Human Molecular Genetics, 2015, Vol. 24, No. 23 6849–6860

doi: 10.1093/hmg/ddv379
Advance Access Publication Date: 16 September 2015
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity

Sara Hägg[1,2], Andrea Ganna[3,4], Sander W. Van Der Laan[5], Tonu Esko[6,7,8,9], Tune H. Pers[6,7,8,10,11], Adam E. Locke[12], Sonja I. Berndt[15], Anne E. Justice[16], Bratati Kahali[13,14], Marten A. Siemelink[5], Gerard Pasterkamp[5,18], the GIANT Consortium, David P. Strachan[19], Elizabeth K. Speliotes[13,14], Kari E. North[16,17], Ruth J.F. Loos[20,21,22,23], Joel N. Hirschhorn[6,7,8], Yudi Pawitan[1] and Erik Ingelsson[2,24,25,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [2]Molecular Epidemiology and Science for Life Laboratory, Department of Medical Sciences, Uppsala University, SE-751 41 Uppsala, Sweden, [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, [5]Laboratory of Experimental Cardiology, Division of Heart and Lungs, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands, [6]Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115, USA, [7]Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA, [8]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA, [9]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia, [10]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark, [11]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark, [12]Center for Statistical Genetics, Department of Biostatistics, [13]Department of Internal Medicine, Division of Gastroenterology, [14]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [15]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA, [16]Department of Epidemiology, [17]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [18]Laboratory of Clinical Chemistry and Hematology, Division Laboratories & Pharmacy, UMC Utrecht, Utrecht, The Netherlands, [19]St George's, University of London, London SW17 0RE, UK, [20]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK, [21]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [22]The Genetics of Obesity and Related Metabolic Traits Program, [23]The Mindich Child Health and Development Institute, [24]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK and [25]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

*To whom correspondence should be addressed. Tel: +46 707569422; Fax: +46 18515570; Email: erik.ingelsson@medsci.uu.se

Received: April 30, 2015. **Revised and Accepted:** September 10, 2015

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Abstract

To date, genome-wide association studies (GWASs) have identified >100 loci with single variants associated with body mass index (BMI). This approach may miss loci with high allelic heterogeneity; therefore, the aim of the present study was to use gene-based meta-analysis to identify regions with high allelic heterogeneity to discover additional obesity susceptibility loci. We included GWAS data from 123 865 individuals of European descent from 46 cohorts in Stage 1 and Metabochip data from additional 103 046 individuals from 43 cohorts in Stage 2, all within the Genetic Investigation of ANthropometric Traits (GIANT) consortium. Each cohort was tested for association between ~2.4 million (Stage 1) or ~200 000 (Stage 2) imputed or genotyped single variants and BMI, and summary statistics were subsequently meta-analyzed in 17 941 genes. We used the 'VErsatile Gene-based Association Study' (VEGAS) approach to assign variants to genes and to calculate gene-based *P*-values based on simulations. The VEGAS method was applied to each cohort separately before a gene-based meta-analysis was performed. In Stage 1, two known (*FTO* and *TMEM18*) and six novel (*PEX2*, *MTFR2*, *SSFA2*, *IARS2*, *CEP295* and *TXNDC12*) loci were associated with BMI ($P < 2.8 \times 10^{-6}$ for 17 941 gene tests). We confirmed all loci, and six of them were gene-wide significant in Stage 2 alone. We provide biological support for the loci by pathway, expression and methylation analyses. Our results indicate that gene-based meta-analysis of GWAS provides a useful strategy to find loci of interest that were not identified in standard single-marker analyses due to high allelic heterogeneity.

Introduction

Genome-wide association studies (GWASs) and large-scale collaborations have revolutionized the field of complex disease genetics. Body mass index (BMI) is commonly used to assess obesity and >100 loci have been discovered to date in studies of individuals of primarily European descent (1,2). Nevertheless, a large proportion of the genetic variance remains undiscovered, implying that more loci influencing BMI are yet to be found (3). Additional explanations for the missing heritability include, among others, incomplete tagging of causal variants and allelic heterogeneity—multiple alleles in the same gene region that affect the same trait (4,5). One approach to detect regions that display substantial allelic heterogeneity is to use gene-based tests of association. Gene-based tests can improve statistical power in the presence of allelic heterogeneity by combining single variants from GWAS into a gene-based score, which substantially reduces the burden of multiple testing and combines signals from multiple associated variants (6–8). One of several methods where significance levels from single variant associations are used to calculate a gene-based test statistic is the VErsatile Gene-based Association Study (VEGAS) approach (8). VEGAS performs equivalently to other gene-based approaches (8), but it is superior to other methods when used in meta-analyses because it uses external reference panels to simulate a null distribution, i.e. it can be run with summary association statistics and does not require individual-level genotypic data.

A gene-based test can also improve the ability to replicate loci displaying high between-study heterogeneity. Differences in the underlying linkage disequilibrium (LD) patterns could lead to different tag single nucleotide polymorphisms (SNPs) being linked to the causal variants in different study populations, a problem that is lessened in gene-based tests as they are combined into the same gene signal. Thus, using gene-based methods should increase statistical power for GWAS regions with allelic and between-study heterogeneity, and applying these techniques could implicate novel loci without increasing sample sizes or collecting new data. Hence, the aims of the present study were (1) to identify new loci associated with BMI and (2) to evaluate our gene-based meta-analysis as a valid approach using a set of sensitivity analyses and comparisons. Toward these aims, we used existing GWAS data from the GIANT (Genetic Investigation of ANthropometric Traits) consortium. To enhance power by allowing for allelic heterogeneity between studies, we applied the VEGAS summary statistics algorithm on each GWAS study separately before combining the results in a single gene-based meta-analysis.

Results

Stage 1 identified six novel loci associated with BMI

We first analyzed the Stage 1 data set comprising 46 studies including up to 123 865 individuals who had undergone GWAS of BMI, as detailed elsewhere by Speliotes *et al.* (1). In brief, the GWAS was done assuming an additive genetic model with inverse normal transformation of BMI. The VEGAS method was applied to each of the GWAS result files from the 46 studies individually, before a gene-based meta-analysis using the Fisher method was conducted on 17 941 genes (Fig. 1; Supplementary Material, Fig. S1). Each gene was defined with symmetric boundaries ±50 kb (the default in VEGAS); however, results from VEGAS are in general robust across different gene boundaries chosen (9). A SNP could contribute to multiple gene signals if the genes are overlapping. Thus, a locus could include multiple neighboring or overlapping genes that could partly rely on the same SNP signals.

Applying a Bonferroni-corrected gene-wide significance threshold ($P < 2.8 \times 10^{-6}$ for 17 941 gene tests) for gene-based association with BMI, we identified six loci that have previously not been implicated in relation to BMI (at least 1 Mb from published genome-wide SNPs); the first locus contained four significant genes in high LD (Supplementary Material, Fig. S2): thioredoxin domain containing 12 (*TXNDC12*); KTI12 homolog, chromatin associated (*KTI12*); basic transcription factor 3-like 4 (*BTF3L4*) and zinc finger, FYVE domain containing 9 (*ZFYVE9*). The other five significant loci harbored only one gene each: peroxisomal biogenesis factor 2 (*PEX2*); mitochondrial fission regulator 2 (*MTFR2*); sperm-specific antigen 2 (*SSFA2*); isoleucyl-tRNA synthetase 2 (*IARS2*) and centrosomal protein 295 kDa (*CEP295*), also known as *KIAA1731*. In addition, two well-known BMI-associated loci were confirmed in our analyses; the first locus contained the fat mass- and obesity-associated gene (*FTO*) and the second locus the gene transmembrane protein 18 (*TMEM18*). Details about the associations are given in Table 1 and regional plots are shown in Supplementary Material, Figure S3A–H.

Stage 2 validated the BMI loci found in Stage 1

In order to validate loci that reached gene-wide significance in the Stage 1 analyses, we performed Stage 2 analyses in 43

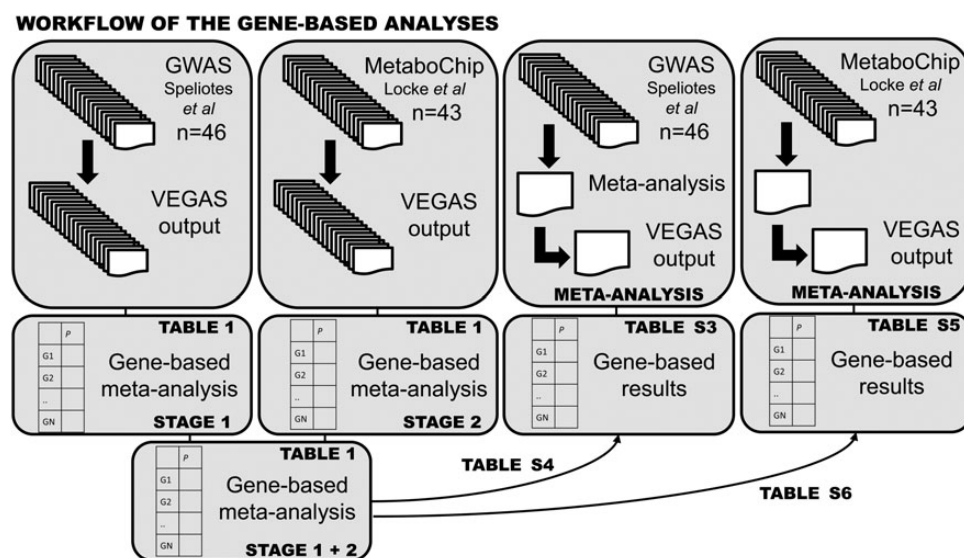

Figure 1. Workflow describing the process of gene-based analyses applied on the different data sets. The gene-based meta-analyses were carried out in a Stage 1 material of 46 different studies from the Speliotes *et al.* paper. Results were then validated in a Stage 2 phase including 43 studies from the MetaboChip in the Locke *et al.* paper depicted in the figure. A joint gene-based meta-analysis of Stage 1 and Stage 2 data sets were subsequently done. The two right panels refer to the gene-based analyses carried out of the summary statistics from Speliotes *et al.* and Locke *et al.* For clarity, the corresponding tables in the paper are illustrated in the figure.

Table 1. Gene-wide significant BMI-associated loci from gene-based meta-analyses of Stage 1, 2 and 1 + 2 data sets

Locus	Gene	Chr	Position Start	Stop	Stage 1 *P*-value	Stage 2 *P*-value	Stage 1 + 2 *P*-value
1	*TXNDC12*	1	52 258 391	52 293 635	1.3E−14	2.3E−14	5.5E−27
	KTI12	1	52 270 364	52 272 060	1.2E−12	9.3E−13	2.0E−23
	BTF3L4	1	52 294 560	52 326 650	1.8E−12	3.2E−12	9.7E−23
	ZFYVE9	1	52 380 633	52 584 946	9.7E−07	1.5E−03	1.6E−08
2	*PEX2*	8	78 055 048	78 075 079	3.4E−13	1.7E−07	1.1E−18
3	*FTO*	16	52 295 375	52 705 882	1.6E−10	3.1E−13	1.1E−21
4	*MTFR2*	6	136 593 860	136 613 142	9.7E−08	4.5E−07	4.9E−13
5	*TMEM18*	2	657 972	667 439	1.8E−07	1.9E−08	4.8E−14
6	*SSFA2*	2	182 464 926	182 503 706	6.3E−07	3.1E−05	2.0E−10
7	*IARS2*	1	218 334 077	218 388 006	2.1E−06	1.1E−08	3.8E−13
8	*CEP295*	11	93 034 463	93 103 170	2.3E−06	8.1E−05	1.7E – 09
Additional gene-wide significant hits in Stage 2							
9	*MC4R*	18	58 038 564	58 040 001	2.2E−03	6.4E−07	5.1E−08
10	*ADCY3*	2	25 042 038	25 142 055	2.2E−01	9.3E−07	5.5E−05
Additional gene-wide significant hits in Stage 1 + 2							
11	*CBX4*	17	75 421 549	75 427 808	1.1E−05	3.9E−06	4.6E−10
12	*SPAG6*	10	22 674 404	22 746 545	3.9E−05	4.5E−05	1.6E−08
13	*SPG11*	15	42 642 185	42 743 168	2.4E−05	2.2E−04	4.5E−08
8	*TAF1D*	11	93 108 745	93 114 310	1.2E−04	5.2E−04	4.9E−07
14	*KHDRBS1*	1	32 252 077	32 282 059	2.0E−05	3.0E−03	5.5E−07
15	*RPL23AP74*	17	56 501 992	56 502 554	2.3E−04	7.5E−04	1.3E−06

additional cohorts including 103 046 individuals with data from the Metabochip array published in Locke *et al.* (2) using the same gene-based approach (Fig. 1; Supplementary Material, Fig. S1). The Metabochip array does not have a dense genome-wide coverage of SNPs; however, the gene coverage as judged by the VEGAS results is almost as complete (17 833 genes in Stage 2 compared with 17 941 in Stage 1). Moreover, analyzing the SNPs that contributed to the VEGAS signals in Stage 1 showed that many SNPs were common between Stages 1 and 2 (Supplementary Material, Fig. S4). Thus, all eight loci replicated in Stage 2 with a significance well below a Bonferroni-corrected threshold ($\alpha = 0.00625$ for eight tests) (Table 1).

Although replication of the significant genes was the primary goal of Stage 2, we also report additional gene-wide significant hits that arise when using these data to the full extent. Hence, two loci, both previously established, including the genes melanocortin 4 receptor (*MC4R*) and adenylate cyclase 3 (*ADCY3*), respectively, were identified in Stage 2 alone (Table 1). Further, in a joint Stage 1 + 2 analysis, in addition to all genes in the eight loci from Stage 1, another five loci also reached the gene-wide threshold and was represented by the genes chromobox homolog 4 (*CBX4*), sperm associated antigen 6 (*SPAG6*), spastic paraplegia 11 (autosomal recessive) (*SPG11*), KH domain containing, RNA binding, signal transduction associated 1 (*KHDRBS1*) and ribosomal protein L23a pseudogene 74 (*RPL23AP74*) (Table 1).

Pathway analyses of BMI-associated genes

Using data from Stage 1, we performed pathway analyses including all genes with nominally significant ($P < 0.05$) association with BMI ($n = 773$) using ConsensusPathDB (10). We observed an enrichment for pathways *disease* (Q-value <0.003) and *synthesis of bile acids and bile salts via 24-hydroxycholesterol* (Q-value <0.05). Combining Stages 1 and 2, pathway analyses of the 942 top-ranking genes (again $P < 0.05$ for association with BMI) revealed enrichment in *peptide hormone biosynthesis* (Q-value <0.002), *osteopontin signaling* (Q-value <0.004), *glycoprotein hormones* (Q-value <0.004) and *regulation of cytoplasmic and nuclear SMAD2/3 signaling* (Q-value <0.05).

We further performed pathway and network analysis using Ingenuity software (11) (Qiagen, fall 2014 version) with the top genes from the gene-based association analysis, also including the established Stage 2 hits *MC4R* and *ADCY3* to allow for a more stable network. The most significant canonical pathway was *Glutathione Redox Reactions II*, $P = 1.96 \times 10^{-3}$. The top network not only involved *FTO, MTFR2, ADCY3* and *MC4R* but also the novel genes *SSFA2* and *ZFYVE9*. The nuclear *FTO* was an important hub with many connections centrally regulating many of the other genes (Fig. 2). Strikingly, one of these genes was leptin (*LEP*), a hormone known to be involved in obesity signaling (12). This finding is especially interesting in the light of recent important functional studies disentangling the causal variants and genes in the *FTO* locus, which have highlighted not only *IRX3* and *IRX5* (13,14) but also *RPRIP1L* (15) and *FTO* (16).

Gene expression analyses of identified genes

To provide biological support for and to further validate the BMI-associated genes identified in Stage 1, we explored those genes in relevant expression data sets. To accomplish this aim, we downloaded all data related to obesity from *blood* or *adipose tissue* in human, mouse or rat found in the Gene Expression Omnibus (GEO) data repository. All data sets were independently cross-examined by two different investigators (S.H. and E.I.) and underwent quality control steps before further inclusion. Altogether, 53 selected experiments including 335 obese cases and 308 controls were finally used for analyses.

In each data set, although our primary goal was to test the 11 genes identified in Stage 1, all genes were screened for differential expression between obese cases and controls, and ranked by significance. This allowed for a null distribution to be generated for each experiment by permuting the ranking of the gene. We then evaluated if the 11 genes were significantly deviating from the null distribution. Using a Bonferroni correction (threshold P-value = 0.0045) for the 11 genes tested, we found 2 genes (*TXNDC12*, $P = 0.00021$ and *SSFA2*, $P = 0.00051$) to be significantly differentially expressed (having a lower rank than expected based on the null) between obese cases and controls in blood and adipose tissue combined (Fig. 3). As judged from the heat map, both *TXNDC12* and *BTF3L4* from the same loci with borderline significance of $P = 0.0066$ had higher gene expression (positive fold change) in obese cases versus controls in most of the studies, whereas *SSFA2*, on the contrary, showed lower expression (negative fold change) in obese in most studies.

Expression and methylation quantitative trait loci analyses

To further investigate functional aspects of the loci identified in Stage 1, we performed both expression quantitative trait loci (eQTL) and methylation QTL (mQTL) analyses (Supplementary Material, Table S7), including any SNPs from the loci identified in Stage 1. Such analyses can identify functional SNPs further emphasizing a role for certain genes in obesity development. We analyzed the largest number of samples available for either *blood* or *adipose tissue*, and additionally using a smaller data set from *carotid plaques*. At first, we used the Genotype-Tissue Expression (GTEx) portal (17) to look for significant *cis*-eQTL associations within 1 Mb window of the genes' start/stop positions in *whole-blood* samples ($n = 168$). We found four significant *cis*-eQTLs (based on permutation) within the gene *TXNDC12* (Supplementary Material, Fig. S5). Further, to calculate gene significance from eQTL data, a gene-specific empirical P-value was generated, and the corresponding Q-value for that gene was calculated (where Q-value <0.05 is indicating significance). In this analysis, the *TXNDC12* gene was found to be significant (Q-value = 0.0065).

Then, we further analyzed *cis*-eQTLs from *adipose tissue* (±1 Mb gene boundaries) by the Genevar (GENe Expression VARiation) platform (18) using default settings with P-value cutoff of <0.001. The samples ($n = 856$) were collected from the TwinsUK cohort with healthy females of European descent (19). Again, *TXNDC12* was found to have two significant eQTLs (Supplementary Material, Fig. S6), and from the same locus, *ZFYVE9* had four significant eQTLs from two different Illumina expression probes (Supplementary Material, Fig. S7). Two other genes also revealed significant eQTLs: *PEX2* with 36 significant eQTLs (Supplementary Material, Fig. S8) and *FTO* with three significant eQTLs (Supplementary Material, Fig. S9).

For the mQTL analyses, we again used the Genevar platform to characterize all genes identified in Stage 1 in the same *adipose tissue* samples from female twins ($n = 856$). The degree of DNA methylation at specific CpG sites was assessed by the Illumina HumanMethylation450 bead chip. The maximum distance to probe start was set to 0.1 Mb, and the P-value cutoff level was <0.001. We found strong enrichment of *cis*-mQTLs in the locus of the *PEX2* gene; two different CpG sites were identified, which were associated with 28 and 82 mQTLs, respectively (Supplementary Material, Fig. S10). The *FTO* gene also showed strong enrichments of multiple *cis*-mQTLs. We found six different CpG sites to be enriched, with 26, 23, 16, 16, 30 and 83 mQTLs, respectively (Supplementary Material, Fig. S11). Lastly, we found five CpG sites with significant SNPs within *TMEM18*. Again, the enrichment was very strong in at least four of the five sites and all together, we identified 41, 17, 9, 38 and 101 mQTLs, respectively (Supplementary Material, Fig. S12).

Finally, we decided to do mQTL follow-up analyses, including association tests with BMI of identified *cis*-mQTLs (13 CpGs from *PEX2*, *FTO* and *TMEM18*) in participants of the Dutch Athero-Express cohort using *carotid plaque* ($N = 488$) and *blood* ($N = 92$) samples. All three genes had at least one significant mQTL in *plaques* (Supplementary Material, Table S1 and Fig. S13) and in *blood* (Supplementary Material, Table S2 and Fig. S13), when adjusting for multiple testing using false discovery rate (FDR < 0.05). Methylation levels in *plaques* at the promoter-associated cg01780754 located within the *PEX2* gene, a probe that was significant in all tissue-specific mQTL analyses, was also significantly associated with BMI (FDR-adjusted P-value = 0.009). Specifically, our data show that the C allele of rs9643429 located upstream of *PEX2* (in LD with the best mQTL SNP) is linked to higher methylation at cg01780754 in *plaques*, and when methylated, cg01780754 is associated with an increase in BMI. The combined functional effects found for *PEX2* is summarized in Supplementary Material, Figure S14.

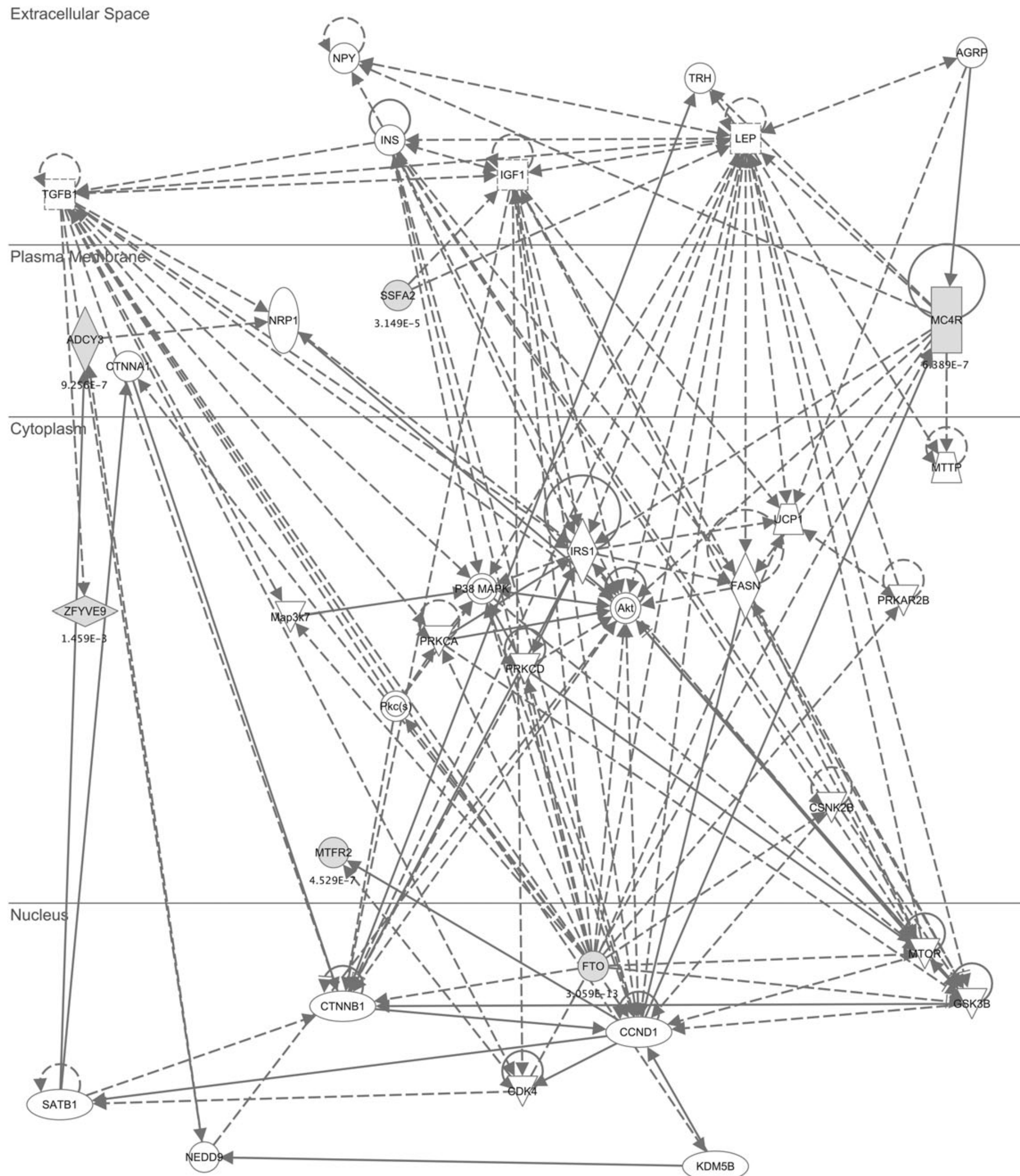

Figure 2. The gene network created with the nuclear gene *FTO* at center, directly and indirectly regulating the other genes in this network. Ingenuity software was used to establish a gene network including the top genes identified in Stage 1 and Stage 2 analyses of the gene-based meta-analyses. The network involved not only the genes *FTO*, *MTFR2*, *ADCY8* and *MC4R* but also the genes *SSFA2* and *ZFYVE9*. The nuclear *FTO* was found to be one of the hubs in the network together with the well-established BMI gene *LEP*.

Evaluation of our gene-based meta-analysis approach

As a secondary aim of this study, we anticipated to evaluate other possible approaches that could have been undertaken, as well as investigating plausible biases from using VEGAS. Our a priori hypothesis was that performing gene-based analyses in each individual study as we did would be a more efficient approach to pick up additionally unknown loci, likely by catching more allelic and between-study heterogeneity, than an approach where gene-based analysis was applied once to the summary statistics from the meta-analysis of SNP-based results. To test this

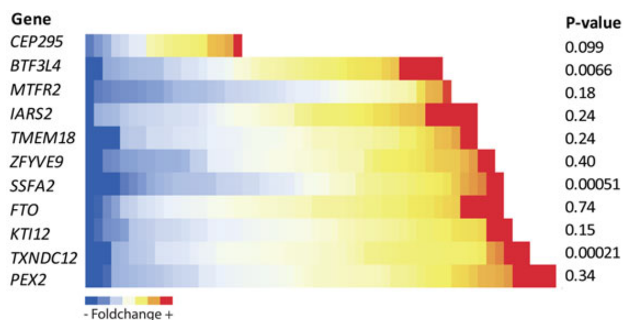

Figure 3. Gene expression validation of genes identified in Stage 1. We downloaded expression data from obese individuals/mice/rats in whole blood or adipose tissue from the GEO repository to look for differential expression of the 11 genes found in Stage 1. Using a rank-based permutation method to generate the null distribution, we concluded that the genes *TXNDC12* and *SSFA2* were significantly differentially expressed between obese cases and controls using a Bonferroni-corrected threshold of 0.0045. The color bar describes the $\log_2$ fold change value for each gene in each study, where blue represents lower expression and red higher expression in obese cases versus controls.

Table 2. Gene length in relation to P-value from VEGAS meta-analysis

P-value cutoff	N Top genes	Mean gene length (bp) Top genes	Mean gene length (bp) Other genes	P-value[a]
<E−6	10	90 603	70 571	0.65
<0.00001	10	90 603	70 571	0.65
<0.0001	16	73 688	70 582	0.51
<0.001	31	59 548	70 612	0.47
<0.01	145	54 417	70 772	0.23
<0.05	530	49 916	71 489	0.06
<0.10	1007	46 644	72 655	0.0052
<0.15	1479	46 391	73 789	3.3E−05
<0.20	1967	46 228	75 072	8.4E−08
<0.30	2912	46 576	77 771	4.8E−12
<0.40	3934	46 818	81 326	4.3E−22
<0.50	4922	49 233	84 205	3.0E−24
<0.60	5982	50 339	88 783	9.8E−29
<0.70	7180	51 681	95 459	1.6E−34
<0.80	8487	52 974	106 608	5.2E−45
<0.90	9972	55 084	128 585	3.8E−57

[a]t-Tests were done on the logarithm of gene length because of the skewed distribution.

assumption empirically, we performed VEGAS gene-based analysis on the summary statistics of the single-SNP meta-analysis of BMI by Speliotes *et al.* (1), which included 123 865 individuals from 43 studies (Fig. 1). The gene-based analysis based on the summary statistics generated a list of 20 significant loci, corresponding to 59 genes, all of which have been reported in the original publication (Supplementary Material, Table S3). We then compared the loci identified in our Stage 1 analysis with this gene-based analysis of the summary statistics from Speliotes *et al.* and found that none of the newly identified loci were gene-wide significant (Supplementary Material, Table S4) in the latter analysis, whereas the previously established loci (*FTO* and *TMEM18*) were represented among the top findings, reaching gene-wide significance.

The same approach was undertaken to examine the gene-based analysis of the summary statistics data from Locke *et al.* (2), which was the data source used for our Stage 2 analyses (Fig. 1). The result from the VEGAS analysis revealed 33 gene-wide significant loci containing 110 genes in total (Supplementary Material, Table S5). All gene-wide significant loci, except for two, were already reported as being associated with BMI in the Locke *et al.* paper. The two newly identified loci were golgin A2 (*GOLGA2*) and secretory carrier membrane protein 4 (*SCAMP4*). When we then compared these results with the gene-based loci identified in our Stage 1, the only common loci were again the established ones (*FTO* and *TMEM18*) (Supplementary Material, Table S6). Thus, our hypothesis that applying the VEGAS software on summary statistics would render fewer novel loci compared with our approach of performing VEGAS in each substudy before meta-analyzing seemed to be true. The most likely reason for this observation is that allelic heterogeneity in the gene-based meta-analysis (Stage 1 and 2 results), where many different SNP signals rather than a strong single driver, gave rise to the gene-wide significant hits. This explanation was further supported by the variety of top SNPs that contributed to the VEGAS gene-based signals in Stages 1 and 2 (Supplementary Material, Figs S3 and S4). However, there is yet to be proven if the novel BMI loci are good biological targets.

Next, we decided to investigate whether the length of the gene could bias our observed results. The VEGAS method corrects for the number of SNPs within each gene and also for gene length as has been shown before (20). Nonetheless, using Stage 1 data, the mean gene length of all genes was 70 582 bp, whereas the median gene length was 28 333 bp (range: 2 304 483 bp), indicating a skewed distribution. At first, we analyzed the mean gene length of a varying list of top genes ($P < 10^{-6}$ to $P < 0.05$ for inclusion) comparing with the rest of the genes, but found no significant differences (Table 2). However, as the top gene list increased when relaxing the P-value threshold even more, there was a significant difference in mean gene length between the groups, with shorter genes among the top genes indicating that VEGAS tends to over-adjust for gene length (Table 2). To further support this notion, we performed correlation analysis and found a weak but significant correlation between increased gene length and higher P-value (Spearman rank $\rho = 0.14$; $P < 10^{-16}$).

Finally, we performed sensitivity analyses in the three most promising novel loci (*PEX2*, *TXNDC12* and *SSFA2*) in Stage 1 cohorts. For each locus and cohort, we deleted the best SNP as reported by VEGAS and consequently reran VEGAS on all data sets. The difference in the gene-based P-value from the two runs (with and without the best SNP) in each cohort and locus is presented in Supplementary Material, Figure S15. The mean difference in gene-based P-value in each cohort ($P_{Sensitivity} - P_{Normal}$) was 0.015 for *PEX2*, 0.020 for *SSFA2* and 0.022 for *TXNDC12*. The overall conclusion when summing up the Fisher χ^2 statistics over all cohorts was that the final gene-based P-value changed one 10-fold to the higher for all three loci, still well below gene-wide detection for *PEX2* and *TXNDC12*, and reaching borderline significance level for *SSFA2*.

Discussion

In this study, we present a novel approach utilizing previous GWAS results by applying gene-based methods on individual study results followed by meta-analysis in order to find new loci associated with BMI. The study samples used were all from the GIANT consortium including Stage 1 (1) and Stage 2 data (2). We identified six new loci associated with BMI, including several genes with high potential for further functional follow-up studies (e.g. *PEX2*, *TXNDC12* and *SSFA2*), and could confirm two established loci (*FTO* and *TMEM18*). All loci were replicated in

Stage 2, and we provide further biological support from a range of bioinformatics approaches, such as expression analyses, as well as pathways, and eQTL and mQTL analyses (Supplementary Material, Table S7).

The concept of allelic heterogeneity

Allelic heterogeneity occurs when multiple alleles in the same locus are associated with the same trait. The phenomenon of allelic heterogeneity is being reported for more and more loci and traits, and it is a well-known phenomenon in long-established loci, such as *BRCA1* and *BRCA2*. It is easy to imagine several causal variants that affect the same gene transcripts through changes in expression levels or gene structure and function, and thereby in the end affect the same biological mechanism. Loci that display high allelic heterogeneity can be missed by conventional GWAS of individual variants, as the statistics are based on one variant at the time. Gene-based methods on the other hand combine all signals for individual variants within the same gene to a single gene-based signal, hence accounting for allelic heterogeneity. Moreover, by applying a gene-based meta-analysis on the study-level data (instead of performing the analyses on the meta-analysis summary statistics), the allelic heterogeneity is better accounted for, and moreover, between-study heterogeneity is captured if existing. By comparing the results of the gene-based analysis done on a cohort level (Table 1) with a gene-based analysis based on summary statistics (Supplementary Material, Tables S3 and S5), we could conclude that very different set of genes are being identified using the two approaches. Thus, this observation supports our notion that the heterogeneity is more accurately picked up by VEGAS when run on cohort level data rather than directly on the GWAS meta-analysis summary data. The likely reason is that all SNPs and cohorts contribute to the result, not just a single significant SNP in a large cohort, a suggestion that is further supported by the replication in Stage 2 (Table 1), sensitivity analyses without the best SNP (Supplementary Material, Fig. S15) and the variety of lead SNPs from different studies in the significant loci (Supplementary Material, Figs S3 and S4).

In performing a gene-based association test, one presumes that the contributions of different causal variants will be associated with the outcome via the same downstream disease mechanism. However, an important feature in Fisher's method for meta-analysis is that it allows effect sizes in both positive and negative directions, so there is no restriction in the types of allelic heterogeneity within the gene.

Multivariate methods where independent SNPs are combined within the same locus have improved the total variance explained, suggesting that allelic heterogeneity could be a major contributor (4,5). However, independent SNP signals within a locus are often difficult to identify because standard GWASs are not designed to do so and there are computational limitations of multivariate studies. On the contrary, the gene-based approach has the potential to detect loci where independent SNP signals contribute to the same gene-based signal. Thus, loci with strong secondary signals are prioritized, and as a consequence, we identified some of the established BMI loci with known secondary signals among our top findings. *FTO* and *TMEM18* (identified in Stage 1), as well as *MC4R* (identified in Stage 2), have been reported to harbor multiple SNPs associated with BMI (21).

The loci with allelic heterogeneity are also more likely to contain functional SNPs, as judged by software such as PolyPhen and by eQTL mapping (4,5). In light of this, it is not surprising that the QTL investigations in this study concluded that 5 of the 11 genes (4 of the 8 loci) that we discovered in Stage 1 showed evidence of enrichment in eQTLs and/or mQTLs (Supplementary Material, Table S7). When also considering the expression and pathway analyses, we found functional evidence in 7 of the 11 genes (6 of the 8 loci) (Supplementary Material, Table S7).

VEGAS in comparison with other gene-based methods

There is a growing number of gene-based bioinformatics tools available. Many of them use participant-level genotype data as input, perform a gene-based association test using permutations and summarize the results in a list of gene-to-trait associations (6,22). These methods usually perform well in terms of limiting false-positive rates; however, they have a great disadvantage being computationally demanding and needing access to participant-level genotype data, which usually is impossible in the setting of large genetic consortia. Other methods, including VEGAS, only use a list of SNP *P*-values as input and then consider the underlying SNP–SNP correlation pattern using HapMap Project's genotype data or other data sources (7,8,23). These methods are computationally less burdensome and use summary result files, which is crucial when working within large international consortia.

The method in VEGAS that we used performs well compared with other gene-based analyses considering different gene boundaries (9). Basically, VEGAS is stable across different boundaries chosen and remains powerful even with the inclusion of non-significant SNPs (9). We used the default gene boundary in VEGAS (±50 kb), which is appropriate as GWAS hits are enriched within this distance (24).

GWASs are known to prioritize SNPs in the close vicinity to or within genes of long length (20,24). The VEGAS method, however, has been shown to compensate for the gene length bias in the analysis (20). Our study investigated this further and found that VEGAS tends to be too stringent when compensating for gene length, and as a consequence, long genes are less likely to be significant. The fact that we did not identify any genes related to behavior and appetite, which are usually found in genetic studies of BMI, could be a result of this phenomena, as brain genes tend to be longer than other genes (25).

Functional aspects of newly identified genes

The protein encoded by *TXNDC12* is characterized by a conserved active motif called the thioredoxin fold that catalyzes disulfide bond formation and isomerization (26). The *txndc12* knockout model in mice does not show any phenotype changes. However, a study in grouper fish concluded that TXNDC12 is an important antioxidant with potential physiological properties (27). They further noticed that it is predominately expressed in liver, muscle and brain, and in addition, we found it to be differentially expressed in both blood and adipose tissue (which they did not test). Another interesting candidate gene from the same locus is *ZFYVE9*, which encodes a double zinc finger protein that interacts with SMAD proteins that are signal transducers in the transforming growth factor β (TGF-β) signaling pathway. The relation to TGF-β is also demonstrated in the network analysis (Fig. 2). A mediated knockdown model of *zfyve9a* in zebrafish inhibits the formation of the liver in early embryogenesis (28). In addition, we found significant eQTLs in adipose tissue strengthening the functional aspect of this gene. The four genes located in this locus at Chromosome 1 are found in a region with strong LD (Supplementary Material, Fig. S2). Using VEGAS, we cannot

compensate for including the same SNPs in several genes; thus, many SNPs are common between genes in a region like this (Supplementary Material, Fig. S16). Based on the functional data we present, *TXNDC12* seems to be the best candidate for involvement in adiposity among the four, followed by *ZFYVE9* that also had some functional validation and did not share any of the most strongly associated BMI SNPs with *TXNDC12*.

The protein encoded by *PEX2* belongs to a family of peroxisomal membrane proteins where gene mutations lead to peroxisome biogenesis disorders, such as Zellweger syndrome with various abnormalities seen in the central nervous system (29). The *pex2* knockout mouse dies before weaning at ∼13 days of age, which is in line with humans suffering from the Zellweger syndrome who usually die at adolescence. Decreased levels of triglycerides and high-density lipoprotein cholesterol in plasma are seen before dying (30), and PEX2 has also been shown to decrease angiogenesis in ischemia and wound healing (31). We found significant eQTLs in the *PEX2* locus and strong enrichment of mQTLs in adipose tissue, carotid plaques and blood. Moreover, DNA methylation in carotid plaques was associated with BMI, indicating an intrinsic SNP–methylation–obesity relationship (Supplementary Material, Fig. S14). Thus, we highlight a functional role of *PEX2* through gene expression and in particular methylation in multiple tissues.

SSFA2 was first considered a cancer-related gene as its expression level is altered in colon cancer cells (32), but other findings have described its role in metabolic disturbances through regulation of energy homeostasis and the exocrine system with implications for obesity and diabetes (33–36). *SSFA2* is expressed in brown adipose tissue, liver, pancreas, stomach and kidney. The *ssfa2* gene knockout mouse model exhibit decreased body weight and adipose tissue, resistance to diet-induced obesity and increased food intake and metabolic rate. Moreover, abnormal glucose homeostasis and hormone levels are detected (33,36) and it is connected to *LEP* in the gene network (Fig. 2). Our gene expression analysis further concluded that *SSFA2* is associated with obesity (Fig. 3).

The *MTFR2* gene is not well known but is involved in aerobic respiration in the mitochondria. The Ingenuity analysis highlighted *MTFR2* in the cytoplasm where it is regulated by nuclear genes important for obesity (Fig. 2). *IARS2* encodes a tRNA synthetase for isoleucine found in mitochondria. Isoleucine is the predominant amino acid in human mitochondria and small changes of the aminoacylation properties may be of great importance (37). Point mutations found in this gene have been shown to cause fatal cardiomyopathy, growth retardation, hearing deficiency, neuropathy and skeletal dysplasia (38,39). *CEP295*, also called *KIAA1731*, is a putative centrosomal protein containing an ALMS motif, also seen in the gene Alström syndrome 1 (*ALMS1*). As the name imply, mutations in *ALMS1* cause Alström syndrome, a rare progressive condition characterized by neurosensory degeneration and metabolic defects (40).

Strengths and limitations

A main strength of this study was the very large sample size used for the VEGAS analyses, and that we could do gene-based analyses in each participating study before meta-analyzing (instead of using the summary statistics from the meta-analysis to do gene-based analyses). Our approach allowed for the identification of six novel loci associated with BMI. Moreover, as no new samples were interrogated in this study, this approach highlights the potential of the gene-based meta-analysis method as such for reuse of existing data. The soundness of this approach was confirmed by replication of all loci in an independent validation stage (Stage 2) as well as by confirmation of well-known BMI GWAS loci with documented strong secondary signals (*FTO*, *TMEM18* and *MC4R*). Furthermore, by using pathway analyses, gene expression data and eQTL and mQTL analyses, we were able to functionally validate genes in six of eight loci found. This also illustrates the possibility of using VEGAS to prioritize among candidate GWAS genes as the top genes from VEGAS are more likely to harbor functional variants.

Limitations of the study include the assumptions needed for the gene-based analysis, such as the one-directional effect size. It is also worth mentioning that the VEGAS method could be run with other settings, for example using a different LD panel, which would most likely render slightly different results. We further concluded that VEGAS overadjusts for gene length, which may have had an impact on the results. Moreover, the functional validation of the significant genes was only concentrated to those specific genes; no additional genes in the close vicinity were considered to limit the extent of multiple testing; hence, further functional studies are warranted to conclude whether the identified genes are truly linked to obesity. The gene expression analysis was restricted to blood and adipose tissue as these were the tissues frequently available in the GEO repository; investigating brain tissue in relation to obesity status would most likely have rendered a different result. Finally, as all results are based on analyses of individuals of European descent, the generalizability to other ethnicities is unknown.

Conclusions

In this study, we present a novel approach for conducting gene-based meta-analysis to reutilize large GWAS data sets in order to identify new genes involved in complex diseases. Moreover, we perform extensive bioinformatic functional validation of the identified genes illustrating the potential of using methods like VEGAS to prioritize among candidate genes and to detect loci with high allelic heterogeneity. Using this approach, we highlight several new promising gene candidates for further validation in functional studies, such as *TXNDC12*, *PEX2* and *SSFA2*. Further attempts using gene-based meta-analyses of existing data for other phenotypes using the methods outlined are encouraged.

Materials and Methods

Study samples

The Stage 1 data set comprised 46 studies with in total 123 865 individuals of European descent. The sample sizes varied from 276 to 26 799 individuals per study, and they were genotyped using Affymetrix or Illumina technology as described elsewhere (1). BMI levels were inverse normally transformed and adjusted for age, age^2 and study-specific covariates such as population stratification by principal components. All genotype data were imputed to the HapMap CEU panel and badly imputed SNPs ($\hat{r}^2 < 0.3$ in MACH, observed/expected dosage variance <0.3 in BimBam or proper_info <0.4 in IMPUTE), and those with a minor allele count <6 per cohort-specific strata were removed. Analysis using additive models for association between genotype and BMI was performed on cohort level, and each GWAS was then corrected with lambda genomic control. For the meta-analysis on individual-variant level, the inverse variance method in METAL software was used [analyses performed as previously reported (1)].

The Stage 2 data set included 43 additional studies of European ancestry with genotype data available from the Illumina iSelect Metabochip, the only additional data available to us for replication at the time of analysis, covering 196 725 SNPs selected for cardiac and metabolic phenotypes, including 5055 SNPs selected for associations with BMI (2). The Metabochip data were not imputed, but corrected with lambda genomic control in each study separately.

The VEGAS method

To yield lists of gene-based *P*-values for association with BMI, the VEGAS software (http://gump.qimr.edu.au/VEGAS/) was applied to each cohort separately. The VEGAS algorithm has been described elsewhere (8), but in brief, it takes the full set of markers within a gene and accounts for LD by simulations from HapMap LD structures. A gene-based test statistic is calculated as the sum of all χ^2-converted SNP *P*-values within that gene. The null distribution is then calculated using Monte Carlo simulation taking into account the LD structure that results in a large number of multivariate normal vectors with the same distribution as our gene-based test statistic. The simulation approach is computationally faster than using permutations but performs equally well (8). The assignment of SNPs to genes was done with the UCSC genome browser hg18 assembly rendering 17 941 genes. Gene boundaries were defined as ±50 kb of 5′ and 3′ untranslated regions, which is the default setting in VEGAS. However, a study investigating different boundaries in relation to the gene-based result in several methods found the VEGAS sum method to be stable for different lengths chosen (9). To limit the computational burden, we set the maximum number of simulations to 1 000 000, meaning that the lowest *P*-value that could be estimated was 1×10^{-6}. The rationale for this number of simulations was the Bonferroni-corrected gene-wide significance threshold of $P < 2.8 \times 10^{-6}$ (correcting for 17 941 gene tests). Due to this truncation, the *P*-values from this gene-based approach appear higher than those from single variant-based GWAS for some cohorts (for example, for the *FTO* locus).

Meta-analysis using Fisher's method

In gene-based association analysis, an underlying assumption is that all causal effects within a gene are having the same direction. Therefore, when performing a meta-analysis on gene level, a one-sided χ^2 distribution as used in the Fisher method is preferred. The Fisher formula for meta-analysis is as follows:

$$\chi^2 = -2\sum_{i=1}^{k} \log_e(p_i)$$

Pathway analysis

Enrichment analysis for pathways was carried out using the web-based tool ConsensusPathDB (10) (http://cpdb.molgen.mpg.de/CPDB/). Genes were searched for overrepresentation among different pathways, and Q-values were calculated from corrected *P*-values using FDR from the hypergeometric distribution. Ingenuity Pathway Analysis (Qiagen, fall 2014 version) was used for visualizing and analyzing the gene-based results of BMI (11). We only focused on genes available in the Ingenuity Knowledge Base. The confidence filter was set to consider only experimentally observed direct and indirect relationships in mice, rats and humans. A maximum of 35 nodes per network and a maximum of 25 networks were generated. Ingenuity calculates the significance values for canonical pathways by Fisher's exact test right tailed.

Gene expression analysis methods

Data sets for functional validation of the BMI-associated genes from Stage 1 were identified in GEO (http://www.ncbi.nlm.nih.gov/geo/) using the following search term: ('blood' OR 'adipose tissue') AND ('adiposity' OR 'obesity' OR 'obese' OR 'body-mass index' OR 'body mass index' OR 'BMI') AND (Homo sapiens [Organism] OR Mus musculus[Organism] OR Rattus norvegicus [Organism] AND (gds[Entry Type] OR gse[Entry Type])), which rendered 322 hits. Two investigators (S.H. and E.I.) went through the list independently and selected those data sets that fulfilled the criteria of having obese cases versus non-obese controls within similar experimental settings. The final number of data sets that were selected via consensus for further analysis was 76. The data sets were downloaded from the GEO repository using R with the GEOquery package and analyzed using the limma package. Each data set could sometimes be split into multiple subsets depending on experimental designs, e.g. mouse models with different time intervals, as long as the case–control criteria were met. All data sets were quantile normalized and $\log_2$-transformed if not done already, and quality control of the final differential analysis was done using volcano plots. Bad quality data, microRNA-chip data and lack of control samples were common reasons for exclusions of data sets during the analysis procedure. The final data set included 335 cases and 308 controls from 53 experiments. Differential expression analysis was conducted within each experiment, and the transcript with the lowest *P*-value within the same gene, if multiple transcripts existed, was selected. Genes were compared between different array platforms using the official human HGNC gene symbol list downloaded from NCBI. All gene aliases and orthologous genes in mouse and rat were consequently translated to the official human symbol and accounted for in the search. Every gene in each data set was then ranked according to *P*-value, and the final rank was scaled to the same distribution (rank = position/number_of_genes). The average ranking for each gene was calculated across all 53 data sets where missing genes in any data set were assigned the rank 0.5. To calculate the null distribution, each data set was permuted 300 times and the average gene ranking was computed 300 times for each gene. Then, the 300 permuted gene ranks were compared with the original gene rank for each gene separately to generate a *P*-value. The final *P*-value was further Bonferroni corrected taking 11 tests into account for the 11 genes tested.

QTL methods

Enrichments of QTL were tested using three approaches. First, the GTEx portal (17) (http://www.gtexportal.org/home/) was used via their web interface for screening of eQTLs in whole blood. Expression data were collected from Affymetrix arrays or Illumina RNA sequencing, and genotype data from Illumina arrays. We only tested for *cis*-eQTLs allowing SNPs to reside in a window of 1 Mb from start or stop of the corresponding gene. The overall enrichment significance for a gene was calculated with FDR using permutation and presented in an adjusted Q-value. The number of significant *cis*-eQTLs within each gene was dependent on the permutation-based *P*-value threshold for that gene. The samples ($n = 168$) used for whole-blood analysis were collected through the Cancer Human Biobank (caHUB) of

the National Cancer Institute and came from adult individuals of any sex and ancestry, taken either during surgery or postmortem. Specifications on sample quality or data preprocessing can be found online. The second software used for QTL analyses was the Genevar platform (18) (http://www.sanger.ac.uk/resources/software/genevar/) developed at the Wellcome Trust Sanger Institute. The software is a standalone Java application, which was downloaded to the computer and once running, it connects to the online databases needed. Again, we defined *cis*-eQTLs as SNPs within ±1 Mb of gene boundaries and used the default settings for significance with cutoff for *P*-value <0.001. The corresponding numbers for the *cis*-mQTLs were ±0.1 Mb gene boundaries and a *P*-value cutoff level <0.001. The Genevar software does not present any overall gene scores as GTEx does. The samples were from adipose tissue collected from the TwinsUK cohort with 856 healthy female twins of European descent run on Illumina HumanMethylation450 bead chip (19,41).

Lastly, mQTL analyses were performed in Athero-Express, a cohort including patients from two Dutch hospitals undergoing carotid endarterectomy. DNA was isolated from carotid artery plaque samples of 500 patients and ethylenediaminetetraacetic acid whole-blood samples of 100 matched patients and checked for purity and concentration using a Nanodrop system (Thermo Scientific, Massachusetts, USA). DNA concentrations were equalized, bisulphate converted and used to measure DNA methylation using the HumanMethylation450 bead chip (Illumina, San Diego, USA), according to the manufacturer's protocol. Data were processed using the 'MethylAid' in R (42) with normalization and batch-effect correction using 'Functional Normalization' with eight principal components of control probes. Invariable probes (β range: 0–0.1 and 0.9–1) and outliers at >3 SD from median were removed from each probe. After quality control, 488 plaque samples, 92 blood samples and 443 872 probes (41 640 probes were excluded) remained for further analysis. Genotyping, available in 444 plaques and 92 blood samples, has previously been performed and is described elsewhere (43). For the association with BMI, linear regression modeling was performed with covariates age and sex. mQTL analysis was performed in SNPtest (v2.5) using a model with covariates age, sex and SNP array type. The Benjamini–Hochberg method was used to account for multiple testing, where a FDR <5% was considered statistically significant.

Gene length analyses

The information on gene length was extracted and downloaded from the table browser at UCSC Genome Browser website. The genes were sorted based on significance from the VEGAS Stage 1 analysis, and the top genes were compared with the complementary gene list using different cutoff values ($P < 10^{-6}$, $<10^{-5}$, <0.0001, <0.001, <0.01, <0.05, <0.10, <0.15, <0.20, <0.30, <0.40, <0.50, <0.60, <0.70, <0.80, <0.90). A two samples *t*-test was used in the calculation and applied to the log-transformed values of gene lengths due to the skewed distribution of gene length. In addition, a Spearman rank correlation test between gene length and *P*-value was performed.

Sensitivity analyses

The best SNP as judged from the lowest *P*-value within each gene is reported by VEGAS for each individual cohort. The stability of the gene-based *P*-value can be investigated if the best SNP is deleted from the data set. Consequently, for each of the three loci (*PEX2*, *TXNDC12* and *SSFA2*), the best SNP as reported for each cohort and locus was deleted and VEGAS analysis was again undertaken. The difference between the originally reported gene-based *P*-value and the newly generated gene-based *P*-value (without the best SNP in the data set) was calculated and evaluated.

Statistical software

The VEGAS software was used in an offline command version and additional analyses were carried out using Unix, Perl, R 3.0.0 or 3.1.1 and Mathematica version 8. The LD pattern plot was created using SNAP software version 2.2 (SNP Annotation and Proxy Search; http://www.broadinstitute.org/mpg/snap/ldplot.php) with SNP data set 1000 Genomes Pilot 1 and population panel CEU. The union of all SNPs contributing to the gene-based signal from the *TXNDC12* locus with four genes (*TXNDC12*, *BTF3L4*, *KTI12* and *ZFYVE9*) was submitted ($n = 90$ SNPs). Regional plots were made by the Locus Zoom online web tool (http://csg.sph.umich.edu/locuszoom/) where SNPs contributing to the gene-based signals and the corresponding *P*-value from each GWAS were uploaded for each locus separately. The best SNP in the locus was chosen for each cohort, and if several cohorts had the same SNP, it was plotted multiple times. The reference panel was 1000 Genomes build hg18 with population panel CEU.

Ethics statement

We are only using summary statistics from each cohort within GIANT, and the data have already been published and described elsewhere (1,2). Thus, this study does not need any additional ethical permit.

Supplementary material

Supplementary Material is available at *HMG* online.

Acknowledgements

The authors would like to acknowledge all contributors to the Genetic Investigation of ANthropometric Traits (GIANT) Consortium and all research participants of studies within GIANT for their very kind contributions toward science.

Conflict of Interest statement: None declared.

Funding

The study was supported by the Foundation for Geriatric Diseases, Karolinska Institutet.

References

1. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R. *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937–948.
2. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S. and Buchkovich, M.L. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.
3. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R.,

Chakravarti, A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
4. Ehret, G.B., Lamparter, D., Hoggart, C.J., Whittaker, J.C., Beckmann, J.S. and Kutalik, Z. (2012) A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am. J. Hum. Genet*., **91**, 863–871.
5. Wood, A.R., Hernandez, D.G., Nalls, M.A., Yaghootkar, H., Gibbs, J.R., Harries, L.W., Chong, S., Moore, M., Weedon, M. N., Guralnik, J.M. *et al*. (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet*., **20**, 4082–4092.
6. Huang, H., Chanda, P., Alonso, A., Bader, J.S. and Arking, D.E. (2011) Gene-based tests of association. *PLoS Genet*., **7**, e1002177.
7. Li, M.X., Gui, H.S., Kwan, J.S. and Sham, P.C. (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet*., **88**, 283–293.
8. Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P. M., Martin, N.G. *et al*. (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet*., **87**, 139–145.
9. Petersen, A., Alvarez, C., DeClaire, S. and Tintle, N.L. (2013) Assessing methods for assigning SNPs to genes in gene-based tests of association using common variants. *PLoS One*, **8**, e62161.
10. Kamburov, A., Stelzl, U., Lehrach, H. and Herwig, R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*., **41**, D793–D800.
11. Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K. *et al*. (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1032–1037.
12. Munzberg, H. and Morrison, C.D. (2015) Structure, production and signaling of leptin. *Metabolism*, **64**, 13–23.
13. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V. *et al*. (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med*., **373**, 895–907.
14. Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gomez-Marin, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F. *et al*. (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, **507**, 371–375.
15. Stratigopoulos, G., Martin Carli, J.F., O'Day, D.R., Wang, L., Leduc, C.A., Lanzano, P., Chung, W.K., Rosenbaum, M., Egli, D., Doherty, D.A. *et al*. (2014) Hypomorphism for RPGRIP1L, a ciliary gene vicinal to the FTO locus, causes increased adiposity in mice. *Cell Metab*., **19**, 767–779.
16. Merkestein, M., Laber, S., McMurray, F., Andrew, D., Sachse, G., Sanderson, J., Li, M., Usher, S., Sellayah, D., Ashcroft, F.M. *et al*. (2015) FTO influences adipogenesis by regulating mitotic clonal expansion. *Nat. Commun*., **6**, 6792.
17. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S. and Hasz, R. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet*., **45**, 580–585.
18. Yang, T.P., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E., Deloukas, P. and Dermitzakis, E.T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.
19. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A. *et al*. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet*., **44**, 1084–1089.
20. Mirina, A., Atzmon, G., Ye, K. and Bergman, A. (2012) Gene size matters. *PLoS One*, **7**, e49093.
21. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J. *et al*. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet., 44*, **369–375**, S361–S363.
22. Moskvina, V., O'Dushlaine, C., Purcell, S., Craddock, N., Holmans, P. and O'Donovan, M.C. (2011) Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genet. Epidemiol*., **35**, 861–866.
23. Hong, M.G., Pawitan, Y., Magnusson, P.K. and Prince, J.A. (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet*., **126**, 289–301.
24. Leslie, R., O'Donnell, C.J. and Johnson, A.D. (2014) GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, **30**, i185–i194.
25. Huang, Y., Xie, C., Ye, A.Y., Li, C.Y., Gao, G. and Wei, L. (2013) Recent adaptive events in human brain revealed by meta-analysis of positively selected genes. *PLoS One*, **8**, e61280.
26. Alanen, H.I., Williamson, R.A., Howard, M.J., Lappi, A.K., Jantti, H.P., Rautio, S.M., Kellokumpu, S. and Ruddock, L.W. (2003) Functional characterization of ERp18, a new endoplasmic reticulum-located thioredoxin superfamily member. *J. Biol. Chem*., **278**, 28912–28920.
27. Wei, J., Ji, H., Guo, M. and Qin, Q. (2012) Isolation and characterization of a thioredoxin domain-containing protein 12 from orange-spotted grouper, *Epinephelus coioides*. *Fish Shellfish Immunol*., **33**, 667–673.
28. Liu, N., Li, Z., Pei, D. and Shu, X. (2013) Zfyve9a regulates the proliferation of hepatic cells during zebrafish embryogenesis. *Int. J. Dev. Biol*., **57**, 773–778.
29. Mignarri, A., Vinciguerra, C., Giorgio, A., Ferdinandusse, S., Waterham, H., Wanders, R., Bertini, E., Dotti, M.T. and Federico, A. (2012) Zellweger spectrum disorder with mild phenotype caused by PEX2 gene mutations. *JIMD Rep*., **6**, 43–46.
30. Kovacs, W.J., Shackelford, J.E., Tape, K.N., Richards, M.J., Faust, P.L., Fliesler, S.J. and Krisans, S.K. (2004) Disturbed cholesterol homeostasis in a peroxisome-deficient PEX2 knockout mouse model. *Mol. Cell. Biol*., **24**, 1–13.
31. Nedeau, A.E., Gallagher, K.A., Liu, Z.J. and Velazquez, O.C. (2011) Elevation of hemopexin-like fragment of matrix metalloproteinase-2 tissue levels inhibits ischemic wound healing and angiogenesis. *J. Vasc. Surg*., **54**, 1430–1438.
32. Inokuchi, J., Komiya, M., Baba, I., Naito, S., Sasazuki, T. and Shirasawa, S. (2004) Deregulated expression of KRAP, a novel gene encoding actin-interacting protein, in human colon cancer cells. *J. Hum. Genet*., **49**, 46–52.
33. Fujimoto, T., Miyasaka, K., Koyanagi, M., Tsunoda, T., Baba, I., Doi, K., Ohta, M., Kato, N., Sasazuki, T. and Shirasawa, S. (2009) Altered energy homeostasis and resistance to diet-induced obesity in KRAP-deficient mice. *PLoS One*, **4**, e4240.
34. Fujimoto, T. and Shirasawa, S. (2011) KRAS-induced actin-interacting protein: a potent target for obesity, diabetes and cancer. *Anticancer Res*., **31**, 2413–2417.
35. Fujimoto, T. and Shirasawa, S. (2012) Identification of KRAP-expressing cells and the functional relevance of KRAP to

the subcellular localization of IP3R in the stomach and kidney. *Int. J. Mol. Med.*, **30**, 1287–1293.
36. Miyasaka, K., Fujimoto, T., Kawanami, T., Takiguchi, S., Jimi, A., Funakoshi, A. and Shirasawa, S. (2011) Pancreatic hypertrophy in Ki-ras-induced actin-interacting protein gene knockout mice. *Pancreas*, **40**, 79–83.
37. Degoul, F., Brule, H., Cepanec, C., Helm, M., Marsac, C., Leroux, J., Giege, R. and Florentz, C. (1998) Isoleucylation properties of native human mitochondrial tRNAIle and tRNAIle transcripts. Implications for cardiomyopathy-related point mutations (4269, 4317) in the tRNAIle gene. *Hum. Mol. Genet.*, **7**, 347–354.
38. Perli, E., Giordano, C., Tuppen, H.A., Montopoli, M., Montanari, A., Orlandi, M., Pisano, A., Catanzaro, D., Caparrotta, L., Musumeci, B. *et al.* (2012) Isoleucyl-tRNA synthetase levels modulate the penetrance of a homoplasmic m.4277T>C mitochondrial tRNA(Ile) mutation causing hypertrophic cardiomyopathy. *Hum. Mol. Genet.*, **21**, 85–100.
39. Schwartzentruber, J., Buhas, D., Majewski, J., Sasarman, F., Papillon-Cavanagh, S., Thiffaut, I., Sheldon, K.M., Massicotte, C., Patry, L., Simon, M. *et al.* (2014) Mutation in the nuclear-encoded mitochondrial isoleucyl-tRNA synthetase IARS2 in patients with cataracts, growth hormone deficiency with short stature, partial sensorineural deafness, and peripheral neuropathy or with Leigh syndrome. *Hum. Mutat.*, **35**, 1285–1289.
40. Knorz, V.J., Spalluto, C., Lessard, M., Purvis, T.L., Adigun, F.F., Collin, G.B., Hanley, N.A., Wilson, D.I. and Hearn, T. (2010) Centriolar association of ALMS1 and likely centrosomal functions of the ALMS motif-containing proteins C10orf90 and KIAA1731. *Mol. Biol. Cell*, **21**, 3617–3629.
41. Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M. *et al.* (2013) Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.*, **93**, 876–890.
42. van Iterson, M., Tobi, E.W., Slieker, R.C., den Hollander, W., Luijk, R., Slagboom, P.E. and Heijmans, B.T. (2014) MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*, **30**, 3435–3437.
43. Azghandi, S., Prell, C., van der Laan, S.W., Schneider, M., Malik, R., Berer, K., Gerdes, N., Pasterkamp, G., Weber, C., Haffner, C. *et al.* (2015) Deficiency of the stroke relevant HDAC9 gene attenuates atherosclerosis in accord with allele-specific effects at 7p21.1. *Stroke*, **46**, 197–202.