal candidates to identify genetic determinants of complex traits including diseases and drug response. Clearly this approach warrants serious consideration given the potential to affect our understanding of human health.

## METHODS SUMMARY

Subjects used in the present study were of Caucasian descent. They were recruited as dense three-generation pedigrees, and comprehensively scored for multiple phenotypes including biometric traits related to obesity. Peripheral blood ($N = 1{,}002$) and subcutaneous fat ($N = 673$) were collected, and DNA and RNA extracted. The RNA samples (a total of 1,765 samples), including reference pools, were hybridized to a single custom-made human array containing 23,720 unique oligonucleotide probes. We estimated the differential expression, heritability, *cis* and *trans* eQTL, and association signals for each gene expression trait in each tissue. For the genetics of gene expression analysis, all subjects in these cohorts were genotyped at 1,732 microsatellites. A subset of 150 unrelated subjects, donating both blood and adipose tissue, was genotyped at 317,000 SNPs. Multiple testing for significance was taken into account through the use of an FDR procedure. The expression and clinical data were adjusted for standard covariates including age and sex for all analyses. The gene–gene co-expression network was constructed from the human adipose tissue expression data and compared to a similarly constructed adipose tissue network from an experimental mouse cross. Finally, expression variation markers (eSNPs) mapping to a core network module identified in human adipose tissue and found to be conserved in mice and previously shown to be enriched for genes in a causal relationship with obesity were tested jointly for association to obesity-related traits in humans using multiple regression analysis.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 23 July 2007; accepted 28 January 2008.**
**Published online 16 March 2008.**

1. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302 (2003).
2. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
3. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141–2144 (2003).
4. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409,** 922–927 (2001).
5. Welsh, J. B. *et al.* Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA* **98,** 1176–1181 (2001).
6. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
7. Schadt, E. E., Sachs, A. & Friend, S. Embracing complexity, inching closer to reality. *Sci. STKE* **2005,** pe40 (2005).

8. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105,** 363–374 (2004).
9. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
10. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genet.* **37,** 225–232 (2005).
11. Chesler, E. J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genet.* **37,** 233–242 (2005).
12. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75,** 1094–1105 (2004).
13. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430,** 743–747 (2004).
14. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genet.* **37,** 1224–1233 (2005).
15. Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701–703 (2005).
16. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437,** 1365–1369 (2005).
17. Ranganathan, P. *et al.* Expression profiling of genes regulated by TGF-β: differential regulation in normal and tumour cells. *BMC Genom.* **8,** 98, doi:10.1186/1471-2164-8-98 (2007).
18. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102,** 1572–1577 (2005).
19. Hubbard, T. *et al.* Ensembl 2005. *Nucleic Acids Res.* **33,** D447–D453 (2005).
20. Whitney, A. R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA* **100,** 1896–1901 (2003).
21. Storey, J. D. & Tibshirani, R. Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol. Biol.* **224,** 149–157 (2003).
22. Di Gregorio, G. B. *et al.* Expression of CD68 and macrophage chemoattractant protein-1 genes in human adipose and muscle tissues: association with cytokine expression, insulin resistance, and reduction by pioglitazone. *Diabetes* **54,** 2305–2313 (2005).
23. Lumeng, C. N., Bodzin, J. L. & Saltiel, A. R. Obesity induces a phenotypic switch in adipose tissue macrophage polarization. *J. Clin. Invest.* **117,** 175–184 (2007).
24. Neels, J. G. & Olefsky, J. M. Inflamed fat: what starts the fire? *J. Clin. Invest.* **116,** 33–35 (2006).
25. Wellen, K. E. & Hotamisligil, G. S. Obesity-induced inflammatory changes in adipose tissue. *J. Clin. Invest.* **112,** 1785–1788 (2003).
26. Steemers, F. J. & Gunderson, K. L. Illumina, Inc. *Pharmacogenomics* **6,** 777–782 (2005).
27. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** Article17 (2005).
28. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297,** 1551–1555 (2002).
29. Ghazalpour, A. *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2,** e130 (2006).
30. Lum, P. Y. *et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J. Neurochem.* **97** (suppl. 1), 50–62 (2006).
31. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* doi:10.1038/nature06757 (this issue).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** The authors acknowledge the participating families and the staff at the Clinical Research Centre for their cooperation. Genotyping service was provided at the deCode Genetics genotyping facilities.

**Author Contributions** V.E., E.E.S., K.S. and G.T. wrote the paper. G.T., E.E.S., A.K., D.G. and F.Z. performed statistical analysis. Tissue sampling and/or molecular profiling was carried out by H.G.G., T.S., B.G.L., G.H.E., S.C., M.M., Aslaug Jonasdottir, Adalbjorg Jonasdottir, G.B. and K.K. V.E., J.Z., U.T., A.S.L., A.H., B.Z., G.B.W., S. Gunnarsdottir, S. Gretarsdottir, K.P.M., V.S., I.R., A.H., U.S., H.S., R.F., J.R.G., K.S., M.L.R. and J.R.L. performed the genetic analysis and/or data-mining. K.S. and E.E.S. contributed equally to this work.

**Author Information** All the gene expression data generated for this study have been deposited into the GEO database under accession numbers GSE7965 and GPL3991. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.S. (kari.stefansson@decode.is) or E.E.S. (eric_schadt@merck.com).

## METHODS

**Human study populations and sample processing.** The subjects, ranging in age from 18 to 85 years old, in the IFB and IFA cohorts were clustered into multi-generational families on the basis of relatedness of individuals in the Icelandic genealogy database[32]. For the IFB cohort, 1,002 Icelandic subjects were recruited, and for the IFA cohort, 673 subjects were recruited. All participants in the IFA and IFB cohorts were scored for various clinical traits related to obesity, including height, weight, waist circumference, hip circumference and percentage body fat (PBF) measured by bioimpedance. In addition to the IFA and IFB cohorts, 85 (43 males and 42 females) Icelandic individuals were recruited to generate a blood RNA reference pool for the IFB cohort. Furthermore, ten (six females and four males) additional Icelandic individuals being operated on for abdominal hernia were recruited to construct an adipose reference RNA pool for the IFA cohort. Ethical approval for the present study was granted by the National Bioethics Committee (NBC 01-033) and the Icelandic Data Protection Authority (DPA). All participants in the study signed informed consent. All personal identifiers associated with tissue samples, clinical information and genealogy were encrypted by the DPA, using a third-party encryption system in which DPA maintains the code[32]. The RNA and DNA sample preparation, microarray hybridization and expression analysis are described in the MIAME checklist that is provided in the Supplementary Information.

**Identifying differentially expressed genes.** To assess whether a gene in a given sample was differentially expressed, we used a previously described and validated error model for testing whether the mean log ratio of the intensity measurements between the experiment and reference channels was significantly different from zero[33,34]. On the basis of this error model, we obtained *P* values for each of the individual gene expression measures in each sample as described previously[33]. We then computed the standard deviation of $-\log_{10}$ of the *P* value for each gene expression measure over all samples profiled for a given tissue, and then rank-ordered all of the genes profiled in each tissue on the basis of this standard deviation value (rank-ordered in descending order). Genes that fall at the top of this rank-ordered list can be considered to be the most differentially expressed or variable genes in the study. We have previously shown that this type of ordering approach well captures the most active genes in a set of samples[33]. To demonstrate the number of genome-wide significant eQTL and eSNPs as a function of differential gene expression, we binned the expression traits into quartiles (Q1 → Q4) on the basis of the rank-ordered gene list, with each bin containing 5,939 genes and the bins increasing in significance with respect to differential expression, from Q1 to Q4.

**Heritability, genome-wide linkage and association analysis.** All subjects were clustered into families in which each proband is related to at least one other proband within five meiotic events; members of the IFB cohort were clustered into 209 families with 938 contributing individuals, and those from IFA were clustered into 124 families with 570 contributing individuals. Individuals in these cohorts were genotyped with 1,732 microsatellites uniformly distributed across the human genome, as described previously[35]. Each gene expression trait was treated as a quantitative trait. For the heritability calculations, linkage analysis and association to genetic markers, the expression trait values were first adjusted for relevant covariates such as sex, age, blood cell count and BMI using multiple linear regression analysis as (age + age$^2$ + neutrophil + monocyte + lymphocyte) × sex + trait in blood and as (age × age$^2$ + log(BMI)) × sex + trait for IFA. Traits were then standardized by mapping the distributions of the inverse normal transformation to each of the expression traits onto a normal distribution with a mean of 0 and a variance of 1. This was done to eliminate the effect of outliers on all subsequent analyses. To calculate the heritability, a polygenic model was fitted to determine how much of the variation in the trait was caused by genetic effects. To carry out these calculations, we used SOLAR 2.0, a publicly available software package for human genetic analysis[36].

Linkage analysis and the calculation of IBD matrices used in the heritability calculations were carried out using the program Allegro[37]. The linkage analysis was based on a locally most-powerful score statistic for a gaussian variance component model with an additive variance component and assuming heritability for each trait was known. Significance was assessed using the exponential tilting method[38], which has previously been demonstrated to give accurate type I error rates[39]. The accuracy of type I error rates was verified for the present score statistic by carrying out extensive simulation analysis, including simulations that assumed various deviations from the gaussian variance component model[40].

Multiple testing for significance was taken into account through the use of FDR procedures[21]. The software QVALUE was used in the calculations[21]. The proportion of significant tests $\eta$ was estimated as $\eta = 1 - \pi_0$, where $\pi_0$ is the estimate of the overall proportion of true null hypotheses. In estimating $\pi_0$, the pi0.meth = "bootstrap" option in the QVALUE software was used.

**Controlling for multiple testing in the genome-wide association scans.** To control for multiple testing in the genome-wide association scans carried out on the gene expression traits in a subset of the IFA cohort, we used simulations to adjust the *P* values for each trait for the number of SNPs tested. In each simulation, we permuted the gene expression trait values for the 150 individuals and recalculated the association test for all SNPs in the 2 Mb window centred at the probe sequence location for the corresponding gene. This was repeated up to 50,000 times depending on the significance of the original *cis* association identified for the expression trait in question. More specifically, if we define the Bonferroni-adjusted *P* value, $P_{\mathrm{Badj}}$, as $P \times N$, where *P* is the unadjusted *P* value and *N* is the number of SNPs tested, the number of permutations[41] performed for each trait was selected as $100/P_{\mathrm{Badj}}$. The minimum number of permutations performed for any given expression trait was 500. This was sufficient for roughly 70% of the traits. An adjusted *P* value was then calculated as the fraction of simulations that produced an association for any SNP tested that was at least as significant as the most significant *cis* association observed in the original data set. For the X chromosome, the permutations were done preserving the sex of the individuals. The permutation test was applied to those traits where the strongest *cis* association corresponded to a $P > 0.000001$, whereas for traits with more significant *cis* associations a simple Bonferroni correction was used to calculate the adjusted *P* values. The Bonferroni adjustment was applied to approximately 10% of the traits. Multiple testing for significance was then taken into account through the use of the FDR procedures[21]. Here, the calculated *P* values (as described above) were used as the input to estimate the overall FDR.

**Assessing the significance of *trans*-acting eQTL signals.** We defined a linkage for a gene expression trait as being *trans*-acting (distal to the physical location of the probe) if the associated LOD score curve peak was located on a different chromosome to the physical location of corresponding probe sequence. To assess the significance of the observed *trans*-acting eQTL signals, we created 10 sets of simulated genotypes for all of the 1,732 microsatellite markers using drop-down simulations, under the assumption of no linkage anywhere in the genome, for the same family structure as that used in the linkage analysis. For each marker, the simulated genotypes matched the original genotypes both in terms of missing genotypes and in terms of the frequency distribution of the genotypes for each marker. We then ran the linkage analysis on each of the simulated data sets for all of the 20,877 uniquely mapped traits. From each linkage run, we identified the strongest *trans*-acting eQTL for each gene expression trait. Combining the results for all 20,877 gene expression traits over all 10 simulated data sets yielded a reference distribution of 208,770 of the strongest *trans*-acting eQTL detected in the simulated data. By comparing the observed *trans*-acting eQTL distribution to this reference distribution, we assigned empirical *P* values to the *trans*-acting eQTL signals observed in the original analysis.

**Assessing the significance of eQTL hotspots.** Given the strong correlation structure among gene expression traits, if one expression trait falsely links to a given genomic region then it is possible that many other expression traits highly correlated with this expression trait may also falsely link to the given genomic region. To assess whether eQTL hotspots were artefacts driven by false-positive eQTL of highly correlated expression traits, we again used the eQTL results generated on the 10 simulated data sets described above. Using the same linkage threshold as that used in the observed data to detect *trans* eQTL, we examined whether hotspots were detected that were of similar magnitude or greater than what was detected in the observed data. In all of the simulated data sets, we observed hotspots of similar or greater magnitude than the hotspots we detected in the observed data (see Supplementary Fig. 1), suggesting that the hotspots detected in the observed data could be due to false linkages of highly correlated gene sets.

**Construction of the adipose co-expression network.** A previously described weighted gene co-expression network reconstruction algorithm was used to reconstruct the human and mouse co-expression networks[27]. The weighted network reconstruction algorithm involved first constructing a matrix of Pearson correlations between all gene expression pairs. The correlation matrix was then transformed into an adjacency matrix using a power function $f(x) = x^{\beta}$. The adjacency matrix defines the weighted co-expression network. The parameter $\beta$ of the power function was determined such that the resulting adjacency matrix was approximately scale-free based on a previously proposed model-fitting index[27]. This index is defined as the coefficient of determination ($R^2$) of the linear model constructed by regressing $\log(p(k))$ onto $\log(k)$, or by regressing $\log(p(k))$ onto $\log(k) + k$, where $k$ represents the number of edges connecting to the given node and $p(k)$ is the frequency distribution of the degree $k$ in the co-expression network. The model-fitting index of a perfect scale-free network is 1. The exponent of the power function, $\beta$, was chosen to be the smallest value such that the co-expression network exhibited the scale-free property, that is, the model-fitting index $R^2 \geq 0.8$.

The adjacency matrix was further transformed into a topological overlap matrix to more readily identify modules of highly co-regulated genes. The topological overlap captures not only the direct interaction between two genes but also their indirect interactions through all the other genes in the network. Traditionally, the connectivity of a node is defined as the sum of its connection strengths in the adjacency matrix with all other genes in the network. Here we extended the definition to the topological overlap matrix and derived a topological overlap connectivity map. Module identification was conducted through a dynamic programming procedure to search the topological overlap matrix ordered by hierarchical clustering for maximum sets of inter-connected genes[28].

**Testing the association of *cis* variants to obesity traits.** To test the association of *cis* variants for the genes in the MEMN module to the obesity traits BMI (or PBF), we tested the difference between two regression models: model 1 is $\log(\mathrm{BMI}_j) \sim \mathrm{sex}_j \times (\mathrm{age}_j + \mathrm{age}_j^2)$, where subscript $j$ refers to individual $j$, and model 2 is $\log(\mathrm{BMI}_j) \sim \mathrm{sex}_j \times (\mathrm{age}_j + \mathrm{age}_j^2) + g_{ji} + \ldots + g_{jn}$, where $g_{ji}$ is the minor allele count for individual $j$ and *cis* variant $i$. To adjust for relationship and for associations occurring by chance, we simulated genotypes for all of the *cis* variants through the genealogy of 708,683 Icelanders. Here, for each of the 20,000 simulated sets of genotypes we constructed, we repeated the association tests between the *cis* variants and the obesity traits. We then calculated adjusted $P$ values as the fraction of simulations that yielded equally or more significant association between a particular trait and the corresponding *cis* variants. These adjusted $P$ values are summarized in Table 2 in the main text.


32. Gulcher, J. R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8,** 739–742 (2000).
33. He, Y. D. *et al.* Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics* **19,** 956–965 (2003).
34. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
35. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
36. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62,** 1198–1211 (1998).
37. Gudbjartsson, D. F., Jonasson, K., Frigge, M. L. & Kong, A. Allegro, a new computer program for multipoint linkage analysis. *Nature Genet.* **25,** 12–13 (2000).
38. Kong, A. & Cox, N. J. Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61,** 1179–1188 (1997).
39. Badner, J. A., Gershon, E. S. & Goldin, L. R. Optimal ascertainment strategies to detect linkage to common disease alleles. *Am. J. Hum. Genet.* **63,** 880–888 (1998).
40. Amos, C. I. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54,** 535–543 (1994).
41. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971 (1994).