Integrative Systems Biology: Elucidating Complex Traits

Pers, Tune Hannes; Brunak, Søren

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Integrative Systems Biology: Elucidating Complex Traits

Tune H Pers

31st March 2011

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS
DTU Systems Biology
Department of Systems Biology

REGION H
Institute of
Preventive Medicine

More is different.

P. W. Anderson, 1972

Preface

The questions that motivated this Thesis originates from my interest in exploring the unknown. Standing on the shoulders of previous hard working scientists, be it laborants or Nobel laureates, we now have the unprecedented chance to understand (patho)-biology, by assembling well-studied pieces of knowledge into descriptions of biological systems.

The work presented in this PhD Thesis has been carried out between April 1 2008 and March 31 2011 at the Center for Biological Sequence Analysis at the Technical University of Denmark with *Søren Brunak* as my supervisor. Besides Søren, *Thorkild I. A. Sørensen* from the Institute of Preventive Medicine, Copenhagen University Hospitals, has been my co-supervisor.[1]

I wish to express special thanks to Thorkild for hiring me as a student-researcher during my Master's studies at the University of Copenhagen in 2006, and for always motivating me to systematically explore the unknown. Thorkild has steered me through several projects and always encouraged me to write up work — also in case of negative results ("Husk at klare 'nul-resultater' ofte - og især hvis hypotesen er velbegrundet - er mere værdifulde end vage positive fund. De klare nuller gør arbejdet færdigt så ingen derefter behøver bruge mere krudt på dem")[2].

I am very thankful to Søren and *Kasper Lage* for paving the way in a research field, which not until recently has started moving from reductionist-dominated biology towards systems biology. Søren has never said no to a chat, nor a short pitch of a new idea. He has supported me in everything I have done, and I have learned a lot from his pragmatic way of being. Kasper has been a key friend and strongly inspiring researcher during my time as a PhD student, and my research stays in Boston.

My thanks go to *Trey Ideker*, *Rohith Srivas* and *Gregory Hannum* at the University of California, San Diego. Trey welcomed me whole-heartedly, and my 6-month research stay in his group was extremely inspiring and joyful.

I would like to thank *Majken K. Jensen* for almost daily exchanges on thoughts about genome-wide association data analysis techniques, and a lot of other things; *Niclas Tue Hansen* for conceiving the original idea for the MetaRanker method (presented in this Thesis) and his immense amount of work on that paper; *Daniel Edsgärd* and *Nils Weinhold* for daily chats on data integration methodologies; *Piotr Dworzynski* for helping me to implement MetaRanker as an efficient web-tool; *Peter Wad Sackett*, *John Damm Sørensen*, *Kristoffer Rapacki*, and *Hans Henrik Stærfeldt* for never hesitating to help me with technical assistance; *Olga Rigina* for help on InWeb-related issues; all *Integrative Systems Biology group members* for fruitful discussions; *Søren Mørk*, *Agata Wesolowska*, and *Rachita Yadav* for commenting on the Thesis; and *Anders Krogh* for a great time at the Bioinformatics Centre of the University of Copenhagen during my Master's studies. In addition I would like to thank *Annette V. Uldall*, *Lone Boesen*, *Dorthe Kjærsgaard*, *Marlene Beck*, *Louise Juul Hansen* for always helping me trough practicalities, never being impatient, and always smiling; and all others at CBS for a really great working atmosphere.

Benedicte and I have developed a perfect balance between science and other aspects in life. I am deeply thankful to her for her beautiful way of being.

[1] The Institute of Preventive Medicine contributed with one third of the funding.

[2] Mail fra 5.12.2010

Table of contents

Preface i
Table of contents iii
Summary iv
Dansk resumé v
Papers included in the Thesis vii
Related papers not included in the Thesis viii

1 Introductory remarks **1**

2 Human genetic variation and genome-wide association studies **3**
2.1 Human genetic variation 3
2.2 Genome-wide association analysis 4
2.3 Pros and cons of genome-wide association studies 8
2.4 Probing rare genetic variation 10

3 Approaches for integrative analyses of genome-wide association data **11**
3.1 Motivation for integrative systems biology approaches 11
3.2 Integrative approaches for genome-wide association data analysis . . . 12
3.3 Paper I - A method for protein complex-based risk gene mapping . . 32
3.4 Paper II - A method for evidence layer-based risk gene mapping . . . 53

4 Integrative analyses of genetic variation in obesity **69**
4.1 Common genetic variation in obesity 69
4.2 Studies on the biology of the FTO gene 70
4.3 Integrative analyses of body-mass index 76
4.4 Paper III - The ASIP gene's putative association with extreme overweight 77

5 Integrative analyses based on metabolic network reconstructions **93**
5.1 Metabolic network reconstructions 93
5.2 Integration of metabolic reconstructions with gene expression data . . 95
5.3 Paper IV - A method for metabolic biomarker discovery 95
5.4 Paper V - Predicting 6-month weight maintenance 109

6 Concluding remarks **129**

Bibliography **131**

Appendices **157**

Summary

Risk-phenotypes and diseases are often caused by perturbed cellular networks, as biological processes depend on an overwhelming number of heavily intertwined components. The impact of a genetically altered gene may ripple through its molecular neighborhood instead of being confined to the gene product itself. My doctoral studies have been focused on the development of integrative approaches to identify *systemic* risk-modifying and disease-causing patterns. They have been rooted in the hypothesis that data integration of complementary data sets may yield additional etiologic insights compared to analyses conducted within a single type of data.

The first line of research presented here outlines two integrative methodologies designed to identify etiological pathways and susceptibility genes. In **Paper I**, my coworkers and I present an integrative approach that interrogates protein complexes for enrichment in incident coronary heart disease (CHD) associations from genome-wide association (GWA) data. We show that integration of a moderately powered GWA data with protein-protein interaction (PPI) data successfully identifies candidate susceptibility genes for incident CHD. In **Paper II**, we present an integrative method that combines heterogeneous data from GWA studies, PPI screens, disease similarities, linkage studies, and gene expression experiments into a multi-layered evidence network, which can be used to prioritize the protein-coding part of the genome according to a particular indication. We applied the method to bipolar disorder and type 2 diabetes, and validated it by replicating a single-nucleotide polymorphism (SNP) within a novel bipolar disorder susceptibility gene.

Next, I present the avenue of my research that has been focused on the analysis of genetic variation in obesity. In section 4.2, I outline results from our bioinformatics-based analysis of the FTO locus. Genetic variation within the FTO locus provides the hitherto strongest association between common SNPs and obesity, but the mechanisms leading to this association are still unknown. In **Paper III**, we demonstrate that body-mass index associated gene products coalesce onto distinct protein complexes, and show that these putative risk modules incriminate novel candidate obesity-susceptibility genes.

The last overall line of research presented here, provides examples on how networks of human metabolism may serve as a data integration framework for differential gene expression data. In **Paper IV**, we present a method that can be used to identify metabolically-related sets of enzymes, which exhibit modest but concordant changes in gene expression. In **Paper V**, we used that approach to identify metabolites as biomarkers for weight maintenance upon dietary-induced weight loss.

The approaches presented in this PhD Thesis provide integrative methodologies for the aggregation of multiple, functionally relevant data types. Together they represent a novel bioinformatics-based toolbox for analyses of genetic variation in human traits and disease.

The Thesis is structured as follows. Chapter 1 presents a few introductory remarks to integrative systems biology, and Chapter 2 gives a brief description of human genetic variation and GWA analysis. Chapters 3-5 present the main topics in the Thesis (integrative methodologies for the analysis of GWA data, integrative analyses of genetic variation in obesity, and integrative analyses based on metabolic networks). Chapter 6 summarizes the Thesis with a few concluding remarks.

Dansk resumé

Komplekse sygdomme og lidelser er i mange tilfælde forårsaget af dysfunktion i de underliggende netværk af cellulære komponenter. Da kroppens celler består af et stort antal indbyrdes forbundne biologiske processer, vil effekterne af fejl i kodningen af et givent gen sjældent begrænses til de proteiner, genet koder for, men derimod spredes til de andre cellulære komponenter, som disse proteiner interagerer med. Mit PhD-studie har været fokuseret på at udvikle dataintegrationsbaserede metoder, der kan identificere sådanne fejl-regulerede cellulære netværk. Min hypotese har været, at integration af komplementære datatyper øger muligheden for at identificere de kausale netværk af faktorer, som fører til sygdom.

I min afhandling vil jeg starte med at præsentere to overordnede dataintegrationsbaserede metoder, der ved hjælp af data fra genetiske associationsstudier og andre relevante datatyper kan identificere kausale gener i komplekse sygdomme. I **Paper I** præsenterer vi en metode, som kan opspore proteinkomplekser, hvis underliggende genvarianter associerer med hjertekarsygdom. Vi har identificeret et specifikt proteinkompleks, hvis underliggende gener associerer med individers risiko for at udvikle hjertesygdommen. I **Paper II** præsenterer vi en metode, der integrerer data fra en bred vifte af sygdoms-specifikke datatyper, såsom genetiske ensartetheder mellem sygdomme, genetiske associationsstudier og genekspressionsstudier. Vi benytter metoden til at lave analyser af genetisk variation i type 2 diabetes og maniodepressivitet. Vi finder, at specifikke genvarianter af YWHAH-genet medfører en øgning i risikoen for at udvikle sidstnævnte lidelse.

Dernæst præsenterer jeg resultater af den del af mit PhD-forløb, som har været fokuseret på analysen af genetisk variation i fedme. I afsnit 4.2 skitserer jeg resultater af vores bioinformatiske analyse af FTO genet. Analysen har været motiveret af, at denne region udgør den hidtil mest signifikante association med fedme – dog er de specifikke mekanismer, som øger genvariantanlægsbærernes risiko for fedme, stadig ukendte. I **Paper III** viser vi, at proteiner fra kendte fedmegener er beriget i specifikke proteinkomplekser og tydeliggør, at kendskabet til disse proteinkomplekser er nyttigt i forhold til at finde nye genvarianter, der øger risikoen for at udvikle fedme.

Det sidste overordnede forskningsområde jeg præsenterer resultater af, beskæftiger sig med, hvordan man kan benytte rekonstruerede netværk af det menneskelige stofskifte som platform til bedre at kunne forstå komplekse fænotyper og sygdom. I **Paper IV** præsenterer vi en metode, som kan benyttes til at identificere grupper af enzymer, som udviser moderate, men samstemmende ændringer i genekspression. I **Paper V** benytter vi denne metode til at identificere metabolitter, der fungerer som biomarkører for vægtvedligeholdelse efter et vægttabsinterventionsstudie.

Metoderne, som jeg præsenterer i denne afhandling, illustrerer, hvordan genetiske datasæt med succes kan integreres med anden sygdomsspecifik evidens. Fremgangsmådernes fleksible karakter gør dem til nyttige redskaber i fremtidens analyser af komplekse fænotyper og sygdomme.

Afhandlingen introduceres med indledende kommentarer om systembiologi og dataintegration (afsnit 1). I afsnit 2 gives en kort introduktion til genetisk variation og genetiske associationsstudier. Hernæst gennemgår afsnit 3-5 hovedemnerne i mit PhD-studie (dataintegrationsbaserede analyser af genetisk data, integrationsbaserede analyser af genetisk variation i fedme og integrationsbaserede analyser baseret på netværk af stofskiftet). Afsnit 6 afrunder afhandlingen ved en kort perspektivering samt konklusion.

Papers included in the Thesis

Paper I Jensen MK*, **Pers TH***, Dworzynski P, Girman C , Brunak S, and Rimm EB. Protein interaction-based genome-wide analysis of incident coronary heart disease. Circulation: Cardiovascular Genetics *(Accepted)*

Paper II **Pers TH***, Hansen NT*, Lage K, Koefoed P, Dworzynski P, Miller ML, Flint TJ, Mellerup E, Dam H, Andreassen OA, Djurovic S, Melle I, Børglum AD, Werge T, Purcell S, Ferreira MA, Kouskoumvekaki I, Workman CT, Hansen T, Mors O, Brunak S. Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genetic Epidemiology 2011; 35: 318-332

Paper III **Pers TH**, Nitsch D, Moreau Y, co-workers from the NUGENOB study, co-workers form the GOYA study, Brunak S*, Sørensen TIA.* Protein complex analysis associates SNPs in ASIP with extreme overweight. *(In preparation)*

Paper IV Zelezniak A*, **Pers TH***, Soares S*, Patti ME, Patil KR. Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. PLoS Computational Biology 2010; 6(4)

Paper V Mutch DM, **Pers TH**, Temanni MR, Pelloux V, Marquez-Quioones A, Holst C, Martinez JA, Babalis D, A van Baak M , Handjieva-Darlenska T, Walker CG, Astrup A, Saris WHM, Langin D, Viguerie N, Zucker JD, and Clément K on behalf of the Diogenes Project. A distinct adipose tissue gene expression response to caloric restriction predicts 6-month weight maintenance in obese subjects. American Journal of Clinical Nutrition *(Accepted)*

* These authors contributed equally

Related papers not included in the Thesis

Paper VI **Pers TH**, Martin FP, Verdich C, Holst C, Johansen JV, Astrup A, Polak A, Martinez JA, Rezzi S, Black EE, Saris WHM, Kochar S, Macdonald IA, Sørensen TIA, Ramadan Z. Prediction of Fat Oxidation Capacity using 1H-NMR and LC-MS Lipid Metabolomic Data combined with Phenotypic Data. Chemometrics and Intelligent Laboratory Systems 2008; 93: 34-42

Paper VII Zimmermann E, Kring SII, Berentzen TL, Holst C, **Pers TH**, Hansen T, Pedersen O, Sørensen TIA, Jess T. Fatness-associated FTO gene variant increases mortality independent of fatness - in cohorts of Danish men. PLoS One 2009; 4(2)

Paper VIII Capel F, Klimcakova E, Viguerie N, Roussel B, Vitkova M, Kovacikova M, Polak J, Kovacova Z, Galitzky J, Maoret JJ, Hanacek J, **Pers TH**, Bouloumie A, Stich V, and Langin D. Macrophages and Adipocytes in Human Obesity - Adipose Tissue Gene Expression and Insulin Sensitivity During Calorie Restriction and Weight Stabilization. Diabetes 2009; 58:1558-1567

Paper IX **Pers TH**, Albrechtsen A, Holts C, Sørensen TIA, and Gerds T. The Validation and Assessment of Machine Learning: A game of prediction from high dimensional data. PLoS One 2009; 4(8)

Paper X Dalgaard MD, Weinhold N, Edsgärd D, Silver JD, **Pers TH**, Jørgensen N, Juul A, Gerds TA, Giwercman A, Giwercman YL, Cedermark GC, Virtanen HE, Toppari J, Daugaard G, Jensen TS, Brunak S, Rajpert-De Meyts E, Skakkebæk NE, Leffers H and Gupta R. A genome-wide association study of men with symptoms of testicular dysgenesis syndrome and its network biology interpretation. *(In review, Journal of Medical Genetics)*

A brief comment on the related papers that were not included in the Thesis. These papers either constitute a basis for the papers presented in the Thesis, or they represent work that incorporated methodologies, which I developed during my doctoral studies. They are cited throughout the Thesis.

Introductory remarks 1

Throughout human history, maps have been used to help charter unfamiliar territory by representing vast amounts of information in a manageable format. The surface of our planet was extensively mapped during the 18th century, the human genome was sequenced and mapped through the last decade, and now scientists are systematically mapping variation that shapes traits and cause disease. Genome-wide association (GWA) analysis is one of the first unbiased approaches that followed the completion of the human genome sequence and led to new insights into common diseases. When I started my PhD studies in April 2008, the FTO gene locus was the only body-mass index associated locus reported by GWA analysis. Now, three years later, the number of body-mass index associated loci has grown to 32.

Rapid advancements in sequencing technologies were driving the completion of the human genome sequence. Importantly this development was paralleled by the advancement of technologies allowing large-scale quantification of molecular components such as messenger RNA, proteins, and metabolites (referred to as -omics approaches). Collectively these new technological frontiers in biology spurred a research paradigm commonly referred to as *systems biology*. In this new paradigm the term *system* is broadly and loosely defined as a set of components and their mutual relationships.

Despite the fact that most biological systems are still incompletely understood, delineation of higher level cellular organization such as protein-protein, DNA-protein, and lipid-protein interactions constantly provide more and more refined reconstructions of cellular networks - another major achievement of the recent years of research.

Biological data generation at hospitals and research institutions increases rapidly. Clinical and cohort-based system-wide data (e.g. GWA data) may be overlaid with cellular networks, as for instance protein-protein interaction (PPI) data, to build multidimensional disease models. Genome-wide (and other systems-wide) data sets are per definition unbiased towards known susceptibility loci and hence provide excellent starting points for generation of novel risk-phenotype specific hypotheses about etiological genes and pathways. Towards that end, systematic data integration is a necessity, and modeling of systems of inter-connected cellular components often is of advantage.

As we are increasing the detail of the evolving networks of molecular pathobiology, a new map is being drawn offering a first glimpse into our molecular past and clinical future.

Human genetic variation and genome-wide association studies

2

Family history is a strong indicator of human disease. Consequently inherited genetic variation is believed to play a major role in the development of human diseases and phenotypic traits in general. Decades of intense investigations including recent large-collaborative efforts such as the International HapMap Project, the large community gathering around GWA analyses, and the 1000 Genomes Project, have provided a first comprehensive view into the genetic foundation of common traits and diseases, although large parts of the genetic architecture still remains incompletely understood. Below, I describe systematic approaches that have been applied to elucidate genetic variation, and identify common susceptibility sites in the human genome.

2.1 Human genetic variation

In 2003 the HapMap Project paved the way for GWA study analysis by publishing a database containing 1.3 million human genome single nucleotide polymorphisms (SNPs) [International HapMap Consortium, 2005]. The consortium categorized most single-base pair inherited variants that are prevalent at a minor allele frequency greater than five percent.

The HapMap Project was based on the observation that linkage disequilibrium between adjacent bases in the genome allows one to select SNPs that tag larger haplotype blocks in the human genome. Linkage disequilibrium describes the non-random association of two alleles and structures genomes into haplotype blocks consisting of a number of correlated alleles within each region. Through generations, haplotype blocks are only slowly changed by *de novo* mutations and recombination events. In 2010 the HapMap Consortium published the most recent version of the database (version 3), which holds more than 3.1 million common tag SNPs [Altshuler et al., 2010] surveyed across 11 populations and 1,184 individuals.

The extent of variation uncovered in the first phase of the HapMap Project was larger than first expected [Hardy and Singleton, 2009]. It appeared that on average a polymorphic site with a minor allele frequency >1% exists every 3,000 base pairs (bps) in the human genome (in total 10 million SNPs) [International HapMap Consortium, 2003]. In 2001, when the HapMap Project was initiated the price for sequencing a

million bases in an individual was >$10, but since than prices have dropped sharply, and currently mapping of one million bases of DNA sequence costs <$0.3 (**Fig. 2.1**).[1]

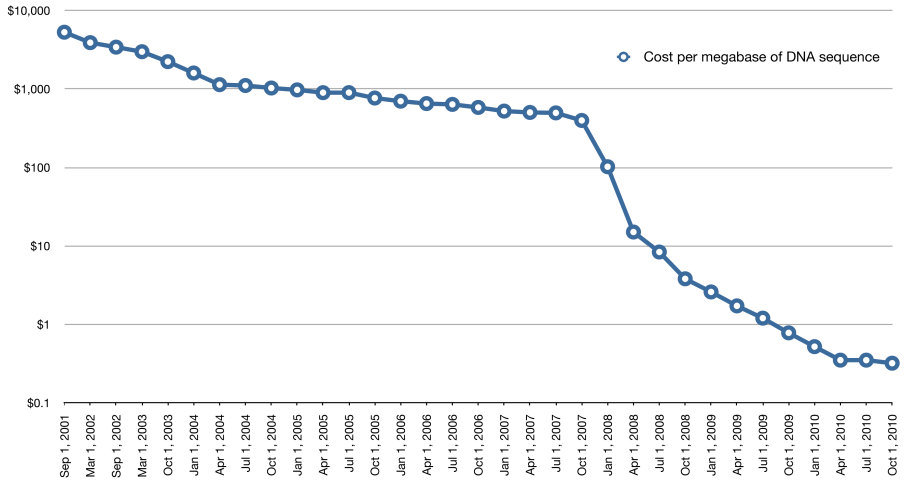


Figure 2.1: Overview of the cost per megabase (1 million bases) of DNA sequence. The cost per megabase DNA sequence has fallen rapidly since mid 2007. Note that the y-axis is on a logarithmic scale. Data source: The National Human Genome Research Institute, USA, http://www.genome.gov/sequencingcosts

The HapMap Project was partly driven by the common disease – common variant (CDCV) hypothesis [Reich and Lander, 2001], which states that common diseases are caused by a few common SNPs (> 5% minor allele frequency) with relatively large effect sizes. The CDCV hypothesis became the foundation for GWA analysis, as it suggests that it is sufficient to investigate common SNPs to understand the genetic architecture of complex traits. The CDCV hypothesis is now increasingly being challenged, as common SNPs seem to capture a relative small part of the genetic variation in traits and diseases. However, in 2007 it was rapidly adopted as heritability of complex traits in this way could be analyzed by a near-complete set of common SNPs that could fit onto a single genotyping microarray.

2.2 Genome-wide association analysis

During the last years, linkage studies and candidate gene approaches have been succeeded by GWA studies as the main approach to identify DNA variation that associates with common traits. The first GWA study was published on age-related macular degeneration in 2005 [Klein et al., 2005]. In 2007 the Wellcome Trust Case Control Consortium published a landmark paper presenting GWA studies on 7 case control studies of common diseases [Wellcome Trust Case Control Consortium, 2007], among others type 2 diabetes and bipolar disorder – data sets that we used in the integrative analyses carried out during my doctoral studies. The Wellcome Trust Case Control

[1] Currently, a complete human genome can be sequenced for <$29,000.

Consortium's study marked merely the beginning; by March 2011 546 GWA studies have reported a total of 2,420 associations.[2]

Data cleaning and population substructure adjustments

The methodology of GWA studies follows established protocols. After DNA has been extracted and hybridized to microarrays, SNPs are called using the manufacturers' or the open source community's software tools. Once alleles have been assigned to genotyped SNPs across all individuals, a range of quality control measures are used to exclude specific individuals, and to remove low-confidence SNPs. Individuals may either be excluded due to more similar than expected genetic backgrounds (in case they are siblings), due to ethnical divergence (as can be identified by principal component analysis on the genotypes), or due to a large amount of SNPs that could not be called by the genotyping software. SNPs may be excluded across all individuals in case they have call-rates below a given cut-off (typically 99%), have minor allele frequencies below a given cut-off (typically 1-5%), or are in Hardy-Weinberg disequilibria[3]. The minor allele frequency and Hardy-Weinberg equilibrium calculations are assessed in controls only, since they, in case of association, per definition will differ between cases and controls.

At least a slight degree of population substructure is expected in most cohorts. In association analysis, population substructure may lead to spurious inflation of SNPs; if some individuals originate from a genetically slightly different region and by coincidence are categorized as cases, then SNPs, which frequencies differed due to ancestry, will falsely associate with case-control status and generate false-positive associations (type 1 errors). Population substructure can be accounted for by including the first three to ten vectors from a principal component analysis done on the individuals' genotype vectors as covariates in the statistical associations model.[4] The NUGENOB Study on dietary induced weight loss – an ongoing project of mine – provides an illustrative example on population substructure as individuals were recruited from 7 centers across Europe (**Fig. 2.2**).[5]

Association analysis

In GWA analysis each SNP is tested for potential association with the given phenotype. In that respect the phenotype is either coded as a dichotomous variable (e.g. lean versus obese individuals), or as a continuous variable (e.g. the kilos of weight lost in a diet intervention study). Several different study designs do exist, but will not be reviewed in this Thesis. Most often covariates such as age, gender, and ancestry are included to adjust for non-trait related phenotypic heterogeneity in the cohort. The genotype may

[2] Data for this calculation came from the National Human Genome Research Institute GWA Study Catalog [Hindorff et al., 2009]. The reported numbers include SNPs that were reported in several publications, and SNPs that are in linkage disequilibrium. Therefore the unique number of independently associated loci will be smaller.

[3] The Hardy-Weinberg principle states that both allele and genotype frequencies remain constant in the absence of selection, drift, gene flow, mutation, non-random mating, and with an infinite population size.

[4] Note that the phenotype-causing genetic variation is expected to be subtle compared to the variation caused by different genetic backgrounds, and therefore is not expected to reside in some of the first principal components [Price et al., 2006].

[5] A finding which replicates a previous report on population substructure across Europe [Novembre et al., 2008, Lao et al., 2008].

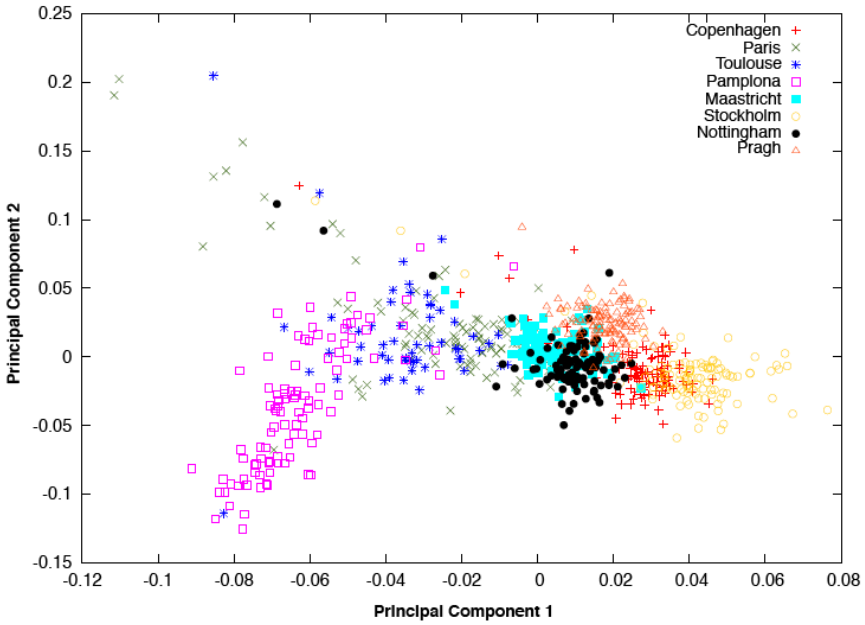


Figure 2.2: Population substructure in the NUGENOB cohort [Sørensen et al., 2006]. The first two principal components of a principal component analysis were plotted against each other. By correlating the longitude and latitude of the European centers we found that latitude explained 66% and longitude explained 48% of the variation in the first principal component. Including both latitude and longitude explained 69% of the variation in the first principal component.

be coded as an additive, recessive, dominant model, or by use of a indicator variable whereby no specific genetic model is assumed.[6]

Multiple testing correction and replication

As the same null-hypothesis is tested for each SNP in a GWA study (SNP x is not associated with the trait), multiple testing correction needs to be applied to adjust for spurious correlations that are expected to be incurred due to the large number of repetitive tests. Therefore, multiple testing procedures such as Bonferroni correction are used to calculate a threshold for when to deem a given SNP as significant.[7] However, Bonferroni correction is an overly conservative framework since SNPs are correlated, and the *effective* number of SNPs and statistical tests employed therefore is smaller than the observed number of SNPs. During my doctoral studies, I have implemented a method that calculates the number of effective tests, which is presented in section 3.2.

To increase the fraction of genuine associations identified, GWA studies often involve a discovery and replication phase. In the initial discovery phase, GWA analysis is

[6] For more information on statistical considerations on GWA analysis please refer to [Wang et al., 2005].

[7] For instance only SNPs with a p-value less than 0.05 / 10^8 = 5x10^{-8} are called significant when 1 million SNPs are tested on a Affymetrix 6.0 microarray with 1 million SNPs.

used to identify associations. In the replication phase promising top SNPs from the discovery analysis are genotyped in independent cohorts. Finally, SNPs are meta-analyzed across the discovery- and replication cohorts, and any SNPs with p-values below the Bonferroni threshold are trusted as *bona fide* associations.[8]

Statistical power considerations

There is an import trade-off between decreasing the cost of a given study (minimizing the study cohort size) and increasing the statistical power to detect causal variants. Among the factors that demand an increase in the cohort size are:

- Chance-correlations
- Genotyping errors
- Phenotype misclassification
- Low effect sizes
- Low risk-allele frequencies
- Low linkage disequilibrium between the tag SNP and the true causal variant
- Genetic heterogeneity caused by non-additive interactions between genes or gene-environment interactions.

Ideally the researcher estimates the different factors prior to the GWA study to determine the optimal cohort size. However, it will not always be possible since large-scale recruitment may prove easier for some traits and more difficult for others. For instance body-mass index and height are more feasible to measure than weight loss in a clinical intervention study or coronary heart disease incidences in prospective studies.

Imputation

In imputation haplotypic information from for instance the HapMap database is used to infer alleles of non-genotyped SNPs residing on the same haplotype block as genotyped SNPs. Imputation increases power and, importantly, enables meta-analysis across different experimental platforms and versions of the same microarray.[9]

Single nucleotide polymorphisms impact on biology

For most associations found in GWA studies, the associated SNPs are not the causative variants [Robinson, 2010]. Instead, they act as signposts for the real variants, by constituting surrogate markers tagging the real functional polymorphisms. Consequently, re-sequencing (also referred to as fine-mapping) and functional studies, are needed to find the causal variants. Several explanations on how alleviating or aggravating variants may impact downstream biology are imaginable: deregulation of the nearest gene's (or more distal gene's) gene-product; alteration of protein function caused by changes in

[8] Please refer to [Ioannidis et al., 2009] for a review.

[9] Imputation will not be described further in this Thesis. For more information please refer to [de Bakker et al., 2008].

protein structure due to altered amino acid sequences (missense or nonsense mutations); or other functional complications such as small in-frame insertions or deletions, or small frameshift insertions or deletions. While amino acid changing effects are supposed to be more severe, regulatory perturbations are suggested to be less severe but more widespread [Kasowski et al., 2010]. In a recent study, 88% of the previously identified trait-associated variants were reported to be within intronic regions, and to be significantly underrepresented in intergenic regions [Hindorff et al., 2009]. In addition, trait-associated SNPs were found to be overrepresented in non-synonymous variants and 5 kilo bases (kb) promoter regions. However, merely 12% of trait-associated SNPs indentified by GWA studies are in strong linkage disequilibrium with protein-coding regions [Manolio, 2010]. The latter has been proposed as one of the reasons why only minor fractions of the heritability of complex traits is accounted for by GWA analyses. However, several other explanations for this 'missing heritability'[10] have been proposed and are discussed in the next section.

2.3 Pros and cons of genome-wide association studies

The median per allele odds ratio associated SNPs found by GWA analysis is 1.33 [Manolio, 2010], a finding emphasizing that GWA studies mostly have identified common low-risk alleles. Additionally, for most complex traits, common SNP associations account for at most 10% of the genetic variation [Frazer et al., 2009] (exceptions are age-related macular degeneration, Crohn's disease, and several endophenotypes such as lipid levels and metabolite concentrations [Manolio et al., 2009, Teslovich et al., 2010, Illig et al., 2010]). This fraction might increase as causal SNPs still need to be identified for most of the established associations, and current effect size estimates hence may be underestimated [Ku et al., 2010]. Concerns about the absence of high-risk alleles and the low amount of genetic variability accounted for has been expressed as a criticism to GWA studies and its key CDCV hypothesis. The CDCV hypothesis is challenged by the multiple rare variant hypothesis, which states that a given disease is caused by numerous variants that need not to be the same among individuals with the particular disease. Rare variants have not been analysed in GWA studies, as current genotyping microarrays have been densely filled with common SNPs, and since there generally is low statistical power to identify rare SNPs by association analysis. Rare variants and the multiple rare variant hypothesis are the currently most widely accepted single explanation for the relatively low explanatory power of GWA findings [Cirulli and Goldstein, 2010]. However, there are other possible reasons, the most pronounced being:

Epistasis between genetic variants, i.e. non-additive interactions between to alleles, are known to confer large effects on phenotypes [Cordell, 2009]. However, due to the large number of possible SNP-SNP combinations they are difficult to identify.

Parent-of-origin effects may influence risk for disease. A recent study found that allelic relative risk differed based on whether the risk allele was inherited from the paternal or maternal line [Kong et al., 2009].

Copy number variants have been shown to play significant roles in mental disorders such as schizophrenia [Stefansson et al., 2008, McCarthy et al., 2009]. However, a recent study by the Wellcome Trust Case Control Consortium showed that *common* copy number alterations most likely will not contribute significantly

[10] Coined by Maher in [Maher, 2008].

to the majority of complex traits (of the copy number variants that associated strongly with disease, p-value < 10^{-8}, only 13% had odds ratios above 2 and 1% had odds ratios above 10) [Wellcome Trust Case Control Consortium et al., 2010].

Gene-environment interactions are not systematically accounted for in current GWA studies [Thomas, 2010]. A related issue is the hitherto almost entirely unknown interaction between genetic diversity in our gut bacteria and human genetic variation, and its impact on metabolism [Qin et al., 2010].

Epigenetics, *viz.* the passing on of transgenerational effects that are not manifested through variation in the DNA sequence, but rather in the way how the DNA is condensed, is another mechanism that is likely to impact heritability of traits [Petronis, 2010].

Future research will show whether any of the above avenues will add significantly to the inherited risk for complex diseases. Currently, they are used as alternative explanations, and in some cases, arguments against analyses of common variation. In the GWA research field, the major hope was that findings could be used for genetic counseling. However, the relatively low effect sizes of causal variants have moved that goal into a unforeseeable future. Let me illustrate that by a small anecdote. It turns out that I am a heterozygous carrier of the FTO rs1421085 risk variant C (passed to me through my maternal line), which by Dina *et al* was found to be associated with a significantly elevated risk of being obese [Dina et al., 2007]. In their work, they report a per allele odds ratio of 1.56 (95% confidence interval: 1.40-1.75), which is relatively high for a common variant association[11]. In my case, a per allele odds ratio of 1.56 can be translated into a 56% higher risk of being obese compared to the cohort without the variant. Obviously, this value is too small to be useful for genetic counseling and more genetic and lifestyle information is needed to predict my risk of becoming obese.

However, it is important to keep in mind that relatively small effect sizes do not preclude biological insight as exemplified by a recent GWA analysis for body-mass index in which SNPs with per allele effect sizes for the rs29941 SNP as low as 0.06 body-mass index points (kg/m^2) were identified through GWA meta-analysis [Speliotes et al., 2010].[12] One of the major insights drawn from GWA analyses is that common variation significantly contributes to complex traits (albeit with low effect sizes in some cases). Despite that the fraction of the overall trait-variability remains relatively low, much has been learned as to which genes and pathways are implicated in shaping particular traits and diseases.[13] One such example is body-mass index and the central hypothalamus pathways as outlined in section 4.1 on page 69. There are odd examples, too, one of them being that the KITLG gene associates with testicular carcinoma on the one hand [Rapley et al., 2009], and hair color on the other [Sulem et al., 2007], or findings showing that the ORMDL3 gene associates with both childhood asthma [Moffatt et al., 2007] and Crohn's disease [Barrett et al., 2008].

Another biological insight from GWA studies is that the majority of associated SNPs affect adjacent genes instead of having trans-effects of more distal genes [Heid

[11] The minor allele frequency of the rs1421085 SNP is 46% in the European population.

[12] I have inherited two copies of the risk allele (G) of that locus. Given that 'worst-case scenario' these findings imply that 2 * 217 = 434 grams of my 77kg body weight theoretically results from that variant.

[13] Even though the causal relationship as to which gene is affected by a particular SNP remains unclear in almost all cases, and even though the mechanisms by which SNPs perturb gene expression and or protein function still have to be found.

et al., 2010]. Another unexpected and challenging finding is that gene-poor areas are harboring associations to traits [Manolio et al., 2008], one such example being the to-date strongest association for myocardial infarction heart disease and SNPs in the 9p21 region with 150kb to the nearest protein-coding genes [Mcpherson et al., 2007] or a SNP on 8q24 that associates with colorectal and prostate cancer but is >300 kb away from the closest gene (MYC) [Pomerantz et al., 2009].

In summary, GWA studies have proven useful in systematically identifying common low-risk variants, which for most traits account for at most 10% of the genetic variation. At the same time GWA studies have provided novel insight as to which pathways are underlying a number of complex traits. Apart from DNA sequencing, which is briefly discussed in the following section, integrative systems biology analyses of multiple trait-specific evidence sources may augment genetic analyses as discussed in the remainder of the Thesis.

2.4 Probing rare genetic variation

In 2010, the 1000 Genomes Project published a detailed catalog detailing more than 95% of all SNPs with >1% minor allele frequency in the non-coding part of the human genome, and SNPs with minor allele frequencies down to >0.1% minor allele frequencies in coding regions [1000 Genomes Project Consortium et al., 2010]. By sequencing 179 human genomes selected from each of the five major population groups (Europe, Americas, East Asia, West Asia and West Africa), the consortium reported 8 million previously unknown SNPs. The 1000 Genomes Database on genetic variation is now being used to impute existing GWA studies, to facilitate the simultaneous analysis of common and rare variants (> 0.1% minor allele frequency). However, re-sequencing is still needed to identify variants with minor allele frequencies below the ones cataloged by the 1000 Genomes Project. Sequencing of the diploid genome of Craig Venter showed that there might be around 3.2 million SNPs within a single genome [Levy et al., 2007] and recent exome sequencing studies have shown that 17,000, or 0.5%, of these may reside within coding sequences [Ng et al., 2009]. The increased number of rare SNPs to be analyzed will require novel analysis techniques that alleviate the multiple testing problem, for instance by use of bioinformatics approaches as discussed in the concluding remarks (Chapter 6). To summarize, the question whether rare variants contribute significantly to common diseases presents literately an open chapter.

Approaches for integrative analyses of genome-wide association data

3

Decades of research in molecular biology have exposed layer upon layer of molecular complexity. Complex networks glue these layers together, and sub-networks within these shape traits, and, in some cases, disease. Systems biology approaches are being developed to identify these non-obvious systemic patterns. In the following, I will discuss (1) the motivation for integrative systems biology approaches, (2) general integrative methodologies for the analysis of GWA data, (3) issues concerning scoring of genes based on GWA data, and (4) definition and delineation of biological pathways.

3.1 Motivation for integrative systems biology approaches

Systems-based approaches are needed to capture the molecular dynamics driving the complex development of traits and deregulation of disease at the pathway-level. Cellular components such as genes, RNAs, proteins and metabolites are connected through intertwined pathways forming complex networks. Genetic alteration most often will not be restricted to the directly affected gene product, but may ripple through the gene product's physical neighborhood [Barabási et al., 2011]. As a result the origin and development of traits are believed to arise from networks of genes, proteins and/or metabolites, which may be hard to identify through non-systemic approaches.

Systems-based approaches may yield more robust results. Within a given phenotype-causing pathway, the pathogenic components might differ between individuals or cohorts [Cantor et al., 2010]. When different components of the *same* etiological pathway are perturbed between individuals or between cohorts, significance testing at the pathway-level may yield more sensitive and robust results.[1]

Systems-based approaches are more sensitive to modest but coordinated associations with the trait. Biologically significant genes do not necessarily harbor SNPs with large effect sizes. Consequently, GWA studies have identified far from all susceptibility loci, especially for traits for which large-collaborative meta-analyses have been difficult to accomplish. Based on molecular network reconstructions, systems-based approaches can help to detect central etiological pathways that exhibit enrichment in SNPs with

[1] Please refer to Chuan *et al* for an example [Chuang et al., 2007].

modest, but consistently larger than under the null hypothesis expected effect sizes [Wang et al., 2009].

Systems-based approaches preserve the desired unbiased nature of large-scale data sets. Most often integrative system biology approaches are data-driven with no *a priori* assumptions as to which components are involved in pathogenesis.

Integration-based approaches are more likely to capture pathobiological processes. Complex traits may be caused by several pathobiological processes across distinct biological domains, which are not captured by a single type of technology. Data integration across complementary disease-specific evidence sources increases the chance of identifying patterns that predispose to disease.

The major hypothesis underlying integrative systems biology approaches is that heterogenic layers of evidence for a particular trait will coalesce on a few molecular pathways. In the following, I will briefly outline integrative approaches that rely on integration of both trait-specific evidence sources and pathway organization data to leverage GWA analyses.

3.2 Integrative approaches for genome-wide association data analysis

Common to most integrative analyses is that they, to some extent, rely on re-analysis of already available data sets. This trend is increasing as data generation is growing faster than the manpower available for data analysis, as mentioned in a recent editorial in Nature Genetics [Editorial, 2010]. In the following sections, I will give some examples on integrative methods. Many of these methods rely on *a priory* defined biological information from databases. Databases most commonly used for retrieval of information on functionally-related gene sets, PPIs, and metabolic networks are listed in **Table 3.1**. Databases that provide valuable phenotype-specific information are listed in **Table 3.2**. Note that the following enumerations are not exhaustive as the field is growing rapidly. Also, please be aware that the methods and databases described here, solely focus on intracellular molecular networks. The research on molecular networks that glue together cells, tissues, and organs still is in its very infancy and therefore not a subject in this Thesis. For reviews on the challenges and limitations in pathway-based approaches please refer to [Cantor et al., 2010, Elbers et al., 2009].

Database	Information	Website	Reference	Comment
Biocarta	Pathways	www.biocarta.com	-	The database comprises a total of 350 canonical pathways.
Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway	Pathways	www.genome.jp/kegg/pathway.html	[Ogata et al., 1999]	The database comprises a large collection of biological pathways annotated through literature-based searches.
Protein Analysis Through Evolutionary Relationships (PANTHER) Pathway	Pathways	www.pantherdb.org/pathway	[Thomas et al., 2003]	The database comprises >165 signaling pathways.
Reactome	Pathways	www.reactome.org	-	The database contains 1,116 pathways.
Molecular Signatures Database (MSigDB)	Pathways and gene sets	www.broadinstitute.org/gsea/msigdb	[Subramanian et al., 2005]	The database contains co-expressed gene sets across a large list of traits and conditions (3,272 curated gene sets, and 880 canonical pathways).
Gene Ontology (GO)	Gene sets	www.geneontology.org	[Ashburner et al., 2000]	The database contains a controlled vocabulary for gene product annotations. For instance the molecular function ontology can be used to extract all genes annotated with a specific function.
Homo sapiens Reconstruction 1 (Recon 1)	Metabolic reactions	http://bigg.ucsd.edu	[Duarte et al., 2007]	A high-confidence network of human metabolism. See Table 5.1.

Continued on next page…

Table 3.1 -- Continued

Edinburgh Human Metabolic Network (EHMN)	Metabolic reactions	www.ehmn.bioinformatics.ed.ac.uk	[Ma et al., 2007]	A high-confidence network of human metabolism. See Table 5.1.
InWeb	PPIs	-	[Lage et al., 2007]	A PPI meta-database comprising several other PPI databses: BIND [Bader et al., 2001], BioGRID [Stark et al., 2006], CORUM [Ruepp et al., 2008], DIP [Salwinski et al., 2004], IntAct [Hermjakob et al., 2004], HPRD [Peri et al., 2003], MINT [Chatr-aryamontri et al., 2007], MPact [Güldener et al., 2006], MPPI [Pagel et al., 2005] and OPHID [Brown and Jurisica, 2005].
Interaction Reference Index database (iRefWeb)	PPIs	http://wodaklab.org/iRefWeb	[Turner et al., 2010]	A PPI meta-database integrating data from 10 predominantly experimental databases (BIND, BIND_TRANSLATION, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MPact, MPPI, and OPHID).
PINA	PPIs	http://csbi.ltdk.helsinki.fi/pina	[Wu et al., 2008]	A PPI meta-database integrating data from 7 experimental databases (IntAct, MINT, BioGRID, DIP, HPRD, MIPS/MPact).

Continued on next page…

Table 3.1 -- Continued

STRING	PPIs	http://string-db.org	[Szklarczyk et al., 2011]	The database contains experimentally derived and predicted PPIs. The latter are referred to as indirect PPIs as they do not need to denote physical interactions.
ConsensusPathDB	PPIs, metabolic, signaling and regulatory networks	http://cpdb.molgen.mpg.de	[Kamburov et al., 2011]	A meta-database comprising interaction data from 20 public interaction databases and literature mining.

Table 3.1: Overview of databases that provide information on pathways, gene sets, protein-protein interactions (PPIs), metabolic reactions, and signaling networks.

Database	Information	Website	Reference	Comment
NCBI *Gene Expression Omnibus* (GeO)	Gene expression datasets	www.ncbi.nlm.nih.gov/geo	-	Gene expression experiments in standardized formats.
The European Bioinformatics Institute *ArrayExpress* database	Gene expression datasets	www.ebi.ac.uk/arrayexpress	-	Gene expression experiments in standardized formats. Overlapping with GeO.
The Human Protein Atlas	Protein expression	www.proteinatlas.org	[Uhlén et al., 2005]	Protein expression for the gene products of 10,118 genes across various tissues and cell lines.
NCBI *database for Genotypes and Phenotypes* (dbGaP)	Genotypic datasets	www.ncbi.nlm.nih.gov/gap	[Mailman et al., 2007]	A comprehensive archive of genotype data from GWA and sequencing studies.
European Genome-phenome Archive (EGA)	Genotypic datasets	www.ebi.ac.uk/ega	-	SNP and copy-number variation genotype data from GWA studies and genotyping done with re-sequencing methods.
National Genome Research Institute (NHGRI) *Catalog of Published GWA Studies*	SNP-phenotype associations	www.genome.gov/GWAStudies	[Hindorff et al., 2009]	GWA study associations from 819 publications comprising 4,008 SNPs.[a]
Genetic Association Database (GAD)	Gene-phenotype relationships	http://geneticassociationdb.nih.gov	[Becker et al., 2004]	Associations from GWA studies and candidate gene approaches.

Continued on next page…

Table 3.2 -- Continued

NCBI *Online Mendelian Inheritance in Man* (OMIM)	Gene-phenotype relationships	www.ncbi.nlm.nih.gov/omim	-	Manually curated archive of inherited Mendelian disorders and their associated genes.
GeneCards	Gene-phenotype relationships	www.genecards.org	[Safran et al., 2010]	The *GeneCards* category *Disorders & Mutations* reports associations between genes and disease keywords based on text-mining of PubMed abstracts.
Human Genome Epidemiology Network (HuGE)	Gene-phenotype relationships	http://hugenavigator.net/HuGENavigator	[Yu et al., 2008]	The HuGE *Phenopedia* web tool facilitates lookup of associations for a particular trait or disease. The HuGE *Genopedia* web tool facilitates lookup of associations for a particular gene.

Table 3.2: Overview of databases that provide trait and disease-specific evidences sources, such as gene expression, protein expression, SNP-phenotype associations, and gene-phenotype relationships. Information from these databases may serve as phenotype-specific evidence layers in integrative approaches as they provide useful information on genes' and proteins' expression levels in various tissues and under a plethora of conditions, and summarize the accumulated disease-specific evidence for a particular gene or locus. Abbreviations: NCBI, National Center for Biotechnology Information.
[a] Accessed March 28 2011

Approaches relying on pre-defined pathways

Many methods use pre-annotated gene sets to search for pathways that are enriched in gene-products with SNPs exhibiting associations with the trait of interest (**Tab.** 3.3). The first and most highly cited method is a SNP set enrichment analysis approach developed by Wang *et al* in 2007 [Wang et al., 2007]. Since then, several similar methods have been published, among others the MAGENTA method [Segrè et al., 2010]. Similar to the Wang *et al* method, MAGENTA is based on the algorithm used in traditional gene set enrichment analysis (GSEA) of gene expression data [Subramanian et al., 2005]. In short, the GSEA framework uses a ranksum statistic to assess whether predefined gene sets are enriched in the top or bottom of a sorted list of genes based on differential expression or fold change in gene expression analysis, or SNPs' significance in GWA studies.

An alternative approach is to assess the enrichment of pre-annotated pathways based on *z*-scores. The original approach was presented by Ideker *et al* in 2002 [Ideker et al., 2002], and we used this approach to score protein complexes for their enrichment in GWA associations, as presented in **Paper I** on page 32.

The advantage of these approaches is that specific canonical pathways can be assessed for enrichment of a given exposure as for instance SNPs. Disadvantages are that in many cases, pathways have not been delineated and thus are not within databases, and that pre-defined gene sets in many cases are computationally derived and thereby more likely to be error prone. For instance 95% of all gene annotations in the Gene Ontology database (GO)[2] [Ashburner et al., 2000] is computationally derived and the proportion of genes that have a least one experimental annotation is as low as 1% for some organisms [Yon Rhee et al., 2008]. Thus, this type of approach strongly depends on the specific gene sets tested for enrichment.

[2] http://www.geneontology.org

Method	Integrated data types	Software or web tool	Reference	Comment
GRAIL	GWA data and text-mining	www.broadinstitute.org/mpg/grail (Web tool)	[Raychaudhuri et al., 2009]	The method identifies pairs of SNPs, within genes that are co-mentioned in PubMed abstracts.
INTERSNP	GWA data and KEGG pathways	http://intersnp.meb.unibonn.de (*C/C++* command line tool)	[Herold et al., 2009]	The method searches for SNP-SNP interactions and confines its search space by use of predefined biological information.
ALIGATOR	GWA and GO gene sets	http://x004.psycm.uwcm.ac.uk/~peter (Fortran command line tool)	[Holmans et al., 2009]	The method identifies GO gene sets that are enriched in GWA signal.
SNP ratio test	GWA data and KEGG pathways	https://sourceforge.net/projects/snpratiotest (Perl command line tool)	[O'Dushlaine et al., 2009]	The method assesses overrepresentation of GWA signal within pre-specified pathways.
GSEA-SNP	GWA data and MSigDB gene sets	http://www.nr.no/pages/samba/area_emr_smbi_gseasnp (*R package*[b])	[Holden et al., 2008]	The method uses the GSEA algorithm to assess enrichment of GWA signal in pre-defined gene sets.
GeSBAP	GWA data, GO gene sets, KEGG and Biocarta pathways	http://bioinfo.cipf.es/gesbap (Web tool)	[Medina et al., 2009]	The method assesses predefined gene sets for enrichment in GWA signal.
GSEA for SNPs	GWA and PPI data	www.openbioinformatics.org/gengen (Perl command line tool)	[Wang et al., 2007]	The method uses the GSEA algorithm to assess enrichment of GWA associations in predefined gene sets.

Continued on next page…

Table 3.3 -- Continued

MAGENTA	GWA data and pathways from KEGG, Ingenuity[a], PANTHER, Reactome, BioCarta and GO	www.broadinstitute.org/mpg/magenta (Matlab scripts) [Segrè et al., 2010]	The method uses the GSEA algorithm to assess enrichment of GWA associations in predefined gene sets.

Table 3.3: Integrative approaches based on GWA data and pre-defined pathways. Common to these methods is that they assess significance of GWA signal based on pre-defined gene sets. Some of them use PPI data to reduce the search space in which to assess SNP-SNP interactions. Only methods that provide publicly available tools are listed.
[a] www.ingenuity.com;
[b] www.r-project.org

Network-search approaches

Another set of approaches circumvent the reliance on canonical pathways and instead search through gene networks to identify sub-networks enriched for a given trait (**Tab. 3.4**) - a general methodology pioneered by Ideker *et al* [Ideker et al., 2002]. These approaches are based on cellular networks as for instance predicted PPIs from the STRING database [Szklarczyk et al., 2011], or experimental PPI meta-databases like iRefWeb or InWeb [Lage et al., 2007]. Each gene in the network (represented as a network node) is for instance scored based on the GWA data, and a heuristic algorithm is used to search for sub-networks that are enriched in genes with better than expected scores. The advantage of these methods is that they do not require any assumptions about pathway delineation. A disadvantage is that oftentimes large sub-networks are found to be enriched, a critical point as it complicates interpretation and validation. In addition, no optimal solution is guaranteed as the search problem is known to be NP-hard, which means that it cannot be solved in polynomial time. The first network-search approach to be applied on GWA data was published by Torkamani *et al* in 2008 [Torkamani et al., 2008].

Analysis type	Integrated data types	Software or web tool	Reference	Comment
Analysis of schizophrenia	GWA and PPI data	*jActiveModules* [Ideker et al., 2002] (*Cytoscape*[a] plugin)	[Baranzini et al., 2009]	The authors used the *Cytoscape* plugin *jActiveModules* to search for PPI sub-networks that were enriched in GWA signal.
dmGWAS	GWA and PPI data	http://bioinfo.mc.vanderbilt.edu/dmGWAS.html (*R* package)	[Jia et al., 2011]	The method identifies PPI sub-networks that are enriched in GWA associations.
Analyses of 4 common diseases	GWA and PPI data	http://www.daimi.au.dk/~memily/BiRC/Software.html (*C/C++* command line tool)	[Emily et al., 2009]	The method identifies SNP-SNP interactions by constraining its search space to pairs of SNPs that reside in genes, which gene-products are characterized by mutual physical interaction.
Analyses of 7 common diseases	GWA and PPI data	*jActiveModules* [Ideker et al., 2002] (*Cytoscape* plugin)	[Torkamani et al., 2008]	The authors used the *Cytoscape* plugin *jActiveModules* to search for PPI sub-networks that were enriched in GWA signal.

Table 3.4: Network-based approaches for the analysis of GWA data. Common to these methods is that they do not rely on any *a priori* defined gene sets, but rather search through PPI networks to identify sub-networks that are enriched in GWA signal. Only methods that provide or were based on publicly available tools are listed. For more information on the specific databases and evidence layers, please refer to Tables 3.1 and 3.2.
[a] [Shannon et al., 2003], www.cytoscape.org

Layer-based approaches

A third type of approaches is not using pre-defined biological contexts for integration (**Tab.** 3.5). Instead, these approaches rely on trait-specific evidence sources to produce genome-wide evidence layers, which are summarized to a single meta rank that prioritizes genes' significance across all evidence layers. These methods assume that the causal genes are within the top percentiles of a significant number of evidences layers (not necessarily all), and thereby will show up in the top of the final meta rank. Examples of these types of methods are CANDID [Hutz et al., 2008] and MetaRanker (**Paper II** p. 53). An advantage of these methods is that they are flexible and allow integration of several heterogeneous data types. Disadvantages are that they require that evidence layers are indeed phenotype-specific, since otherwise, the signal expected be emerge across layers will be diluted and the causal genes will not appear in the top of the final meta rank.

All of the approaches outlined in **Tables** 3.3-3.5 rely on a SNP to gene mapping step, a gene scoring step, and some sort of pathway representation. The former two points can be accomplished in several ways, and the latter likewise requires likewise careful considerations, as outlined in the following sections.

Method	Integrated data types	Software or web tool	Reference	Comment
Biofilter	GWA data; structural information from the *Protein Families Database* (PFAM)[a]; GO gene sets; DIP, *Netpath*[b], KEGG, and Reactome pathways	http://chgr.mc.vanderbilt.edu/ritchielab/method.php (Pre-processed data files containing models)	[Bush et al., 2009]	The method detects epistatic SNP-SNP interactions by confining the search space based on various evidence sources.
SPOT	GWA data, genes specified by the user, and *PolyPhen*[c] predictions	https://spot.cgsmd.isi.edu (Web tool)	[Saccone et al., 2010]	The method prioritizes SNPs according to their occurence in user-specified genes and foretold functional effects based on *PolyPhen* predictions.
Path	GWA data, OMIM, KEGG, PharmKB[d], GAD, and the Innate Immune Database (IIDB)[e]	http://genapha.icapture.ubc.ca/PathTutorial (Java software)	[Zamar et al., 2009]	The method identifies SNP-SNP interactions by use of the evidence layer to constrain the search space.
CANDID	GWA data, text-mining, protein domains, sequence conservation, gene expression, PPI data, linkage data, and custom user-specified data	https://dsgweb.wustl.edu/hutz/candid.html (Web tool)	[Hutz et al., 2008]	The method prioritizes all human protein-coding genes based on user-specified combinations of evidence layers. The user can assign weights to layers.

Continued on next page…

Table 3.5 -- Continued

MetaRanker	GWA data, PPI data, user-specified disease genes, linkage data, expression data, and phenotype-similarities established by text-mining of the GeneCards database	www.cbs.dtu.dk/services/metaranker (Web tool)	[Pers et al., 2011] (**Paper II**)	The method prioritizes all human protein-coding genes based on user-specified combinations of evidence layers. All layers are treated on an equal footing.

Table 3.5: Layer-based approaches based on multiple (>2) evidence layers of different types. Only methods that provide publicly available tools are listed. For more information on the specific databases and evidence layers, please refer to Tables 3.1 and 3.2. [a] [Finn et al., 2008], http://pfam.sanger.ac.uk;
[b] [Kandasamy et al., 2010], www.netpath.org;
[c] [Ramensky et al., 2002], http://genetics.bwh.harvard.edu/pph;
[d] [Klein et al., 2001], http://www.pharmgkb.org;
[e] http://db.systemsbiology.net/cgi-bin/GLUE/U54/IIDBHome.cgi

SNP to gene mapping

Findings from GWA studies have shown that most trait-associated SNPs perturb regulatory mechanisms that impact transcriptional or translational efficiency [Hardy and Singleton, 2009].[3] Especially variants controlling expression levels of adjacent genes are found to be overrepresented among trait-associated loci [Nica et al., 2010, Nicolae et al., 2010, Allen et al., 2010] (also referred to as in-*cis* expression quantitative trait loci, as opposed to in-*trans* expression quantitative trait loci, which denote variants that control distal genes' expression levels)[4]. Whereas this evidence often originates from analyses carried out in cell lines, a recent meta-analyses on waist-hip ratio provided preliminary evidence from population-based cohorts [Heid et al., 2010]. The authors reported that several of the waist-hip ratio associated SNPs were significantly associated with expression levels of adjacent genes across a range of relevant tissues.

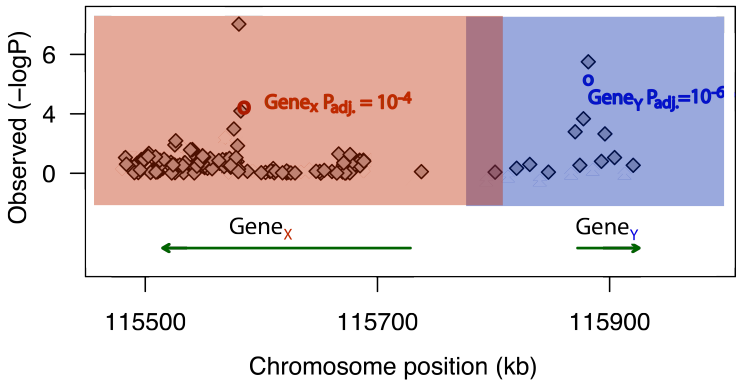


Figure 3.1: Mapping of SNPs to genes and subsequent scoring of genes. The red and blue squares mark all SNPs (diamonds) that have been mapped to Gene$_X$ and Gene$_Y$ (green arrows), respectively. Note that SNPs are allowed to map to more than one gene. A given gene is scored by assigning the lowest p-value of all SNPs mapped to it as its gene p-value. Subsequently, this p-value is adjusted for the number of independent SNPs mapped to the gene to yield an adjusted gene p-value (denoted by red and blue circles for Gene$_X$ and Gene$_Y$, respectively. Abbreviations: logP, logarithm with base 10 of the p-value; P$_{adj.}$, adjusted p-value;kb, kilo bases.

These observations can be formulated into a parsimonious nearest-gene mapping framework (**Fig. 3.1**); SNPs are mapped to the genes in their neighborhood (i.e. within a pre-defined distance), and are allowed to map to several genes. The advantage of such an approach is that it is simple and guarantees mappings for a large fraction of all SNPs and genes. For instance 423,450 (57%) percent of the Affymetrix 6.0 SNPs are mapped to 21,800 genes (using 70 kb upstream and 20 kb downstream flanking regions). One of the approach's inherent limitations is its inability to capture long-range,

[3] For instance variation in transcription factor bindings sites is known to play a major role in phenotype diversity [Kasowski et al., 2010].

[4] Briefly, expression quantitative trait loci analysis is a technique to assess whether a given variant influences genes expression levels. Studies have shown that every gene has an associated expression quantitative loci in a given tissue under specific conditions [Nica et al., 2010]. For more information on expression quantitative trait loci analysis please refer to [Quigley and Balmain, 2009].

i.e. *trans*-regulatory, SNP-gene relationships as for instance exemplified by the previously mentioned example on the association between the common colorectal cancer pre-disposition SNP at 8q24 and its recently identified function as a long-range enhancer of the MYC oncogene.

Ideally the boundaries of genes' flanking regions would be gene-specific and based on expression quantitative trait study data within the relevant tissues and under the right conditions. However, for most traits this is currently not feasible. Nevertheless, the approach has been used by most integrative GWA approaches, albeit with varying length of the flanking regions extending the most extreme transcripts of a given gene. The first pathway-based GWA analysis method by Wang *et al* used 500 kb as flanking regions, since most enhancers of a gene are located within that distance, and most linkage disequilibrium block are less than 500 kb [Wang et al., 2007]. However, in a later pathway-based landmark study Wang *et al* used 20 kb as flanking regions without any particular argument for doing so [Wang et al., 2009]. Segre *et al* used 110 kb and 40 kb in the MAGENTA tool and argued that their boundaries reflected the 99th percentile of all *cis*-expression quantitative trait loci from nearest gene's start and end sites [Segrè et al., 2010].

Gene scoring

The parsimonious way to score genes is to consider the most significant SNP mapped to a given gene, as the p-value of that gene. However, longer genes tend to harbor more SNPs, and thereby have an increased likelihood to incur SNPs that by chance correlate with trait of interest across the individuals investigated (denoting so-called chance-correlations). In addition, Bonferroni or Sidak correction cannot simply be applied on the number of SNPs mapped to a given gene, since linkage disequilibrium ties together adjacent SNPs into non-independent patterns. **Figure** 3.2 shows how this parsimonious way of scoring genes would be strongly biased towards longer genes (**Fig.** 3.2**a**) and genes with more SNPs mapped to them (**Fig.** 3.2**b**).[5] I will briefly discuss three possible ways used to alleviate these two biases in gene scoring frameworks.

One approach is to use permutation analysis to calculate SNP-count and linkage disequilibrium adjusted gene p-values. First each gene is assigned a p-value that equals the p-value of the most significant SNP mapped to the given gene (in the following referred to as $P_{gene,raw}$). Thereafter all individuals' phenotype-genotype relationships are randomized and the GWA analysis is re-computed a large number of times (>1000). In each randomization a new $P_{gene,raw}$ is calculated for each gene and subsequently stored into gene-specific background distribution of $P_{gene,raw}$ values. Upon completion of the randomization analysis, the background distribution is used to compute the adjusted gene p-value for each gene (by counting the number of random $P_{gene,raw}$ values that are lower than the observed $P_{gene,raw}$). This approach is used in for instance [Wang et al., 2007]. A drawback of this method is that it requires large computing resources since a total GWA study with 1 million SNP takes minutes to compute, which means that 10,000 permutations would require weeks of computation time if not parallelized.

Another approach is to calculate the number of effective tests a given set of linked SNPs corresponds to. In other words, this framework takes the linkage disequilibrium between say k SNPs into account to estimate the number of independent tests

[5] Data for these figures is taken from the NUGENOB intervention study GWA analysis (that was briefly mentioned in the above description of population substructure, section 2.2 p. 4.), which is one of my ongoing projects and not further described here [Sørensen et al., 2006].

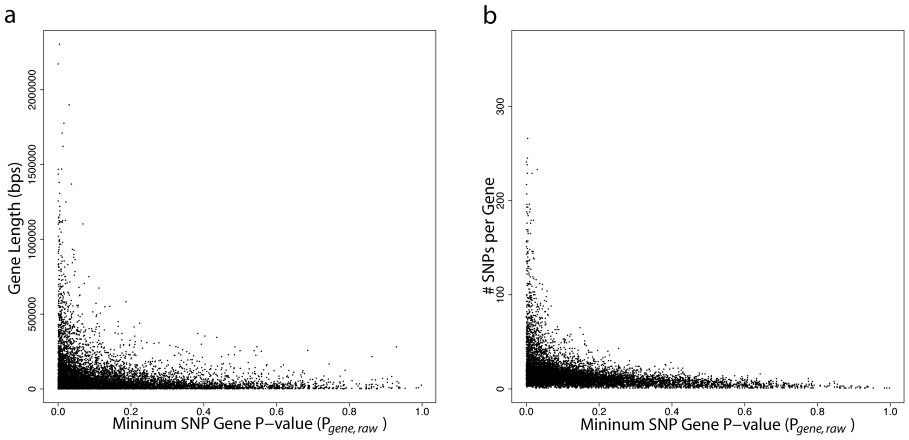


Figure 3.2: Bias of the parsimonious gene scoring approach, in which a gene is scored based on the lowest SNP p-value of all SNPs mapped to the given gene. Panel (**a**) shows how longer genes are more likely to have lower minimum SNP gene p-values (referred to as $P_{gene,raw}$ in the text), and panel (**b**) illustrates that the same bias holds true for the number of SNPs mapped to the given gene. (As expected, as the gene length and number of SNPs mapped to a given gene are highly correlated.)

that are actually being tested (letting pairs of SNPs that are in high linkage disequilibrium only count once), and in most cases will be substantially smaller than k. Galwey developed such a framework that relies on eigenvalue decomposition of the genotype data to calculate the number of independent tests for a given set of SNPs and afterwards applies Sidak correction to adjust $P_{gene,raw}$ with that number [Galwey, 2009]. Similar to the permutation-based approach, genotype data is needed to accomplish this correction. In case genotype data for the study cohort or another control cohort is not available genotype data from the HapMap or 1000 Genomes projects can be used. Based on Galwey's proposed scheme, I have implemented a software tool that enables researchers to perform the effective test gene scoring and used it in **Papers I - III**. To test the tool, I used the permutation-based correction method to calculate permutation-based p-values for all human protein-coding genes[6] (**Fig. 3.3a-b**). Hereafter, I correlated them with the gene p-values derived by my implementation of the Galwey approach, and confirmed that the Galwey correction framework actually was alleviating the SNP-count bias, as the permutation-based p-values and Galwey adjusted p-values were highly correlated, r^2=0.98 (**Fig. 3.3d**). This analytical correction approach is faster than the permutation-based approach as it does not rely on a vast number of permutation rounds.

Finally, I would like to mention a third gene scoring approach. This approach was used as part of the MAGENTA method [Segrè et al., 2010] and implements a linear regression framework to regress gene length in kb, number of SNPs per kb, independent SNPs per kb, number of recombination hotspots per kb, genetic distance in centi-Morgan per kb on the $P_{gene,raw}$ to adjusted obtain gene p-values. It does not rely on genotype data and is computationally faster than the two above-mentioned approaches.

[6] Using GWA data from the NUGENOB diet intervention study.

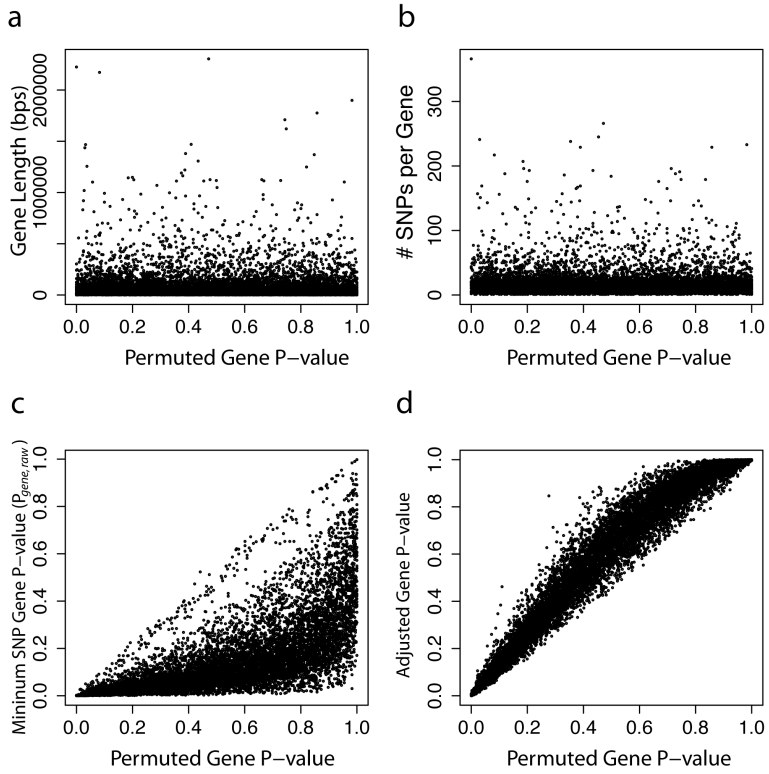


Figure 3.3: Gene p-values derived by permutation. Panel (**a**) and (**b**) show that the permutation-based gene p-values are not correlated with neither gene length nor the number of SNPs mapped to a given gene. Panel (**c**) illustrates that minimum SNP gene p-values (referred to as $P_{gene,raw}$ in the text) tend to be lower than the corresponding permutation-based p-values. Panel (**d**) shows that the Galwey correction-based p-values and permutation-based p-values are highly correlated (r^2=0.98), even though the Galwey correction method seems to slightly over-correct gene p-values.

In a recent GWA meta-analysis of human height, Allen *et al* found that allelic heterogeneity may be a frequent feature within polygenic traits [Allen et al., 2010]. That finding may induce a refinement of the current minimum scoring gene scoring methods, to incorporate the possibility of multiple independent common GWA signals within the same gene. In summary, both SNP to gene mapping and gene scoring approaches leave room for improvements. Tissue and condition-specific mapping approaches will become available in the near future as more systematic coupled gene variation and gene expression data will be available.

Definition and delineation of molecular pathways

Proteins that physically interact are very likely to be implicated in the same risk phenotype [Ideker and Sharan, 2008,Goh et al., 2007]. This premise on modular organization of cellular biology [Hartwell et al., 1999], also referred to as the guilt-by-association hypothesis, underlies most pathway-based approaches. However, while clinical and

cohort-based data sets are literately piling up in large data-centers, the mapping of molecular networks is still at a very early stage. It has been estimated that 10^5 to 10^6 of protein-protein and protein-DNA interactions occur within a single cell [Heard et al., 2010], a number which is still more than three times the number of all known PPIs from large meta-databases (<300,000 interactions).[7] While most molecular components are known, their spatial and functional relationships remain mostly elusive. Add to that the interactions that are specific to certain conditions or developmental stages, and it becomes even more clear that much still has to be learned.

Many canonical pathways from the KEGG and Biocarta databases resemble textbooks by presenting linear and independent representations of pathways. But, unlike simplified figures in textbooks of molecular components' topology, cellular networks are interwoven and condition-specific. Thus, pathways constructed based on manually annotated published text will necessarily be confined by the state of human knowledge and prone to false-negatives [Kraft and Raychaudhuri, 2009]. Experimentally-derived PPIs produced by both large- and small-scale screenings techniques, partly alleviates this bias. However, high-throughput experimentally-derived PPI also pose challenges since they may comprise many false-positive interactions. Generally one must be aware of these limitations, since the overall pathway-based analysis will only be as good as the PPI data is (in terms of accuracy and comprehensiveness). In the remainder of this section I will discuss possible biases, when analyzing GWA data in the context of protein complexes.

Protein-protein interaction meta-databases. There has been a steady growth of interactions in PPI databases (**Fig.** 3.4). However, PPI databases are often prone to a large number of both false-negatives and false-positives. Whereas additional experiments are the only way to improve the former, there are several measures that can be used to confine the latter. Some of them are

i) to score PPIs based on the number of independent publications citing the PPI,

ii) to score PPIs based on the type of technology by which the given PPI was reported,

iii) to assign PPIs from small-scale studies higher confidence than PPIs from large-scale studies, and

iv) to score PPIs based on their surrounding neighborhood's network topology. If the non-shared interaction partners of a pair of interacting proteins are interacting too, it is more likely that the primary interaction is genuine.

The four major PPI meta-databases are the Center for Biological Sequence Analysis in-house database InWeb [Lage et al., 2007], iRefWeb[8] [Turner et al., 2010], ConsensusPathDB[9] [Kamburov et al., 2011], and STRING [Szklarczyk et al., 2011] (see **Tab.** 3.1). In the InWeb database, we map out false-positive interactions by aggregating the above-mentioned confidence measures (i - iv) into a single score that can be

[7] Experimental techniques for the discovery of PPIs are not a subject in this Thesis. Briefly, the most often used techniques to identify PPIs are purification methods that identify protein complexes, such as immunoprecipitation [Phizicky and Fields, 1995] and affinity purification followed by mass spectrometry [Gavin et al., 2006], and the yeast to hybrid technique to detect binary interactions [Ito et al., 2001].

[8] http://wodaklab.org/iRefWeb

[9] http://cpdb.molgen.mpg.de

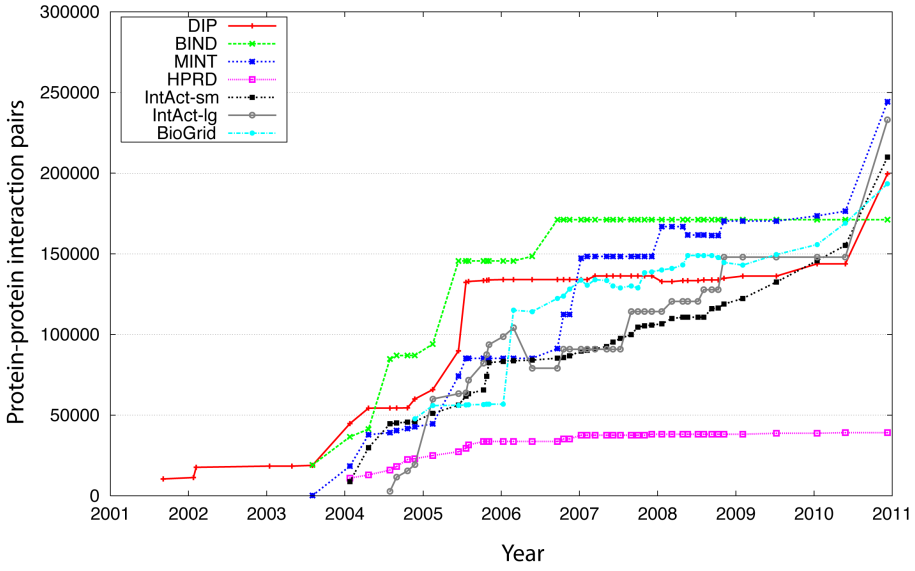


Figure 3.4: Growth in major human protein-protein interaction (PPI) databases. Please refer to Table 3.1 on page 15 for references on the various databases. Abbreviations: sm, small-scale experiments; lg, large-scale experiments.

used to discard low-confidence interactions. In addition, as most PPIs are derived from model organisms, the InWeb database extends its coverage by inferring PPIs from high-throughput PPI screens in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Escherichia coli* and *Caenorhabditis elegans*.[10] The InWeb database is not publicly available and comprises approximately 13,000 non-redundant PPIs consolidated from 40,085 original publications. Consequently, it harbors more interactions than the largest human interaction database HPRD [Peri et al., 2003]. iRefWeb integrates data from 10 predominantly experimental source databases. The current version 7.0 contains 263,479 non-redundant PPIs consolidated from 49,255 original publications. In addition to PPIs, ConsensusPathDB integrates metabolic, signaling and gene regulatory interaction networks, too. The database currently comprise 43,512 molecular components and 162,152 physical interactions. STRING integrates both experimentally-derived and predicted PPIs. The latter type of PPIs are predicted by signifying co-occurrence of gene (or protein) pairs. In most cases, they denote indirect interactions, i.e. functional rather than physical interaction.

As the coverage of PPI databases is growing and the number of falsely annotated PPIs is decreasing, it is my opinion that PPI data provides an unbiased and more powerful resource to identify etiological pathways compared to in-complete manually annotated pathways.

Confounding factors. Genes with similar functions may reside adjacent to each other on the same chromosome; the histone cluster 1 genes on human 6p21.33 is an often used example. When scoring gene sets, which for instance have been assembled based

[10] Protein-protein interactions have been shown to be conserved between species [Butland et al., 2005].

on PPI data, based on GWA data, chromosomal co-localization may become a confounder. In the case where a given SNP with a non-random association with the trait of interest is mapped to several genes that all are within the same gene set, the overall score for the gene set is inflated because the assumed independence of gene scores is violated. One solution is to filter out genes that have a co-localizing partner within a given gene set. We applied this correction on the complexes used in **Paper I** and [Dalgaard et al., 2011].

3.3 Paper I - A method for protein complex-based risk gene mapping

In the following paper, Majken K. Jensen and I meta-analyzed GWA data from two prospective studies of incident coronary heart disease (CHD), and subsequently developed a pathway-based analysis technique to search for protein complexes that are enriched in GWA study signal. We identified a protein complex centered on the ADBR1 gene, that was significantly enriched in proteins, which underlying genes had associations with CHD. We validated our findings in independent phenotypic data sets from mouse and human studies, which were not used in our discovery GWA analysis. By use of several sampling steps, we showed that the top complex indeed associates with CHD.

Protein interaction-based genome-wide analysis of incident coronary heart disease

Majken K Jensen, PhD*; Tune H Pers, PhD*; Piotr Dworzynski, BSc; Cynthia J. Girman, DrPH; Søren Brunak, PhD, and Eric B. Rimm, ScD

* Dr's Jensen and Pers contributed equally to the manuscript

Short title: Protein interaction-based GWAS of CHD

From the Departments of Nutrition (MKJ, EBR) and Epidemiology (EBR), Harvard School of Public Health, and the Channing Laboratory (EBR), Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Merck Research Laboratories (CJG), North Wales, PA; Department of Systems Biology (THP, PD), Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark; Institute of Preventive Medicine (THP), Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark and Novo Nordisk Foundation Center for Protein Research (SB), University of Copenhagen, Copenhagen, Denmark

Address for correspondence:
Dr. Majken K. Jensen, Department of Nutrition, Harvard School of Public Health, 655 Huntington Avenue, Boston, 02115 MA (mkjensen@hsph.harvard.edu; tel 1-617-312-3216; fax 1-617-432-2435) or Dr. Tune H. Pers, Center for Biological Sequence Analysis, Dept. of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark (tune.pers@cbs.dtu.dk; tel 1-857-210-6257)

Word count: Text 7,863; abstract 249

Background – Network-based approaches may leverage genome-wide association (GWA) analysis by testing for the aggregate association across several pathway members. We aimed to examine if networks of genes that represent experimentally determined protein-protein interactions are enriched in genes associated with risk of coronary heart disease (CHD).

Methods and Results – GWA analyses of ~700,000 SNPs in 899 incident CHD cases and 1,823 age- and sex-matched controls within the Nurses' Health and the Health Professionals Follow-Up Studies were used to assign gene-wise p-values. A large database of protein-protein interactions (PPI) was used to assemble 8,300 unbiased protein complexes and corresponding gene-sets. Superimposed gene-wise p-values were used to rank gene-sets based on their enrichment in genes associated with CHD. After correcting for the number of complexes tested, one gene-set was overrepresented in CHD-associated genes (p-value=0.002). Centered on the beta-1-adrenergic receptor gene (*ADRB1*), this complex included 18 protein interaction partners that, so far, have not been identified as candidate loci for CHD. Five of the 19 genes in the top-complex are reported to be involved in abnormal cardiovascular system physiology based on knock-out mice (4-fold enrichment; p-value, Fisher's exact test= 0.006). Ingenuity pathway analysis revealed that especially canonical pathways related to blood pressure regulation were significantly enriched in the genes from the top complex.

Conclusion – The integration of a GWA study with PPI data successfully identifies a set of candidate susceptibility genes for incident CHD that would have been missed in single-marker GWA analysis.

Genome-wide association (GWA) studies provide a unique opportunity for the unbiased exploration of novel genetic variation of importance to phenotypic traits. The first series of GWA studies of coronary heart disease (CHD) and more broadly defined cardiovascular disease (CVD) phenotypes elucidated DNA sequence variations at the 9p21.3 locus as a robustly replicated risk-conferring region,[1, 2 3] but through a series of larger GWA study consortia about 10 susceptibility loci have been reported.[4-6] The recent publication of results from the multi-ethnic Coronary Artery Disease (C4D) Genetics Consortium,[7] the first Han Chinese GWA study,[8] and the CARDIoGRAM consortium with more than 20,000 coronary artery disease cases,[9] yielded an additional 18 new loci. However, the complexity of the phenotype,[10] small effect sizes, and between-study differences may complicate the identification of many true associations in meta-analyses that necessarily assumes homogeneity across the individual studies. Most GWA studies to date have focused on the identification of the strongest single-locus associations, but the identification of combined effects of many weakly associated variants is especially appealing for complex diseases, such as CHD, that is likely not caused by single variants or by a single biological pathway. Thus, another suggested approach for reducing the noise inherent in moderately powered high-density data collected within internally homogenous populations, is the integration of additional biological data on pathway organization through the use of a protein-protein interaction (PPI) database.[11-16] By enabling tests of sets of single nucleotide polymorphisms (SNPs) within physically interacting gene products (direct or indirect), PPI data can augment GWA analysis since a set of SNPs, each with a moderate, but genuine association, in aggregate may have improved statistical significance. Although several databases provide gene-sets that resemble well-known canonical pathways, high-confidence PPI data may to a larger degree mimic the unbiased nature of GWA studies due to its increased coverage and detail of even non-canonical pathways.[11, 17] Initial approaches have proven useful to suggest novel genes and gene-networks involved in other complex phenotypes such as obesity,[18] type 2 diabetes,[13] breast and pancreatic cancer,[19] multiple sclerosis,[20] and Crohn's disease,[21] that were not identified in the traditional GWA analysis. The completeness of such integrative analysis relies strongly on the gene-sets tested. We aimed to examine if networks of genes that represent experimentally determined protein-protein interactions are enriched in genes associated with risk of incident CHD. To leverage our GWA analysis of CHD within two homogenous American prospective cohorts including 899 incident cases collected through more than 10 years of follow-up, we used our PPI database *InWeb*,[14] which covers ~13,000 human proteins and 173,500 *high-confidence* experimentally-derived protein-protein interactions based on 11 publicly available PPI-databases.

Methods

Study population

The Nurses' Health Study (NHS) enrolled 121,701 female nurses aged 30 to 55 who returned a mailed questionnaire in 1976 regarding lifestyle and medical history. The Health Professionals Follow-up Study (HPFS) enrolled 51,529 males aged 40 to 75 who returned a similar questionnaire in 1986. Participants of both cohorts have received follow-up questionnaires biennially to record newly diagnosed illnesses. Detailed descriptions of the study cohorts have been published previously.[22, 23]

Blood collection and DNA extraction in nested case-control study

Between 1989 and 1990, a blood sample was requested from all active participants in NHS and collected from 32,826 women. Similarly, blood samples were requested between 1993 and 1995 and obtained from 18,225 HPFS participants. For details on storage of blood samples, please see the online supplement.

In both cohorts, nested case-control studies were designed using incident CHD, with non-fatal myocardial infarction (MI) and fatal CHD as the outcome. Diagnosis of MI was confirmed on the basis of the criteria of the World Health Organization (symptoms plus either diagnostic electrocardiographic changes or elevated levels of cardiac enzymes). Deaths were identified from state vital records and the National Death Index or reported by the participant's next of kin or the postal system. Fatal CHD was confirmed by an examination of hospital or autopsy records, by the listing of CHD as the cause of death on the death certificate, if CHD was the underlying and most plausible cause, and if evidence of previous CHD was available. Among participants who provided blood samples and who were free of diagnosed cardiovascular disease or cancer at blood draw, we identified 474 women and 454 men with incident CHD between blood draw and June, 2004. Using risk-set sampling,[24] controls were selected randomly and matched in a 1:2 ratio on age, smoking, and month of blood return, among participants who were free of cardiovascular disease at the time CHD was diagnosed in the case. In this study design, a control for an early case may be included again if the person develops CHD during follow-up, thus after counting such converters only once (as cases), the total number of samples sent for genotyping were 1354 HPFS samples and 1521 NHS samples.

The present study was approved by the institutional review boards at Brigham and Women's Hospital and Harvard School of Public Health.

Genotyping and Quality Control

Details on the protocol for DNA extraction has been included in the online supplement. Genotyping was done using the Affymetrix Genome-Wide Human 6.0 array and the Birdseed calling algorithm.[25] Genotypic data for a total of 1,330 HPFS samples (98%) passed laboratory technical quality control criteria and missing call <0.05. Likewise, 96% of the NHS samples were successfully genotyped. A subset of 312 NHS samples were not genotyped together with the remaining CHD case-control set as they overlapped with previous GWA studies of breast cancer (Illumina 550) and type 2 diabetes (Affymetrix 6.0). These samples were processed and subjected to quality control as part of the earlier GWAS (leaving n=272 samples with available data) and SNPs also present on the Affymetrix 6.0 platform were subsequently merged with the cleaned CHD data. Details on methods for data cleaning and assesment of population structure in the datasets are included in the online supplement. Due to very few samples with substantial evidence of non-European genetic ancestry, these samples were excluded from subsequent analysis (n=24). SNPs that were monomorphic, had a missing call rate ≥2%, a HWE p-value <1×10-4, or a MAF <0.02 were excluded, leaving a total of 724,881 in HPFS and SNPs that passed quality control in HPFS and 721,316 in NHS for analysis of called genotypes. Imputation of ~2.5 million SNPs was performed using MACH software (v1.0.16) with HapMap CEU phased II data (Release 22) as the reference panel.

Genome-wide association analysis of coronary heart disease

To analyze the association between each SNP (coded as counts of minor alleles) and risk of CHD, we ran logistic regression analysis using PLINK software.[26] We adjusted for matching factors used in the design of the nested case-control study (age and smoking) and the top three eigenvectors. We also analyzed the MACH dosage files of the imputed SNPs (with MAF ≥0.05) in logistic regression models (adjusting for same covariates as above) using the ProbABEL package from the ABEL set of programs.[27] Fixed-effects meta-analysis was performed to combine the study-specific β-estimates using the METAL package.[28]

Systems biology-based approaches that integrate data on protein interactions are necessarily restricted to the protein-coding part of the genome. We mapped all GWA SNPs that passed quality control to 21,800 protein-coding genes (423,450 mapped SNPs, ~57% of all SNPs on the Affymetrix 6.0 arrays) (Figure 1a). This process is gene-centric such that SNPs that are not within genes or their 70 kb upstream and 20 kb downstream flanking regions were discarded. SNPs were allowed to map to more than one gene.

Each gene was assigned a p-value based on the SNP with the lowest GWA p-value within the gene transcript(s) and its flanking regions. Subsequently, the Šidàk correction was applied to adjust the p-value for each gene by the number of effective tests (uncorrelated number of SNPs within each gene and its flaking regions, as per Galwey 2007).[16, 29]

Protein-Protein Interactions and CHD-specific protein complexes

Protein-protein interactions comprise both transient interactions (e.g. phosphorylation events) and stable interactions (e.g. the cytoskeleton). Our comprehensive, experimentally derived database of protein-protein interactions InWeb (version 2.9) covers ~13,000 human proteins and 350,029 protein-protein interactions of which 173,500 can be regarded as high-confidence interactions (as described below).[14] The database is updated on a monthly basis with interactions retrieved from all major experimental PPI databases (details available in online supplement). Strengths of the InWeb database include the relative high coverage (4-fold increase in number of interaction compared the Human Protein Reference Database, HPRD)[30] and a quantitative assessment of confidence in the reported interactions. The (continuous) confidence score (ranging from 0 [low support] to 1 [strong support]) is assigned by taking into account a) the number and quality of the publications reporting each of the interactions and b) the number of shared interaction partners of two interacting proteins.[14] The assembly of 8,531 gene-sets was accomplished by iteratively assigning each protein in the database and its first-order interaction partners to a protein complex (Figure 1, Step1b). As such a construction of protein complexes results in a relatively large number of overlapping complexes, complexes that were more than 80% similar (similarity of gene-sets assessed by the Jaccard Index) were merged. After superimposing the gene-wise p-values from the GWA analysis onto the network, we used a modified version of an approach published by Ideker *et al.* to iteratively assess whether any of the gene-sets that were derived from the protein complexes were enriched in CHD-associated genes.[31] Given a gene-set of size k, this was accomplished by (1) converting all k gene p-values to z-scores using the inverse normal cumulative distribution function, (2) weighting them with the interaction confidence score of the protein-protein interaction with the central hub protein (a step that was not part of the original algorithm), (3) calculating a gene-set score by summing the weighted z-scores, and then (4) subtracting the sum of an average gene-set of size k (calculated based on 100,000 randomized gene-set scores), and dividing by the standard deviation of an average sub-network of size k. Formally, step 1 can be formulated as $z_i = F^{-1}(1 - p_i), i \in \{1,\ldots,k\}$, steps 2-3 as $S_{gene-set} = \frac{1}{k}\sum cs^i_{hub} z_j$, and step 4 as $Z_{gene-set} = \frac{\left(S_{gene-set} - \mu_k\right)}{\sigma_k}$, where p_i denotes the p-value of gene i, z_i denotes the z-score of gene i, F^{-1} denotes the inverse normal cumulative distribution function, cs^i_{hub} the confidence score for the protein-protein interaction between gene product i and the central hub gene product, $S_{gene\text{-}set}$ denotes the score of the gene-set after steps 1-4, μ_k and σ_k denote the mean and standard deviation of 100,000 randomized gene-set scores, and $Z_{gene\text{-}set}$ denotes the final gene-set z-score. Using this methodology, all gene-sets were ranked based on their computed z-scores (Figure 1c). Because SNPs were allowed to map to several overlapping genes, some gene-sets may be assigned artificially inflated scores if they comprise genes that overlap on a given chromosome and are scored based on the same low SNP p-value. To avoid this potential bias we discarded one of the genes in any overlapping gene pair in a given complex (genes were considered to overlap if their transcripts were closer than 200kb to each other). This approach should be considered as conservative as it avoids inflated complex scores, but in some cases may reduce significance of truly associated complexes that comprise co-localizing gene-products with independent associations. In our present analysis, the top complex remained the same with or without discarding overlapping genes (and for different exclusion thresholds). We assessed the significance of our observed top scoring complex by comparing its score with a background distribution of 100 scores generated under the null hypothesis that the complex is not associated with CHD case control status. The background distribution was estimated on the basis of 100 permutations of our GWA meta-analysis (randomizing the case control status) and re-computations of the gene scoring- and complex scoring step for each permutation. An ideal scenario would include up to 1 million permutations but the aggregate computing times for the GWA analysis, the gene scoring step, and the complex enrichment analysis did not allow for this.

After identification of the top-ranking complex we searched the literature to see if the genes were known as human CVD candidate genes. To assess over-representation of known CVD susceptibility genes we used a list of 123 genes reported by Samani *et al*. and updated it with GWA findings in the NIH

Catalog of Genome-Wide Association Studies (Suppl. Table 1).[3, 32] We also tested for overrepresentation of a list of 889 genes found to affect the cardiovascular system physiology (MP:0001544) in knockout mice (Mouse Genome Informatics database; www.informatics.jax.org, Jackson Laboratory, Bare Harbor Maine) (of which 837 were among the gene products in our PPI database). To ensure that the observation that genes from our top complex were overrepresented in the mouse cardiovascular physiology gene-set was not due to chance, we compared the observed enrichment score to a background distribution of 10,000 scores computed based on randomly sampled protein complexes. Each of the random complexes matched the observed complex in size, and each gene-product was sampled with a probability equal to its observed prevalence in the total set of protein complexes. In addition to the enrichment analysis of known human and mice CHD risk genes, we used the Ingenuity Pathway Analysis software tool (IPA, version 9.0, Ingenuity Systems Inc. 2011) to systematically test the complex genes for pathway enrichment.

Results

Characteristics of incident cases and matching controls in the two cohorts are presented in Table 1. The women in the NHS were slightly younger, more likely to smoke, and more likely to report a diagnosis of hypertension or diabetes. GWA analysis of each cohort separately and in meta-analysis did not reveal any markers that exceeded the genome-wide significant threshold (Supplement, fig 1).

Based on the InWeb database, a total of 8,351 protein complexes were assembled based on large-scale proteomics data from human and model organisms. We restrained our analyses to high-confidence protein-protein interactions only, including a subset that we recently validated experimentally in human heart tissue.[45] The resulting protein complexes were tested for enrichment in CHD-associated genes by using the gene-wise p-values from the GWA analysis to create z-scores and ranking the complexes (gene-sets) by their combined z-scores, adjusted for the size of each gene-set, and weighted by the confidence of the interactions between the peripheral gene-products and the central protein of the complex. After correcting for the number of complexes tested, one gene-set was significantly overrepresented in CHD-associated genes from our GWA meta-analysis (p-value=0.002). The gene complex was centered on the known candidate gene for the beta-1-adrenergic receptor (*ADRB1*) (fig. 2). To ensure that the top complex was not merely significantly enriched in genes with low p-values but indeed significantly associated with CHD case control status, we permuted the phenotype-genotype association in the GWA analysis 100 times and re-computed the complex score at each iteration. We found that the score for the observed *ADBR1* complex was superior to any of the scores for the randomized complexes. In Figure 2, the additional 18 genes that were part of the complex of interacting proteins are scaled according to their gene-wise p-values. As shown in more detail in Table 2, the genes; membrane-associated guanylate kinase inverted 1 (*MAGI1*), the protein kinase cAMP-dependent catalytic alpha (*PRKACA*), and the Golgi associated PDZ and coiled-coil motif containing (*GOPC*) were nominally significant after correcting for the number of independent SNPs in each gene, whereas the remainder showed weaker, or no association. In the combined test of a gene-set, all known interaction partners are included regardless of their GWA signal and the strength of the association for the complex relies on the sum of all gene-wise p-values of the interacting genes. Our results did not change when we based our analysis on the imputed GWA data rather than the hard-call genotypes.

Next, we assessed whether the *ADBR1* complex was enriched in known human or mice CVD risk genes. No significant overlap with the list of 123 susceptibility genes reported by Samani *et al.* and the genetic loci identified in GWA studies of CVD was observed (p-value=0.1). [3, 32] To test for enrichment in CHD-specific evidence from mouse studies, we searched for the genes in the top-complex in an *a priori* defined set of genes causing abnormal cardiovascular physiology in knockout mice. Among a total of 889 genes reported for that phenotype, 837 human homologs were among the 12,793 genes included in our analysis, and 5 were part of the 19 genes in the *ADRB1* complex; representing a 4-fold enrichment (p-value, Fisher's exact test =0.006). The five genes also found in mice knockout gene-sets, were *ADRB1*, *ADRA2A*, *ARRB1*, *PDE4D*, and *GRB2* of which all except *PDE4D* were reported to play a role in the regulation of blood pressure, cardiac function, and hypertrophy. Because proteins that are known to interact physically are more likely to have similar functional annotation,[33] possible chance-correlations resulting in a gene with a low p-value could potentially result in a falsely associated complex if the falsely associated gene's annotation resembles the phenotype of interest. To test for this possible bias, we subjected the mouse gene-set enrichment analysis to 10,000 random complexes sampled from the PPI network and found that only 13 out of the 10,000 randomized enrichment scores were lower than our observed score (p-value=0.001).

We used Ingenuity Pathway Analysis to examine whether the annotations of the genes in the *ADRB1* complex were enriched for any particular phenotype. Between 10 and 12 of the 19 genes were reported in cardiovascular, neurological, endocrine, and immunological disorders (Table 3). Moreover, several cardiovascular related pathways were enriched in genes from the complex. The top canonical pathway was cardiac hypertrophy signaling. To better ensure that the observed enrichment was not due to chance, we sampled 100 random gene-sets comprising 19 genes each, and performed IPA analysis based on each set. Only one random gene-set exhibited enrichment in cardiovascular disease genes as strong as the observed enrichment for the *ADBR1* complex gene-set. Thus, we conclude that our top complex was significantly enriched in genes associated with cardiovascular disease (p-value< 0.05). None of the random gene-sets were significantly enriched in the cardiac hypertrophy canonical pathway, suggesting that the *ADBR1* complex gene set was significantly enriched in genes from this pathway too. We confined this IPA permutation analysis to 100 iterations as the software does not allow automation and all runs were done manually.

Discussion

We conducted a protein network-based GWA analysis to leverage our moderately powered GWA study of CHD. Using GWA data from two individually homogeneous studies, we integrated the gene-wise p-values with a large database of protein-protein interactions. By exploiting the complementary nature of genetic variation and biochemical data, we successfully identified a gene complex of 19 candidate genes that may play a role in the etiology of incident CHD. Subsequent pathway enrichment analysis indicated that the top complex was significantly enriched in (a) genes from the canonical cardiac hypertrophy signaling pathway (the highest ranking pathway in the IPA analysis), (b) genes annotated with cardiovascular disease (the second most enriched trait in the IPA analysis), and (c) mouse genes annotated in the cardiovascular system physiology. Our results provide preliminary evidence that known CHD-related genes coalesce onto distinct protein complexes. Most of the genes in the top complex had relatively small effect sizes, making them unlikely findings in traditional single-locus GWA analyses of CHD.

To our knowledge, our study of incident CHD is the first attempt at integration of data on the human interactome with GWA data in relation to incident CHD. As shown in the enrichment analyses, the top complex comprises several genes that previously have been annotated to cardiovascular disease and, in particular, the cardiac hypertrophy signaling pathways. Except for *ADRB1*, these known genes were not nominally significant by themselves but leveraged due to their interaction with genes that comprised SNPs, which exhibited association with CHD in our GWA study. In addition, the top complex was significantly enriched in genes from the Mouse Genetics Initiative database that were annotated in the 'cardiovascular system physiology'. The genes *ADRB1*, *GRB2*, *ADRA2A* were found to overlap between all three *a priori* defined gene-sets (overview provided in Table 4). It is well-known that the β1-adrenergic receptor plays an important role in the regulation of cardiac contractility. In candidate genetic studies, *ADRB1* SNPs have been associated with blood pressure[34] and risk of future CHD, which might be particularly true for individuals with elevated blood pressure.[35] Studies on the adrenergic pathway genes, including *ADRA2A*, that encodes the α2A-adrenergic receptor, have not shown consistent associations. However, recently a polymorphism in *ADRA2A* that caused overexpression of the protein, was shown to strongly reduce insulin secretion from pancreatic cells and be associated with an elevated risk of type 2 diabetes.[36] The *GRB2* gene encodes the growth factor receptor-bound protein 2. So far, information on this genetic locus links it to an important role in lymphocytes and growth cells, but no human genetic epidemiologic studies have investigated this locus in relation to cardiometabolic disorders.

Alternative approaches for augmenting GWA data by testing significance beyond single locus associations include pathway-based approaches, such as methods that search the protein interactome for dense subnetworks enriched in GWA signal[19, 20] and methods that assess pre-defined gene-sets for enrichment in GWA signal,[16, 17, 37] The former class of methods are inspired by early work of Ideker,[31, 381] and employ an heuristic search algorithm to identify subnetworks that are enriched in gene-products that in aggregate associate with the phenotype. The advantage of these methods is that they do not assume any *a priori* delineation of pathways. However, the main drawback is that they rely on user-specified parameters that control the size of the subnetworks identified by the algorithm. In addition, none of them incorporate information on the confidence of the experimentally derived protein-protein interactions. While our approach resembles the recently presented dmGWAS approach,[19] only ours incorporated a score on confidence in the reported interactions. Another strength of our approach is that it is based on a PPI database that, despite its high coverage (our analysis includes twice as many interactions as those used in the dmGWAS method), solely includes high-confidence experimentally derived interactions. While InWeb does not rely on predicted protein-protein interactions, which are more prone to false-positive interactions, it still entails approximately 173,500 interactions from a total of 11 databases. Our integration-based approach has strengths, but limitations as well. One of the inherent limitations is that it only covers roughly 60% of all SNPs present on genotyping platforms. Consequently, SNPs within distal enhancer regions are discarded, as are other long-range regulatory relationships. However, systematic tissue- and condition-specific expression quantitative trait loci analyses are increasingly contributing to the development of more refined SNP to gene mapping schemes. Among other limitations, we had a relative small sample size in our GWA study of incident CHD and were limited to Caucasians. However, the application of the novel PPI approach still allowed us to uncover gene sets that were not otherwise identified. Replication in another prospective study setting is necessary to verify and demonstrate the importance of our identified top complex in incident cardiovascular disease. To test at the gene-level, genome-wide data would be preferable. However, we have not been able to identify a prospective study of CHD with sufficient number of cases where our protein interaction-based analysis could be repeated. Alternative approaches might be

the creation of a score of the top SNPs within the genes of the *ADBR1* complex. However, such an approach might be limited as a single SNP is unlikely to capture the variation at the locus.

In conclusion, our approach suggests that integration of other layers of biological evidence with a moderately powered GWA study of CHD in two homogenous study populations can yield potentially interesting sets of candidate genes that would be missed in traditional statistical GWA analyses. We identified one gene-set, centered on *ADRB1*, that was overrepresented in CHD-associated genes in our GWA study and also enriched in genes involved in the cardiovascular disease phenotype and particularly blood pressure regulation pathways. Our novel approach highlighted 19 genes that warrant further association and functional studies in terms of risk of CHD and blood pressure.

Funding Sources
This study was supported by HL34594, CA87969, HL35464, and CA55075 from the National Institutes of Health, Bethesda, MD, with additional support for genotyping from Merck Research Laboratories, North Wales, PA.

Disclosures
Merck Research Laboratories supported the genotyping of the Nurses' Health and Health Professionals Follow-Up Studies though an unrestricted grant.

Reference List

1. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316:1491-1493.

2. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316:1488-1491.

3. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, Konig IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357:443-453.

4. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fetiveau R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonzin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Spreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burtt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall AS, Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, Konig I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Huge A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, Konig IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Holm H, Thorleifsson G, Thorsteinsdottir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardissino D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41:334-341.

5. Erdmann J, Grosshennig A, Braund PS, Konig IR, Hengstenberg C, Hall AS, Linsel-Nitschke P, Kathiresan S, Wright B, Tregouet DA, Cambien F, Bruse P, Aherrahrou Z, Wagner AK, Stark K, Schwartz SM, Salomaa V, Elosua R, Melander O, Voight BF, O'Donnell CJ, Peltonen L, Siscovick DS, Altshuler D, Merlini PA, Peyvandi F, Bernardinelli L, Ardissino D, Schillert A, Blankenberg S, Zeller T, Wild P, Schwarz DF, Tiret L, Perret C, Schreiber S, El Mokhtari NE, Schafer A, Marz W, Renner W, Bugert P, Kluter H, Schrezenmeir J, Rubin D, Ball SG, Balmforth AJ, Wichmann HE, Meitinger T, Fischer M, Meisinger C, Baumert J, Peters A, Ouwehand WH, Deloukas P, Thompson JR, Ziegler A, Samani NJ, Schunkert H. New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet 2009; 41:280-282.

6. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447:661-678.

7. Peden JF, Hopewell JC, Saleheen D, Chambers JC, Hager J, Soranzo N, Collins R, Danesh J, Elliott P, Farrall M, Stirrups K, Zhang W, Hamsten A, Parish S, Lathrop M, Watkins HC, Clarke R, Deloukas P, Kooner JS, Goel A, Ongen H, Strawbridge RJ, Heath S, Malarstig A, Helgadottir A, Ohrvik J, Murtaza M, Potter S, Hunt SE, Delepine M, Jalilzadeh S, Axelsson T, Syvanen AC, Gwilliam R, Bumpstead S, Gray E, Edkins S, Folkersen L, Kyriakou T, Franco-Cereceda A, Gabrielsen A, Seedorf U, Eriksson P, Offer A, Bowman L, Sleight P, Armitage J, Peto R, Abecasis G, Ahmed N, Caulfield M, Donnelly P, Froguel P, Kooner AS, McCarthy MI, Samani NJ, Scott J, Sehmi J, Silveira A, Hellenius ML, van 't Hooft FM, Olsson G, Rust S, Assmann G, Barlera S, Tognoni G, Franzosi MG, Linksted P, Green FR, Rasheed A, Zaidi M, Shah N, Samuel M, Mallick NH, Azhar M, Zaman KS, Samad A, Ishaq M, Gardezi AR, Memon FU, Frossard PM, Spector T, Peltonen L, Nieminen MS, Sinisalo J, Salomaa V, Ripatti S, Bennett D, Leander K, Gigante B, de FU, Pietri S, Gori F, Marchioli R, Sivapalaratnam S, Kastelein JJ, Trip MD, Theodoraki EV, Dedoussis GV, Engert JC, Yusuf S, Anand SS. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet 2011; 43:339-44.

8. Wang F, Xu CQ, He Q, Cai JP, Li XC, Wang D, Xiong X, Liao YH, Zeng QT, Yang YZ, Cheng X, Li C, Yang R, Wang CC, Wu G, Lu QL, Bai Y, Huang YF, Yin D, Yang Q, Wang XJ, Dai DP, Zhang RF, Wan J, Ren JH, Li SS, Zhao YY, Fu FF, Huang Y, Li QX, Shi SW, Lin N, Pan ZW, Li Y, Yu B, Wu YX, Ke YH, Lei J, Wang N, Luo CY, Ji LY, Gao LJ, Li L, Liu H, Huang EW, Cui J, Jia N, Ren X, Li H, Ke T, Zhang XQ, Liu JY, Liu MG, Xia H, Yang B, Shi LS, Xia YL, Tu X, Wang QK. Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. Nat Genet 2011; 43:345-9.

9. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C, Absher D, Aherrahrou Z, Allayee H, Altshuler D, Anand SS, Andersen K, Anderson JL, Ardissino D, Ball SG, Balmforth AJ, Barnes TA, Becker DM, Becker LC, Berger K, Bis JC, Boekholdt SM, Boerwinkle E, Braund PS, Brown MJ, Burnett MS, Buysschaert I, Carlquist JF, Chen L, Cichon S, Codd V, Davies RW, Dedoussis G, Dehghan A, Demissie S, Devaney JM, Diemert P, Do R, Doering A, Eifert S, Mokhtari NE, Ellis SG, Elosua R, Engert JC, Epstein SE, de FU, Fischer M, Folsom AR, Freyer J, Gigante B, Girelli D, Gretarsdottir S, Gudnason V, Gulcher JR, Halperin E, Hammond N, Hazen SL, Hofman A, Horne BD, Illig T, Iribarren C, Jones GT, Jukema JW, Kaiser MA, Kaplan LM, Kastelein JJ, Khaw KT, Knowles JW, Kolovou G, Kong A, Laaksonen R, Lambrechts D, Leander K, Lettre G, Li M, Lieb W, Loley C, Lotery AJ, Mannucci PM, Maouche S, Martinelli N, McKeown PP, Meisinger C, Meitinger T, Melander O, Merlini PA, Mooser V, Morgan T, Muhleisen TW, Muhlestein JB, Munzel T, Musunuru K, Nahrstaedt J, Nelson CP, Nothen MM, Olivieri O, Patel RS, Patterson CC, Peters A, Peyvandi F, Qu L, Quyyumi AA, Rader DJ, Rallidis LS, Rice C, Rosendaal FR, Rubin D, Salomaa V, Sampietro ML, Sandhu MS, Schadt E, Schafer A, Schillert A, Schreiber S, Schrezenmeir J, Schwartz SM, Siscovick DS, Sivananthan M, Sivapalaratnam S, Smith A, Smith TB, Snoep JD, Soranzo N, Spertus JA, Stark K, Stirrups K, Stoll M, Tang WH, Tennstedt S, Thorgeirsson G, Thorleifsson G, Tomaszewski M, Uitterlinden AG, van Rij AM, Voight BF, Wareham NJ, Wells GA, Wichmann HE, Wild PS, Willenborg C, Witteman JC, Wright BJ, Ye S, Zeller T, Ziegler A, Cambien F, Goodall AH, Cupples LA, Quertermous T, Marz W, Hengstenberg C, Blankenberg S, Ouwehand WH, Hall AS, Deloukas P, Thompson JR, Stefansson K, Roberts R, Thorsteinsdottir U, O'Donnell CJ, McPherson R, Erdmann J, Samani NJ. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 2011; 43:333-338.

10. Kitsios GD, Dahabreh IJ, Trikalinos TA, Schmid CH, Huggins GS, Kent DM. Heterogeneity of the phenotypic definition of coronary artery disease and its impact on genetic association studies. Circ Cardiovasc Genet 2011; 4:58-67.

11. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 2009; 33:419-431.

12. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 2010; 86:6-22.

13. Perry JR, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, Frayling TM. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 2009; 58:1463-1467.

14. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 2007; 25:309-316.

15. Brorsson C, Hansen NT, Lage K, Bergholdt R, Brunak S, Pociot F. Identification of T1D susceptibility genes within the MHC region by combining protein interaction networks and SNP genotyping data. Diabetes Obes Metab 2009; 11 Suppl 1:60-66.

16. Pers TH, Hansen NT, Lage K, Koefoed P, Dworzynski P, Miller ML, Flint TJ, Mellerup E, Dam H, Andreassen OA, Djurovic S, Melle I, Borglum AD, Werge T, Purcell S, Ferreira MA, Kouskoumvekaki I, Workman CT, Hansen T, Mors O, Brunak S. Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genet Epidemiol 2011; 35:318-32.

17. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 2007; 81:1278-1283..

18. Liu YJ, Guo YF, Zhang LS, Pei YF, Yu N, Yu P, Papasian CJ, Deng HW. Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. Obesity 2010; 18:2339-2346.

19. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics 2011; 27:95-102.

20. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 2009; 18:2078-2090.

21. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van LJ, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet 2009; 84:399-405.

22. Colditz GA, Manson JE, Hankinson SE. The Nurses' Health Study: 20-year contribution to the understanding of health among women. J Womens Health 1997; 6:49-62.

23. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of a expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am J Epidemiol 1992; 135:1114-1126.

24. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika 1978; 65:153-158.

25. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008; 40:1253-1260.

26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559-575.

27. Aulchenko YS, Struchalin MV, Van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics 2010; 11:134.

28. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, vey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008; 40:161-169.

29. Galwey NW. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet Epidemiol 2009; 33:559-568.

30. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003; 13:2363-2371.

31. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002; 18 Suppl 1:S233-S240.

32. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009; 106:9362-9367.

33. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999; 402:C47-C52.

34. Johnson AD, Newton-Cheh C, Chasman DI, Ehret GB, Johnson T, Rose L, Rice K, Verwoert GC, Launer LJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, Caulfield M, Van Duijn CM, Ridker PM, Munroe PB, Levy D. Association of Hypertension Drug Target Genes With Blood Pressure and Hypertension in 86 588 Individuals. Hypertension 2011; 57: 903-10.

35. Leineweber K, Heusch G. Beta 1- and beta 2-adrenoceptor polymorphisms and cardiovascular diseases. Br J Pharmacol 2009; 158:61-69.

36. Rosengren AH, Jokubka R, Tojjar D, Granhall C, Hansson O, Li DQ, Nagaraj V, Reinbothe TM, Tuncel J, Eliasson L, Groop L, Rorsman P, Salehi A, Lyssenko V, Luthman H, Renstrom E.

Overexpression of alpha2A-adrenergic receptors contributes to type 2 diabetes. Science 2010; 327:217-220.

37. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 2010; 6 (8). pii: e1001058.

38. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007; 3:140.

Table 1. Baseline characteristics of women and men in whom coronary heart disease developed during follow-up and matched controls in the Nurses' Health Study (NHS) and the Health Professionals Follow-Up Study (HPFS).*

	HPFS		NHS	
Characteristic	Cases	Controls	Cases	Controls
N	425	878	464	945
Age, years	64.5 (8.6)	64.2 (8.5)	60.2 (6.3)	59.8 (6.3)
Women, %	0%	0%	100%	100%
Hypertension, † %	37.2%	29.0%	50.2%	27.3%
Diabetes, †%	9.0%	3.8%	14.4%	6.24%
Current smoker, %	9.7%	8.7%	27.8%	26.1%
Total cholesterol, mg/dL	5.5 (1.0)	5.2 (1.0)	6.1 (1.0)	5.9 (1.0)
HDL cholesterol, mg/dL	1.1 (0.3)	1.2 (0.3)	1.3 (0.4)	1.6 (0.4)
Triglyceride, mg/dL	1.8 (1.5)	1.5 (2.2)	1.6 (1.0)	1.3 (0.7)
BMI, kg/m^2	26.0 (3.2)	25.6 (3.3)	26.0 (6.6)	24.5 (5.8)

*Age and smoking were matching factors. Values are means and standard deviation of continuous covariates (except triglyceride levels which is reported as median and IQR) or percentages. Triglyceride levels were log-transformed before analysis and only reported in fasting participants (HPFS= 65%, NHS= 79%).
† Self-reported diagnosis before blood draw.

Table 2. Genes and primary SNPs in the top-ranking protein complex based on the GWA meta-analysis of risk of CHD in the Nurses' Health and Health Professionals Follow-Up Studies.

Gene	Gene p- value*	Top SNP	MAF	OR	Top SNP, raw p-value	# SNPs in gene	# independent SNPs in gene
MAGI1	7.8E-04	rs7620106	0.40	1.30	9.1E-06	251	86
PRKACA	0.004	rs40282	0.46	1.20	0.002	2	2
GOPC	0.028	rs12664183	0.28	1.23	0.001	100	27
ADRB1	0.041	rs17653278	0.06	0.70	0.003	41	12
MAGI3	0.073	rs4839312	0.26	1.21	0.005	62	14
MAGI2	0.086	rs2065198	0.46	1.22	0.001	579	149
GRB2	0.107	rs7223674	0.05	0.72	0.014	32	8
DLGAP2	0.143	rs7836020	0.45	1.18	0.005	100	33
ARRB1	0.217	rs2279129	0.08	0.75	0.013	34	19
DLG4	0.251	rs5412	0.16	1.15	0.069	7	4
GNAL	0.277	rs2848465	0.22	0.83	0.009	85	36
GIPC1	0.304	rs4926215	0.47	0.89	0.042	15	8
DLG1	0.335	rs7616531	0.26	1.17	0.020	56	20
GPRASP1	0.348	rs17340189	0.11	1.15	0.090	6	5
ADRA2A	0.355	rs7908645	0.34	1.13	0.056	15	8
SH3GL3	0.441	rs8025427	0.42	1.15	0.018	68	31
GNAS	0.508	rs1022697	0.43	1.13	0.032	50	21
SH3GL2	0.562	rs10810813	0.16	0.83	0.019	162	43
PDE4D	0.677	rs17799450	0.08	1.34	0.015	312	74

*adjusted for the number of independent SNP within loci (see last column, independent SNPs in gene). Full gene names available in online supplement.

Table 3. Diseases and Disorders, and canonical pathways enriched in genes from top complex, identified by Ingenuity Pathway Analysis

IPA Disease/Disorder	**P-value for enrichment**	**# genes**
Respiratory Disease	2.34E-07 - 5.00E-02	3
Cardiovascular Disease	1.12E-05 - 4.31E-02	12
Neurological Disease	2.81E-05 - 3.51E-02	12
Endocrine System Disorders	3.63E-05 - 2.94E-02	10
Immunological Disease	3.63E-05 - 1.56E-02	11
IPA canonical pathway	**P-value for enrichment**	**Ratio** (# genes in top complex/ total # genes in pathway)
Cardiac Hypertrophy Signaling	1.62E-07	0.024 (6/246)
G Beta Gamma Signaling	3.28E-06	0.034 (4/117)
cAMP-mediated Signaling	5.75E-06	0.023 (5/216)
PTEN Signaling	6.83E-06	0.033 (4/123)
Cardiac b-adrenergic Signaling	1.43E-05	0.026 (4/151)

Table 4. Overview of genes* in the identified top complex and their implication in the IPA cardiovascular disease set (CVD), the cardiac hypertrophy signaling pathway (Hypertrophy) and the mouse knock out models of abnormal cardiovascular physiology (MGI).

Gene	SNP	CVD	Hypertrophy	MGI
MAGI1	rs7620106	yes	No	no
PRKACA	rs40282	no	Yes	no
GOPC	rs12664183	no	No	no
ADRB1	rs17653278	yes	yes	yes
MAGI3	rs4839312	yes	no	no
MAGI2	rs2065198	yes	no	no
GRB2	rs7223674	yes	yes	yes
DLGAP2	rs7836020	yes	no	no
ARRB1	rs2279129	no	no	yes
DLG4	rs5412	no	no	no
GNAL	rs2848465	yes	no	no
GIPC1	rs4926215	no	yes	no
DLG1	rs7616531	no	no	no
GPRASP1	rs17340189	no	no	no
ADRA2A	rs7908645	yes	yes	yes
SH3GL3	rs8025427	yes	no	no
GNAS	rs1022697	yes	yes	no
SH3GL2	rs10810813	yes	no	no
PDE4D	rs17799450	yes	no	yes

*Full gene names available in online supplement.

Figure Legends:

Fig 1. Conceptual framework for the integration of GWA data with protein-protein interaction data. The approach consists of three overall steps. First, GWA meta-analysis , SNPs are mapped to genes, genes are scored based on its most significant SNP, and the gene scores are adjusted by the number of independent SNPs mapped to the given gene. Second, protein complexes are assembled based on experimentally derived protein-protein interactions. Finally, the gene-sets underlying the protein complexes are scored based on their genes' p-values and their protein-protein interaction confidence scores.

Fig 2. Top-ranking protein complex from the genome-wide analysis of coronary heart disease in the Nurses' Health and the Health Professionals Follow-Up Studies. The gene products (nodes) are scaled in size according to their significance (larger indicates smaller p-value). Edges between the nodes denote experimentally-derived protein-protein interactions. Red nodes denote genes in the complex with corrected gene-wise p-values < 0.05.
Full gene names available in online supplement.

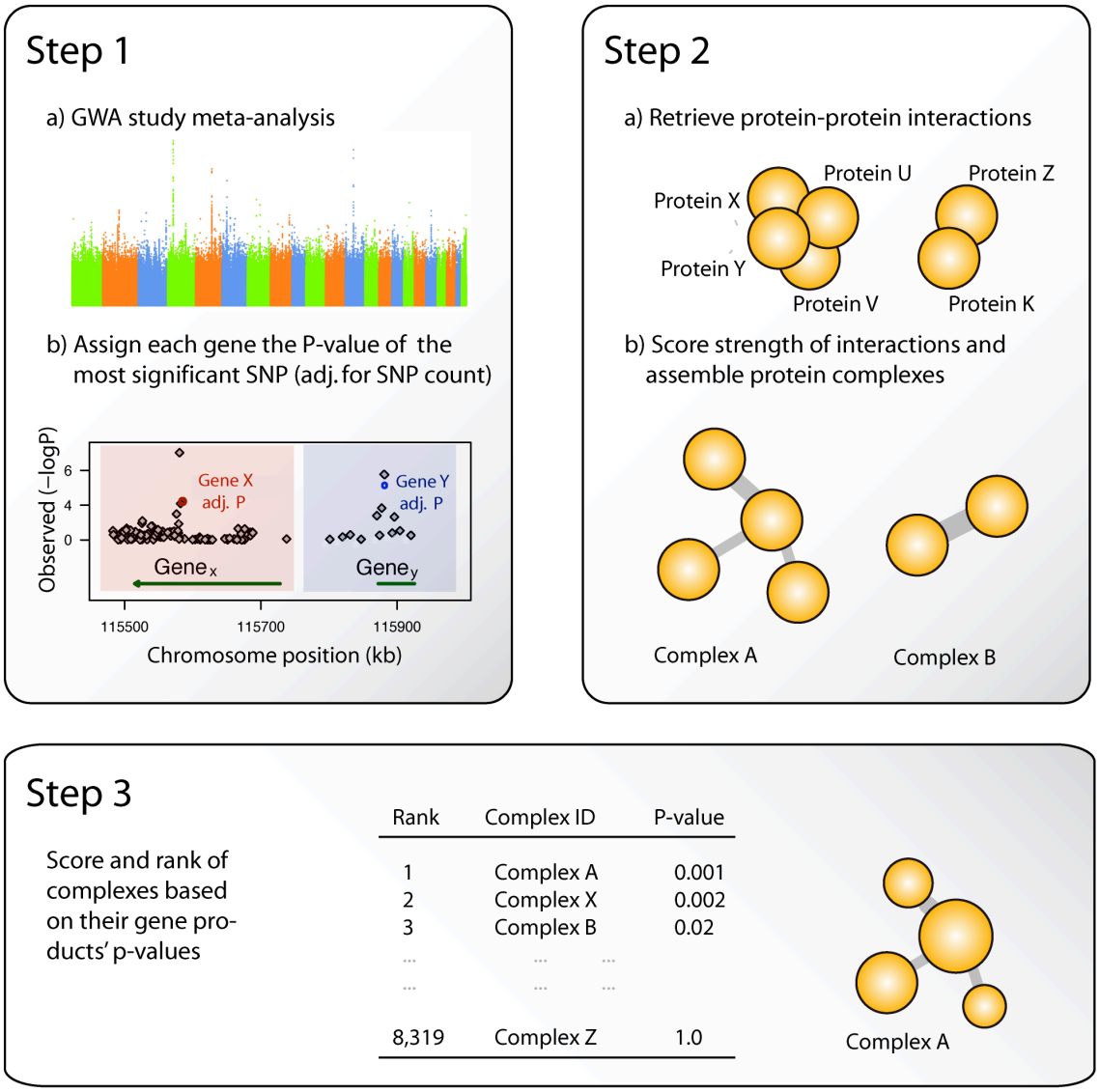

Step 1
a) GWA study meta-analysis
b) Assign each gene the P-value of the most significant SNP (adj. for SNP count)
Observed (−logP)
6
4
0
Gene X
adj. P
Gene Y
adj. P
Gene x
Gene y
115500
115700
115900
Chromosome position (kb)
Step 2
a) Retrieve protein-protein interactions
Protein X
Protein U
Protein Z
Protein Y
Protein V
Protein K
b) Score strength of interactions and assemble protein complexes
Complex A
Complex B
Step 3
Score and rank of complexes based on their gene products' p-values
Rank
Complex ID
P-value
1
Complex A
0.001
2
Complex X
0.002
3
Complex B
0.02
8,319
Complex Z
1.0
Complex A

Figure 1.

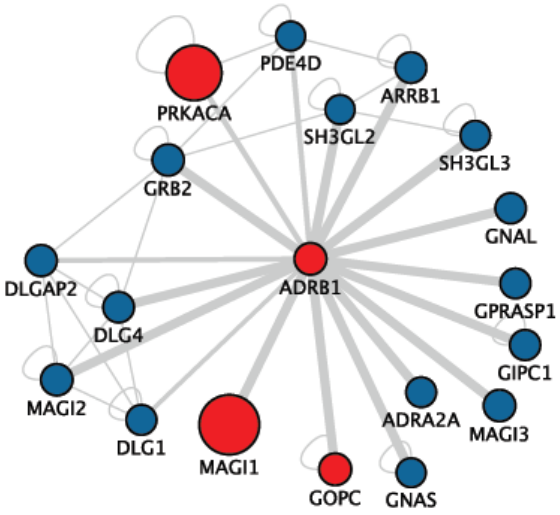


Figure 2.

[Supplementary Files and Tables can be found in online supplement of the paper]

3.4 Paper II - A method for evidence layer-based risk gene mapping

The following paper presents a method that maps candidate disease genes based on a broad range of relevant evidence sources, such as GWA data, propensity to interact with proteins encoded by known disease genes, data from linkage studies, genetic evidence from similar phenotypes, and differential gene expression data. Our framework is very general and can incorporate essentially any source of data the researcher finds relevant.

We apply the methodology to bipolar disorder, a complex psychiatric disorder for which GWA studies have only been moderately successful, and follow up with an experimental validation of the top candidate by genotyping 5 polymorphisms in the YWHAH gene in 640 patients and 1,377 controls in two independent cohorts. We thereby prove association between the rs1049583 polymorphism and bipolar disorder.

One of the important features of the method is that it directly points at causal relationships of the highest scoring susceptibility genes. This is illustrated in the paper by discussion of the precise, biochemical context of top ranking genes, which is a key problem needed in current GWA study follow-up, where no immediate link to functional implications result from the pure statistical association.

Rare coding mutations with potentially large effects are typically missed in modestly powered GWA studies. These potentially missed variants are of high value, because they can help decipher the molecular basis of the disease, since they are more straightforwardly transferred to model systems for subsequent biological studies. This fact stresses the need to evolve general methods for genome-scale candidate prioritization, complementary to GWA studies, to identify a more complete range of potential susceptibility variants in highly polygenic common human traits.

Genetic Epidemiology 35:318–332 (2011)

Meta-Analysis of Heterogeneous Data Sources for Genome-Scale Identification of Risk Genes in Complex Phenotypes

Tune H. Pers,[1,2] Niclas Tue Hansen,[1] Kasper Lage,[1,3–6] Pernille Koefoed,[7,8] Piotr Dworzynski,[1] Martin Lee Miller,[1] Tracey J. Flint,[9] Erling Mellerup,[7,8] Henrik Dam,[8] Ole A. Andreassen,[10] Srdjan Djurovic,[10] Ingrid Melle,[10] Anders D. Børglum,[9,11] Thomas Werge,[12] Shaun Purcell,[5,13] Manuel A. Ferreira,[5,13] Irene Kouskoumvekaki,[1] Christopher T. Workman,[1] Torben Hansen,[14,15] Ole Mors,[9] and Søren Brunak[1,6*]

[1]*Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark*
[2]*Institute of Preventive Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark*
[3]*Pediatric Surgical Research Laboratories, MassGeneral Hospital for Children, Massachusetts General Hospital, Boston, Massachusetts*
[4]*Harvard Medical School, Boston, Massachusetts*
[5]*Broad Institute of Harvard and MIT, Cambridge, Massachusetts*
[6]*Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark*
[7]*Department of Neuroscience and Pharmacology, Laboratory of Neuropsychiatry, University of Copenhagen, Copenhagen, Denmark*
[8]*Center of Psychiatry, Rigshospitalet, Copenhagen, Denmark*
[9]*Center for Psychiatric Research, Aarhus University Hospital, Risskov, Denmark*
[10]*Top-project, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway*
[11]*Institute of Human Genetics, University of Aarhus, Aarhus, Denmark*
[12]*Research Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospital, Roskilde, Denmark*
[13]*Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts*
[14]*Hagedorn Research Institute, Gentofte, Denmark*
[15]*Marie Krogh Center for Metabolic Research, University of Copenhagen, Copenhagen, Denmark*

Meta-analyses of large-scale association studies typically proceed solely within one data type and do not exploit the potential complementarities in other sources of molecular evidence. Here, we present an approach to combine heterogeneous data from genome-wide association (GWA) studies, protein-protein interaction screens, disease similarity, linkage studies, and gene expression experiments into a multi-layered evidence network which is used to prioritize the entire protein-coding part of the genome identifying a shortlist of candidate genes. We report specifically results on bipolar disorder, a genetically complex disease where GWA studies have only been moderately successful. We validate one such candidate experimentally, *YWHAH*, by genotyping five variations in 640 patients and 1,377 controls. We found a significant allelic association for the rs1049583 polymorphism in *YWHAH* (adjusted $P = 5.6e{-3}$) with an odds ratio of 1.28 [1.12–1.48], which replicates a previous case-control study. In addition, we demonstrate our approach's general applicability by use of type 2 diabetes data sets. The method presented augments moderately powered GWA data, and represents a validated, flexible, and publicly available framework for identifying risk genes in highly polygenic diseases. The method is made available as a web service at www.cbs.dtu.dk/services/metaranker. *Genet. Epidemiol.* 35:318–332, 2011.

Key words: genome-wide association; meta-analysis; data integration; bipolar disorder; type 2 diabetes

Additional Supporting Information may be found in the online version of this article.
Tune H. Pers and Niclas Tue Hansen to be considered joint first authors.
Contract grant sponsor: Danish Research Council for Technology and Production Sciences; Contract grant number: 274-06-0301; Contract grant sponsors: Villum Kann Rasmussen Foundation; Danish Strategic Research Council; Danish Medical Research Council; Stanley Medical Research Institute; Lundbeck Foundation; Novo Nordisk Foundation; Research Council of Norway; South-Eastern Norway Health Authorities; EU grant PsychGene; Contract grant number: PIAP-GA-2008-218251.
*Correspondence to: Søren Brunak, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark. E-mail: brunak@cbs.dtu.dk
Received 20 October 2010; Revised 8 February 2011; Accepted 10 February 2011
Published online 11 April 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20580

INTRODUCTION

Generally, efforts to find major risk factors for complex, polygenic diseases using GWA studies only have been moderately successful [Maher, 2008; Ropers, 2007]. A widespread explanation is that complex diseases, unlike rare and monogenic diseases, may be caused by multiple individual susceptibility alleles with low effect sizes [Kryukov et al., 2007; Pritchard, 2001; Purcell et al., 2009; Sklar et al., 2008]. The modest contributions of rare risk alleles combined with massive statistical chance-correlations in the genome-wide association (GWA) analyses require large cohorts and refined single nucleotide polymorphism (SNP) selection

strategies to make replicable discoveries [Lettre et al., 2008; Sandhu et al., 2008; Zeggini et al., 2008].

A common strategy to alleviate the problems of detecting common variants is to expand the study cohorts. Crohn's disease and lipid levels are examples where large collaborative meta-analysis has uncovered between 20 and 50% of the genetic variance [Barrett et al., 2008; Teslovich et al., 2010]. However, currently less than 10–20% of the variation observed in highly heritable traits such as human height and type 2 diabetes (T2D) has been accounted for by such efforts [Lango Allen et al., 2010; Voight et al., 2010]. Thus, a large part of the scattered and rare genetic risk factors in complex diseases still remains to be identified [Altshuler and Daly, 2007; Couzin and Kaiser, 2007; Shriner et al., 2007].

We propose a widely applicable approach to identify susceptibility genes based on meta-analysis of heterogeneous molecular data sets, which is complementary to GWA-based meta-analyses that combine data of the same type. We present a flexible method that augments modestly or underpowered GWA data and prioritizes the genome in relation to the phenotype of interest by integrating potentially complementary evidence layers of heterogeneous data sources for a given risk phenotype or disease. They include:

(1) SNP to phenotype associations from GWA studies which represent a rapidly growing resource of unbiased genome-wide associations of common risk-alleles.
(2) Interacting pairs of candidate proteins and proteins encoded by known phenotype susceptibility genes—a type of data which also targets rare alleles. Two proteins involved in the same biological (dys)function often interact. This trend has been confirmed across several species [Gavin et al., 2006; Giot et al., 2003; Li et al., 2004; van Driel et al., 2006] and compared to other types of networks, protein-protein interactions appear to be an excellent data source for phenotypic enrichment [Fraser and Plotkin, 2007].
(3) Data from linkage studies capturing co-segregation of chromosomal regions and disease-specific phenotypes in families is another methodology complementary to GWA analyses capable of identifying regions harboring rare disease-specific variants.
(4) Quantitative data on disease similarities, which may add information that cross normal disease definition barriers [Allan et al., 2008]. Integration of genes involved in diseases similar to the phenotype of interest may supplement the phenotype-specific evidence identified in the above layers 1–3.
(5) Gene expression levels may be affected directly or indirectly by polymorphisms associated with disease. Differential expression between cases and controls may also add important tissue-specific information.

The strategy allows for the integration of complementary data sources in a single meta-analysis leading to a prioritization of protein-coding genes from the entire genome in relation to one particular indication. Although we use these specific data types here, any number and combination of evidence layers can be used. The data sources can be perceived as layers that are collapsed into an integrative meta-layer providing an informed selection of new candidates (Fig. 1). The integration provides a list of high-ranking candidate genes with robust support from the different evidence layers, where a small number of genes subsequently can be subjected to further experimental analysis.

The method suggested here makes use of and includes GWA data on a par with other data types and produces an independent evidence layer from this source of data as well. Previous gene-prioritization approaches have either (a) used one particular data type as a scaffold for integration of other data types, thus constraining the joint ranking of complementary disease associated evidence [Baranzini et al., 2009; Elbers et al., 2009; Emily et al., 2009; Herold et al., 2009; Holden et al., 2008; Holmans et al., 2009; O'Dushlaine et al., 2009; Pan, 2008; Pattin and Moore, 2008; Torkamani et al., 2008], (b) relied solely on one or two molecular data types [Bush et al., 2009; Medina et al., 2009; Saccone et al., 2008; Zamar et al., 2009], or (c) not used GWA data for integrative purposes at all [Adie et al., 2006; Aerts et al., 2006; Ala et al., 2008; Calvo et al., 2006; Franke et al., 2006; Freudenberg and Propping, 2002; Gaulton

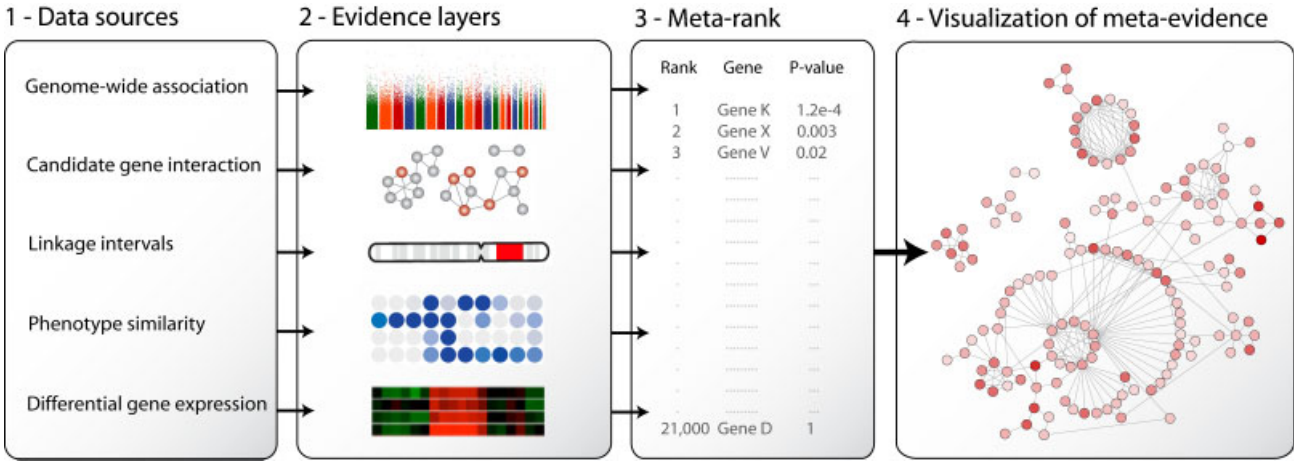


Fig. 1. Integrative approach to gene prioritization for a given disease. In this example, five data sources represent evidence layers which are converted into rank distributions. The evidence layers are subsequently integrated into a single meta-evidence rank, quantifying the likelihood of genes being involved in the disease. The meta-evidence based rank can subsequently be visualized using a protein-protein interaction network. Note that all data types are treated similarly and symmetrically when computing the meta-rank.

et al., 2007; George et al., 2006; Ghazalpour et al., 2006; Goehler et al., 2004; Hristovski et al., 2005; Ideker and Sharan, 2008; Kohler et al., 2008; Lage et al., 2007a; Lesnick et al., 2007; Lim et al., 2006; Linghu et al., 2009; Lopez-Bigas and Ouzounis, 2004; Ma et al., 2007; Perez-Iratxeta et al., 2005; Pujana et al., 2007; Rossi et al., 2006; Sharma et al., 2010; Tiffin et al., 2005; Turner et al., 2003; van Driel et al., 2005; Vanunu et al., 2010; Wood et al., 2007; Wu et al., 2008; Xu and Li, 2006; Yu et al., 2008]. When integrated, most commonly, GWA data have been combined with one or two molecular data types only, thus not taking advantage of the full spectrum of genetic evidence for a specific disease. These approaches include integration of GWA data with biological pathway information from KEGG [Kanehisa et al., 2008], pathways in general [Bush et al., 2009; Elbers et al., 2009; Medina et al., 2009; O'Dushlaine et al., 2009; Zamar et al., 2009], gene ontology databases [Bush et al., 2009; Franke et al., 2006; Holmans et al., 2009; Medina et al., 2009; Wang et al., 2007], a priori known disease susceptibility pathways [Lesnick et al., 2007; Wilke et al., 2008], BioCarta pathways [Medina et al., 2009], protein interaction networks [Baranzini et al., 2009; Elbers et al., 2009; Emily et al., 2009; Franke et al., 2006; Linghu et al., 2009; Pan, 2008; Pattin and Moore, 2008; Torkamani et al., 2008], OMIM [Hamosh et al., 2002], linkage data [Saccone et al., 2008], or a priori defined gene sets [Holden et al., 2008]. Again, as these approaches proceed solely within a few data types, they do not take full advantage of the broad spectrum of genetic evidence for at specific disease. One of the exceptional methods allowing GWA data to be integrated with other genetic evidence sources is the CANDID software tool [Hutz et al., 2008].

We demonstrate and benchmark our approach (denoted MetaRanker) by making an integrative meta-rank analysis of bipolar disorder (BD) and validate a new candidate by investigating a top-ranking gene that has very strong support when integrating BD specific evidence from GWA data, linkage data, known candidate genes, genes from similar diseases, and expression data. In our validation a common allele in 14-3-3eta (*YWHAH*) gene strongly associated to BD, implicating serotonin biosynthesis in the etiology of this common psychiatric disorder, and confirms the strength of conducting meta-analyses not only within one type of data, but also across multiple data sources. As discussed below, *YWHAH* has also been found in another recent study, why our finding must be considered as replication. Together, the genotyping further strengthens the validity of our approach. In addition we apply and benchmark MetaRanker on T2D to showcase the generality of the approach. MetaRanker is available as a web service at www.cbs.dtu.dk/services/metaranker.

METHODS

DATA SOURCES

From WTCCC's website (http://www.wtccc.org.uk/) we downloaded the summary statistics for the study of BD and T2D on the Affymetrix 500k platform. The bipolar study comprised genotypes from 1,868 bipolar patients and 2,938 controls [Wellcome Trust Case Control Consortium, 2007]. The T2D study comprised genotypes from 2,000 diabetic subjects and 2,938 controls.

We reviewed the literature on major depression, mania, and BD to find seed genes associated with these phenotypes. Of particular relevance were the molecular mechanisms related to monoamines, stress response, neurodevelopment, lithium treatment, neuronal signaling, and circadian control. We selected genes critical to one or more of these molecular systems resulting in a list of 34 seed genes (Supplementary Table I). For the T2D analysis we obtained seed gene sets from two recent reviews [Doria et al., 2008; Florez, 2008] (Supplementary Table IX). Protein-protein interactions were retrieved from the InWeb interactome, which is a human protein-protein interaction network based on experiments in both humans and model organisms [Lage et al., 2007a]. InWeb is the outcome of an integrative pipeline assembling and reducing experimental data from BIND [Bader et al., 2001], DIP [Salwinski et al., 2004], BioGRID [Stark et al., 2006], HPRD [Peri et al., 2003], IntAct [Kerrien et al., 2007], MPact [Guldener et al., 2006], MPPI [Mewes et al., 2006], DOMINO [Ceol et al., 2007], Corum [Ruepp et al., 2010], PDZBase [Beuming et al., 2005], and MINT [Chatraryamontri et al., 2007]. The final interactome contained 173,500 unique scored protein-protein interactions derived from 40,085 articles covering 13,000 proteins. All interactions were scored and benchmarked against a gold standard to ensure that we only used high-confidence interactions in the analysis.

We retrieved records related to BD from the OMIM database. The records describing major affective disorder (bipolar disorder) cited more than 150 genetic studies of the disease. In consensus, the major affective disorder entries highlight seven genetic regions being particularly important for the disease (Supplementary Table III). Linkage peaks used in the T2D analysis where retrieved from a recent review [Lillioja and Wilton, 2009] (Supplementary Table X).

The GeneCards encyclopedia [Rebhan et al., 1998] (http://www.genecards.org) is a comprehensive resource for gene-related information. We mined this resource for relationships between genes and BioAlma disease keywords (download date 11/12/2009). All relationships were extracted into a standardized format.

We downloaded the data from the BD gene expression study [Ryan et al., 2006] via the Gene Expression Omnibus database [Edgar et al., 2002]. The data set comprised two separate sets of post-mortem samples: 61 samples from the dorsolateral prefrontal cortex and 21 samples from the orbitofrontal cortex. We pooled the two data sets to get a total of 40 bipolar samples and 42 controls. The T2D expression data sets were downloaded from the Diabetes Genome Anatomy Project's website (http://www.diabetesgenome.org). The skeletal muscle data set [Mootha et al., 2003] comprised 17 normal glucose tolerance controls and 18 subjects with T2D. The pancreatic islet gene expression data set [Gunton et al., 2005] comprised 7 normal glucose tolerance controls and 5 subjects with T2D. Each data set had measurements of hybridization levels of at least 22,283 probe sets.

CONSTRUCTION OF EVIDENCE LAYERS

The GWA-layers were constructed by calculating association *P*-values using Fisher's test on a 2×3 table assuming an additive genetic model. If any cell had five or fewer observations, we omitted that SNP. We mapped each SNP to an Ensembl gene identifier using 70 kb base pair upstream and 20 kb downstream flanking regions and

the genes were scored by considering the most significant associated SNP. Each gene score was then adjusted for the effective number of independent SNPs in the given gene by use of an approach described in [Galwey, 2009] and implemented by us into a efficient C++ method (available upon request). We used HapMap phase III genotypes as input for that method. Finally, the GWA evidence layers were created by ranking all genes based on their adjusted score.

The candidate gene interaction layers were constructed by counting the number of proteins each InWeb protein interacted with, and how many of these proteins were products of BD or T2D seed genes. Using a cumulative hypergeometric distribution, we calculated the likelihood of a protein interacting with that number or more seed gene products given its number of interaction partners. The evidence layer was produced by ranking all proteins based on the significance of their interaction with the known seed proteins.

The linkage layers were assembled by retrieving all protein-coding Ensembl genes within each of the linkage regions associated with BD or T2D. In a given linkage region with LOD score l and k genes, all genes were assigned a score equal to l/k. Hence, genes within a region with relatively few genes and a high LOD score were assigned a higher prior probability of being true disease genes than genes within longer regions or regions with lower LOD scores. If no LOD scores were available all linkage regions we assigned a LOD score of one. Based on these probabilities, we ranked the genome into the BD and T2D specific linkage evidence layers.

The disease similarity layers were constructed by counting the number of genes labeled with each BioAlma keyword. For each keyword, we counted the number of genes that both were associated with that keyword and ''bipolar disorder'' (in the T2D analysis ''diabetes mellitus non-insulin-dependent'' or ''diabetes mellitus''). The significance of the co-occurrence of a keyword with these disease terms was calculated using a cumulative hypergeometric distribution. The phenotype association layers were constructed by rank ordering all genes based on the significance of co-occurrence between their associated keywords and the BD and T2D disease terms.

The expression layers were assembled by normalizing the original gene expression studies using the robust multiarray average method [Irizarry et al., 2003] and applying a Student's t-test to calculate differential expression between cases and controls. Subsequently, we mapped probe sets to Ensembl gene identifiers and produced the evidence layers by ranking the genes based on the differential expression levels.

INTEGRATION OF EVIDENCE LAYERS

Each evidence layer was produced by ranking all genes based on their probability of being associated with the studied phenotype given the data, i.e. ranking the individual data sources after P-values in increasing order and giving them rank scores equalling their rank divided by the total number of genes in the specific experiment. If genes were missing in specific layers they were added and obtained a rank score equal to 1. Next, we combined the GWA, candidate gene interaction, disease similarity, linkage, gene expression layers into a single meta-rank by multiplying each gene's rank score from all five layers. In algebraic terms the final gene score S_x is written as $S_x = \prod_{i=1}^{j} rank_x^i / n_{genes}$, where S_x is gene x's meta-rank score, j is the number of evidence layers, $rank_x^i$ is the rank of gene x in evidence layer i, and n_{genes} is the number of genes in evidence layer i. We calculated permutation-based P-values by comparing the observed meta-rank scores to a distribution of 10^7 randomized meta-rank scores, thus giving an estimate of the probability of achieving a given score by chance alone. The randomized meta-rank distribution was calculated by shuffling gene ranks in each input layer and recording of the best score in the subsequent meta-score calculation 10^7 times.

ENRICHMENT OF META-RANK TOP RESULTS IN SKLAR ET AL. GWA STUDY

Sklar et al. [2008] performed a large GWA study of BD in parallel to the WTCCC. All SNPs in their data set were mapped to genes using 5,000 base pair flanking regions, and the genes were ranked according to the most significant associated SNP. We defined the list of candidates as the 63 genes getting higher than expected P-values in the integrative genome-wide rank. Using Fisher's method for combining P-values, we calculate the combined significance of the most associated SNPs in the candidate genes. We randomly generated 100,000 sets of 63 genes in order to estimate the likelihood of achieving the combined P-value or better.

VISUALIZATION OF RESULTS

In order to visualize the proteins interacting with candidates involved in molecular mechanisms theoretically related to BD, we constructed a first-order protein-protein interaction network extending from all seed gene products. Input proteins with more than 20 interaction partners without interactions with other nodes in the network were removed. The network was visualized using Cytoscape [Cline et al., 2007] and colored according to ranks from the evidence layers.

EXPERIMENTAL VALIDATION

We included two independent case-control cohorts: one from Denmark and one from Norway. The sample characteristics are shown in Table I. *The Danish sample* (DK) included 421 unrelated BD patients. Of these 256 patients were recruited by the Danish Psychiatric Biobank from the psychiatry departments in the Copenhagen area.

TABLE I. Characteristics of the two case-control samples

		Controls		Cases	
		Male	Female	Male	Female
Danish sample	n	572	577	184	237
	Mean age	40 (±11)	40 (±12)	43 (±14)	45 (±14)
	AFA			34 (±13)	32 (±13)
Norwegian sample	n	104	124	101	118
	Mean age	41 (±10)	39 (±10)	41 (±13)	43 (±13)
	AFA			31 (±12)	30 (±11)

The table includes the number of subjects (n), the mean age at the time of last assessment for each sample and the age at first admission (AFA).

All patients had been clinically diagnosed with BD according to the ICD-10 disease classification system. The rest of cases included 165 unrelated patients collected in Denmark. All cases were diagnosed with SCAN [Wing et al., 1998] interviews fulfilling a best estimate diagnosis of bipolar affective disorder according to the ICD-10-DCR [WHO, 1993] and the DSM-IV [American Psychiatric Association, 1994] criteria for bipolar I disorder. The healthy control group ($n = 1{,}149$) was recruited among 15,000 healthy blood donors from the Danish Blood Donor Corps collected in the Copenhagen area in 2005. The donor corps includes more than 5% of the Danish population that donate blood on a voluntary and unpaid basis. Apparent behavioural abnormality was an exclusion criterion. All cases and controls were of Danish Caucasian origin. *The Norwegian sample* (NO) included 219 unrelated patients diagnosed with bipolar I disorder (127) or bipolar II disorder (79), or bipolar NOS (13) according to DSM-IV using structural clinical interviews for DSM-IV (SCID). The 228 controls were randomly selected from statistical records of persons from the Oslo area born in Norway and underwent screening interviews. All cases and controls were of Norwegian Caucasian origin. The Danish Scientific Committees, the Danish Data Protection Agency, the Norwegian Scientific-Ethical Committees and the Norwegian Data Protection Agency approved the study. All patients and controls have given written informed consent prior to inclusion into the project.

We used data from the HapMap CEU population website (www.hapmap.org; HapMap Data Rel 22/phaseII Apr07) to identify tag SNPs that covered most of the common variants within the *YWHAH* gene region (including 3 kb upstream of the 5′ end and 1 kb downstream of the 3′ end of the gene). Tag SNP selection was performed using pair-wise tagging, with $r^2 \geq 0.8$ and minor allele frequency ≥ 0.05. We identified 5 tag SNPs covering all 17 informative SNPs in the *YWHAH* gene with an average $r^2 = 0.951$. The selected SNPs were: (1) rs3761432 at the 5′ end (tagging rs3827334 and rs933226); (2) rs929036 at the 5′ end; (3) rs2267172 at the 5′ end; (4) rs2858753 located in intron 1 (tagging: rs3747158, rs4820059, rs2246704); and (5) rs1049583 located in the 3′UTR (tagging: rs2301415, rs7290696, rs8141011, rs7291050, rs2858750, rs2853884, rs2853887). We extracted genomic DNA from whole blood and analyzed the five tag SNPs using TaqMan genotyping assays (Applied Biosystem, Lincoln, CA) following the manufacturer's instructions on an ABI7900HT system (Applied Biosystem). The assay verifications were as follows: (1) rs3761432 accuracy was 99.7% (21.1% rerun of the samples) and a success rate of 99.3%; (2) rs929036 accuracy was 100% (7.2% rerun of the samples) and success rate was 97.9%; (3) rs2267172 accuracy was 100% (13.1% rerun of the samples) and success rate was 99.9% (this SNP was only analyzed in the Danish sample); (4) rs2858753 accuracy was 100% (8.1% rerun of the samples) and a success rate of 99.6%; (5) rs1049583 accuracy was 99.5% (8.2% rerun of the samples) and success rate was 99.4%. All genotypes were tested for Hardy-Weinberg equilibrium ($P < 0.01$ was considered to be in Hardy-Weinberg disequilibrium).

The linkage disequilibrium block structure was assessed with Haploview 4.1 using the default setting for all three samples and in the combined sample. One block was identified between rs2858753 and rs1049583 (D′ on 0.98 and r^2 on 0.53). Haplotypes between these two markers were estimated in Haploview (Supplementary Tables VI–VIII).

Pearson's chi-square tests were used to compare genotype distributions using the computer software SPSS version 15 (data not shown). We adjusted all *P*-values for multiple testing using Bonferroni correction. Allele frequencies and haplotypes were analyzed using Haploview, and all *P*-values for allele frequencies and haplotypes were adjusted for multiple testing using the permutation test (100,000 estimates) in Haploview. Odds ratios were calculated using "Calculator for confidence intervals of odds ratio in an unmatched case control study" (http://www.hutchon.net/ConfidOR.htm). A recent report found no population stratification between our two Scandinavian subsamples [Kahler et al., 2009].

COMPARISON WITH CANDID

The web server tool CANDID (https://dsgweb.wustl.edu/hutz/candid.html) was used for benchmarking (CANDID database version 6 from March 2010). For the BD comparison we specified the same BD-specific keywords for the CANDID literature layer as we used for finding the seed genes for our analysis (bipolar disorder, monoamines, stress response, neurodevelopment, lithium treatment, neuronal signaling, circadian control). For the expression layer we selected whole brain as tissue (code 29). For the association layer we used the same GWA results as used in our approach. For the linkage layer we uploaded all genes from the same linkage regions as used in our approach. We included both the domain and conservation layer in the final CANDID analysis and all layers were weighted equally.

For the T2D comparison keywords selected for the literature layer were: T2D, insulin resistance, insulin deficiency, and beta cell failure. We used skeletal muscle, pancreas, pancreatic islets, and liver (codes 75, 57, and 45, respectively) as tissue codes. For the linkage layer we uploaded all genes from the same linkage regions as used in our T2D analysis. Again we included both the domain and conservation layer in the final CANDID analysis and all layers were weighted equally. We compared the CANDID results to our T2D analysis in which we used the skeletal muscle expression data set as the differential gene expression layer.

RESULTS

We evaluated our approach by applying it to BD, a psychiatric disease for which a recent GWA meta-analysis [Ferreira et al., 2008] covering 10,500 subjects reported two disease loci that reached genome-wide significance. In addition we have applied the method to T2D, which has a higher prevalence than BD and to date has approximately 37 significant common variant associations [Voight et al., 2010].

The evidence layers for BD were constructed from five different disease-specific data sources: GWA data, known bipolar candidate genes, linkage regions associated with BD, known susceptibility genes from diseases similar to BD, and gene expression data from post-mortem brain samples of bipolar patients. From the WTCCC website, we downloaded summary statistics of genotypes from 1,868 bipolar and 14,311 combined controls [Wellcome Trust Case Control Consortium, 2007]. We selected a set of well-established genes from each of the known BD susceptibility pathways, which resulted in a list of 34 seed genes

(*Methods* and Supplementary Table I). In order to score all proteins according to the likelihood that they interacted with any of the seed genes, we used an updated version of the previously introduced InWeb [Lage et al., 2007a] protein-protein interaction network, comprising 173,500 physical interactions between 13,000 human proteins. From OMIM [Hamosh et al., 2002] we retrieved seven cytogenetic bands associated with BD. We mined the GeneCards resource for 3,785 BioAlma disease terms which mapped to a total of 13,891 genes in order to identify genes co-occurring with diseases similar to BD. Finally, we retrieved a gene expression data set with 82 samples from post-mortem brains of which 40 were bipolar patients and 42 were controls [Ryan et al., 2006].

CONSTRUCTION OF EVIDENCE LAYERS FOR BIPOLAR DISORDER

GWA layer. The summary statistics from the WTCCC study of BD included 459,293 SNPs. These SNPs were mapped to 20,049 protein-coding genes. Each gene was scored based on the SNP with the lowest *P*-value within the most extreme gene transcript including 70 kb upstream and 20 kb downstream flanks. Afterwards each gene score was adjusted for the number of independent SNPs mapped to it (*Methods*) and all genes were ranked according to their adjusted scores.

Candidate gene interaction layer. We ranked each protein in the InWeb interactome based on its hypergeometric enrichment of known seed gene products in its first-order protein interaction neighborhood. The proteins most significantly enriched for interactions with the BD seed gene products were corticotropin releasing hormone (*CRH*), urocortin precursor (*UCN*), and the aromatic-L-amino-acid decarboxylase (*DDC*).

Linkage layer. OMIM associates the linkage regions 2q22-q24, 6q23-q24, 16p12, 18p, 21q22.13, 22q12, and Xq28 to major affective disorders including BD. These regions span 946 genes resulting in an average of 135 genes in each region (see details in Supplementary Table III). We ranked all genes based on their occurrence in, and size of, the linkage regions.

Disease similarity layer. Within the GeneCards respository 46 genes were labeled with the ''bipolar disorder'' term. Out of 3,875 BioAlma terms, 164 terms co-occurred at least once with ''bipolar disorder.'' The most significantly co-occurring disease terms were schizophrenia, mood disorders, and involutional depression (for details see Supplementary Table II). In total, we ranked 7,555 genes which were associated with 164 disease terms co-occurring with ''bipolar disorder.''

Differential gene expression layer. Differential expression was calculated based on normalized expression of 22,283 probe sets in the gene expression study by Ryan et al. [2006] and used to rank a total of 13,068 genes.

INTEGRATION OF EVIDENCE LAYERS FOR BIPOLAR DISORDER

We combined the ranks from each of the five evidence layers to produce a meta-rank that could be used to prioritize candidates for BD. An overview of the evidence may be obtained using the protein-protein interaction network created using the 34 seed genes for BD. Figure 2A shows the 20 best candidates from the integrative approach, whereas Figure 2B depicts the structure of the protein-protein interaction network comprising 247 proteins and 291 interactions. Note that the protein interaction network is used here only as a scaffold for visualization of the joint significance of the genes across all ranks. The interaction data enters the ranking as phenotype association evidence on an equal footing with all other data types.

The most significant new candidate that had the best support across all five evidence layers was *YWHAH* (permutation $P = 0.04$). Figure 3 summarizes the data supporting the rank of *YWHAH* in each layer. Figure 3A shows the slightly skewed allele distribution of the rs6518758 SNP, which was not found in the original GWA study since it was nonsignificant after correction for multiple testing (unadjusted $P = 0.05$). The protein *YWHAH* interacts with tryptophan hydroxylase 1 (*TPH1*), which is the rate-limiting step in the serotonin biosynthesis [Ichimura et al., 1995; Serretti et al., 2001] and with the glucocorticoid receptor (*NR3C1*), which has been associated to major depression and BD [Kim et al., 2005; Spiliotaki et al., 2006; Wakui et al., 1997] (Fig. 3B). The *YWHAH* gene is located in the cytogenetic band 22q12.3 that was associated to BD [Kelsoe et al., 2001]. This is the most gene rich region of the seven cytogenetic bands that have been linked to BD (Fig. 3C). *YWHAH* has previously been associated to schizophrenia, but these studies did not investigate the association in bipolar patients [Bell et al., 2000; Toyooka et al., 1999]. Schizophrenia is the disease term that most significantly co-occurs with ''bipolar disorder'' in the phenotype association evidence layer (Fig. 3D). Figure 3E shows a histogram of the expression levels of *YWHAH* mRNA for bipolar patients and controls. *YWHAH* is among the 5% most significantly differentially expressed genes and clearly tends to have lower expression in bipolar patients.

Interestingly, a validating observation of the meta-rank (Fig. 2A) shows that SNPs in the top-scoring genes tend to be more significantly associated to BD than randomly selected genes in an independent BD GWA study [Sklar et al., 2008] ($P = 0.055$).

EXPERIMENTAL VALIDATION OF BIPOLAR DISORDER GENE

To investigate the association of variants within *YWHAH* to BD, we genotyped six markers in one Danish and one Norwegian sample in a total of 640 BD patients and 1,377 controls. Table I characterizes the studied individuals. We selected five tag SNPs covering the *YWHAH* gene region (3 kb 5′ and 1 kb 3′ of the gene region, respectively). The genotype distributions were in Hardy-Weinberg equilibrium among controls both within each cohort, and in the combined sample.

Table II summarizes the allele distributions of the five selected SNPs in each of the two case-control samples. We found significant differences in minor allele frequencies between cases and controls for the marker rs1049583 in both samples and in the combined sample ($P = 5.0e{-}4$) with an odds ratio of 1.29 [1.12–1.48] (adjusted for multiple testing using Bonferroni correction, $P = 5.6e{-}3$).

We analyzed the markers for haplotypes using the Haploview software (with default settings) and identified two haplotype blocks, one between rs3761432 and rs929036, and one between marker rs2858753 and rs1049583 (block

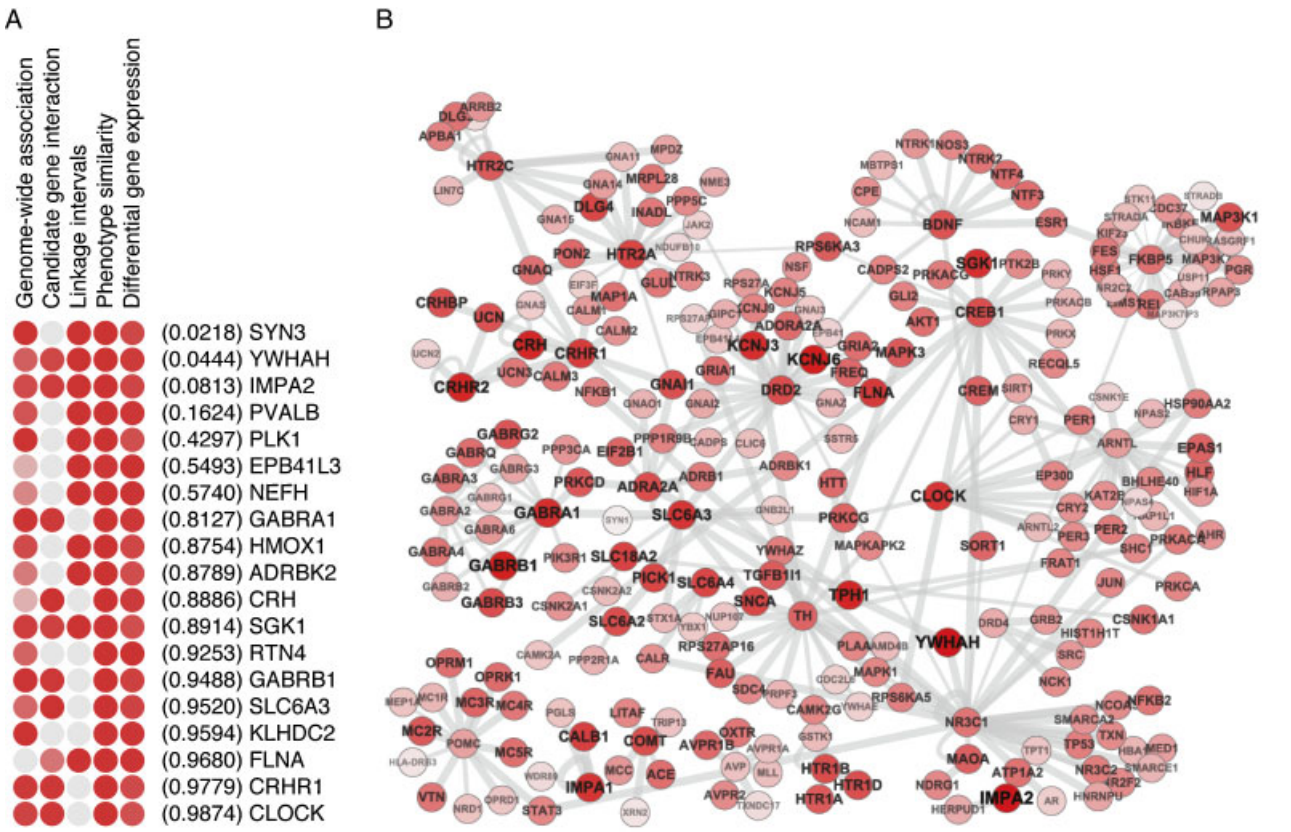


Fig. 2. Integration of genome-wide association data, candidate gene interaction, linkage intervals, disease similarity, and differential gene expression data for bipolar disorder. (A) List of the top 20 candidates for bipolar disorder. The matrix shows the contribution from the individual layers of evidence. Permutation based *P*-values are noted in parentheses. (B) Protein-protein interaction network based on seed genes for bipolar disorder visualizing the meta-rank. The color scheme goes from strong evidence of association (red) to no evidence of association (light gray).

two). The CG haplotype in first block was associated with BD in the Norwegian sample (adjusted $P = 0.018$), but this was not confirmed in the Danish sample (Supplementary Table VI). The haplotype between the two minor alleles, GA, in the latter block was significantly associated with BD in both samples and in the combined sample (adjusted $P = 0.0059$) (Supplementary Table VI).

INTEGRATION OF EVIDENCE LAYERS FOR T2D

We repeated the analysis for T2D, a complex heterogenic disease that is based on a complex interplay between both rare and common variants, which when exposed to certain environmental factors cause T2D. To construct a GWA-based evidence layer, we used genotype data from the WTCCC T2D study. The candidate gene interaction layer was again based on the InWeb protein-interaction database now using a recently published list of 42 T2D susceptibility genes seeding the analysis [Doria et al., 2008]. This seeding gene set contained both monogenic and common T2D susceptibility genes (Supplementary Table IX). To assess the sensitivity and impact of the seed genes on the final gene rank, we also used a list of 20 genes at or nearby loci detected in recent T2D GWA studies (Supplementary Table IX) [Florez, 2008]. The linkage evidence layer was constructed from 18 T2D linkage regions reported in at least 10 linkage reports as well as their LOD scores as included in a recent review (Supplementary Table X) [Lillioja and Wilton, 2009]. Genes within linkage intervals were both weighted based on the number of genes within the given interval and the LOD score for that interval. The disease similarity layer was constructed by mining GeneCards for all BioAlma terms co-occurring with ''diabetes mellitus'' or ''diabetes mellitus non-insulin-dependent.'' The most highly co-occurring terms were ''diabetes mellitus insulin-dependent,'' ''insulin resistance,'' and ''insulin sensitivity'' (Supplementary Table XI). Finally, the gene expression layer was build using a highly cited skeletal muscle expression study carried out in T2D subjects and healthy controls [Mootha et al., 2003].

After constructing the five T2D evidence layers and collapsing all genome ranks into a single meta-rank, we identified solute carrier family 2 member 4 (*SLC2A4*), as the top-ranking gene (Supplementary Table XII). It was followed by glycogen synthase 1 (*GYS1*) the and transcription factor 7 like 2 (*TCF7L2*) genes, the latter being expected as the *TCF7L2* gene was discovered in the GWA study used in our analysis and among the seed genes for the candidate gene interaction layer.

In order to assess the robustness of the meta-rank in relation to changes in the individual seed data sets we repeated the analysis several times changing the different data sets in the individual evidence layers. First, we seeded the candidate gene evidence layer with the

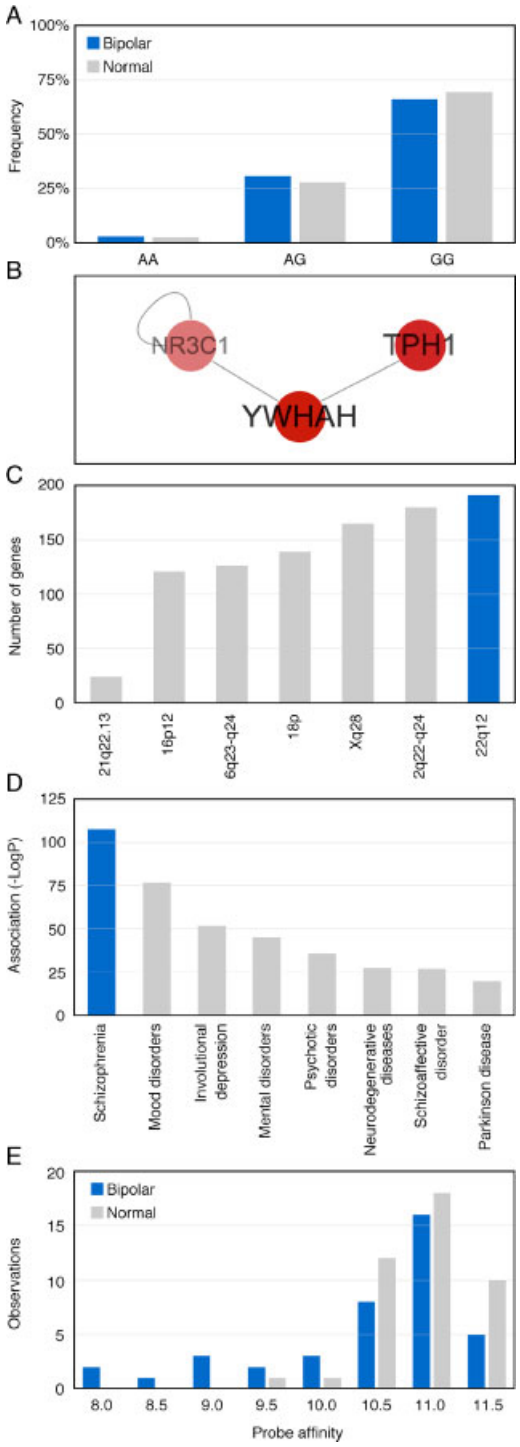


Fig. 3. Summary of the evidence indicating involvement of *YWHAH* in bipolar disorder. (A) Distribution of genotypes of the SNP rs9609396 in bipolar and normal subjects. (B) The local protein-protein interaction network around *YWHAH*. Proteins are color-coded according to their position in the meta-rank. (C) The number of genes in each linkage region associated with bipolar disorder. *YWHAH* is located in the cytogenetic band 22q12.3. (D) The disease terms most significantly co-occurring with bipolar disorder. (E) Expression of the *YWHAH* gene in post-mortem brain samples of bipolar and normal subjects.

GWA-based list of just 20 genes extracted from a review by Florez [Florez, 2008] and found that *GYS1*, *TCF7L2*, and *SLC2A4* again were the highest ranking genes. Second, we exchanged the expression data with a highly cited pancreatic islets data set [Gunton et al., 2005]. In the analysis based on the Doria et al. seed gene set, the insulin receptor gene (*INSR*), *TCFL2*, and the hepatocyte nuclear factor 4 alpha (*HNF4A*) were the highest ranking genes. Reassuringly, in the analysis based on the Florez seed gene set, these genes also ranked among the top genes (ranks 2, 1 and 5, respectively), demonstrating that the top findings are indeed robust across a number of data sets.

COMPARISON WITH AN INDEPENDENT T2D GWA STUDY

In order to justify that an integrated approach yields more robust results and novelty compared to other modestly powered GWA studies, we compared the final rank of our meta-analysis to the results from the Diabetes Genetics Initiative GWA for T2D [Saxena et al., 2007]. As shown in Figure 4A, we found that for all thresholds our method had considerably superior performance in finding known T2D susceptibility genes (for benchmark set construction see Methods) than the gene rank produced solely by the original Diabetes Genetics Initiative GWA study. Note that when disabling the phenotype-similarity layer our integrative method exclusively relies on experimental evidence layers, and does not include text mining to rediscover known genes. The analysis again shows that the diversity of data types is useful, as compared to applying just a single evidence type.

COMPARISON WITH THE CANDID GENE-PRIORITIZATION METHOD

Gene prioritization methods are difficult to compare and benchmark as they often use widely different types of information to seed the analysis. As described above, another problem is that many methods using GWA data work from a limited, predefined set of pathways or interaction networks, thus not extending the search for novel disease genes to the entire protein-coding genome as in the method described in this paper. Availability of methods is another serious constraint.

One of the most generally available and user friendly method is the Endeavour gene-prioritization software tool [Aerts et al., 2006], which can be seeded by a user-selected gene set as in our method, but it does not incorporate GWA data. To our knowledge the only genome-wide gene prioritization method, which allows researchers to upload their GWA data to a publicly available software tool is CANDID—an integrative method that uses genetic data sources along with text mining of PubMed abstracts [Hutz et al., 2008]. For benchmarking, we collected 186 BD-susceptibility genes from a very recent review [Luykx et al., 2010] and used the HuGE Navigator [Yu et al., 2008] to extract 86 T2D susceptibility genes that have at least been associated with T2D in 10 independent publications (Supplementary Table XIII). Based on these genes we constructed two benchmark gene sets used as gold standards in the comparison (genes used to seed our analyses were excluded). In the case of BD, where we undertook genotyping of a high ranking gene, the 34 seed genes were selected around 2 years prior to the publication

TABLE II. Summary of five tag SNPs in the YWHAH gene in two Scandinavian samples

		Danish sample			Norwegian sample			Combined sample		
SNP	Allele minor/major	MAF case control	*P* (*P**)	OR 95% CI	MAF case control	*P* (*P**)	OR 95% CI	MAF case control	*P* (*P**)	OR 95% CI
rs3761432	C/T	18.9	NS	ND	23.8	0.002	1.70	20.6	0.023	1.21
		18.0	(NS)		15.6	(0.011)	(1.21–2.38)	17.6	(NS)	(1.02–1.43)
rs929036	G/A	40.8	NS	ND	40.6	NS	ND	40.8	0.054	0.87
		44.6			41.2			44.0	(NS)	(0.76–1.00)
rs2267172	C/G	5.6	NS	ND	ND					
		6.2			ND					
rs2858753	G/C	45.9	0.043	1.18	48.6	NS	ND	46.8	0.009	1.19
		41.9	(NS)	(1.01–1.38)	45.4			42.4	(NS)	(1.05–1.37)
rs1049583	A/G	34.5	0.028	1.21	38.4	0.0088	1.45	35.9	0.0005	1.28
		30.4	(NS)	(1.02–1.43)	30.0	(0.043)	(1.10–1.91)	30.3	(0.0056)	(1.12–1.48)

MAF, minor allele frequency; *P*, *P*-value; OR, odd ratio; CI, confidence interval; NS, non significant; ND, not done.
**P*-value corrected with 100,000 permutations using Haploview 4.1.

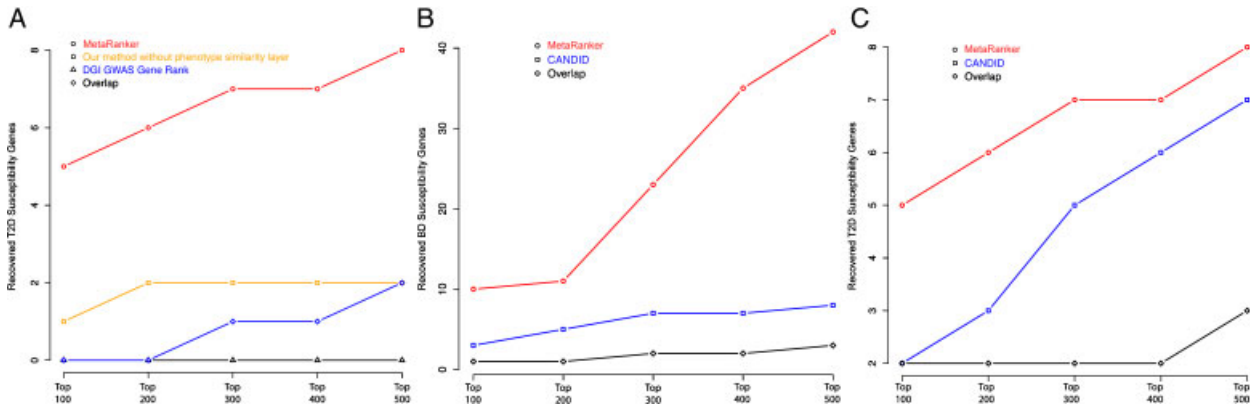


Fig. 4. Overview of comparison between the number of benchmark genes recapitulated by our MetaRanker method (red curves), the DGI GWA study (blue curve in panel A), and the CANDID method (blue curves in panels B and C). The plot in panel (A) shows that MetaRanker predicts more correct benchmark genes compared to a prediction solely based on the DGI GWA study, also when the phenotype similarity layer is left out (orange curve). The number of BD (B) and T2D (C) benchmark genes recapitulated among the top hits returned by MetaRanker and CANDID illustrate how our method increasingly identifies more benchmark genes than CANDID. The black curves reflect the overlap of correctly predicted genes between the different approaches.

of the above mentioned review. As can be seen on Figure 4B and C our method was for both BD and T2D able to detect a considerably larger proportion of benchmark genes (see *Methods* for additional benchmark details).

DISCUSSION

In this paper, we present a systems biology approach to integrate heterogeneous data sources ranging from prior knowledge of the molecular disease models, high-throughput expression data to disease-related protein complexes. The different data sources represent evidence layers symmetrically, providing a prioritization of the genome that enables informed selection of candidates for thorough subsequent genotyping. The in-depth experimental replication of one BD risk gene confirms the feasibility of conducting such integrative meta-analyses, thus adding value to the data produced in the numerous GWA studies completed so far.

Our approach has the advantage that it, in contrast to GWA studies, points directly to risk genes, and places them in a functional context hinting at the molecular etiology of the phenotype in question. (Please refer to Supplementary Note 6 for a discussion on *YWHAH's* putative role in BD.) Androgen and estrogen receptors have *YWHAH* binding motifs [Zilliacus et al., 2001] providing encouraging in-depth analysis of gender specificity of the association between the gene and BD. We therefore also analyzed the gender-specific samples independently and found that in males the allelic association for both rs2858753 and rs1049583 were highly significant (adjusted $P = 0.0027$ and $P = 3.0e{-}5$) (Supplementary Tables IV and V) with an OR of 1.44 and 1.63, respectively. These findings indicate that the alleles of *YWHAH* associated with BD might be interacting with alleles of steroid hormone receptors. Our findings are further supported by a recent study carried out in parallel

with our work, which also showed an association between BD and a *YWHAH* variant located 3,7 kb upstream of our rs1049583 association [Grover et al., 2009].

In the T2D analysis we identified the ectonucleotide pyrophosphatase/phosphodiesterase 1 (*ENPP1)* gene as one of the putative T2D susceptibility genes. The gene product of *ENPP1* is a candidate gene in insulin resistance [Maddux et al., 1995], but has not been identified in T2D GWA studies so far, possibly due to the genetic heterogeneity of T2D [Prudente et al., 2009]. Notably, in our analysis *ENPP1* received support from four of our five evidence layers: In the GWA-based layer the gene was located above the 94.4 percentile within gene rank distribution having a SNP-count and linkage disequilibrium adjusted *P*-value of 0.05; in the candidate gene interaction layer it ranked above the 99.8 percentile since its only interacting protein was the gene product of the known T2D susceptibility gene *INSR*; in the phenotype similarity layer it ranked above the 99.9 percentile as it has previously been annotated to insulin resistance, which is closely related to T2D measured in terms of co-occurring BioAlma disease terms; and finally, in the linkage layer it was located above the 99.7 percentile as it is located within a T2D linkage region replicated through 10 independent studies. Our finding of *ENPP1* indicates that even for a common complex disease like T2D, simultaneous and symmetric consideration of several different disease-related evidence layers may uncover promising findings left unidentified by GWA meta-analysis.

The approach presented here has strengths, but also disease-specific limitations. The data available for integration depends strongly on the disease in question, as does their quality and relevance. This arbitrariness leaves it up to the user to find the optimal combination. There is no guaranty that the benchmarks presented above will generalize to other diseases. In the ideal case the integration would be based on well-powered GWA study data, a tissue-specific SNP to gene mapping originating from tissue-specific expression quantitative trait loci data, high-confidence candidate gene interaction layer seed genes, differential expression data from a meta-analysis of relevant expression studies, copy-number aberrations, and rare variant data. The complementarities between these data are obviously also disease specific.

Currently, most GWA studies are carried out strictly within specific disease categories. The method we have presented here exploits evidence from overlapping phenotypes which can be used to improve the identification of important candidate genes. *YWHAH* alleles have been previously been associated to schizophrenia, and we show that one SNP in the gene has strong associations to BD in two independent samples. A recent gene expression study showed that *YWHAH* is differentially expressed in post-mortem brains of patients with major depression as well [Kang et al., 2007]. In combination, this should encourage a more dynamic view of disease definitions that would permit overlapping phenotypic traits to have common genetic origins in the relatively gene-poor human genome. Additionally, our genome-wide prioritization of bipolar candidates allows targeted analysis of SNPs in functionally relevant regions of high-scoring genes that are not included on commercial arrays and therefore are missed in most recent GWA studies of BD.

In conclusion, we have performed an integrative meta-analysis of highly different and heterogeneous molecular-level sources of information to identify and rank disease genes according to their likely phenotypic relationships. In the online-available version of the method users can upload GWA study *P*-values, known susceptibility genes, linkage regions, differential expression *P*-values, and keywords for their disease or risk-phenotype (www.cbs.dtu.dk/services/metaranker). In this study we have shown that our method enriches subsequent experimental validation in, for instance, GWA studies for likely disease susceptibility genes by producing a small shortlist of promising candidates. This type of approach seems to be well suited to complex diseases harboring multiple individual susceptibility alleles with low effect sizes since it interrogates both information on common associations and information on rare gene-disease associations simultaneously. Thus, it is likely to identify additional susceptibility genes in situations where no single data type is likely to reveal the complete picture.

ACKNOWLEDGMENTS

We thank B. Bennike and B. Hansen for their excellent laboratory work, and B. Daigle from Stanford University and M.K. Jensen from Harvard School of Public Health for useful comments on the integrative methodology and overall shape of the manuscript. We thank P. Sklar from Massachusetts General Hospital, Harvard Medical School for providing data for validation of the method. We thank patients and controls for their participation in the study, and the health professionals who facilitated our work.

REFERENCES

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. 2006. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22:773–774.

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. 2006. Gene prioritization through genomic data fusion. Nat Biotechnol 24:537–544.

Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F. 2008. Prediction of human disease genes by human-mouse conserved coexpression analysis. PLoS Comput Biol 4: e1000043.

Allan CL, Cardno AG, McGuffin P. 2008. Schizophrenia: from genes to phenes to disease. Curr Psychiatry Rep 10:339–343.

Altshuler D, Daly M. 2007. Guilt beyond a reasonable doubt. Nat Genet 39:813–815.

American Psychiatric Association. 1994. Diagnostic and Statistical Manual of Mental Disorders. Washington: American Psychiatric Association.

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. 2001. BIND—The Biomolecular Interaction Network Database. Nucleic Acids Res 29:242–245.

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. 2009. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18:2078–2090.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van

Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

Bell R, Munro J, Russ C, Powell JF, Bruinvels A, Kerwin RW, Collier DA. 2000. Systematic screening of the 14-3-3 eta (eta) chain gene for polymorphic variants and case-control analysis in schizophrenia. Am J Med Genet 96:736–743.

Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. 2005. PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics 21:827–828.

Bush WS, Dudek SM, Ritchie MD. 2009. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput 368–379.

Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, Mootha VK. 2006. Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet 38:576–582.

Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G. 2007. DOMINO: a database of domain-peptide interactions. Nucleic Acids Res 35:D557–D560.

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. 2007. MINT: the Molecular INTeraction database. Nucleic Acids Res 35:D572–D574.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. 2007. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2: 2366–2382.

Couzin J, Kaiser J. 2007. Genome-wide association. Closing the net on common disease genes. Science 316:820–822.

Doria A, Patti ME, Kahn CR. 2008. The emerging genetic architecture of type 2 diabetes. Cell Metab 8:186–200.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210.

Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. 2009. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33:419–431.

Emily M, Mailund T, Hein J, Schauser L, Schierup MH. 2009. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet 17:1231–1240.

Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, Smoller JW, Grozeva D, Stone J, Nikolov I, Chambert K, Hamshere ML, Nimgaonkar VL, Moskvina V, Thase ME, Caesar S, Sachs GS, Franklin J, Gordon-Smith K, Ardlie KG, Gabriel SB, Fraser C, Blumenstiel B, Defelice M, Breen G, Gill M, Morris DW, Elkin A, Muir WJ, McGhee KA, Williamson R, Macintyre DJ, Maclean AW, St Clair D, Robinson M, Van Beck M, Pereira AC, Kandaswamy R, McQuillin A, Collier DA, Bass NJ, Young AH, Lawrence J, Nicol Ferrier I, Anjorin A, Farmer A, Curtis D, Scolnick EM, McGuffin P, Daly MJ, Corvin AP, Holmans PA, Blackwood DH, Gurling HM, Owen MJ, Purcell SM, Sklar P, Craddock N. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 40:1042–1044.

Florez JC. 2008. Clinical review: the genetics of type 2 diabetes: a realistic appraisal in 2008. J Clin Endocrinol Metab 93: 4633–4642.

Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78:1011–1025.

Fraser HB, Plotkin JB. 2007. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol 8:R252.

Freudenberg J, Propping P. 2002. A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18:S110–S115.

Galwey NW. 2009. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet Epidemiol 33:559–568.

Gaulton KJ, Mohlke KL, Vision TJ. 2007. A computational system to select candidate genes for complex human traits. Bioinformatics 23:1132–1140.

Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631–636.

George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA. 2006. Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res 34: e130.

Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2:e130.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley Jr RL, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. 2003. A protein interaction map of *Drosophila melanogaster*. Science 302: 1727–1736.

Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, et al. 2004. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell 15: 853–865.

Grover D, Verma R, Goes FS, Mahon PL, Gershon ES, McMahon FJ, Potash JB. 2009. Family-based association of YWHAH in psychotic bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 150B: 977–983.

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V. 2006. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34:D436–D441.

Gunton JE, Kulkarni RN, Yim S, Okada T, Hawthorne WJ, Tseng YH, Roberson RS, Ricordi C, O'Connell PJ, Gonzalez FJ, Kahn CR. 2005. Loss of ARNT/HIF1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. Cell 122: 337–349.

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55.

Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. 2009. INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 25:3275–3281.

Holden M, Deng S, Wojnowski L, Kulle B. 2008. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics 24:2784–2785.

Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N. 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet 85:13–24.

Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. 2005. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74:289–298.

Hutz JE, Kraja AT, McLeod HL, Province MA. 2008. CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol 32:779–790.

Ichimura T, Uchiyama J, Kunihiro O, Ito M, Horigome T, Omata S, Shinkai F, Kaji H, Isobe T. 1995. Identification of the site of interaction of the 14-3-3 protein with phosphorylated tryptophan hydroxylase. J Biol Chem 270:28515–28518.

Ideker T, Sharan R. 2008. Protein networks in disease. Genome Res 18: 644–652.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

Kahler AK, Otnaess MK, Wirgenes KV, Hansen T, Jonsson EG, Agartz I, Hall H, Werge T, Morken G, Mors O, Mellerup E, Dam H, Koefod P, Melle I, Steen VM, Andreassen OA, Djurovic S. 2009. Association study of PDE4B gene variants in scandinavian schizophrenia and bipolar disorder multicenter case-control samples. Am J Med Genet B Neuropsychiatr Genet 153B:86–96.

Kanehisa M, Goto S, Hattori M, oki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480–D484.

Kang HJ, Adams DH, Simen A, Simen BB, Rajkowska G, Stockmeier CA, Overholser JC, Meltzer HY, Jurjus GJ, Konick LC, Newton SS, Duman RS. 2007. Gene expression profiling in postmortem prefrontal cortex of major depressive disorder. J Neurosci 27:13329–13340.

Kelsoe JR, Spence MA, Loetscher E, Foguet M, Sadovnick AD, Remick RA, Flodman P, Khristich J, Mroczkowski-Parker Z, Brown JL, Masser D, Ungerleider S, Rapaport MH, Wishart WL, Luebbert H. 2001. A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. Proc Natl Acad Sci USA 98:585–590.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. 2007. IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35: D561–D565.

Kim YS, Jang SW, Sung HJ, Lee HJ, Kim IS, Na DS, Ko J. 2005. Role of 14-3-3 eta as a positive regulator of the glucocorticoid receptor transcriptional activation. Endocrinology 146:3133–3140.

Kohler S, Bauer S, Horn D, Robinson PN. 2008. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82:949–958.

Kryukov G, Pennacchio L, Sunyaev S. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727–739.

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. 2007a. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotech 25:309–316.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, MosleyJr TH, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838.

Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM. 2007. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 3:e98.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa L, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann HE, Mohlke KL, Hirschhorn JN. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40:584–591.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain P-O, Han J-DJ, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual J-F, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, van den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. 2004. A map of the interactome network of the metazoan C. elegans Science 303:540–543.

Lillioja S, Wilton A. 2009. Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies. Diabetologia 52:1061–1074.

Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY. 2006. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125:801–814.

Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C. 2009. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10:R91.

Lopez-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32:3108–3114.

Luykx JJ, Boks MP, Terwindt AP, Bakker S, Kahn RS, Ophoff RA. 2010. The involvement of GSK3beta in bipolar disorder: integrating evidence from multiple types of genetic studies. Eur Neuropsychopharmacol 20:357–368.

Ma X, Lee H, Wang L, Sun F. 2007. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23:215–221.

Maddux BA, Sbraccia P, Kumakura S, Sasson S, Youngren J, Fisher A, Spencer S, Grupe A, Henzel W, Stewart TA, et al. 1995. Membrane glycoprotein PC-1 and insulin resistance in non-insulin-dependent diabetes mellitus. Nature 373:448–451.

Maher B. 2008. Personal genomes: the case of the missing heritability. Nature 456:18–21.

Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, Al-Shahrour F, Dopazo J. 2009. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res 37:W340–W344.

Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V. 2006. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res 34:D169–D172.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34:267–273.

O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A. 2009. The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 25:2762–2763.

Pan W. 2008. Network-based model weighting to detect multiple loci influencing complex diseases. Hum Genet 124:225–234.

Pattin KA, Moore JH. 2008. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Hum Genet 124:19–29.

Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. 2005. G2D: a tool for mining genes associated with disease. BMC Genet 6:45.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13:2363–2371.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Prudente S, Morini E, Trischitta V. 2009. Insulin signaling regulating genes: effect on T2DM and cardiovascular risk. Nat Rev Endocrinol 5:682–693.

Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M. 2007. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 39:1338–1349.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 14:656–664.

Ropers H-H. 2007. New perspectives for the elucidation of genetic disorders. Am J Hum Genet 81:199–207.

Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S. 2006. TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34: W285–W292.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res 38:D497–D501.

Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, Bahn S. 2006. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatry 11:965–978.

Saccone SF, Saccone NL, Swan GE, Madden PA, Goate AM, Rice JP, Bierut LJ. 2008. Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. Bioinformatics 24:1805–1811.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449–D451.

Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, Song K, Yuan X, Johnson T, Ashford S, Inouye M, Luben R, Sims M, Hadley D, McArdle W, Barter P, Kesaniemi YA, Mahley RW, McPherson R, Grundy SM, Bingham SA, Khaw KT, Loos RJ, Waeber G, Barroso I, Strachan DP, Deloukas P, Vollenweider P, Wareham NJ, Mooser V. 2008. LDL-cholesterol concentrations: a genome-wide association study. Lancet 371: 483–491.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C,

Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336.

Serretti A, Lilli R, Lorenzi C, Lattuada E, Cusin C, Smeraldi E. 2001. Tryptophan hydroxylase gene and major psychoses. Psychiatry Res 103:79–86.

Sharma A, Chavali S, Tabassum R, Tandon N, Bharadwaj D. 2010. Gene prioritization in Type 2 Diabetes using domain interactions and network analysis. BMC Genomics 11:84.

Shriner D, Vaughan LK, Padilla MA, Tiwari HK. 2007. Problems with genome-wide association studies. Science 316:1840–1842.

Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, Nimgaonkar VL, McQueen MB, Faraone SV, Kirby A, de Bakker PI, Ogdie MN, Thase ME, Sachs GS, Todd-Brown K, Gabriel SB, Sougnez C, Gates C, Blumenstiel B, Defelice M, Ardlie KG, Franklin J, Muir WJ, McGhee KA, MacIntyre DJ, McLean A, VanBeck M, McQuillin A, Bass NJ, Robinson M, Lawrence J, Anjorin A, Curtis D, Scolnick EM, Daly MJ, Blackwood DH, Gurling HM, Purcell SM. 2008. Whole-genome association study of bipolar disorder. Mol Psychiatry 13:558–569.

Spiliotaki M, Salpeas V, Malitas P, Alevizos V, Moutsatsou P. 2006. Altered glucocorticoid receptor signaling cascade in lymphocytes of bipolar disorder patients. Psychoneuroendocrinology 31:748–760.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34:D535–D539.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Shin Cho Y, Jin Go M, Jin Kim Y, Lee JY, Park T, Kim K, Sim X, Twee-Hee Ong R, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Hua Zhao J, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RY, Wright AF, Witteman JC, Wilson JF, Willemsen G, Wichmann HE, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJ, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruokonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BW, Pedersen NL, Pattaro C, Parker AN, Pare G, Oostra BA, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, Meitinger T, McPherson R, McCarthy MI, McArdle W, Masson D, Martin NG, Marroni F, Mangino M, Magnusson PK, Lucas G, Luben R, Loos RJ, Lokki ML, Lettre G, Langenberg C, Launer LJ, Lakatta EG, Laaksonen R, Kyvik KO, Kronenberg F, Konig IR, Khaw KT, Kaprio J, Kaplan LM, Johansson A, Jarvelin MR, Janssens AC, Ingelsson E, Igl W, Kees Hovingh G, Hottenga JJ, Hofman A, Hicks AA, Hengstenberg C, Heid IM, Hayward C, Havulinna AS, Hastie ND, Harris TB, Haritunians T, Hall AS, Gyllensten U, Guiducci C, Groop LC, Gonzalez E, Gieger C, Freimer NB, Ferrucci L, Erdmann J, Elliott P, Ejebe KG, Doring A, Dominiczak AF, Demissie S, Deloukas P, de Geus EJ, de Faire U, Crawford G, Collins FS, Chen YD, Caulfield MJ, Campbell H, Burtt NP, Bonnycastle LL, Boomsma DI, Boekholdt SM, Bergman RN, Barroso I, Bandinelli S, Ballantyne CM, Assimes TL, Quertermous T, Altshuler D, Seielstad M, Wong TY, Tai ES, Feranil AB, Kuzawa CW, Adair LS, Taylor Jr HA, Borecki IB, Gabriel SB, Wilson JG, Holm H, Thorsteinsdottir U, Gudnason V, Krauss RM, Mohlke KL, Ordovas JM, Munroe PB, Kooner JS, Tall AR, Hegele RA, Kastelein JJ, Schadt EE, Rotter JI, Boerwinkle E, Strachan DP, Mooser V, Stefansson K, Reilly MP, Samani NJ, Schunkert H, Cupples LA, Sandhu MS, Ridker PM, Rader DJ, van Duijn CM, Peltonen L, Abecasis GR, Boehnke M, Kathiresan S. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713.

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33:1544–1552.

Torkamani A, Topol EJ, Schork NJ. 2008. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92:265–272.

Toyooka K, Muratake T, Tanaka T, Igarashi S, Watanabe H, Takeuchi H, Hayashi S, Maeda M, Takahashi M, Tsuji S, Kumanishi T, Takahashi Y. 1999. 14-3-3 protein eta chain gene (YWHAH) polymorphism and its genetic association with schizophrenia. Am J Med Genet 88:164–167.

Turner FS, Clutterbuck DR, Semple CA. 2003. POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4:R75.

van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G. 2005. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33:W758–W761.

van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. 2006. A text-mining analysis of the human phenome. Eur J Hum Genet 14:535–542.

Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. 2010. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6:e1000641.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Bostrom K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, GreenT, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jorgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589.

Wakui H, Wright AP, Gustafsson J, Zilliacus J. 1997. Interaction of the ligand-activated glucocorticoid receptor with the 14-3-3 eta protein. J Biol Chem 272:8153–8156.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81: 1278–1283.

Wellcome-Trust-Case-Control-Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

WHO. 1993. The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research. Geneva: World Health Organization.

Wilke RA, Mareedu RK, Moore JH. 2008. The pathway less traveled: moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. Curr Pharmacogenomics Person Med 6: 150–159.

Wing JK, Sartorius N, Üstün TB, editors. 1998. Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN. Geneva: World Health Organization.

Wood JG, Rogina B, Lavu S, Howitz K, Helfand SL, Tatar M, Sinclair D. 2007. The genomic landscapes of human breast and colorectal cancers. Science 318:1108–1113.

Wu X, Jiang R, Zhang MQ, Li S. 2008. Network-based global inference of human disease genes. Mol Syst Biol 4:189.

Xu J, Li Y. 2006. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22: 2800–2805.

Yu S, Tranchevent LC, De Moor B, Moreau Y. 2008. Gene prioritization and clustering by multi-view text mining. BMC Bioinformatics 11:28.

Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. 2008. A navigator for human genome epidemiology. Nat Genet 40:124–125.

Zamar D, Tripp B, Ellis G, Daley D. 2009. Path: a tool to facilitate pathway-based genetic association analysis. Bioinformatics 25:2444–2446.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638–645.

Zilliacus J, Holter E, Wakui H, Tazawa H, Treuter E, Gustafsson JA. 2001. Regulation of glucocorticoid receptor activity by 14-3-3-dependent intracellular relocalization of the corepressor RIP140. Mol Endocrinol 15:501–511.

Integrative analyses of genetic variation in obesity

4

Part of my doctoral research was focused on the analysis of genetic variation in obesity. In the following chapter, I will briefly outline findings from GWA studies on obesity risk phenotypes, preliminary findings from a sequence-based analysis of the FTO gene, and results from integrative analyses of genetic variation in body-mass index.

4.1 Common genetic variation in obesity

Obesity is adding substantially to morbidity and mortality due to adverse consequences on human health [Kopelman, 2007]. In 2030 the prevalence of obesity, the weight in kilograms divided by the square of the height in meters above 30, is predicted to be 36.2% for westernized countries and 19.7% world-wide [Kelly et al., 2008]. The obesity epidemic is, among other factors, driven by a change in our near environment (as e.g. increased availability of energy dense, palatable and inexpensive food), but individuals respond differently, as genetic variation determines how susceptible a given person is towards becoming obese. Some obesity statistics have reported that the prevalence of obesity in the USA have stabilized during the last decade [Yanovski and Yanovski, 2011]. This observation would indicate that susceptible individuals already have become obese, while the resistant individuals stay lean despite the 'obesogenic' environment. However, a general trend for all westernized societies is that those who are susceptible are steadily becoming more obese and a growing fraction is becoming morbidly obese (body-mass index >40) [Yanovski and Yanovski, 2011]. Genetic factors have been estimated to account for 40-70% of the population variation in obesity [Speliotes et al., 2010], and hence a major factor in the susceptibility landscape of obesity.

Prior to GWA studies, candidate gene approaches and linkage studies have identified several genes (among others leptin gene, LEP, the leptin receptor, LEPR, proopiomelanocortin, POMC, and the melanocortin 4 receptor, MC4R) that segregated in Mendelian patterns and led to extreme forms of overweight [McCarthy, 2010]. In 2007, gene variants in the FTO gene were the first common SNPs to be associated with body-mass index [Frayling et al., 2007, Scott et al., 2007, Dina et al., 2007] (as described in the next section). Since then, GWA analyses of body-mass index [Willer et al., 2008b, Thorleifsson et al., 2008, Speliotes et al., 2010] and obesity-related risk-phenotypes such as waist circumference [Lindgren et al., 2009], waist-hip circumference [Lindgren et al., 2009, Heid et al., 2010], weight [Johansson et al., 2010] and early onset obesity [Meyre

et al., 2009] have identified >45 independent susceptibility loci for obesity-related risk phenotypes.

Even though recent GWA meta-analysis findings have provided some evidence to the hypothesis that risk of being obese is increased by deregulation of central energy expenditure regulating pathways in the hypothalamus [Fawcett and Barroso, 2010], the specific biological pathways causal to obesity remain obscure. A recent meta-analysis for fat distribution showed that genetic variation within genes expressed in adipose tissue contribute to common forms of obesity [Heid et al., 2010], showing that pathways related to fatty acid metabolism may play important roles in obesity. For instance, excess storage of triglycerides in fat cells and these cells' inertia may cause metabolic dysfunction [Sørensen et al., 2010].

Despite the relative large number of novel obesity-associated loci, the known associations account for <2% percent of the genetic variation in obesity-related risk phenotypes. The large fraction of unaccounted genetic variability in current studies indicates that common variants do not provide the complete picture of the susceptibility landscape in obesity. Consequently, research focus has turned towards complementary analyses techniques, such as systematic investigations of copy number variations' impact on body-mass index [Walters et al., 2010] and role in early onset obesity [Bochukova et al., 2009], and integrative systems biology analyses that integrate a variety of relevant evidence sources. The copy number variation studies found large de-novo rearrangements at 16p11.2, a region comprising several genes including a gene (SH2B1) that has also has been identified in GWA studies. However, both authors concluded that the contribution of copy number variations to obesity is relatively low. In summary, despite large-collaborative efforts and systematic analysis of the role of common variation in obesity, the percent variability explained remains low, compared to estimates of the heritability of obesity.

Before I return to the role of integrative approaches for the analyses of genetic variation in obesity (Section 4.3), I will summarize the analyses of the FTO gene that my co-workers and I did during the first year of my doctoral studies.

4.2 Studies on the biology of the FTO gene

In 2007 several GWA studies reported strong associations between variants in the FTO (fat mass and obesity associated) gene and body-mass index [Dina et al., 2007, Frayling et al., 2007, Scuteri et al., 2007]. Shortly hereafter, we reported an association between the FTO risk allele rs9939609 and an obesity-independent increase in all cause mortality [Zimmermann et al., 2009], a finding that further encouraged us to analyze genetic variation in the FTO gene.[1] The increase in all cause mortality seemed to resemble other findings from mouse studies, namely that loss-of-function mutations in mouse Fto causes autosomal-recessive lethality [Boissel et al., 2009]. FTO was originally discovered in mouse mutants, where homozygosity of a 1.6-Mb deletion comprising at least six genes including Fto and Ftm (RPGRIP1L), led to loss of genetic control of left-right symmetry in the brain, defects in brain morphogensis and to death early in development [van der Hoeven et al., 1994].

Now, almost 3 years after the FTO associations were discovered, association studies have shown that variants in FTO may lead to increased energy intake and reduced satiety [Tung and Yeo, 2011]. As associations themselves do not give any mechanistic insights, various research groups have conducted bioinformatics- and *in vitro* bio-

[1] The rs9939609 variant is still the FTO SNP that exhibits the strongest association to obesity.

chemical analyses on the FTO gene and its protein sequence. These analyses suggest that the FTO gene product functions as a 2-oxoglutarate dependent demethylase or dioxygenase [Gerken et al., 2007, Sanchez-Pulido and Andrade-Navarro, 2007]. In addition, mouse models provided evidence that the Fto gene and not the neighboring Ftm (homolog of the human RPGRIP1L gene) associates with obesity, as FTO loss of function [Fischer et al., 2009] and partial loss-of-function [Church et al., 2009] result in reduced fat mass and increased energy expenditure in mice. In addition, over-expression of Fto has been shown to cause obesity [Church et al., 2010] in mice.

These interesting pleiotropic effects of the FTO locus inspired us[2] to look deeper into the putative cellular mechanisms linking the sequence variation to obesity and possibly increased all cause mortality. As the FTO-risk alleles are located within a 45 kb haplotype block in the first FTO intron (and 90 kb promoter region of the reverse strand RPGRIP1L gene), and introns are known to harbor regulatory control elements, such as small nucleolar RNAs, miRNAs [Cheng et al., 2005] and other intronic non-coding RNAs [Ashe et al., 1997, Mattick, 2004], we hypothesized that the FTO risk-alleles influenced the function of an unknown non-coding RNA putatively regulating FTO and/or RPGRIP1L expression. Examples on sequence variants disrupting non-coding RNAs are rare, but may have profound effects [Iwai and Naraba, 2005].

Bioinformatics sequence-based analysis of the FTO locus

First, we scanned all SNPs in linkage disequilibrium (r^2>0.50) with the rs9939609 reported in Frayling *et al* [Frayling et al., 2007], and Zimmermann et al. [Zimmermann et al., 2009] for overlap with predicted functional non-coding RNAs based on genome-wide predictions from the Evofold [Pedersen et al., 2006] and RNAz prediction tools [Washietl et al., 2005b, Washietl et al., 2005a]. Based on the genome-wide RNAz screen (assessing only predictions with a probability > 0.9, yielding a sensitivity of 75% and specificity of 98%), we identified a SNP rs1421085 with high linkage disequilibrium with the rs9939609 (r^2 = 0.91), which overlapped with a non-coding RNA prediction (**Fig.** 4.1).

Interestingly, the SNP was one of the four most significantly associated SNPs suggested in the work by Dina *et al* [Dina et al., 2007] and located within one of the most highly conserved regions of the human genome, being conserved in several vertebrates including chicken (**Fig.** 4.2). Highly conserved non-coding regions of the genome are known to be strongly enriched for non-coding RNAs [Siepel et al., 2005].

The candidate RNA had a predicted length of 239 bps, but the actual length of the putative non-coding RNA could extend the predicted size, since only a small part of it may have been detected in the sliding window approach employed by the prediction algorithm. We used the RNAfold and RNAalifold tools[3] to predict the structure of the candidate RNA based on its sequence (**Fig.** 4.3**a**), and based on the multiple alignment of 12 vertebrate species (**Fig.** 4.3**b**). The structure predictions resembled each other, which confirmed the original RNAz algorithm finding that the candidate RNA sequence is able to fold into a stable secondary structure.

To assess how the risk allele affected the secondary structure of the candidate RNA (as measured by the change in minimum free energy) we again used the RNAfold and RNAalifold RNA structure predictions tools to calculate the minimum free energy based on the human sequence and an alignment of the candidate RNA sequence in 12

[2] Asli Silahtaroglu, Claus Hansen and Niels Tommerup from the University of Copenhagen, my supervisors, and me.

[3] http://rna.tbi.univie.ac.at

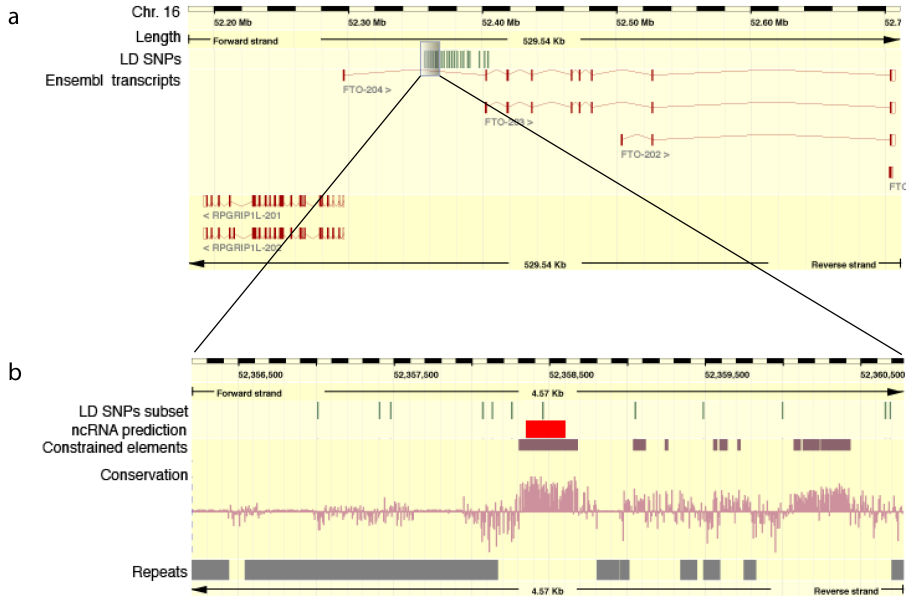


Figure 4.1: Overview of the FTO and RPGRIP1L region. Panel (**a**) shows the SNPs within the FTO intron 1 haplotype block that associates with body-mass index. Panel (**b**) gives a more detailed overview of the SNPs within and in close proximity to the candidate RNA. The Ensembl Genome Browser (www.ensembl.org) was used to generate genomic parts of the figure. Abbreviations: Chr, chromosome; LD, linkage disequilibrium; ncRNA, non-coding RNA.

vertebrate sequences. We found that the risk allele decreased minimum free energy of the candidate RNA structure based on the multiple alignment, and that the change in minimum free energy was largest for the structure folded from the sequence of the forward strand (**Tab.** 4.1).

We identified six compensatory double substitutions and several single compatible substitutions in the stem harboring the rs1421085 polymorphism, which increased our confidence in the prediction. The candidate non-coding RNA is located in the latter part of the first FTO intron and in the promoter on the antisense RPGRIP1L gene (63kb upstream from the RPGRIP1L transcription start site, Ensembl genome browser[4]). Cap analysis of gene expression in humans and mouse suggests that the candidate RNA is included in RPGRIP1L transcripts expressed in cecum and cerebrum [Kawaji et al., 2006]. In addition, the genomic region between RPGRIP1L and FTO harbors a CpG island and may function as a bi-directional promoter [Engström et al., 2006].

Expression analysis of the FTO locus

To assess whether the candidate RNA was expressed, and possibly co-expressed with the FTO and RPGRIP1L gene, we collaborated with experimentalists from the Tommerup Laboratory from the University of Copenhagen.

[4] www.ensembl.org

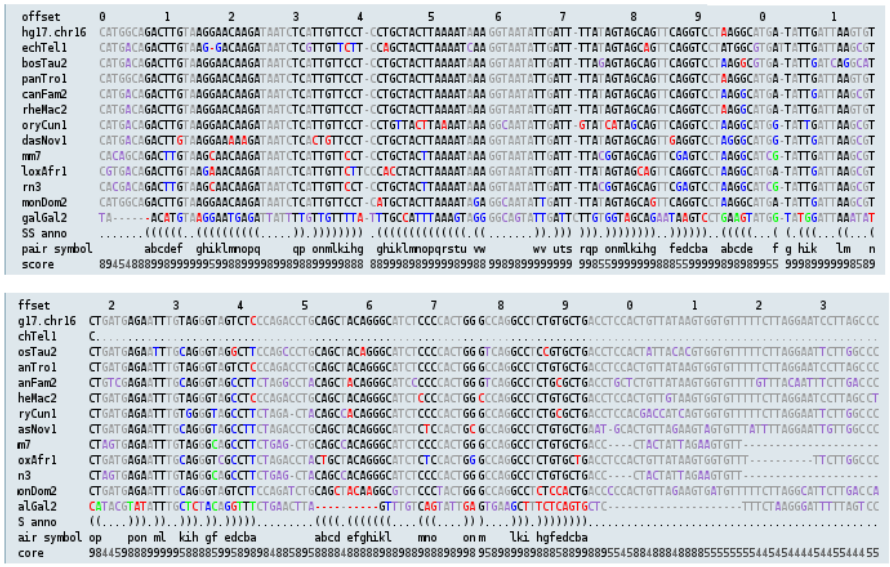


Figure 4.2: Alignment of candidate RNA sequence in 12 vertebrate species. Abbreviations: SS anno, secondary structure annotation. Species abbreviations:
echTel1, *Echinops telfairi*, Lesser Hedgehog Tenrec
bostau2, *Bos taurus*, Domestic Cow
panTro1, *Pan troglodytes*, Common Chimpanzee
canFam2, *Canis familiaris*, Domestic Dog
rheMac2, *Macaca mulatta*, Rhesus Macaque
oryCun1, *Oryctolagus cuniculus*, European Rabbit
dasNov1, *Dasypus novemcinctus*, Nine-Banded Armadillo
mm7, *Mus musculus*, House Mouse
loxAfr1, *Loxodonta africana*, African bush elephant
rn3, *Rattus norvegicus*, brown rat
MonDom2 *Monodelphis domestica*, opossum
galGal2, *Gallus gallus*, chicken

Quantitative real time polymerase chain reaction (qPCR) expression analysis of the candidate RNA, FTO and RPGRIP1L showed that these three genes were co-expressed across all tissues (temporal lobe, parietal lobe, occipital lobe, frontal lobe, diencephalon, cerebellum right, and cerebellum left) in our human brain cDNA panel (**Fig. 4.4a**). However, the average crossing point values (Ct values) indicated that the expression of the candidate RNA was low, with average Ct values above 30 (average 31.4±0.8).[5] Expression levels of FTO and RPGRIP1L were much higher with average

[5] In qPCR analysis the Ct value denotes the number of cycles needed to detect a expression signal that

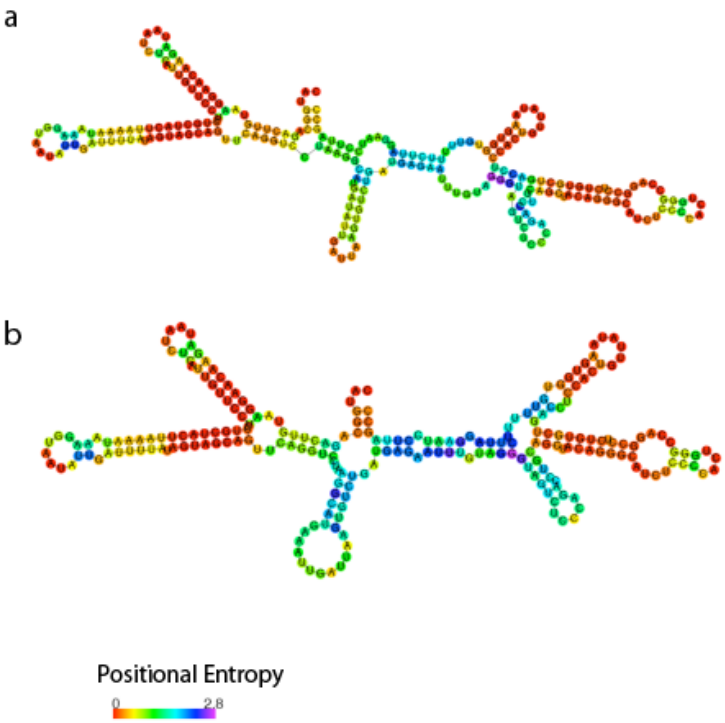


Figure 4.3: The predicted structure of the candidate RNA based on its sequence (**a**) and based on a multiple alignment of 12 vertebrates (**b**). The colors denote positional entropy, where dark red denotes basepairing characterized by low entropy, i.e. strong binding.

Ct values of 24.7±0.8 and 24.3±0.6, respectively. Based on the difference in Ct value averages between the candidate RNA, FTO, and RPGRIP1L, we estimated that FTO and RPGRIP1L were >100 fold more abundant in the brain tissues than the candidate RNA.

Expression analysis on a larger human cDNA panel revealed that the candidate RNA was expressed at detectable levels in a majority of the tissue samples (**Fig.** 4.4**b**). Among the brain tissues, the expression of the candidate RNA and FTO and RPGRIP1L was largest in the *total brain* sample. Major expression peaks of non-neural origin included fetal liver, heart, skeletal muscle, and colon. The average expression levels of the candidate RNA were low judged on Ct values (average 30.6±1.5). For FTO and RPGRIP1L, average Ct levels were consistently higher compared to the candidate RNA, averaging 23.8±1.5 (FTO) and 26.7±2.1 (RPGRIP1L). This indicated that these genes were approximately 40 fold more abundant than the candidate RNA transcript in these tissues.

To determine where the candidate non-coding RNA localized in the brain and which strand it is expressed from, Dr. Silahtaroglu performed *in situ* hybridizations

exceeds the background level, i.e. there is a negative correlation between Ct levels and target nucleic acid in the sample.

	Minimum Free Energy	
	rs1421085 major allele	rs1421085 risk allele
Human sequence minimum free energy		
Forward strand, kcal/mol	-81.40	-83.24
Reverse strand, kcal/mol	-82.22	-82.51
Multiple alignment minimum free energy		
Forward strand, kcal/mol	-95.61	-103.34
Reverse strand, kcal/mol	-97.83	-97.95

Table 4.1: Based on the structure prediction of the human candidate RNA sequence, the minimum free energy of the structure with the rs1421085 major allele had a larger minimum free energy than the predicted minimum free energy from the candidate RNA sequence containing the risk allele. However, structure prediction based on the multiple alignment of 12 vertebrate species resulted in stronger structure (lower minimum free energy) for the risk allele sequence. For both the human sequence and the multiple alignment the structure prediction based on the forward strand resulted in a slightly stronger structure. Abbreviations: kcal, kilo calories.

on frozen adult mouse brain sections. Two sets of probes were placed within the candidate RNA where one pair included the rs1421085 polymorphism. One set was placed outside the candidate RNA region (**Supplementary Figures** 1 and 2 pp. 160-161). The *in situ* hybridization using the probes within the candidate RNA showed expression signal from both strands, with the strongest signal coming from the forward strand.

Concluding Remarks

Even though our preliminary findings are interesting, as no mechanisms has been identified yet, they are challenging, too. Further work is needed to verify that the expression of the candidate RNA is not caused by non-functional intron sequences.

Similar to previous expression studies [Gerken et al., 2007, Frayling et al., 2007], we found that expression levels of FTO and RPGRIP1L is relatively high in various brain tissues. The candidate RNA seemed to be co-expressed with these genes, even though absolute expression levels of the candidate RNA are relatively low. Also low expression levels often are observed for regulatory non-coding RNAs [Mattick and Makunin, 2006], care must be taken as genuinely low expression levels are difficult to differentiate from noise in qPCR measurements.

Further, as the length of the candidate RNA is unknown, its structure prediction may change drastically if the real length should differ from the predicted length. The predicted length is 239 bps, while single precursor miRNA are generally approximately 100 bps in length, and non-coding RNAs such as the mouse Air RNA occupies approximately 108 kb of genomic DNA [Saunders et al., 2007].

In addition, our preliminary findings cannot exclude that other mechanisms are responsible for mediating the effects of the SNPs on downstream phenotypes.

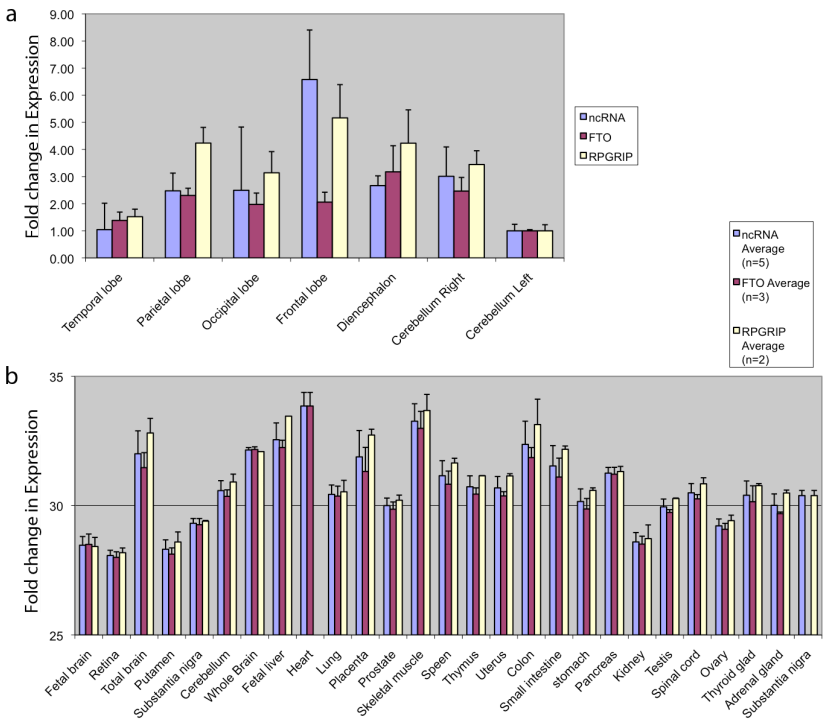


Figure 4.4: Quantitative real time polymerase chain reaction (qPCR) expression analysis of the candidate RNA, FTO and RPGRIP1L. Panel (**a**) illustrates the expression of the candidate RNA (abbreviated ncRNA in the figure), the FTO gene, and the RPGRIP1L gene across a human brain cDNA panel. Panel (**b**) depicts the expression of the candidate RNA and the two genes in a larger human cDNA panel, which includes non-brain tissues as well. The candidate RNA's and the two genes' expression fold changes (compared to the least expressed transcript in the given sample) are similar across all tissues in panel (**a**) and (**b**). However, the absolute expression levels between the two genes and the candidate RNA on average differed by a factor 100 in in panel (**a**) and with a factor 40 in panel (**b**). Abbreviations: ncRNA, candidate non-coding RNA.

4.3 Integrative analyses of body-mass index

Integrative pathway-based approaches have so far yielded limited insights as to which pathway may cause obesity [Liu et al., 2010, Speliotes et al., 2010, Heid et al., 2010]. In the currently largest GWA study for body-mass index, researchers used the MAGENTA method [Segrè et al., 2010] to search for overrepresented pathways among genes within 300 kb flanking regions of the 32 associated SNPs. Based on gene sets annotated in the KEGG Pathway, Ingenuity, PANTHER, and Gene Ontology databases, they identified a couple of enriched pathways (platelet-derived growth factor signaling, translation elongation, hormone or nuclear-hormone receptor binding, homeobox transcription, regulation of cellular metabolism, neurogenesis and neuron differentiation, protein phosphorylation, and numerous other pathways related to growth, metabolism, immune and neuronal processes) [Speliotes et al., 2010]. However, interpretation of these

results is challenging due to the large number of genes within the enriched gene sets (714 genes in the largest significantly enriched gene set). In addition, the pathways will need to be validated in independent cohorts or other phenotypic data sets to verify that they are not results of chance-correlations.

In another large GWA meta-analysis, constituting the largest GWA study of the obesity risk-phenotype fat distribution, Heid *et al* searched for pathways enriched in waist-hip ratio associations [Heid et al., 2010]. They relied on SNPs from the discovery phase in their meta-analysis with p-values $< 10^{-5}$ (48 independent SNPs in 95 genes) and only identified a slight overrepresentation of developmental processes.[6]

The limited success in identifying etiologic pathways adds to the notion that obesity is a highly heterogeneous trait, and that the pathogenic processes are inadequately captured by solely proceeding within GWA data and currently existing integrative methods. Instead of assessing *canonical* pathways for enrichment in GWA signal, we hypothesized that gene-products within implicated body-mass index associated loci may physically interact with hitherto unknown obesity susceptibility gene products. To assess this hypothesis, we integrated GWA data from the Genetics of in Obesity of Young Adults (GOYA) study with PPI data, and are currently following up upon our results (**Paper III**).

4.4 Paper III - The ASIP gene's putative association with extreme overweight

Very recently the GIANT consortium (Genomewide Investigation of ANThropometric measures) identified 32 loci, which robustly associated with body-mass index [Speliotes et al., 2010]. Based on these implicated loci, I assembled a list of putative body-mass index susceptibility genes, and tested whether any of the protein complexes from our PPI meta-database was significantly enriched in these. The most significant complex was centered on the known susceptibility gene MC4R. We found 13 other genes in the complex and validated one of them, ASIP, in an independent GWA-based cohort for early onset obesity.

The following manuscript is written in a short format (following guidelines for Brief Communications and Reports) and is still under preparation. Currently, we are seeking to replicate the SNPs in the ASIP gene in additional independent cohorts. Until the current findings have been replicated in these they should be interpreted with care.

[6] They used the PANTHER tool [Thomas et al., 2003], which assesses over-representation of gene ontology gene sets, and the GRAIL tool [Raychaudhuri et al., 2009].

Protein complex analysis associates SNPs in ASIP with extreme overweight

Tune H Pers[1,2], *working group in preliminary order* (Daniela Nitsch[3], Yves Moreau[3], co-workers from the NUGENOB study, co-workers form the GOYA study), Søren Brunak[1,4,*], Thorkild I A Sørensen[2,*]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

[2] Institute of Preventive Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark

[3] Katholieke Universiteit Leuven, Belgium.

[4] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally

Also 32 loci to date have been robustly associated with body-mass index (BMI), they account for less 2-4% of the genetic variation in BMI. As the heritability of BMI is estimated to be 40-60%, we searched for novel BMI risk variant within genes shown to interact physically with gene products from the BMI-associated loci. Using an independent replication cohort, we found that the rs819163 SNP in the agouti signaling protein (ASIP) gene, which gene product interacts with the MC4R protein, associated with extreme obesity (P = 1.9e-5).

Recently, Speliotes et al. reported 32 loci that convincingly associated with BMI (kg/m^2) [1]. In the Genetics of Overweight Young Adults study (GOYA), we replicated 9 of the found loci in a genome-wide association (GWA) study of ~2,600 extreme overweight cases (mean BMI = 35.3) and a similar number of population-based controls, both drawn from a cohort of 225,000 European women and men [2]. However, despite estimations that heritability of BMI is between 40 and 60%, the current loci explain less than 5% of the genetic variation in BMI, suggesting that several risk variants still need to be identified. SNPs exerting modest relative risks are less likely to be identified in GWA studies due to the large number of SNPs tested. Within that line of reasoning it has been hypothesized that genetic associations are enriched within biological pathways [3]. Also, definition and delineation of pathway organization is constrained to the interactome data present in databases, protein-protein interaction data is increasingly being considered as reliable, high-coverage, and phenotypically unbiased pathway organization source [4],

The premise of our study was that gene products from known BMI-associated loci are over-represented in specific protein complexes that are etiological to obesity. Specifically, we hypothesized that the gene-sets coding for proteins in risk modules incriminate currently unknown BMI susceptibility genes. In this study, we aimed at first identifying putative obesity risk gene sets, and subsequently to test each incurred gene for association with BMI. While the former step is based on already known BMI susceptibility genes and protein complex organization data, the latter should be based on a study cohort that is independent from the one used to identify the known obesity susceptibility genes. The overall aim is to identify genes that have not yet been associated with BMI.

To that end, we first compiled a list of 40 putative BMI susceptibility genes by considering all genes reported by the GIANT consortium (including up to 3 genes from single loci in case multiple genes were reported for that loci) (Suppl. Table 1). Next, we compiled a list of 12,714 protein complexes by extracting high-confidence protein complexes from a protein-protein interaction meta-database used and validated in previous studies (Suppl. Methods) [4,5]. We iteratively calculated the over-representation of BMI-associated genes in each protein complex and identified a complex centered on the MC4R gene as being most highly enriched in BMI-associated genes (P = 4.1e-4) (**Figure 1**). To adjust for multiple testing of the large number of protein complexes, we assembled 100,000 random complexes retaining the properties from the observed protein complexes and found that the observed MC4R complex indeed was significantly enriched in BMI-associated genes (P = 2.0e-4).

The MC4R complex comprised 14 gene products, of which MC4R and POMC led to the over-representation of known BMI-associated genes (**Table 1**). Among the 14 genes were NPY and AGRP, which in previous candidate gene approaches have been associated with obesity [6], but also a few other genes (ADRBK1, ASIP, ATRNL1, MC1R, NPY1R, NPY5R, PMCH, and PRKACA) that, to our knowledge, have not been genetically associated with human BMI before. Note, that these findings are solely based on GIANT- and protein-protein interaction data. To validate these genes as likely human obesity susceptibility genes, we retrieved GWA data of 2,633 extreme overweight young adults and an equal number of randomly selected population-based controls from the GOYA study. We used the MetaRanker tool [7] to compute p-values for the 14 genes in the MC4R complex based on their GOYA SNPs associations with extreme BMI (**Table 1**, column 4). Briefly, the MetaRanker method accomplishes this by (a) mapping SNPs to genes, (b) scoring genes based on their most significant SNP, and (c) adjusting genes' p-values based on the

number of independent SNPs at each loci. After further adjusting the gene p-values by the number of genes in the complex (n = 14), we found that the ASIP gene remained highly significant (adjusted P = 2.7e-4) (**Table 1**, column 5). The other genes exhibiting slight significance the previously known BMI susceptibility genes POMC and MC4R (P = 0.01 and P = 0.03, respectively). To assess whether the entire MC4R complex validated in independent GWA data, we used the complete GOYA GWA data to calculate p-values for 21,845 human protein-coding genes, quantifying their association with extreme obesity, and used a one-sided Kolmogorov-Smirnov test to test whether the MC4R complex gene p-values were significantly enriched in the low end of the distribution of all gene p-values (Suppl. Methods). We found that this was indeed the case (P = 2.5e-4). In summary, the MC4R complex validated as being enriched in BMI-associated genetic variation in an independent study cohort (GOYA data was not part of the GIANT meta-analysis), and, in particular, the ASIP stands out as a novel BMI susceptibility gene.

In the ASIP locus the SNP rs819163 was most significant (P = 1.9e-05) with an odds ratio of 0.77 (L95 = 0.68; U95 = 0.87) (**Figure 2**). In the previous GOYA GWA analysis, this SNP was not considered for replication as it was marginally above their replication selection threshold (P < 1.0e-5). To replicate the ASIP loci in an additional cohort, we retrieved unpublished genotype data from the NUGENOB weight loss intervention study [8]. Based on genotype data from 770 obese women and men (mean BMI = 35.6), we computed the association with a highly correlated nearby SNP rs819614 (r^2 = 0.92; D' = 1) with weight loss and postprandial fat oxidataion capacity. Whereas the rs819164 was not significant with respect to weight loss, it reached statistical significance in the association analysis with postprandial fat oxidation capactiy (P = 0.004).

ASIP encodes a 14,515-kDa protein of 132 amino acids that is primarily known to function as a paracrine signaling molecule, which induces hair follicle melanocytes pigmentation changes [10]. Interestingly, the ASIP protein has been shown to act on melancortin signaling by inhibiting MC4R [11], has been observed under respiratory quotient and long term weight change linkage peaks [12,13], and the ASIP mouse homolog, which is 85% similar the human protein, has been implicated in obesity [14].

To evaluate the tissue-specificity of the MC4R complex, we retrieved published gene expression data from a total of 951 healthy subjects across 37 human tissues (Suppl. Methods, Suppl. Table 2) [15]. Using a validated methodology [16], we calculated for each tissue the average Pearson correlation between the expression levels of the MC4R gene and the expression levels of its protein interactions partners, and found that these were most highly correlated in the thyroid (average r^2 = 0.69). By computing the average Pearson correlation coefficient for 100,000 random complexes, we found that the observed thyroid average Pearson correlation coefficient was indeed significant (P = 5e-4) (Suppl. Methods). The only other tissues where the MC4R complex average Pearson correlation coefficient was significant was smooth muscle (average r^2 = 0.33; P = 0.03). MC4R is hypothesized to impact weight regulation particularly by acting in the hypothalamus [17]. However, despite the hypothalamus average Pearson correlation coefficient being above the 90 percentile across all 39 tissues, it was not significant after permutation analysis (average r^2 = 0.33; P = 0.15). The difference between the average correlation coefficients in thyroid and hypothalamus was partly due to differences in co-expression between MC4R and ASIP, which was markedly high in the thyriod (r^2 = 0.89, P = 0.01), and absent in the hypothalamus. Note, that these findings were based on gene expression data from non-obese individuals. Unfortunately,

we could not retrieve any data from human hypothalamic samples, which otherwise would have allowed us to assess if the obese-state introduces MC4R-comlex co-expression changes.

In conclusion, we identified a protein complex centered on MC4R, which exhibited significant over-representation of genes from known BMI-associated loci. Using data from the independent GOYA and NUGENOB cohorts, we validate the ASIP gene as candidate gene for extreme obesity and find indication for an association with postprandial fat oxidation capacity. Finally, we show that the specific MC4R complex, including ASIP, in non-obese individuals is highly co-expressed in the thyroid and hypothalamus. Studies of gene expression across the here-in studied tissues from obese individuals are needed to clarify whether, and in which tissues, the MC4R risk module exhibits altered co-expression, and thereby altered function in obesity.

Table 1

HGNC ID	Ensembl ID	Study	Pubmed ID	Walley et al, NRG 2009 (PMID 19506576)	Hofker & Wijmenga, Nat Genet 2009 (PMID 19174833)	GIANT BMI	Comment	Has Protein-Protein interaction partners
GIPR	ENSG00000010310	Speliotes, NG, 2018	20935630			x	Same locus as QPCTL	1
QPCTL	ENSG00000011478	Speliotes, NG, 2017	20935630			x		4
MTCH2	ENSG00000109919	Willer CJ, Nat Gen. 2009	19079261	x	x	x		na
NUDT3	ENSG00000112664	Speliotes, NG, 2036	20935630			x		9
HMGCR	ENSG00000113161	Speliotes, NG, 2022	20935630			x	Same locus as FlJ35779	77
RBJ	ENSG00000115137	Speliotes, NG, 2010	20935630			x	HGNC synonym DNAJC27	1
POMC	ENSG00000115138	Speliotes, NG, 2012	20935630			x	Same locus as RBJ	14
FANCL	ENSG00000115392	Speliotes, NG, 2026	20935630			x		18
TNNI3K	ENSG00000116783	Speliotes, NG, 2019	20935630			x		7

PTBP2	ENSG00000117569	Speliotes, NG, 2030	20935630			x		7
SEC16B	ENSG00000120341	Thorleifsson G, Nat Gen. 2009	19079260	x	x	x		18
MTIF3	ENSG00000122033	Speliotes, NG, 2031	20935630			x		na
GTF3A	ENSG00000122034	Speliotes, NG, 2032	20935630			x	Same locus as MTIF3	28
TMEM160	ENSG00000130748	Speliotes, NG, 2024	20935630			x		na
ZC3H4	ENSG00000130749	Speliotes, NG, 2025	20935630			x	Same locus as TMEM160	na
FAIM2	ENSG00000135472	Thorleifsson G, Nat Gen. 2009	19079260	x	x	x		2
HMGA1	ENSG00000137309	Speliotes, NG, 2037	20935630			x	Same locus as NUDT3	62
MAP2K5	ENSG00000137764	Speliotes, NG, 2015	20935630			x		30
ADCY3	ENSG00000138031	Speliotes, NG, 2011	20935630			x	Same locus as RBJ	12
SLC39A8	ENSG00000138821	Speliotes, NG, 2020	20935630			x		na
FTO	ENSG00000140718	Willer CJ, Nat Gen. 2009	19079261	x	x	x		na
TMEM18	ENSG00000151353	Willer CJ, Nat Gen. 2009, Thorleifsson G, Nat Gen. 2009	19079261, 19079260	x	x	x		na
FLJ35779	ENSG00000152359	Speliotes, NG, 2021	20935630			x		2
KCTD15	ENSG00000153885	Willer CJ, Nat Gen. 2009, Willer CJ, Nat Gen. 2009	19079261, 19079260	x	x	x		9
GNPDA2	ENSG00000163281	Willer CJ, Nat Gen. 2009	19079261	x	x	x		na
TUB	ENSG00000166402	Speliotes, NG, 2035	20935630			x	Same locus as RPL27A	15

RPL27A	ENSG00000166441	Speliotes, NG, 2034	20935630			x		627
MC4R	ENSG00000166603	Willer CJ, Nat Gen. 2009, Thorleifsson G, Nat Gen. 2009	19079261, 19079260	x	x	x		13
GPRC5B	ENSG00000167191	Speliotes, NG, 2013	20935630			x	Same locus as GPRC5B	na
LRP1B	ENSG00000168702	Speliotes, NG, 2029	20935630			x		139
ZNF608	ENSG00000168916	Speliotes, NG, 2033	20935630			x		5
ETV5	ENSG00000171656	Thorleifsson G, Nat Gen. 2009	19079260	x	x	x		1
NEGR1	ENSG00000172260	Willer CJ, Nat Gen. 2009, Willer CJ, Nat Gen. 2009	19079261, 19079260	x	x	x		5
LRRN6C	ENSG00000174482	Speliotes, NG, 2023	20935630			x	HGNC synonym LINGO2	na
IQCK	ENSG00000174628	Speliotes, NG, 2014	20935630			x		na
CADM2	ENSG00000175161	Speliotes, NG, 2027	20935630			x		na
BDNF	ENSG00000176697	Thorleifsson G, Nat Gen. 2009	19079260	x	x	x		11
SH2B1	ENSG00000178188	Willer CJ, Nat Gen. 2009, Thorleifsson G, Nat Gen. 2009	19079261, 19079260	x	x	x		10
PRKD1	ENSG00000184304	Speliotes, NG, 2028	20935630			x		26
LBXCOR1	ENSG00000188779	Speliotes, NG, 2016	20935630			x	Same locus as MAP2K5	4

Abreviations:
na: Not among protein in InWeb protein-protein interaction database

Figure 1: The blue notes represent the MC4R complex. Edges between nodes denote physical interaction at the protein level. The first-order protein interaction partners of the proteins in the MC4R-complex are colored in gray.

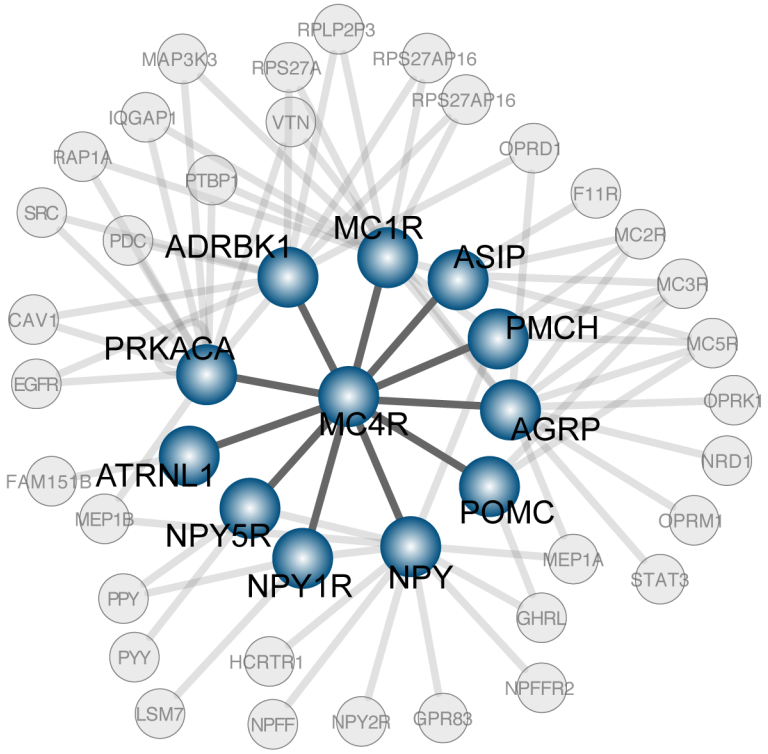
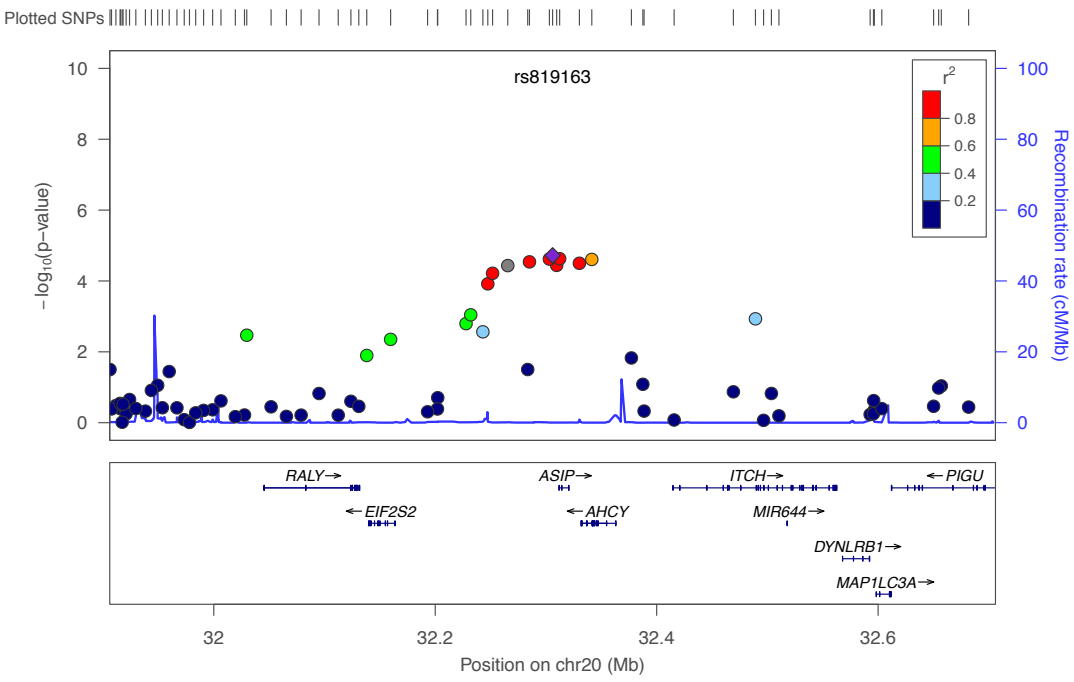


Figure 2: Regional association plot of the ASIP loci.



References

1. Speliotes E, Willer C, Berndt S, Monda K, Thorleifsson G, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics 42: 937-948.
2. Paternoster L, Evans DM, Nohr EA, Holst C, Gaborieau V, et al. (2011) Genome-wide population-based association study of extremely overweight young adults - the GOYA study. PLoS ONE (In review).
3. Hirschhorn J (2009) Genomewide association studies--illuminating biologic pathways. The New England journal of medicine 360: 1699-1701.
4. Lage K, Karlberg O, Storling Z, Olason P, Pedersen A, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology 25: 309-316.
5. Lage K, Mollgard K, Greenway S, Wakimoto H, Gorham J, et al. (2010) Dissecting spatio-temporal protein networks driving human heart development and related disorders. Molecular Systems Biology 6.
6. van Rossum CTM, Pijl H, Adan RAH, Hoebee B, Seidell JC Polymorphisms in the NPY and AGRP genes and body fatness in Dutch adults. International Journal of Obesity aop.
7. Pers T, Hansen N, Lage K, Koefoed P, Dworzynski P, et al. (2011) Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genetic Epidemiology (In press).
8. Sorensen TI, Boutin P, Taylor MA, Larsen LH, Verdich C, et al. (2006) Genetic polymorphisms and weight loss in obesity: a randomised trial of hypo-energetic high- versus low-fat diets. PLoS Clin Trials 1: e12.
9. Petersen M, Taylor MA, Saris WHM, Verdich C, Toubro S, et al. (2006) Randomized, multi-center trial of two hypo-energetic diets in obese subjects: high- versus low-fat content. International Journal of Obesity aop.
10. Jackson IJ (1997) Homologous pigmentation mutations in human, mouse and other model organisms. Hum Mol Genet 6: 1613-1624.
11. Yang YK, Ollmann MM, Wilson BD, Dickinson C, Yamada T, et al. (1997) Effects of recombinant agouti-signaling protein on melanocortin action. Mol Endocrinol 11: 274-280.
12. Norman RA, Tataranni PA, Pratley R, Thompson DB, Hanson RL, et al. (1998) Autosomal genomic scan for loci linked to obesity and energy metabolism in Pima Indians. Am J Hum Genet 62: 659-668.
13. Fox CS, Heard-Costa NL, Vasan RS, Murabito JM, D'Agostino RB, Sr., et al. (2005) Genomewide linkage analysis of weight change in the Framingham Heart Study. J Clin Endocrinol Metab 90: 3197-3201.
14. Mynatt RL, Miltenberger RJ, Klebig ML, Zemel MB, Wilkinson JE, et al. (1997) Combined effects of insulin treatment and adipose tissue-specific agouti expression on the development of obesity. Proc Natl Acad Sci U S A 94: 919-922.
15. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, et al. (2010) A global map of human gene expression. Nature Biotechnology 28: 322-324.
16. Taylor I, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature biotechnology 27: 199-204.
17. Willer C, Sanna S, Jackson A, Scuteri A, Bonnycastle L, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics 40: 161-169.

Supplementary Information

Protein complex analysis associates SNPs in ASIP with extreme overweight

Tune H Pers[1,2], *working group in preliminary order* (Daniela Nitsch[3], Yves Moreau[3], co-workers from the NUGENOB study, co-workers form the GOYA study), Søren Brunak[1,4,*], Thorkild I A Sørensen[2,*]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

[2] Institute of Preventive Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark

[3] Katholieke Universiteit Leuven, Belgium.

[4] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally

Supplementary Methods

Body-mass index loci and genes. To construct a list of body-mass index (BMI) susceptibility genes we mapped all genes reported in Speliotes et al. [1] to Ensembl gene identifiers resulting in a list of 41 genes from a total of 32 BMI-associated loci (Supplementary Table 1).

Protein-protein interaction data and assembly of complexes. Protein-complex assembly was based on our in-house meta-protein-protein interaction (PPI) database, previously described in [3,4]. Briefly, it incorporates PPIs from 11 organisms across 13 databases and incorporates a total of 12,714 proteins and 173,500 PPIs based on 40,085 published articles. We constructed a set of 12,714 protein complexes, by considering each protein and its direct protein interaction partners as a single protein complex entity. Protein complexes were not restricted to a specific size, nor did we merge protein complexes that were highly overlapping. Each protein complex was represented by gene-set, assuming a one-to-one mapping between genes and complexes (a common assumption in PPI-based analysis, as most PPI databases do not report which specific splice product of a given gene was used in the PPI experiment). Complexes were plotted using Cytoscape [5].

Over-representation analysis. Significance of over-representation of BMI susceptibility genes in a given protein complex was tested in the hypergeometric distribution, using parameters *k, n, m, M*. The parameter *k*, denoted the overlap between the protein complex gene-set and BMI susceptibility genes; the parameter *n* denoted the size of the protein complex gene set; the parameter *m* denoted the total number of BMI susceptibility genes found in at least one of all protein complex gene sets (n=29); and parameter *M* denoted the total number of proteins part of the analysis (n=12,714).

Permutation-based protein complex p-values. As we tested 12,714 complexes for over-representation of BMI susceptibility genes we needed to adjust for multiple testing. As the protein complex gene-sets were overlapping and the tests hence not could be regarded as independent instances, we preferred permutation-based p-value computations above Bonferroni correction. We computed a permutation-based p-value for the MC4R complex by sampling 100,000 random complexes of the same size (n=19), while sampling each protein based on its observed likelihood to occur in a protein complex (its prevalence across all observed protein complexes). We computed the overrepresentation of BMI susceptibility genes in each of the random complexes, counted the number of random protein complexes that obtained an overrepresentation p-value equal or better than the observed MC4R complex p-value, and finally calculated a permutation-based p-value by dividing that count with the number of total permutations (n=100,000).

Validation cohorts and gene p-value assignment. We used genome-wide association (GWA) data from the Genetics of Overweight Young Adults study (GOYA) study [6] and the NUGENOB study [7] to validate genes in the MC4R protein complex gene set. As our analysis was gene-centric, we sought to validate each gene at the gene- instead of single-nucleotide polymorphism (SNP)-level. We

accomplished this by uploading each GWA study to the MetaRanker web tool (www.cbs.dtu.dk/services/metaranker) [4], which, among other things, scores the entire protein-coding part of the human genome based on GWA study data uploaded by the user. Briefly, the method (a) uses a predefined mappings scheme to map SNPs to genes, (b) identifies the lowest SNP mapped to a given gene, (c) assigns this p-value as the gene p-values, and (d) adjusts this gene p-value by the number of independent SNPs mapped to the gene.

Enrichment of BMI-associated loci genes in GOYA GWA study. We used a one-sided Kolmogorov-Smirnov test statistic to assess whether the genes from the BMI-associated loci were enriched in the tail of the GOYA GWA study gene p-value distribution (representing genes that are more likely to be associated with BMI). The calculation was done using the statistical software package R (version 2.11).

Gene expression data and average Pearson correlation coefficient analysis. To assess the degree of co-expression between the genes in the MC4R complex, we downloaded a gene expression data set comprising 5,372 human gene expression samples profiled across 369 different cell and tissue types from the ArrayExpress database (www.ebi.ac.uk/arrayexpress; Data set ID, E-TABM-185) [2]. Samples were all from the same microarray platform (Affymetrix U133A) and already normalized. We excluded non-healthy samples and samples derived from cell lines, and retained a set of 951 samples from 37 human tissues for our analyses. We retained only genes, which gene products among our complexes and ended up with at set of 10,457 genes.

To assess to in which tissues and to what extend genes underlying the MC4R complex were co-expressed, we used methodology previously applied within protein complex analysis by Taylor et al. [8]. For a given gene set and tissue, the Pearson correlation coefficient between the central gene (the central hub protein in the given protein complex) and each peripheral gene (the hub protein's physical interaction partners) was computed across all samples and averaged into an average Pearson correlation coefficient (APCC). Hence, the APCC for a given complex specifies the degree of co-expression of the central gene and the peripheral genes in a specific tissue.

To assess whether the APCCs observed for the MC4R complex in the various tissue was deviating significantly from random expectations, we used permutation analysis. For a given tissue with k samples, we re-computed the APCC based on expression data from k samples that were sampled with equal probability from the pool all 951 samples. This procedure was repeated 100,000 times to yield a tissue-specific background distribution of APCC scores, through which the significance of the observed APCC was calculated (by counting the number of random APCC that were equal or higher than the observed APCC for that tissue, and dividing that number by the total number of permutations).

Supplementary Tables

Supplementary Table 1: List of genes assembled based on the loci reported by Speliotes et al. [1].

HGNC ID	Ensembl ID	Comment	Number of PPI partners
GIPR	ENSG00000010310	Same locus as QPCTL	1
QPCTL	ENSG00000011478		4
MTCH2	ENSG00000109919		na
NUDT3	ENSG00000112664		9
HMGCR	ENSG00000113161	Same locus as FlJ35779	77
RBJ	ENSG00000115137	HGNC synonym DNAJC27	1
POMC	ENSG00000115138	Same locus as RBJ	14
FANCL	ENSG00000115392		18
TNNI3K	ENSG00000116783		7
PTBP2	ENSG00000117569		7
SEC16B	ENSG00000120341		18
MTIF3	ENSG00000122033		na
GTF3A	ENSG00000122034	Same locus as MTIF3	28
TMEM160	ENSG00000130748		na
ZC3H4	ENSG00000130749	Same locus as TMEM160	na
FAIM2	ENSG00000135472		2
HMGA1	ENSG00000137309	Same locus as NUDT3	62
MAP2K5	ENSG00000137764		30
ADCY3	ENSG00000138031	Same locus as RBJ	12
SLC39A8	ENSG00000138821		na
FTO	ENSG00000140718		na
TMEM18	ENSG00000151353		na
FLJ35779	ENSG00000152359		2
KCTD15	ENSG00000153885		9
GNPDA2	ENSG00000163281		na
TUB	ENSG00000166402	Same locus as RPL27A	15
RPL27A	ENSG00000166441		627
MC4R	ENSG00000166603		13
GPRC5B	ENSG00000167191	Same locus as GPRC5B	na
LRP1B	ENSG00000168702		139
ZNF608	ENSG00000168916		5
ETV5	ENSG00000171656		1
NEGR1	ENSG00000172260		5
LRRN6C	ENSG00000174482	HGNC synonym LINGO2	na
IQCK	ENSG00000174628		na
CADM2	ENSG00000175161		na
BDNF	ENSG00000176697		11
SH2B1	ENSG00000178188		10
PRKD1	ENSG00000184304		26
LBXCOR1	ENSG00000188779	Same locus as MAP2K5	4

Abreviations: PPI, protein-protein interaction; na, not among protein in PPI database

Table 2: List of the human tissue types used in the co-expression analysis. For more information please refer to Lukk et al. [2].

Human tissue type	Numer of samples
Blood	416
Bone marrow	44
Brain	39
Kidney	38
Brain caudate nucleus	29
Brain cerebellum	26
Lung transplant	25
Placenta basal plate	21
Skin	20
Hypothalamus	20
Lung	20
Atrial myocardium	18
Brain, frontal cortex, primary motor cortex, Brodmanns Area 4	15
Thymus	15
Cardiac ventricle	13
Myometrium	12
Brain, frontal cortex, superior frontal cortex, Brodmanns Area 9	12
Ovary	12
Prostate	11
Lymph node	10
Tonsil	10
Palatine tonsil	9
Skeletal muscle	9
Mesagnium	8
Adipose tissue	8
Smooth muscle	8
Leukocyte	8
Bone	7
Esophagus	7
Thyroid	7
Bronchial epithelia	7
Cord blood	7
Quadriceps muscle	7
Eye	6
Fetal blood	6
Small intestine	6
Hippocampal CA1 tissue	5

References

1. Speliotes E, Willer C, Berndt S, Monda K, Thorleifsson G, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics 42: 937-948.
2. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, et al. (2010) A global map of human gene expression. Nature Biotechnology 28: 322-324.
3. Lage K, Karlberg O, Storling Z, Olason P, Pedersen A, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology 25: 309-316.
4. Pers T, Hansen N, Lage K, Koefoed P, Dworzynski P, et al. (2011) Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genetic Epidemiology (In press).
5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
6. Paternoster L, Evans DM, Nohr EA, Holst C, Gaborieau V, et al. (2011) Genome-wide population-based association study of extremely overweight young adults - the GOYA study. PLoS ONE (In review).
7. Sorensen TI, Boutin P, Taylor MA, Larsen LH, Verdich C, et al. (2006) Genetic polymorphisms and weight loss in obesity: a randomised trial of hypo-energetic high- versus low-fat diets. PLoS Clin Trials 1: e12.
8. Taylor I, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature biotechnology 27: 199-204.

Integrative analyses based on metabolic network reconstructions

5

Inborn errors in metabolism refer to single-gene mutations in enzymes that result in changed metabolite concentrations [Mootha and Hirschhorn, 2010]. Back in 1902 A. E. Garrod predicted that these mutations reflected extreme examples of variation which is present in less severe but more common forms of the disease within the general population [Garrod, 2002]. A recent study has shown that common genetic variation (i.e. inborn variation in metabolism [Mootha and Hirschhorn, 2010]) is responsible for inter-individual differences in metabolite concentrations [Illig et al., 2010]. Other examples on association studies between common genetic variation and endogenous traits, i.e. traits upstream of the classical endpoints like body-mass index, are glucose [Dupuis et al., 2010, Saxena et al., 2010] and lipid traits [Kathiresan et al., 2009, Mohlke et al., 2008, Willer et al., 2008a, Teslovich et al., 2010]. Illig *et al* showed that three of their nine replicated loci reported (within acyl-CoA dehydrogenase genes ACADS, ACADM, and ACADL) associated with fatty acid levels through beta-oxidation pathways [Illig et al., 2010]. Interestingly, reconstructions of human metabolism (the main topic of this chapter), which comprise information on metabolic reactions (enzyme-metabolite relationships), may be used to augment analyses of metabolic diseases. This chapter outlines how reconstructed networks of human metabolism may be used to generate hypotheses as to which metabolite levels are altered due to changes in enzymes' gene expression levels. For phenotypes that have not yet been analyzed by coupled GWA and metabolomics studies, these approaches constitute an elegant way to generate hypotheses about metabolites that may be perturbed in individuals with metabolic risk-phenotypes.

5.1 Metabolic network reconstructions

The human metabolic network is the most well-studied biochemical network [Ma and Goryanin, 2008]. Formalized reconstructed networks are needed to analyze metabolic pathways, as they, like signaling networks, consist of a large interconnected web of reactions. Compared to PPI networks and regulatory networks formed by protein-DNA interactions, metabolic network representations are considered more complete in terms of coverage of molecular components, and more reliable with respect to the interactions contained in the networks [Ma and Goryanin, 2008]. Another difference to

	Enzymes	Reactions	Metabolites	Pathways
EHMN	2,322	2,823	2,671	70
Recon 1	1,496	3,748	1,469	88

Table 5.1: Overview of number of enzymes, metabolic reactions, metabolites and pathways in the *Edinburgh Human Metabolic Network* (EHMN) and *Homo sapiens Reconstruction 1* (*Recon 1*) metabolic network reconstructions.

PPI and regulatory networks is that metabolic networks can comprise up to three high-throughput data layers, *viz.* transcriptomics data measuring enzymes' gene expression levels, proteomics data measuring enzymes' protein concentrations, and metabolomics data measuring metabolite concentrations.

Currently two intracellular human metabolic reconstructions are available; the *Homo sapiens* Recon 1 (*Recon 1*) reconstruction [Duarte et al., 2007] and the Edinburgh Human Metabolic Network (EHMN) reconstruction [Ma et al., 2007]. Both networks consist of a bi-partite graph representation, that is, they comprise networks with two types of nodes namely metabolites and enzymes; a metabolite is connected to an enzyme if it is catalyzed (i.e. produced or consumed) by the particular enzyme. In these metabolic network reconstructions, metabolites are always linked to enzyme nodes and never to each other, and enzyme nodes will never be linked to other enzyme nodes. Both networks are compartmentalized.[1] **Table 5.1** shows the coverage of both reconstructions. The overlap between both reconstructions is hard to quantify, since they use different metabolite nomenclatures that are not easily mapped to each other.

Recon 1 and EHMN reconstructions were assembled in similar manners. First all enzymatic genes in the human genome (along with their correct Enzyme Commission classification number[2] were identified. Next, metabolic reactions were mapped to the enzymes identified in the first step. Annotated information on metabolic reactions was included from several databases, such as the KEGG LIGAND database [Goto et al., 2002], the BioCyc database [Romero et al., 2005], and the Reactome database [Croft et al., 2011]. For the EHMN network assembly manual literature lookups were used as the last step in the construction in order to consolidate inconsistencies in enzyme-reaction mappings and annotations. For the *Recon 1* network, simulations were used to identify gaps in the model, which subsequently were patched by targeted literature searches. Both networks consist solely of interactions with direct physical evidence from literature. Whereas EHMN is mainly aimed at providing a scaffold for data integration, the *Recon 1* model represents a stoichiometrix matrix that can be used for other mathematical analysis like flux balance analyses [Burgard et al., 2004]. Consequently, reconstructions of human metabolism provide relatively complete and high-confidence network models, compared to metabolic databases[3]. In addition to the two metabolic reconstructions described above, there are tissue-specific versions of *Recon 1* [Shlomi et al., 2008] and a liver-specific metabolic network reconstruction [Gille et al., 2010].

[1] The original EHMN was not compartmentalized but recently updated to include compartments [Hao et al., 2010].

[2] http://www.chem.qmul.ac.uk/iubmb/enzyme

[3] For instance approximately 1000 components and reactions in EHMN were not part of the KEGG LIGAND database [Ma and Goryanin, 2008] and several reactions in *Recon 1* were found in literature but not in databases [Duarte et al., 2007].

5.2 Integration of metabolic reconstructions with gene expression data

Metabolic network reconstructions provide scaffolds for integration of high-throughput data, such as gene expression data. They provide a context that enables the direct mapping of measurements to nodes in the network. Differential gene expression analysis per se may provide novel insights at the gene level but miss groups of genes that, only in aggregate, correlate with the trait of interest. For example, in a study of breast cancer [Taylor et al., 2009] the SRC oncogene was not differentially expressed between the cancer patients that were disease-free after follow-up and those who died of cancer. However, when Taylor *et al* compared the co-expression of all the physical interacting gene products of SRC gene, they found that this sub-network was significantly regulated between groups (in terms of coordinated gene expression, and across several independent gene expression studies) [Taylor et al., 2009]. Patil and Nielsen formalized this approach into a framework referred to as Reporter Metabolite Analysis [Patil and Nielsen, 2005]. Whereas Patil and Nielsen applied their method on yeast gene expression data, my colleagues and I applied this method to three different human studies (**Papers IV-V** and [Capel et al., 2009]).

5.3 Paper IV - A method for metabolic biomarker discovery

Cellular metabolic networks are highly interconnected and often tightly regulated; any perturbations at a single node can thus rapidly be conveyed to the rest of the network. Such complexity presents a considerable challenge in pinpointing key molecular mechanisms and biomarkers associated with insulin resistance and type 2 diabetes. In the following paper, we address this problem by using a methodology that integrates gene expression data with the human metabolic reconstructions.

We demonstrate our approach by analyzing gene expression patterns in skeletal muscle. The analysis identified transcription factors and metabolites that represent potential targets for therapeutic agents and future clinical diagnostics for type 2 diabetes and impaired glucose metabolism. In a broader perspective, the study provides a framework for analysis of gene expression datasets from complex diseases in the context of changes in cellular metabolism.

OPEN ACCESS Freely available online

Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes

Aleksej Zelezniak[1☯], Tune H. Pers[2,3☯], Simão Soares[1,4☯], Mary Elizabeth Patti[5], Kiran Raosaheb Patil[1]*

1 Center for Microbial Biotechnology, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, **2** Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, **3** Institute of Preventive Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark, **4** IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, Braga, Portugal, **5** Research Division, Joslin Diabetes Center, Boston, Massachusetts, United States of America

Abstract

Type 2 diabetes mellitus (T2DM) is a disorder characterized by both insulin resistance and impaired insulin secretion. Recent transcriptomics studies related to T2DM have revealed changes in expression of a large number of metabolic genes in a variety of tissues. Identification of the molecular mechanisms underlying these transcriptional changes and their impact on the cellular metabolic phenotype is a challenging task due to the complexity of transcriptional regulation and the highly interconnected nature of the metabolic network. In this study we integrate skeletal muscle gene expression datasets with human metabolic network reconstructions to identify key metabolic regulatory features of T2DM. These features include reporter metabolites—metabolites with significant collective transcriptional response in the associated enzyme-coding genes, and transcription factors with significant enrichment of binding sites in the promoter regions of these genes. In addition to metabolites from TCA cycle, oxidative phosphorylation, and lipid metabolism (known to be associated with T2DM), we identified several reporter metabolites representing novel biomarker candidates. For example, the highly connected metabolites NAD+/NADH and ATP/ADP were also identified as reporter metabolites that are potentially contributing to the widespread gene expression changes observed in T2DM. An algorithm based on the analysis of the promoter regions of the genes associated with reporter metabolites revealed a transcription factor regulatory network connecting several parts of metabolism. The identified transcription factors include members of the CREB, NRF1 and PPAR family, among others, and represent regulatory targets for further experimental analysis. Overall, our results provide a holistic picture of key metabolic and regulatory nodes potentially involved in the pathogenesis of T2DM.

Citation: Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR (2010) Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. PLoS Comput Biol 6(4): e1000729. doi:10.1371/journal.pcbi.1000729

Editor: Christos A. Ouzounis, King's College London, United Kingdom

Received August 26, 2009; **Accepted** March 2, 2010; **Published** April 1, 2010

Funding: MEP appreciates grant support from NIH grants DK062948 and DK060837 and the Graetz Fund. AZ acknowledges support from NOVO scholarship program 2008-9. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: krp@bio.dtu.dk

☯ These authors contributed equally to the work.

Introduction

Type 2 diabetes mellitus (T2DM) is emerging as one of the main threats to human health in the 21st century with an estimated 300 million individuals with T2DM by the year 2025 [1,2]. T2DM is characterized by both insulin resistance (as manifested by reduced insulin-stimulated glucose uptake in skeletal muscle and adipose tissue and inappropriately high hepatic glucose output [3,4]) and reduced insulin secretion by pancreatic β-cells [3,5]. Although the specific molecular pathophysiology remains unclear, many risk factors have been identified for T2DM, including family history of diabetes and prominent environmental factors such as alterations in early life development, excessive food intake, obesity, decreased physical activity and aging [2,3,5]. At the cellular level, multiple regulatory mechanisms and metabolic pathways may contribute to the pathogenesis of insulin resistance, potentially mediated by alterations in insulin signaling [6], mitochondrial oxidative metabolism and ATP production [7–9], fatty acid oxidation [10], or proinflammatory signaling [11]. Similarly, alterations in β-cell development and metabolism [5] may contribute to decreased insulin secretion.

Available human tissue transcriptome data related to T2DM [12,13] provide an opportunity for identification of novel molecular mechanisms underlying the metabolic phenotype of T2DM. This task is challenging due to the need to account for the inherent high connectivity of bio-molecular interaction networks. We have utilized a network-centered methodology to link diabetes-related alterations in gene expression to metabolic hot spots and transcription factors potentially responsible for gene expression changes.

Rationale and methodology

Metabolic phenotypes at a cellular level are essentially characterized by concentrations of metabolites and fluxes through the reactions that make up the metabolic network. Fluxes, in turn, are dependent on metabolite levels, enzyme activities, abundance of effectors and possibly other variables. Measurement of fluxes and metabolite concentrations at the entire metabolic network-scale is, however, a difficult task in humans due to a variety of technological and experimental limitations. By contrast, methods for measurement of expression of genes encoding metabolic enzymes are relatively well-established. Thus, the primary goal of this study is to

Author Summary

Type 2 diabetes mellitus is a complex metabolic disease recognized as one of the main threats to human health in the 21st century. Recent studies of gene expression levels in human tissue samples have indicated that multiple metabolic pathways are dysregulated in diabetes and in individuals at risk for diabetes; which of these are primary, or central to disease pathogenesis, remains a key question. Cellular metabolic networks are highly interconnected and often tightly regulated; any perturbations at a single node can thus rapidly diffuse to the rest of the network. Such complexity presents a considerable challenge in pinpointing key molecular mechanisms and biomarkers associated with insulin resistance and type 2 diabetes. In this study, we address this problem by using a methodology that integrates gene expression data with the human cellular metabolic network. We demonstrate our approach by analyzing gene expression patterns in skeletal muscle. The analysis identified transcription factors and metabolites that represent potential targets for therapeutic agents and future clinical diagnostics for type 2 diabetes and impaired glucose metabolism. In a broader perspective, the study provides a framework for analysis of gene expression datasets from complex diseases in the context of changes in cellular metabolism.

use informatics approaches to integrate available gene expression data with metabolic networks, in order to predict metabolic phenotypes of skeletal muscle linked to the pathogenesis of type 2 diabetes. Such an approach will help not only to gain insight into the organization of transcriptional regulation in human tissues, but also provide guidance for improved design of experimental strategies for obtaining metabolite and flux data, which can be further integrated into metabolic models.

To achieve these goals, we applied an extension of the algorithm described in [14] (for various applications of this algorithm see [14–18]), which enables identification of so-called reporter metabolites, or metabolic hot spots around which transcriptional regulation is centered (Figure 1A). This analysis is based on the assumption that under most conditions of physiological interest, fluxes through enzymes connected to a metabolite are coordinated in order to maintain physiological homeostasis, or to eventually reach a new (pseudo-) steady state. Moreover, transcriptional regulation of expression of genes encoding critical enzymes in metabolic flux pathways facilitates concordance with the metabolic demands of the cell and corresponding stoichiometric and thermodynamic constraints on fluxes. For this analysis, we used two recently published human metabolic network models: i) *Homo sapiens* Recon1 [19], and ii) Edinburgh Human Metabolic Network (EHMN) [20].

We further hypothesized that the observed coordinated changes around reporter metabolites can be, at least in some cases, attributed to common transcriptional regulatory mechanisms. Specifically, we hypothesize that the neighbor enzymes of reporter metabolites may share one or more transcription factor binding sites in the promoter regions of the corresponding genes. In order to identify such potential regulatory players, we tested promoter sequences of the genes associated with the reporter metabolites for enrichment of known transcription factor binding motifs (Figure 1B). Transcription factors identified in this fashion provide

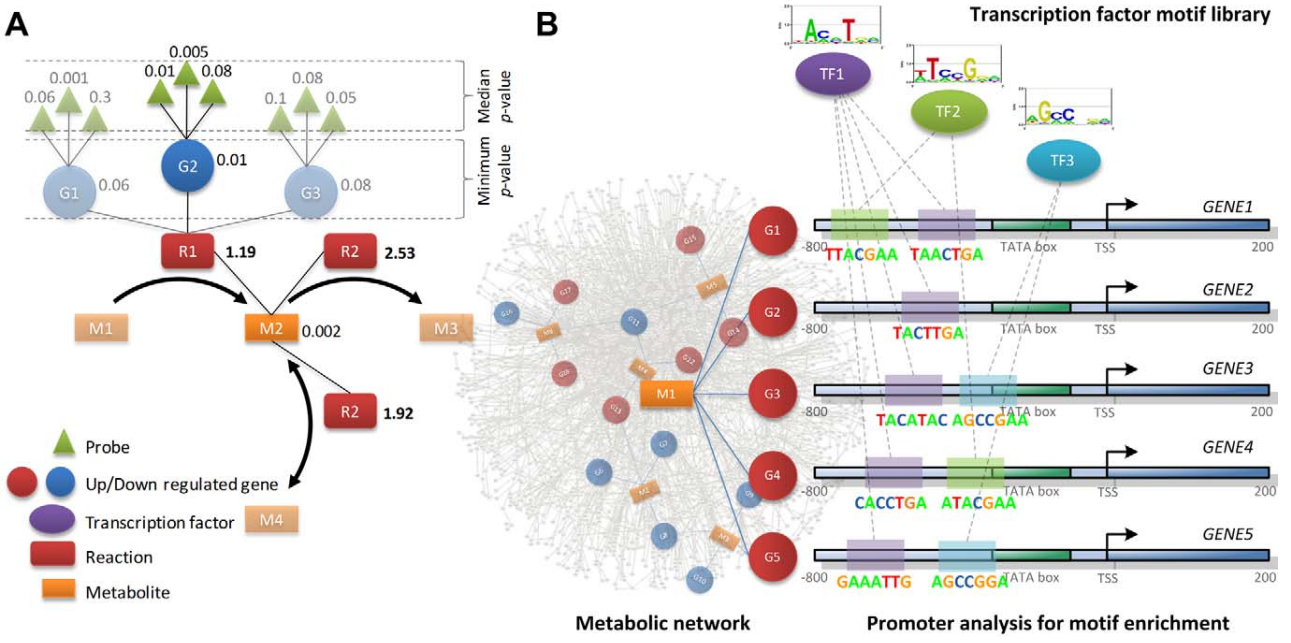


Figure 1. Schematic overview of the methodology used for the identification of reporter metabolites and associated putative regulatory sequence motifs. A) Scoring system for identification of reporter metabolites. Each metabolite is scored based on the scores of the associated enzyme-catalyzed reactions. Each enzyme, in turn, is assigned a score based on median of the p-values of the probes representing the corresponding gene. In case of a reaction catalyzed by an enzyme complex or a set of isozymes, minimum of the p-values of the corresponding enzymes is chosen. Numbers in bold are Z-scores for each reaction, the rest of the numbers represent p-values (significance of differential expression). B) Identification of transcription factor binding motifs. For a reporter metabolite, a set of up/down regulated neighbor (enzyme-coding) genes is selected. Promoter regions, upstream of transcription start site (TSS) of each of the selected genes are assessed for the enrichment of known transcription factor (TF) binding sequence motifs.
doi:10.1371/journal.pcbi.1000729.g001

clues to the regulatory mechanisms that lead to observed gene expression changes in the metabolic network.

Since our goal is to identify reporter metabolites and transcription factors potentially involved in diabetes pathogenesis and progression, we analyzed two independent studies of skeletal muscle transcriptomics in individuals with established type 2 diabetes or insulin resistance [8,9] (Text S1). In the first study [8], biopsies were obtained following insulin stimulation from a cohort of 43 Swedish men of Caucasian ethnicity with a spectrum of glucose tolerance, including 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT), and 18 with established T2DM. The second dataset [9] was derived from a cohort of 15 subjects of Mexican American ethnicity, in whom muscle biopsies were performed in the fasting state. Importantly, this cohort included individuals with not only established diabetes (5 subjects, T2DM), but also individuals with completely normal glucose tolerance but a spectrum of insulin resistance; normal glucose tolerant subjects were subdivided by family history-linked diabetes risk (4 family history positive, more insulin resistant subjects, FH+; and 6 family history negative, more insulin sensitive subjects, FH−). With this approach, the individual contributions of isolated insulin resistance and diabetes risk (in the setting of normoglycemia, FH+), mild elevations in postprandial glucose (IGT), and established diabetes can be individually assessed. Moreover, the possible contribution of family history, potentially mediated by genetics or shared environment, can be assessed. Thus, we predict that analysis of the common patterns resulting from the two datasets will identify regulatory signatures potentially independent of study cohort and design variation but common to the pathophysiology of insulin resistance and diabetes.

Results

In present study, we performed reporter metabolite analysis based on pair-wise comparisons within each dataset; differential expression and its significance were assessed with robust multi-array average (RMA) and empirical Bayes testing. Significance of differential expression for each gene was used as a scoring metric (Materials and Methods). The results are summarized as metabolic signatures (reporter metabolites) and regulatory signatures (transcription factors) for T2DM.

Metabolic signatures of T2DM

Swedish male dataset. Reporter metabolite analysis for three pair-wise comparisons, *viz.*, T2DM *vs* NGT, T2DM *vs* IGT, and IGT *vs* NGT, revealed significant reporter metabolites (p-value$\leq$0.05) participating in lipid metabolism, TCA cycle, oxidative phosphorylation (OXPHOS) and glycolysis (Table 1, Table 2, Table S1 and Table S2). Among reporter metabolites identified for the T2DM *vs* NGT comparison were lipid species 1,2-diacyl-sn-glycerol (DAG), acetoacetyl-CoA, and the sphingolipid sphinganine. These are interesting, as prior studies [3,21–23] have demonstrated that the related lipid molecules diacylglycerols (DAG), long-chain fatty acyl CoAs, and ceramides correlate positively with triglyceride content and inversely with insulin sensitivity [5] and have been shown to induce insulin resistance [3]. Furthermore, given that saturated fatty acids appear to play a particularly important pathogenic role in insulin resistance [24], it is interesting that several metabolites of saturated fatty acids (such as hexanoyl-CoA, palmitoyl-CoA, tetradecanoyl-CoA, lauroyl-CoA, decanoyl-CoA and butanoyl-CoA) were found as reporter metabolites with mostly up-regulated neighboring genes in the IGT *vs* NGT comparison (Table 1 and S2), and thus may serve as potential markers of insulin resistance and IGT.

TCA cycle metabolites citrate and 2-oxoglutarate, with down-regulated neighboring genes, were also uncovered as reporter metabolites in the T2DM *vs* NGT comparison (Table 1, S1 and S2). These results are concordant with a study of human urine metabolome profiles from patients with T2DM [25], in which levels of citrate and 2-oxoglutarate were lower in T2DM compared to healthy controls [26]. Among other mitochondrial metabolites, reduced and oxidized forms of cytochrome c and ubiquinol were identified as reporter metabolites (T2DM *vs* NGT, Table S1) with down-regulated expression of the associated genes.

Impaired glucose tolerance typically reflects an important transition between normoglycemia and overt diabetes, reporter metabolites which are identified in both IGT *vs* NGT and T2DM *vs* NGT, but not significantly different in the T2DM *vs* IGT comparison (e.g. phosphatidylethanolamine, 2-hydroxyglutarate, 2-oxoglutarate, 3′,5′-cyclic AMP, ATP, Table S1 and S2) may be considered novel biomarkers of early-stage glucose intolerance.

Mexican-American dataset. We similarly performed reporter metabolite analysis using both Recon1 and EHMN metabolic models in the Mexican-American dataset. This analysis revealed significant transcriptional regulation in metabolite nodes in TCA cycle, oxidative phosphorylation, and lipid metabolism, for both T2DM *vs* FH− and FH+ *vs* FH− comparisons (Table 2). Similar to the Swedish Caucasian dataset, metabolites involved in oxidative phosphorylation (e.g. ferrocytochrome c, H+, and fumarate) were among the top-ranking reporter metabolites, identified in both the T2DM *vs* FH− and FH+ *vs* FH− comparisons (Table 2, Table S3). Interestingly, urinary levels of fumarate, an important link between the TCA cycle and oxidative phosphorylation, were recently found to be decreased in T2DM patients [25].

Analysis using the EHMN model revealed TCA cycle-related metabolites, including 3-carboxy-1-hydroxypropyl-ThPP, aconitate, succinyl-CoA, malate and fumarate, as significant reporter metabolites (p-value$\leq$0.05), with mostly down-regulated expression of the genes encoding their neighboring enzymes. Ubiquinol was found as reporter metabolite representative of electron transfer chain. Several molecules within β-oxidation pathways, such as 3-cis-dodecenoyl-CoA, glutaryl-CoA, trans-3-decenoyl-CoA, 3-methylbutanoyl-CoA and 3-methylcrotonyl-CoA, as well as in amino acid (leucine, lysine) metabolism were also identified as reporters (Table 2, Table S4). Moreover, glutamate, glycerol derivatives, phosphocreatine, a number of hormone derivatives and many others (Table S3 and S4) were found as significant reporter metabolites in the T2DM *vs* FH− comparison.

Overlapping reporter metabolites between two study populations. In order to determine the extent of overlap between the two study populations, we performed a cluster analysis of the pair-wise comparisons within the Swedish and Mexican-American datasets (Figure 2). Jaccard distance metric between two pair-wise comparisons (e.g. T2DM *vs* FH− and FH+ *vs* FH−) was calculated based on the overlap of reporter metabolites between the two comparisons. Jaccard distance provides a measure of dissimilarity between two sets of reporter metabolites, and is quantified as the fraction of non-overlapping reporter metabolites between the two sets. While similar clustering patterns were observed (Figure 2A and Figure S1A) independent of the use of either EHMN or Recon1 metabolic model, Swedish and Mexican-American studies clustered separately, perhaps related to differences in study population, study design (e.g. fasting studies in Mexican-Americans, insulin-stimulated studies in Swedish) or differences in microarrays used (thus differing in the coverage of metabolic enzymes). We observed substantial overlap between the T2DM *vs* FH− and FH+ *vs* FH− comparisons, suggesting that insulin resistance patterns could contribute to these findings.

Table 1. Reporter metabolites for Swedish male dataset.

Reporter Metabolite	P-values		Enzyme neighbors (Up-regulated:Down-regulated)		
	T2DM/NGT	IGT/NGT	T2DM/NGT	IGT/NGT	
Citrate	**0.047**	0.646	1:0	1:0	TCA cycle
Succinyl-CoA	**0.013**	0.285	2:3	2:3	
2-Hydroxyglutarate*	**0.002**	**0.023**	0:1	0:1	
*2-Oxoglutarate**	**0.049**	**0.047**	8:11	8:11	
Ferrocytochrome C; Ferricytochrome C	**0.006**	**0.032**	1:2	0:3	Oxidative phosphorylation
Ubiquinone-10	**0.017**	0.769	0:5	1:4	
Ubiquinol-10	**0.022**	0.484	0:4	1:3	
Phosphoenolpyruvate*	0.196	**0.037**	1:3	1:3	Glycolysis
D-Glyceraldehyde*	0.083	**0.017**	2:1	3:0	
D-Alanine	**0.016**	0.330	0:3	0:3	Amino acid metabolism
L-Alanine	**0.047**	0.319	3:7	3:7	
3-Methylglutaconyl-CoA†	**0.038**	0.816	0:2	1:1	
*L-Leucine**	**0.047**	0.109	1:3	1:3	
*1,2-Diacyl-sn-glycerol (DAG)**	**0.022**	**0.049**	2:5	2:5	Lipid metabolism
1D-myo-Inositol 1,4-bisphosphate†	0.060	0.151	0:3	2:1	
3-Dehydrosphinganine*	0.232	**0.035**	1:1	2:0	
Acetoacetyl-CoA*	**0.009**	0.462	1:4	2:3	
Butanoyl-CoA†	0.365	**0.038**	0:2	1:1	
*Decanoyl-CoA; Lauroyl-CoA**	0.268	**0.033**	1:2	2:1	
Fatty acid*	**0.021**	0.756	3:4	3:4	
Lophenol*§	**0.007**	0.749	0:1	0:1	
*Palmitoleoyl-CoA**	0.238	**0.019**	1:3	2:2	
*Palmitoyl-CoA**	0.179	**0.014**	3:4	6:1	
Phosphatidyl glycerol phosphate	**0.047**	0.316	0:1	0:1	
Phosphatidylinositol 4,5-bisphosphate	0.097	**0.001**	1:5	2:4	
Propanoyl-CoA*	0.259	**0.016**	2:5	2:5	
Prostaglandin E2	**0.036**	**0.032**	0:3	1:2	
Sphinganine*	**0.038**	0.283	1:3	2:2	
(Gal)3 (GalNAc)1 (Glc)1 (Cer)1*	**0.023**	**0.034**	1:2	1:2	Other
AMP†	**0.041**	0.218	7:17	6:17	
ATP†	**0.003**	**0.010**	28:60	27:60	
cAMP†	**0.033**	**0.049**	2:0	2:0	
CDPcholine	**0.020**	0.122	0:2	0:2	
Choline phosphate	**0.030**	0.573	0:2	1:1	
NAD+*	0.333	**0.020**	29:34	34:34	
Phosphocreatine	**0.025**	0.176	0:1	1:0	
Trichloroethanol*	**0.020**	**0.038**	1:2	3:0	

*Reporter metabolites identified using EHMN metabolic network.
†Reporter metabolites identified in both networks.
§Plant metabolite, likely to be present in the EHMN due to incorrect annotation.
Reporter metabolites with $p \leq 0.05$ in at least one of the comparisons showed in bold. Columns with enzyme neighbors show the number of up- and down-regulated enzyme neighbors in the first condition (e.g. T2DM/NGT up- and down-regulated in T2DM comparing with NGT) for each of comparisons. Reporter metabolites without marks were identified using Recon1 metabolic network. Metabolites written in italics are known to be directly/indirectly related to T2DM, see main text and Table S8.
doi:10.1371/journal.pcbi.1000729.t001

We next examined the overlap of reporter metabolites between the two case studies (Figure 2B, Figure S1B, Table S5 and Table S6). Owing to differences in the metabolite-gene connectivity between EHMN and Recon1, the number of overlapping reporter metabolites is generally higher for the EHMN analysis. To a large extent, this difference is due to the groups of metabolites in EHMN that share the same gene neighbors (whether two metabolites share the same gene neighbors depends not only on the network used, i.e. number of distinct biochemical reactions associated with a particular enzyme, but also on the coverage of genes on the particular microarray chip used). In addition to many other metabolites, phosphocreatine appeared as a significant

Table 2. Reporter metabolites for Mexican-American dataset.

Reporter metabolite	P-values		Enzyme neighbors (Up-regulated:Down-regulated)		
	T2DM/FH−	FH+/FH−	T2DM/FH−	FH+/FH−	
2-Oxoglutarate	**0.001**	**0.001**	2:7	2:7	TCA cycle
L-Malate	0.098	**0.029**	1:4	2:3	
Succinyl-CoA†	**0.011**	**0.009**	0:5	0:5	
Ferrocytochrome C;Ferricytochrome C	**0.008**	**0.007**	0:3	0:3	Oxidative phosphorylation
Fumarate	**0.019**	**0.025**	0:2	0:2	
Ubiquinone-10†;Ubiquinol-10†	**0.040**	**0.021**	1:3	1:3	
2,3-Disphospho-D-glycerate†	**0.021**	**0.004**	0:1	0:1	Glycolysis
2-Phospho-D-glycerate*	**0.038**	**0.006**	0:2	1:1	
beta-D-Fructose*	**0.049**	**0.038**	0:2	0:2	
D-Fructose 2,6-bisphosphate	**0.037**	0.136	0:2	0:1	
D-Fructose 6-phosphate	**0.013**	0.119	4:6	3:7	
D-Glucose*	**0.037**	0.066	0:7	1:5	
D-Glucose 6-phosphate	**0.009**	**0.014**	1:3	1:3	
D-Glycerate 2-phosphate	**0.026**	**0.003**	0:2	1:1	
L-Lactate	**0.048**	0.067	1:2	1:2	
Phosphoenolpyruvate	0.079	**0.048**	2:2	3:1	
Pyruvate	**0.042**	0.202	1:6	1:6	
2-Oxoadipate*	**0.002**	**0.004**	0:1	0:1	Amino acid metabolism
beta-Alanine	**0.031**	**0.027**	1:1	1:1	
L-Glutamate†	**0.025**	**0.009**	1:1	1:1	
(R)-2-Methyl-3-oxopropanoyl-CoA*	**0.043**	0.118	0:2	0:1	Lipid metabolism
*1,2-Diacyl-sn-glycerol (DAG)**	**0.036**	0.117	3:2	5:1	
1D-myo-Inositol 1,4-bisphosphate	**0.025**	0.054	1:2	1:2	
3-cis-Dodecenoyl-CoA*	**0.009**	**0.039**	0:3	0:3	
Acylglycerol*; 2-Acylglycerol*	**0.035**	**0.018**	1:1	1:1	
Glutaryl-CoA†	**0.007**	**0.015**	0:2	0:2	
Glycerol	**0.020**	**0.001**	1:1	1:1	
Glycerol 3-phosphate	0.051	**0.005**	2:1	2:1	
Lipoamide*	**0.014**	**0.006**	0:5	0:5	
Phosphatidylinositol	**0.017**	0.128	1:5	1:5	
trans-3-decenoyl-CoA*	**0.026**	0.076	0:2	0:2	
ADP	**0.047**	0.174	16:31	20:27	Other
CO_2	**0.041**	**0.004**	1:11	3:9	
Coenzyme A†	**0.007**	**0.014**	4:8	3 10	
Creatine;Phosphocreatine†	**0.032**	**0.048**	0:1	0:1	
NAD+†; NADH†	**0.003**	0.095	3:17	17:4	
Trichloroethanol*	**0.021**	**0.006**	2:1	3:0	

*Reporter metabolites identified using EHMN metabolic network.
†Reporter metabolites identified in both networks.
Reporter metabolites with $p \leq 0.05$ in at least one of the comparisons showed in bold. Columns with enzyme neighbors show the number of up- and down-regulated enzyme neighbors in the first condition (e.g. T2DM/FH− up- and down-regulated in T2DM comparing with FH−). Reporter metabolites without marks were identified using Recon1 metabolic network. Metabolites written in italics are known to be directly/indirectly related to T2DM, see main text and Table S8.
doi:10.1371/journal.pcbi.1000729.t002

reporter in both case studies, *viz.*, for T2DM *vs* NGT and T2DM *vs* FH− comparisons. Phosphocreatine is an important energy reservoir metabolite in skeletal muscle, and defects in recovery of phosphocreatine have been identified *in vivo* in humans with insulin resistance [27] and diabetes [28]. Interestingly, low levels of urinary creatine have also been found in patients with T2DM [25].

Regulatory signatures of T2DM

In order to link the identified reporter metabolites to regulatory pathways controlling gene expression, we hypothesized that enzymes associated with reporter metabolites would be regulated by common transcription factors. As potential candidates subjected to such regulation, we selected all reporter metabolites with at least 5

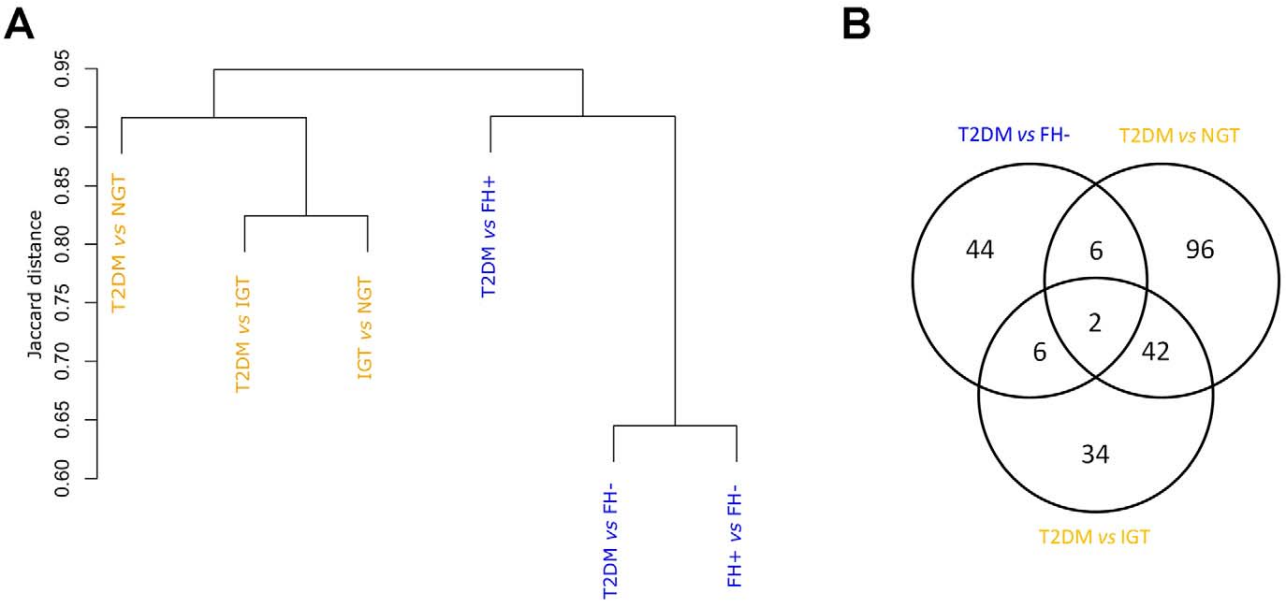


Figure 2. Hierarchical clustering of pair-wise comparisons within the Swedish male and Mexican-American datasets based on the overlapping reporter metabolites (Recon1 model). Comparisons are colored according to the dataset; blue – Mexican-American; orange – Swedish male dataset. A) Dendrogram of reporter metabolites identified in each of the comparisons based on Jaccard distance. B) Venn diagram showing the overlap of the reporter metabolites identified in the different comparisons.
doi:10.1371/journal.pcbi.1000729.g002

up- or down-regulated neighboring genes (Materials and Methods). Up- and down-regulated gene sets were then analyzed separately in order to assess whether their promoter regions were enriched for known transcription factor binding sequence motifs. P-values for enrichment were estimated by using a hypergeometric test, which compared the proportion of promoters from a given gene set containing a particular motif with the frequency of occurrence of that motif in promoter regions of all other metabolic genes. Correction for multiple-testing was done by using q-value [29] and motifs with q-value$\leq$0.05 were considered as significantly enriched.

In accord with our hypothesis, several transcription factor binding sites were overrepresented in the promoter regions of the enzymes associated with reporter metabolites. A summary of the main results from this analysis is illustrated in Figure 3A. Many transcription factors were found to be common across the two case studies (Figure 3B), *albeit* in connection with different reporter metabolites. PPAR family motifs (PPARγ and PPARα:RXRα) were enriched in seven downregulated enzyme sets including ATP. Tax/CREB motifs were enriched in promoters of downregulated enzymes associated with ATP, ADP and phosphate. Additional down-regulated neighbors of ATP were enriched for the binding sites of NF-κB, MEF-2, UF1-H3β, Pax-9 and NKX6.2, while the NRF-1 motif was enriched in the set of up-regulated enzymes neighboring ADP. Another potential regulatory signature was identified around the down-regulated neighbors of phosphatidylinositol and phosphatidylinositol 4,5-bisphospate (important phospholipids which participate in insulin and other signaling reactions), which were significantly enriched for binding sites of p53, PPARγ, SRF, SEF-1, v-Jun, GCNF, AR and many others (Table S7). These and other highly connected reporter metabolites in the metabolite-TF network (Figure 3A) demonstrate the concept that associated metabolic pathways can be transcriptionally regulated in multiple ways in response to environmental stimuli or metabolic perturbation.

Discussion

Maintenance of whole-body glucose metabolism is reliant on a delicately balanced dynamic interaction between tissue sensitivity to insulin (including muscle, adipose and liver) and insulin secretion [5,30]. Unfortunately, the molecular mechanisms responsible for diabetes risk remain unknown. A key metabolic phenotype associated with insulin resistance in humans is inappropriate lipid accumulation in tissues outside of adipose tissue, suggesting defects in fatty acid uptake, synthesis, and/or oxidation. With lipid excess and/or impaired oxidation, as observed in obesity and/or inactivity, flux of long-chain acyl CoAs (LC-CoA) may be redirected into cytosolic lipid species such as diacylglycerols (DAG), triacylglycerols (TG) and ceramides (derivatives of sphingosine and fatty acid metabolism) [5] that are correlated with reductions in insulin signaling and insulin resistance [3,21–23,31]. Whether alterations in mitochondrial oxidative function in humans with insulin resistance and diabetes contribute to, or are a consequence of these defects, remains unclear [32].

Recognizing these important gaps in our knowledge of diabetes pathophysiology, we have integrated transcriptomic data with metabolic networks to systematically identify, in an unbiased fashion, regulatory hot spots (reporter metabolites and associated transcription factors) associated with insulin resistance and T2DM. Our reporter metabolite results provide evidence for transcriptional dysregulation of multiple metabolic pathways in skeletal muscle. Interestingly, many of the reporter metabolites identified in our analysis have been appreciated in prior experimental studies in animal models (metabolites with italic font in Tables 1, 2 and S8). A bird's-eye view of selected metabolic and regulatory nodes identified in our study is depicted in Figure 4.

Key metabolic regulatory nodes in T2DM pathogenesis

Lipid metabolism. In conditions of overnutrition and physical inactivity, availability of cellular fatty acids stimulate

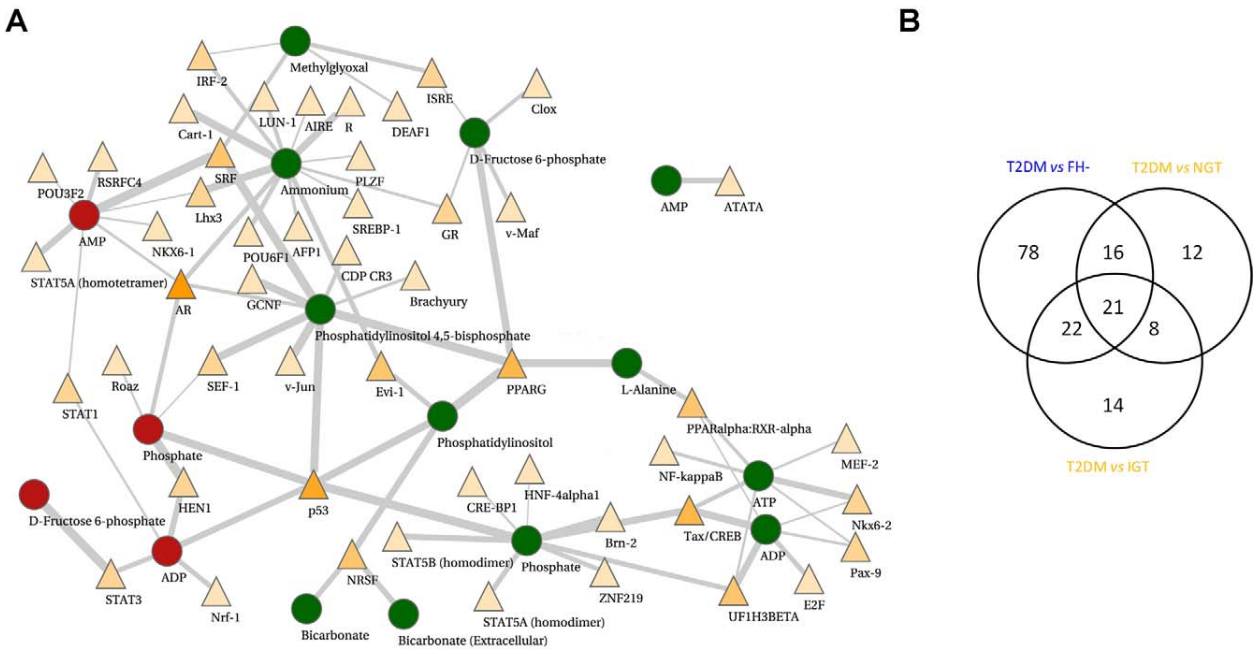


Figure 3. Summary of the main results from the motif enrichment analysis. A) Motif enrichment analysis for the genes associated with reporter metabolites from the T2DM *vs* NGT comparison. Reporter metabolites with up-regulated neighboring gene set are shown as red circles, whereas reporter metabolites with down-regulated neighboring gene set are represented as green circles. Transcription factor binding motifs (shown as triangles) are colored according to the number of enzyme sets in which they are enriched, ranging from light yellow (enriched in few sets) to orange (enriched in as many as 6 sets). Edges are scaled according to q-values signifying the confidence of the motif enrichment. B) Venn diagram showing the overlap of transcription factor binding motifs across the comparisons of T2DM with non-T2DM cases. Comparisons are colored according to the dataset; blue – Mexican-American; orange – Swedish male dataset.
doi:10.1371/journal.pcbi.1000729.g003

ligand–dependent PPARα/δ transcription factors which, in turn, induce transcription of genes responsible for β-oxidation [33,34]. Metabolic byproducts of incomplete β-oxidation, such as acylcarnitines and reactive oxygen species, may accumulate in mitochondria and contribute to insulin resistance [5]. Interestingly, our analysis identified enrichment of PPAR family transcription factor binding motifs in T2DM as compared with insulin sensitive subjects, in both the Swedish and Mexican-American datasets (T2DM *vs* NGT and T2DM *vs* FH−, respectively). Moreover, reporter analysis revealed lipid metabolites (Table S1), known to be natural ligands of PPARγ (prostaglandins) [34].

Another reporter metabolite identified in our analysis is diacylglycerol (DAG), a lipid signaling molecule known to inversely correlate with insulin sensitivity [3,21–23,31]. Our results suggest that perturbations in DAG levels may be accompanied by changes in the adjacent CDP-Choline branch of the Kennedy pathway of phospholipid metabolism (Figure 4). Thus, DAG could potentially affect insulin sensitivity *via* activation of serine/threonine kinases or alterations in phospholipid membrane composition, both of which could lead to defects in insulin signaling, reduced insulin-stimulated glucose uptake, and glycogen synthesis – key metabolic features of diabetes [5] (Figure 4). Together, identification of these lipid-linked regulatory motifs and reporter metabolites known to be involved in type 2 diabetes pathogenesis provides further support for the validity of our approach.

Central carbon metabolism. Using our approach we found several reporter metabolites from the TCA cycle (citrate, 2-oxoglutarate, succinyl-CoA, fumarate and malate) (Figure 4). The down-regulated genes associated with these metabolites support the idea that TCA cycle and/or oxidative phosphorylation flux is reduced in diabetes [9]. It is also interesting that ATP is one of the reporter metabolites, as the majority of cellular ATP is generated *via* respiration. Moreover, significant enrichment of binding motif for NF-κβ in the upregulated ATP neighbors is consistent with the potential role of this transcription factor in mediating oxidative stress responses triggered by by-products of incomplete β-oxidation [35]. Another interesting finding is the enrichment of CREB family and NRF-1 motifs in enzymes associated with ATP and ADP. These results corroborate the role of CREB as an indirect regulator of nuclear-encoded oxidative phosphorylation genes *via* PGC1-α and other regulators linked to nuclear-encoded mitochondrial genes (Figure 4) [9,36,37].

The appearance of highly connected metabolites, such as ATP and NADH, among top-ranking reporter metabolites provides a possible link to the observed network-wide transcriptional changes in IGT and T2DM. Cellular levels of these co-factors are usually constrained within relatively narrow ranges to maintain thermodynamic stability. Oxidative phsophorylation, which is connected to TCA cycle flux *via* succinate and fumarate, accounts for most of the ATP (and NADH) turnover in a respiring cell. Our results suggest reduction in the activity of both TCA cycle and oxidative phosphorylation, in agreement with recent NMR data demonstrating that mitochondrial ATP synthesis is reduced in humans with insulin resistance [38–40]. Another major source of ATP and NADH production in the cell is glycolysis. Reporter metabolites representative of glycolysis (glucose, glucose-6-phosphate, glucose-1-phosphate and pyruvate) also exhibited concordant down-regulation of the neighboring genes.

The concordance between the changes in gene expression levels for glycolysis, TCA cycle and oxidative phosphorylation in IGT and T2DM suggests that transcriptional regulatory mechanisms

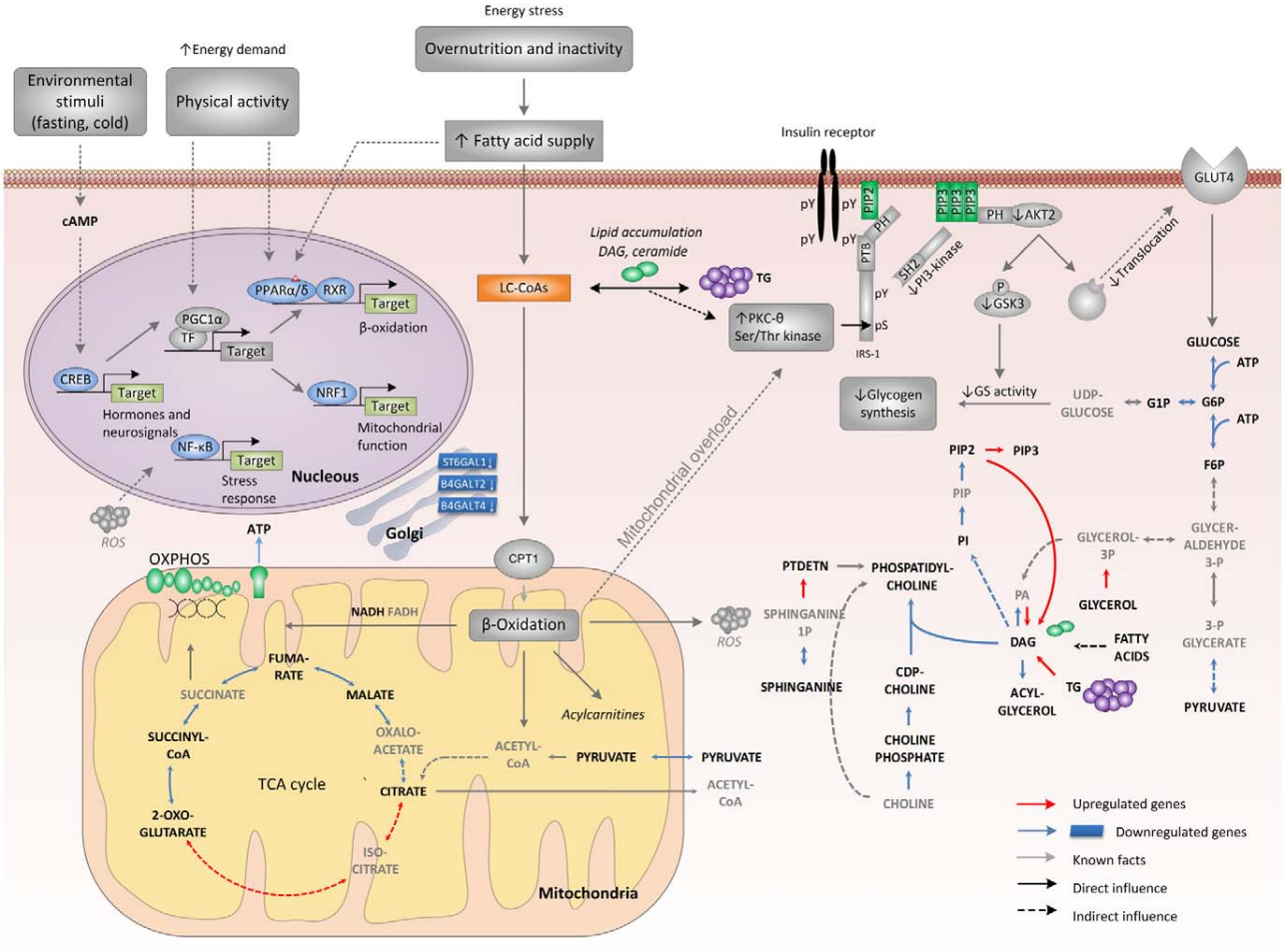


Figure 4. Metabolic and regulatory signatures of type 2 diabetes. Key metabolic and regulatory pathways associated with reporter metabolites identified in this study (T2DM *vs* NGT and T2DM *vs* FH− comparisons) are shown. Metabolites in bold black font are reporter metabolites. Grey shapes and arrows represent facts/hypotheses from previous studies and are not directly based on the results from the present study. Broken lines imply indirect effect while full lines denote direct effect. Chronic overfeeding and physical inactivity increase the influx of fatty acid, which promotes β-oxidation through the activation of PPARα/δ-mediated genes, without coordinated increase in TCA cycle flux. Reporter analysis supports this idea by showing the decreased activity in TCA cycle enzymes associated with reporter metabolites. Eventually, this leads to mitochondrial accumulation of metabolic by-products of incomplete β-oxidation (acylcarnitines ROS). These stresses might lead to mitochondrial overload which together with intracellular lipid-signaling (such as DAG) molecules might trigger serine a serine/threonine (Ser/Thr) kinase (Ser/Thr) cascade initiated by nPKCs. As a result, Ser/Thr phosphorylation of insulin receptor substrate 1 (IRS-1) sites is induced, thereby inhibiting IRS-1 tyrosine phosphorylation and activation of PI 3-kinase, resulting in impeded GLUT4 translocation, reduced glucose transpor, and decreased glycogen synthesis. Increased physical activity/fasting activates PGC1α and CREB (a potent inducer of PGC-1α). These actions combat lipid stress by increasing TCA cycle flux and by coupling ligand-induced PPARα/δ activity with PGC1α-mediated remodeling of downstream metabolic pathways such as respiration and β-oxidation. CDP-choline, cytidine diphosphate choline; DAG, diacylglycerol; G1P, glucose 1-phosphate; G6P, glucose 6-phosphate; GLUT4, glucose transporter-4; GSK3, glycogen synthase kinase-3; IRE1, inositol requiring kinase-1; LC-CoAs, long-chain acyl CoAs; nPKCs, novel protein kinase Cs; PA, phosphatidate; PGC1α, PPARγ co-activator-1α; PH, pleckstrin homology domain;PI, phospatidylinositol; PIP, phospatidylinositol 4-phospate; PIP2, phosphatidylinositol 4,5-bisphospate, PIP3, phospatidylinositol 3,4,5-trisphospate; PI 3-kinase, phosphoinositol 3-kinase; PPARγ, peroxisome proliferator-activated receptor-γ; PTB, phosphotyrosine binding domain; ROS, reactive oxygen species; RXR, retinoid X receptor; SH2, src homology domain; TCA, tricarboxylic acid cycle; TF, transcription factor; CPT1, carnitine palmitoyltransferase-1; PTDETN, phosphatidylethanolamine.
doi:10.1371/journal.pcbi.1000729.g004

may be a response to altered levels of ATP/NADH. Such response may achieve two purposes: (1) regulation of metabolism on global scale, as these co-factors are critical components of many metabolic pathways, and (2) regulation of NADH levels may help in reducing excessive (and potentially deleterious) oxidative stress resulting from sustained oxidation of excessive nutrients [41]. Although the way such regulatory control is mechanistically linked to the corresponding metabolites cannot be deduced from the gene expression data alone, there are several examples where metabolite co-factors are directly involved in regulating gene expression, e.g. NADH(/+) dependent regulation of genes in gram-positive bacteria [42], yeast [43–45] and human [46,47]. NAD+ dependent changes in gene expression levels could also be mediated by the action of PGC-1α and SIRT1 complex, which have important roles in regulation of glucose homeostasis [48]. Additional regulatory links, between glycolytic flux, energy metabolism, TCA cycle flux and fatty acid metabolism are also known in other eukaryotic systems such as baker's yeast [49–51]. Furthermore, several of the enzymes from central carbon metabolism may be regulated to a large extent at the post-transcriptional level [52,53]. Parallels of such regulatory

circuits in human cells may be discovered in the future with the here-identified transcription factors (Table S7) as one of the starting points.

Other pathways. Metabolites involved in protein and lipid glycosylation were found as reporters and characterized by down-regulation of neighboring enzymes (Table S2). Alterations in glycosylation may ultimately cause misfolding of several proteins, a feature previously associated with over-nutrition in hepatocytes [5]. Another reporter metabolite, shared by T2DM *vs* NGT and T2DM *vs* FH− comparison, is trichloroethanol, a metabolite in the cytochrome P450-mediated pathway derived from trichlorethene [54]. Although tricholoethanol or tricholoethene is not an endogenous metabolite in human tissues, it appears that the expression of the cytochrome P450 is altered in T2DM. Interestingly, experimental evidence shows that mouse exposure to trichlorethene leads to PPARα activation and the reprogramming of gene expression, resulting in induction of enzymes mediating β- and ω-oxidation of fatty acids, and increased expression of genes involved in lipid metabolism [55], a pattern similar to the T2DM metabolic phenotype [3].

Reporter metabolites and macroscopic physiological parameters

The identification of reporter metabolites from glycolysis and energy-generation pathways suggests that there may be regulation of certain physiological parameters, such as glucose uptake, at the transcriptional level of the corresponding metabolic pathways. To investigate the extent of such possible regulation, we calculated Pearson correlation coefficients between insulin sensitivity (as measured by either whole-body glucose uptake during the hyperinsulinemic euglycemic clamp or insulin levels achieved during the OGTT) and mean centroid expression levels of genes surrounding reporter metabolites (Swedish dataset) (Materials and methods). A significant linear correlation with whole-body glucose uptake was observed for several reporter metabolites. In most cases, the correlation was significant only for one of the conditions (NGT, IGT or T2DM). For example, significant correlation of transcriptional regulation around dUDP with glucose uptake was found only for NGT samples (Figure 5A). It appears that this potential connection is de-linked under IGT and T2DM conditions. Another example is 1-Phosphatidyl-1D-myo-inositol 3-phosphate (Figure 5B), where significant correlation is observed with insulin level only for IGT. Further investigation of the causal mechanisms behind these observed correlation patterns may help in elucidating the regulatory role of the reporter metabolites in diabetes pathogenesis.

Potential biomarkers and pharmacological targets

A key scientific and clinical challenge is to identify molecular markers of diabetes risk, not only to better understand disease pathophysiology, but also to develop novel therapies for prevention and treatment of established diabetes. In this context, it is interesting that our analysis identified both PPARγ and its potential lipid ligands as regulatory molecules, since PPARγ ligand thiazolidinediones are currently employed as effective therapy for diabetes. We hypothesize that some transcriptional pathways identified in the current analysis, including CREB, NRF-1 and SRF, may be additional novel molecular mediators of the transcriptomic phenotype associated with insulin resistance, and thus potential targets for future intervention strategies. Of course, the potential roles of these pathways will require additional testing in cultured cells and animal models, where their impact on metabolic flux and insulin sensitivity can be fully assessed.

Similarly, reporter metabolites identified in our analysis represent molecules likely to be involved in human skeletal muscle insulin resistance phenoytpes and also novel candidate biomarkers of insulin resistance and diabetes risk. In support of this hypothesis, several of the identified metabolites have known physiological roles in T2DM (Table S8 and Discussion above). Additional molecules have been analyzed either in rodents and/or in other tissues (Table S8) and thus, their appearance as reporter metabolites also strongly implicates their involvement in insulin resistance in human skeletal muscle. Some of the novel metabolites identified in our analysis, including glycolytic and fatty acid oxidation intermediates, are known targets of metformin, a compound effective for diabetes therapy and prevention (Figure 4). We also identified an interesting link between DAG, a reporter metabolite for T2DM, and the CDP-choline branch of the Kennedy pathway of phospholipid metabolism (Figure 4). This pathway has been implicated in cancer development and is being established as anti-tumor drug target [56,57]. Changes in phospholipid metabolism are known to affect the properties of cellular membranes, and subsequently signaling through membrane proteins. Further investigation of the role of phospholipids in T2DM pathogenesis may provide clues to some of the missing links that connect metabolic flux changes with insulin signaling in skeletal muscle cells.

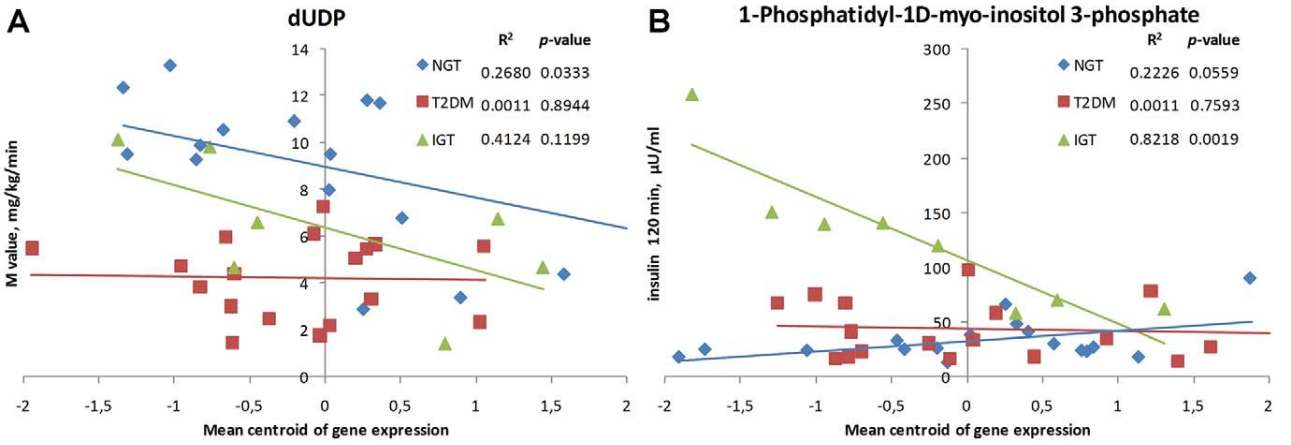


Figure 5. Correlation of glucose uptake and insulin level with mean centroid expression levels of reporter metabolite neighbor genes (Swedish male dataset). M value – whole-body glucose uptake during the hyperinsulinemic euglycemic clamp, Insulin 120 min – insulin levels achieved at the two hour time point of oral glucose tolerance test.
doi:10.1371/journal.pcbi.1000729.g005

Supplementary tables S1, S2, S3, S4 list additional reporter metabolites which are, to our knowledge, not (directly) linked with any of the known metabolic players in T2DM. Our analysis nevertheless suggests them as potential nodes of disruption or as biomarkers. Measurement of the intramyocellular concentration of the reporter metabolites in patients with diabetes risk may help to confirm the role of these metabolites in insulin resistance.

Metabolic hubs as reporters

A particularly interesting finding from our analysis is the identification of highly connected metabolites as reporters, including ATP/ADP and NAD+/NADH. We hypothesize that diverse environmental and genetic risk factors result in insulin resistance when individuals are unable to mediate appropriate compensatory transcriptional and metabolic responses in other parts of the network connected by these hubs. Our results also suggest that alterations in gene expression linked to the highly connected co-factors are likely to be acquired features of established T2DM. Analysis of the transcriptional activity of CREB in the context of ATP concentrations and TCA cycle activity in skeletal muscle may help to elucidate regulatory mechanisms leading to these changes.

Constraints and extension of methodology

Reconstructed human metabolic network models are still evolving, incomplete, and subject to error. Well-annotated pathways such as central carbon metabolism are thereby likely to be over-represented in the reporter analysis. In order to partially compensate for this limitation, we used two reconstructions – Recon1 and EHMN. As network reconstructions will become more complete, it will be possible to better assess the extent of this limitation. Another essential input to our algorithm, in addition to metabolic network, is gene expression data for the genes represented in the network. We would like to note that neither EHMN nor Recon1 network genes were fully represented by the microarray chips used in the two case studies (Text S1). Only 54% and 39% genes from the Recon1 and EHMN, respectively, were represented on the chips used in Mexican-American case study, while this coverage was 85% and 60% in Swedish case study. Interestingly, re-analysis of the Swedish Male dataset by using only a subset of genes from the HG-U133A chip that were represented also on the HuGeneFL (used in Mexican-American case study) showed a large overlap between the two reporter metabolite sets thus obtained (86% for T2DM *vs* NGT comparison and 69% for the rest). The details of this analysis, together with relevant statistical considerations, can be found in Text S1.

Although the present analysis identified common metabolic and regulatory signatures across the two studies, there are several differences in the study designs, and therefore the results must be regarded with certain caution. In addition to relatively low number of subjects in Mexican-American study, the differences include fasting state biopsies in Mexican-American study *vs* post insulin stimulation biopsies in Swedish study. Furthermore, the age and BMI (Body Mass Index) of the individuals participating in the two studies were different and may contribute to the differences in the observed gene expression patterns. To our knowledge, these two case studies represent the only human skeletal muscle transcriptome datasets that were available at the time of here reported computational analysis. Analysis of new datasets which may become available in the future will be useful in obtaining further insight into molecular physiology of skeletal muscle in the context of T2DM. Moreover, emergence of better or new gene expression analysis tools will help to cover parts of metabolic network that are currently inaccessible due to the lack of data.

Extension of the analysis to discover more global regulatory patterns by using additional bio-molecular interaction data [58] such as protein-DNA and protein-protein interactions will definitely be an important step in obtaining a higher resolution picture of T2DM metabolic phenotypes. Availability of such interaction data at the high confidence level of metabolic interactions is the current major bottleneck. Another essential extension of the methodology will require the use of thermodynamic data for metabolic reactions [59–61]. Moreover, since mRNA levels do not necessarily correlate with the protein levels, incorporation of the proteomics data together with the thermodynamic data will allow more accurate interpretation of the reporter metabolites in terms of implications for flux and concentration changes.

Conclusions

We demonstrate the use of a network-guided data integration approach to discover key, physiologically relevant metabolic and regulatory nodes in T2DM pathogenesis. The methodology does not require the use of *a priori* disease-specific knowledge regarding the involvement of specific pathways or metabolites, thereby making it a robust and unbiased analytical framework for studying diseases linked to perturbations in the cellular metabolic network. Our results identify the highly connected metabolites ATP and NAD+ as reporters and potential mediators of the widespread changes in gene expression linked to insulin resistance in muscle. Moreover, our results extend previous knowledge about T2DM pathogenesis at the gene expression level – by reporting additional potential sites of disruption, e.g., TCA cycle and Kennedy pathway of phospholipid metabolism. Several metabolites from other pathways were also found to display significant differential gene expression of the genes around them and we suggest putative regulatory mechanisms behind these alterations. Our results suggest a framework of metabolic disruption observed with insulin resistance and diabetes, which can be used to test the role of specific pathways in mediating disease pathophysiology, and more practically, for the identification of potential biomarkers for preventive and therapeutic monitoring.

Materials and Methods

Gene expression and sequence data

Two datasets used in the study were obtained from the Diabetes Genome Anatomy Project website (http://www.diabetesgenome.org). Brief comparison of microarray platforms from the experimental studies [8,9] used in the current work is presented in the Text S1. Promoter sequences for all genes were obtained from the Ensembl Biomart (http://www.ensembl.org/biomart). The transcriptional start sites (TSSs) were identified based on the annotation of the Ensembl Biomart sequences. Sequences in the −800 to 200 base pairs region of the TSS were deemed as promoter regions for the analysis. Interspersed repeats and low complexity DNA sequences were masked out.

Metabolic networks

Two reconstructions of human metabolic network, *viz*., Recon1 [19] and EHMN [62] were used in this study. The Homo Sapiens Recon1 is a comprehensive literature-based metabolic network reconstruction that accounts for the functions of 1496 ORFs, 2004 proteins, 2766 metabolites and 3311 metabolic and transport reactions. The ENMN (Edinburgh Human Metabolic Model) is a network reconstructed by integrating genome annotation information from different databases and metabolic reaction information from the literature. The network contains nearly 3000 metabolic reactions, which were reorganized into about 70 human-specific pathways according to their functional relationships. The two models mainly

differ in the coverage of reactions and in the accounting of compartmentalization and inter-organelle transport reactions.

Significance of differential gene expression

Preprocessing of the gene expression data was carried out by using the statistical software environment – R (www.r-project.org). The probe intensities were obtained and corrected for background by using robust multi-array average method (RMA) [63] with only perfect-match (PM) probes. Normalization was performed by using the quantiles algorithm. Gene expression values were calculated from the PM probes with the median polish summarization method [63]. All data preprocessing methods were used by invoking them through the *affy* package [64] by using *rma* function. Significance of the differential expression was calculated by using the empirical Bayes test [65]. The probe-sets were grouped into genes, and to each gene the differential expression was defined by choosing the value from the top level probe-set (using the probe-set rank defined by Affymetrix). In case of more than one probe-set present at the top level, the median value was used.

Reporter metabolites

Each metabolite in the metabolic network was scored based on the scores of its k neighbor enzymes (i.e. enzymes catalyzing reactions involving that metabolite, either as a substrate or as a product). Each enzyme was assigned with a p-value for differential expression based on the p-value of the gene encoding for that enzyme. In case of isozymes and enzyme-complexes, genes with most significant expression change were used to score the enzyme (Figure 1). P-values of genes p_i, indicating the significance of differential expression, were converted to Z-scores z_i by using the inverse normal cumulative distribution function (CDF) (θ^{-1}): $Z_i = \theta^{-1}(1-p_i)$. All metabolite nodes were assigned a Z-score, $z_{metabolite}$, calculated as aggregated Z scores of the k neighbor enzymes: $Z_{metabolite} = \frac{1}{k}\sum Z_{ni}$. $z_{metabolite}$ scores were then corrected for the background distribution by subtracting the mean (μ_k) and dividing by the standard deviation (σ_k) of the aggregated Z scores derived by sampling 10000 sets of k enzymes from the network: $Z^{corrected}_{metabolite} = \frac{(Z_{metabolite} - \mu_k)}{\sigma_k}$. Corrected Z-scores were then transformed to p-values by using CDF. Metabolites with p-values less than 0.05 were deemed as reporter metabolites. Detailed information on the reporter scoring can be found in the Text S1 and [14].

Transcription factor binding site enrichment

For all reporter metabolites, we assessed enrichment of known protein-binding sequence motifs in the promoter regions (−800 to 200 base pairs relative to the transcription start site) of the corresponding neighbor genes. In order to obtain robust results, we only considered sets consisting of at least 5 up- or down-regulated genes. For each reporter metabolite, the sequences of all enzyme neighbors were used as the positive sequence set, whereas all other enzymes in the network model were used as the negative (background) set. Known motifs were identified by using position frequency matrices of all known motifs stored in the TRANSFAC database [66]. The motif enrichment analysis tool ASAP [67] was used to scan all TRANSFAC motif matrices against the positive sequence sets of each reporter metabolite. The negative sequence sets were used together with 2nd order background model. A one-tailed Fisher's exact test was used to assess per-sequence over-representation of any known motif, and the threshold used to calculate significance for each TRANSFAC matrix was set to 70% of the highest-scoring sequence motif. The q-value cut-off criteria [29] was used as a post-data measure of statistical significance of all motifs found to be significantly enriched.

Supporting Information

Text S1 Supporting text describing scoring methodology and datasets.
Found at: doi:10.1371/journal.pcbi.1000729.s001 (0.74 MB PDF)

Table S1 Reporter metabolites for Swedish male dataset (Recon1).
Found at: doi:10.1371/journal.pcbi.1000729.s002 (0.05 MB XLS)

Table S2 Reporter metabolites for Swedish male dataset (EHMN).
Found at: doi:10.1371/journal.pcbi.1000729.s003 (0.07 MB XLS)

Table S3 Reporter metabolites for Mexican-American dataset (Recon1).
Found at: doi:10.1371/journal.pcbi.1000729.s004 (0.03 MB XLS)

Table S4 Reporter metabolites for Mexican-American dataset (EHMN).
Found at: doi:10.1371/journal.pcbi.1000729.s005 (0.05 MB XLS)

Table S5 Overlapping reporter metabolites between two case studies (Recon1).
Found at: doi:10.1371/journal.pcbi.1000729.s006 (0.05 MB XLS)

Table S6 Overlapping reporter metabolites between two case studies (EHMN).
Found at: doi:10.1371/journal.pcbi.1000729.s007 (0.06 MB XLS)

Table S7 Results of the motif enrichment analysis.
Found at: doi:10.1371/journal.pcbi.1000729.s008 (0.03 MB PDF)

Table S8 Experimentally studies linking metabolite levels to T2DM pathophysiology
Found at: doi:10.1371/journal.pcbi.1000729.s009 (0.12 MB PDF)

Figure S1 Hierarchical clustering of pair-wise comparisons within the Swedish male and Mexican-American datasets based on the overlapping reporter metabolites (EHMN network).
Found at: doi:10.1371/journal.pcbi.1000729.s010 (0.21 MB TIF)

Acknowledgments

We are thankful to Isabel Rocha for a feedback on the manuscript. We thank Anders Krogh for advising on the choice of motif analysis method. We are grateful to the reviewers for useful comments.

Author Contributions

Conceived and designed the experiments: MEP KRP. Performed the experiments: AZ THP SS. Analyzed the data: AZ THP SS MEP KRP. Wrote the paper: AZ THP MEP KRP.

References

1. Zimmet P, Alberti KG, Shaw J (2001) Global and societal implications of the diabetes epidemic. Nature 414: 782–787.
2. Simpson RW, Shaw JE, Zimmet PZ (2003) The prevention of type 2 diabetes–lifestyle change or pharmacotherapy? A challenge for the 21st century. Diabetes Res Clin Pract 59: 165–180.
3. Shulman GI (2000) Cellular mechanisms of insulin resistance. J Clin Invest 106: 171–176.
4. Pehling G, Tessari P, Gerich JE, Haymond MW, Service FJ, et al. (1984) Abnormal meal carbohydrate disposition in insulin-dependent diabetes. Relative contributions of endogenous glucose production and initial splanchnic uptake and effect of intensive insulin therapy. J Clin Invest 74: 985–991.
5. Muoio DM, Newgard CB (2008) Molecular and metabolic mechanisms of insulin resistance and [beta]-cell failure in type 2 diabetes. Nat Rev Mol Cell Biol 9: 193–205.

6. Saltiel AR, Kahn CR (2001) Insulin signalling and the regulation of glucose and lipid metabolism. Nature 414: 799–806.
7. Kelley DE, He J, Menshikova EV, Ritov VB (2002) Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. Diabetes 51: 2944–2950.
8. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273.
9. Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, et al. (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. Proc Natl Acad Sci U S A 100: 8466–8471.
10. Boden G (1996) Fatty acids and insulin resistance. Diabetes Care 19: 394–395.
11. Ueki K, Kondo T, Tseng YH, Kahn CR (2004) Central role of suppressors of cytokine signaling proteins in hepatic steatosis, insulin resistance, and the metabolic syndrome in the mouse. Proc Natl Acad Sci U S A 101: 10422–10427.
12. Sreekumar R, Halvatsiotis P, Schimke JC, Nair KS (2002) Gene Expression Profile in Skeletal Muscle of Type 2 Diabetes and the Effect of Insulin Treatment. 51: 1913–1920.
13. Yang X, Pratley RE, Tokraks S, Bogardus C, Permana PA (2002) Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant Pima Indians. Diabetologia 45: 1584–1593.
14. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685–2689.
15. Seggewiss J, Becker K, Kotte O, Eisenacher M, Yazdi MR, et al. (2006) Reporter metabolite analysis of transcriptional profiles of a Staphylococcus aureus strain with normal phenotype and its isogenic hemB mutant displaying the small-colony-variant phenotype. J Bacteriol 188: 7765–7777.
16. David H, Hofmann G, Oliveira AP, Jarmer H, Nielsen J (2006) Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans. Genome Biol 7: R108.
17. Capel F, Klimcakova E, Viguerie N, Roussel B, Vitkova M, et al. (2009) Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. Diabetes 58: 1558–1567.
18. Baxter CJ, Redestig H, Schauer N, Repsilber D, Patil KR, et al. (2007) The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. Plant Physiol 143: 312–325.
19. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777–1782.
20. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, et al. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 3: 135.
21. Roden M (2005) Muscle triglycerides and mitochondrial function: possible mechanisms for the development of type 2 diabetes. Int J Obes (Lond) 29 Suppl 2: S111–115.
22. Savage DB, Petersen KF, Shulman GI (2007) Disordered lipid metabolism and the pathogenesis of insulin resistance. Physiol Rev 87: 507–520.
23. Holland WL, Brozinick JT, Wang LP, Hawkins ED, Sargent KM, et al. (2007) Inhibition of ceramide synthesis ameliorates glucocorticoid-, saturated-fat-, and obesity-induced insulin resistance. Cell Metab 5: 167–179.
24. Chavez JA, Summers SA (2003) Characterizing the effects of saturated fatty acids on insulin signaling and ceramide and diacylglycerol accumulation in 3T3-L1 adipocytes and C2C12 myotubes. Arch Biochem Biophys 419: 101–109.
25. Salek RM, Maguire ML, Bentley E, Rubtsov DV, Hough T, et al. (2007) A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiol Genomics 29: 99–108.
26. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, et al. (2009) A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. Cell Metabolism 9: 311–326.
27. Fleischman A, Kron M, Systrom DM, Hrovat M, Grinspoon SK (2009) Mitochondrial Function and Insulin Resistance in Overweight and Normal-Weight Children. J Clin Endocrinol Metab.
28. Phielix E, Schrauwen-Hinderling VB, Mensink M, Lenaers E, Meex R, et al. (2008) Lower intrinsic ADP-stimulated mitochondrial respiration underlies in vivo mitochondrial dysfunction in muscle of male type 2 diabetic patients. Diabetes 57: 2943–2949.
29. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.
30. Bajaj M, Defronzo RA (2003) Metabolic and molecular basis of insulin resistance. J Nucl Cardiol 10: 311–323.
31. Itani SI, Ruderman NB, Schmieder F, Boden G (2002) Lipid-induced insulin resistance in human muscle is associated with changes in diacylglycerol, protein kinase C, and IkappaB-alpha. Diabetes 51: 2005–2011.
32. Patti ME, Corvera S (2009) The Role of Mitochondria in the Pathogenesis of Type 2 Diabetes. Endo Reviews: In press.
33. Koves TR, Li P, An J, Akimoto T, Slentz D, et al. (2005) Peroxisome proliferator-activated receptor-gamma co-activator 1alpha-mediated metabolic remodeling of skeletal myocytes mimics exercise training and reverses lipid-induced mitochondrial inefficiency. J Biol Chem 280: 33588–33598.
34. Kersten S, Desvergne B, Wahli W (2000) Roles of PPARs in health and disease. Nature 405: 421–424.
35. Sen CK, Packer L (1996) Antioxidant and redox regulation of gene transcription. Faseb J 10: 709–720.
36. Scarpulla RC (2008) Nuclear control of respiratory chain expression by nuclear respiratory factors and PGC-1-related coactivator. Ann N Y Acad Sci 1147: 321–334.
37. Scarpulla RC (2006) Nuclear control of respiratory gene expression in mammalian cells. J Cell Biochem 97: 673–683.
38. Szendroedi J, Schmid AI, Chmelik M, Toth C, Brehm A, et al. (2007) Muscle mitochondrial ATP synthesis and glucose transport/phosphorylation in type 2 diabetes. PLoS Med 4: e154.
39. Petersen KF, Dufour S, Shulman GI (2005) Decreased insulin-stimulated ATP synthesis and phosphate transport in muscle of insulin-resistant offspring of type 2 diabetic parents. PLoS Med 2: e233.
40. Petersen KF, Befroy D, Dufour S, Dziura J, Ariyan C, et al. (2003) Mitochondrial dysfunction in the elderly: possible role in insulin resistance. Science 300: 1140–1142.
41. Ristow M, Zarse K, Oberbach A, Kloting N, Birringer M, et al. (2009) Antioxidants prevent health-promoting effects of physical exercise in humans. Proc Natl Acad Sci U S A 106: 8665–8670.
42. Brekasis D, Paget MS (2003) A novel sensor of NADH/NAD+ redox poise in Streptomyces coelicolor A3(2). Embo J 22: 4856–4865.
43. Zhang Q, Piston DW, Goodman RH (2002) Regulation of corepressor function by nuclear NADH. Science 295: 1895–1897.
44. Lin SJ, Defossez PA, Guarente L (2000) Requirement of NAD and SIR2 for life-span extension by calorie restriction in Saccharomyces cerevisiae. Science 289: 2126–2128.
45. Anderson RM, Latorre-Esteves M, Neves AR, Lavu S, Medvedik O, et al. (2003) Yeast life-span extension by calorie restriction is independent of NAD fluctuation. Science 302: 2124–2126.
46. Rutter J, Reick M, Wu LC, McKnight SL (2001) Regulation of clock and NPAS2 DNA binding by the redox state of NAD cofactors. Science 293: 510–514.
47. Agarwal AK, Auchus RJ (2005) Minireview: cellular redox state regulates hydroxysteroid dehydrogenase activity and intracellular hormone potency. Endocrinology 146: 2531–2538.
48. Rodgers JT, Lerin C, Haas W, Gygi SP, Spiegelman BM, et al. (2005) Nutrient control of glucose homeostasis through a complex of PGC-1alpha and SIRT1. Nature 434: 113–118.
49. Cimini D, Patil KR, Schiraldi C, Nielsen J (2009) Global transcriptional response of Saccharomyces cerevisiae to the deletion of SDH3. BMC Syst Biol 3: 17.
50. Raghevendran V, Patil KR, Olsson L, Nielsen J (2006) Hap4 is not essential for activation of respiration at low specific growth rates in Saccharomyces cerevisiae. J Biol Chem 281: 12308–12314.
51. Schuurmans JM, Rossell SL, van Tuijl A, Bakker BM, Hellingwerf KJ, et al. (2008) Effect of hxk2 deletion and HAP4 overexpression on fermentative capacity in Saccharomyces cerevisiae. FEMS Yeast Res 8: 195–203.
52. He J, Watkins S, Kelley DE (2001) Skeletal muscle lipid content and oxidative enzyme activity in relation to muscle fiber type in type 2 diabetes and obesity. Diabetes 50: 817–823.
53. Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MA, de Groot MJ, et al. (2007) The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. Proc Natl Acad Sci U S A 104: 15753–15758.
54. Bruning T, Vamvakas S, Makropoulos V, Birner G (1998) Acute intoxication with trichloroethene: clinical symptoms, toxicokinetics, metabolism, and development of biochemical parameters for renal damage. Toxicol Sci 41: 157–165.
55. Laughter AR, Dunn CS, Swanson CL, Howroyd P, Cattley RC, et al. (2004) Role of the peroxisome proliferator-activated receptor alpha (PPARalpha) in responses to trichloroethylene and metabolites, trichloroacetate and dichloroacetate in mouse liver. Toxicology 203: 83–98.
56. Ramirez de Molina A, Gallego-Ortega D, Sarmentero J, Banez-Coronel M, Martin-Cantalejo Y, et al. (2005) Choline kinase is a novel oncogene that potentiates RhoA-induced carcinogenesis. Cancer Res 65: 5647–5653.
57. Banez-Coronel M, de Molina AR, Rodriguez-Gonzalez A, Sarmentero J, Ramos MA, et al. (2008) Choline kinase alpha depletion selectively kills tumoral cells. Curr Cancer Drug Targets 8: 709–719.
58. Oliveira AP, Patil KR, Nielsen J (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst Biol 2: 17.
59. Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, et al. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2: 50.
60. Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol 2: 2006 0034.
61. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92: 1792–1805.
62. Ma H, Goryanin I (2008) Human metabolic network reconstruction and its impact on drug discovery and development. Drug Discov Today 13: 402–408.

63. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
64. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315.
65. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
66. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
67. Marstrand TT, Frellsen J, Moltke I, Thiim M, Valen E, et al. (2008) Asap: a framework for over-representation statistics for transcription factor binding sites. PLoS ONE 3: e1623.

5.4 Paper V - Predicting 6-month weight maintenance

In the following paper we studied a subset of subjects from the DiOGenes cohort[4] who successfully maintained weight loss after a calorie restricted diet (henceforth referred to as weight maintainers) and subjects who regained the weight they lost after the calorie restricted diet (henceforth referred to as weight regainers). To accomplish this, we examined subcutaneous adipose tissue gene expression and bioclinical markers before and after an 8-week low calorie diet.

We applied the Reporter Metabolite Analysis (described in **Paper IV**) to adipose tissue gene expression profiles, and found that genes coordinating the regulation of fatty acid metabolism, citric acid cycle, oxidative phosphorylation, and apoptosis were regulated differently by the low calorie diet in weight maintainers and weight regainers.

An interesting finding was, that the reporter metabolites reported in **Paper V** resemble our findings from the analysis of the energy restriction phase in the Chapel *et al* study [Capel et al., 2009]. It demonstrates how results obtained at the pathway-level (the metabolite level) may result in more robust findings across studies, than analysis accomplished at the gene level. At the gene level, both studies indeed showed no correlation in genes' differential expression levels at all (**Tab.** 5.2, second column). In contrast, reporter metabolite p-values between both studies were moderately correlated (**Tab.** 5.2, third column).

	Chapel *et al* energy restriction phase [Capel et al., 2009]	
	Correlation at the gene level	Correlation at the reporter metabolite level
Weight maintainer	0.07	0.28
Weight regainer	0.10	0.16

Table 5.2: Correlation at the gene- and reporter metabolite level between the energy restriction phase in the Mutch *et al* analysis (Paper V) and Chapel *et al* analysis [Capel et al., 2009]. Whereas the Spearman correlation coefficients calculated between the two studies is close to zero when based on differential gene expression levels, the correlation coefficients are stronger (moderate correlation) when calculated based on reporter metabolites' p-values. The increase in correlation confirms the premise that systems-based approaches enhance the resemblance in findings between studies (given the findings are likely to be true).

[4] DiOGenes stands for "Diet, Obesity and Genes" and more information is found on www.diogenes-eu.org.

A distinct adipose tissue gene expression response to caloric restriction predicts 6-month weight maintenance in obese subjects

David M Mutch, Tune H Pers, M Ramzi Temanni, Veronique Pelloux, Adriana Marquez-Quiñones, Claus Holst, J Alfredo Martinez, Dimitris Babalis, Marleen A van Baak, Teodora Handjieva-Darlenska, Celia G Walker, Arne Astrup, Wim HM Saris, Dominique Langin, Nathalie Viguerie, Jean-Daniel Zucker, and Karine Clément on behalf of the DiOGenes Project

Corresponding Authors: David M Mutch (dmutch@uoguelph.ca), Karine Clément (karine.clement@psl.aphp.fr)

INSERM, U872, Eq7, Nutriomique, Paris, F-75006 France (DMM, MRM, VP, JDZ, KC); University Pierre et Marie Curie-Paris 6, Cordelier Research Center, UMRS 872, Paris, F-75006 France (DMM, MRM, VP, JDZ, KC); Assistance Publique/Hôpitaux de Paris (AP/HP), Pitié Salpêtrière Hospital, Nutrition and department, Research Center for Human Nutrition (CRNH, Ile de France), Paris, F-75013 France (VP, KC); Department of Human Health & Nutritional Sciences, University of Guelph, Guelph, Ontario, N1G 2W1, Canada (DMM); Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark (THP); Institute of Preventive Medicine, Copenhagen University Hospital, Copenhagen, Denmark (THP, CH); Department of Physiology & Nutrition, University of Navarra, 31008 Pamplona, Spain (JAM); Department of Internal Medicine, University Hospital of Heraklion, Crete, Greece (DB); National Multiprofile Transport Hospital, Sofia, Bulgaria (THD); Medical Research Council, Human Nutrition Research, Cambridge, United Kingdom (CGW); Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Denmark (AA); Nutrition Research centre NUTRIM, Maastricht University, The Netherlands (MAvB, WHMS); INSERM, U858, Obesity Research Laboratory, Rangueil Institute of Molecular Medicine, Toulouse, France (AMQ, DL, NV); University of Toulouse, UPS, Bio-Medical Research Federative Institute of Toulouse, IFR150, Toulouse, France (AMQ, DL, NV); CHU de Toulouse, Biochemistry Laboratory, Biology Institute of Purpan, Toulouse, France (DL, NV).

Clinical Trial Registration: This trial was registered at clinicaltrials.gov as NCT00390637.

Running Title: Predicting 6-month weight maintenance

Funding: This research was funded by the European Community (DiOGenes, Diet, Genes, and Obesity, contract FP6-513946, http://www.diogenes-eu.org/) and the Advanced Food & Materials Network (DMM).

Keywords: functional genomics, diet intervention, subcutaneous adipose tissue, insulin secretion, fatty acid metabolism, metabolic reconstruction analysis.

Abstract

Background: Weight loss has been shown to reduce risk factors associated with cardiovascular disease and diabetes; however, the successful maintenance of weight loss continues to pose a major challenge.
Objective: The present study was designed to assess whether changes in subcutaneous adipose tissue (scAT) gene expression during a low calorie diet (LCD) could be used to differentiate and predict subjects who experience successful short-term weight maintenance from subjects who experience weight regain.
Design: Forty Caucasian women followed a dietary protocol consisting of an 8-week LCD phase followed by a 6-month weight maintenance phase. Participants were classified as weight maintainers (WM; 0-10% weight regain) and weight regainers (WR; 50-100% weight regain) by considering changes in body weight during the two phases. Anthropometric measurements, bio-clinical parameters, and scAT gene expression were studied in all individuals before and after the LCD. Energy intake was monitored during both phases of the protocol.
Results: The LCD resulted in significant decreases in several plasma parameters, such as triglyceride and insulin levels, in WM compared to WR. WR experienced no changes in insulin secretion in response to an oral glucose tolerance test after the LCD, whereas WM had a significant decrease in insulin secretion after the LCD. An ANOVA analysis of scAT gene expression revealed that genes coordinating the regulation of fatty acid metabolism, citric acid cycle, oxidative phosphorylation, and apoptosis were regulated differently by the LCD in WM and WR subjects.
Conclusion: This study suggests that LCD-induced changes in insulin secretion and scAT gene expression may have potential to predict successful short-term weight maintenance.

Introduction

Obesity is associated with an increased risk of cardiovascular disease, diabetes, metabolic syndrome, and a number of cancers; however, weight loss of 5-10% has repeatedly been shown to convey modest to significant reductions in the risk of these downstream complications (1). The most common strategy to promote weight loss in obesity involves modifying lifestyle via changes in dietary and exercise habits. Although reduced caloric intake and increased physical activity favour a reduction in body weight, body fat mass, and improvements in metabolic parameters, one of the greatest difficulties for obesity management is weight maintenance after successful weight loss.

Several meta-analyses have revealed that energy restriction and/or increased physical activity can lead to successful short-term weight loss; however, the long-term effectiveness of these interventions appears challenging (2, 3). Numerous factors have been shown to influence successful weight maintenance, including behaviour (4), physical activity (2), eating habits (5), the length of time an individual has maintained weight loss (6), the degree of energy deficit and consequent weight loss (3, 7), and the influence of altering dietary macronutrient content (i.e. carbohydrate, protein, and fat) (8-11). It is now widely accepted that body weight and body composition are also influenced by a genetic component (which encompasses genetic polymorphisms, epigenetics, and gene transcription); however, our understanding of how these genetic determinants contribute to successful weight maintenance remains limited (12).

Diet-induced weight loss in overweight/obese individuals decreases the expression of genes associated with polyunsaturated fatty acid metabolism, inflammation, and cell death, as well as modifying the expression of genes encoding components of the extracellular matrix (13-17). Previous attempts to predict an individual's response (i.e. high vs. low weight loss) using only subcutaneous adipose tissue (scAT) gene expression profiles appear limited, suggesting that alternate approaches may be required to improve prediction accuracy (18, 19).

Studying changes in gene expression has provided novel insight to help clarify the molecular basis for the metabolic improvements associated with diet-induced weight loss. For example, Capel et al highlighted the interplay between immune cells and adipocytes during the various phases of a weight loss program (caloric restriction and weight stabilization) by monitoring scAT gene expression profiles (20). Recently, Márquez-Quiñones et al focused on the weight maintenance phase of the DiOGenes study and found that unsuccessful participants (i.e. subjects who regained weight following a low calorie diet (LCD) phase) had an increased expression of genes related to cellular growth and differentiation (11).

The present study was designed to further contribute to our understanding of the inter-individual variability regarding successful weight maintenance by determining whether scAT gene expression profiles during an LCD can be used to differentiate and predict subjects who experience successful short-term weight maintenance from participants who experience weight regain. This study provides important knowledge that may prove beneficial in the long-term for the development of personalized strategies to improve successful weight maintenance.

Materials and Methods

Dietary Intervention Study: This study is part of the European Framework project entitled Diet, Obesity, and Genes (DiOGenes). For a thorough description of the overall objective and goals of this dietary intervention, please see Larsen et al (21) and Moore et al (22). Briefly, the project consisted of two phases: an initial weight loss phase and a 6-month weight maintenance phase. 932 overweight and obese Caucasian adults were recruited from across 8 European countries in order to study the effects of dietary macronutrients on weight regain and cardiovascular risk factors. Inclusion and exclusion criteria for study participation were outlined previously (21). Of relevance to the current study, all subjects were weight stable (<3 kg change in body weight) during the 2 month period prior to initiating the study. The initial weight loss phase consisted of an 8 week low calorie diet (LCD; 3300 kJ/d; ~800 kcal; Modifast®, Nutrition et Santé, Revel, France). Only those participants that achieved the targeted weight loss (≥ 8% of initial body weight) were invited to continue the protocol. Subjects were randomized to the weight maintenance phase as described (21). During this weight maintenance phase, participants consumed *ad libitum* one of four low-fat (20-25% energy intake) diets that differed in glycemic index (GI) and protein content (P) (23). More specifically subjects adhered to one of the following diets: low GI (LGI)/low P (LP), high GI (HGI)/LP, LGI/high P (HP), or HGI/HP. Target energy intakes in the LP diets were 10-15% protein and 57-62% carbohydrates, and in the HP diets were 23-28% protein and 45-50% carbohydrates. The goal was to achieve a difference of approximately 15 GI points between the LGI and HGI diets. During this weight maintenance period subjects met with a dietitian at regular intervals.

Ethics: The study was approved by local Ethics committees in the various countries. The protocol was in accordance with the Declaration of Helsinki. All study participants provided written consent.

Blood Sampling: Fasting blood samples were obtained at each of the three clinical investigation days (CIDs) for the analysis of blood metabolites, as outlined in Larsen et al (21). An oral glucose tolerance test (OGTT) lasting 120 minutes was also performed at each CID following the consumption of 75 g of glucose.

Subject Selection: Of the 548 subjects that completed the entire dietary protocol, a subset of 227 women were selected according to the following criteria: age between 20-50 years, non-diabetic (fasting glucose ≤ 7 mmol/L), non-dyslipidemic (fasting total cholesterol ≤ 7 mmol/L and fasting triglycerides ≤ 3.6 mmol/L), availability of a fat biopsy at the required time points, and a complete clinical evaluation during the protocol (SUPPLEMENTARY FIGURE 1). Data at the three distinct time points was necessary: CID1 (prior to commencing the LCD), CID2 (at the end of the LCD phase), and CID3 (at the end of the 6-month weight maintenance phase) (21).

Previously, Márquez-Quiñones et al reported negligible differences in scAT expression profiles between participants in the various weight maintenance diets; therefore subjects in the four different dietary branches were considered all together (see (11)). Subjects were classified according to changes in body weight during the weight maintenance period, which was expressed as a % of weight lost during the LCD period. Subjects who experienced between 0-10% and 50-100% weight regain during the weight maintenance period were classified as “weight maintainers” (WM) and “weight regainers” (WR), respectively. Of the 227 women available, a subset of twenty subjects were randomly selected for each group and matched for the following bio-clinical variables at both CID1 and CID2 time points: body weight (kg), body mass index (BMI), total energy intake (kJ/d), and glucose (mmol/L), insulin (μIU/ml), and insulin resistance (HOMA-IR), and fasting cholesterol (mmol/L), triglycerides (mmol/L), HDL-cholesterol (mmol/L), fructosamin (μmol/L), C-reactive protein (CRP; mg/L), and adiponectin (μg/mL). It is noteworthy to mention that WM and WR subjects were not individually matched, but rather it was the WM and WR groups as a whole that were matched (i.e. using average values for each bio-clinical variable).

Sample preparation and microarray analysis: Subcutaneous adipose tissue (scAT) samples from the periumbilical area were obtained by needle aspiration under local anaesthesia after an overnight fast at each of the time points. For the present prediction study only biopsies at CID1 and CID2 were required. All procedures were standardized between study centres across Europe and biopsy samples were stored at -80°C until analysis. Total RNA was extracted using the RNeasy total RNA Mini kit (Qiagen, Courtaboeuf, France). Total RNA concentration and quality was confirmed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Massy, France). 200 ng of total RNA from each sample was amplified and transcribed into fluorescent cRNA using Agilent's Low RNA Input Linear Amplification kit (Agilent Technologies, Massy, France). Cyanine-5 dye was incorporated into all scAT samples, while the reference pool was labelled with cyanine-3 dye. The reference pool consisted of a commercial mix of human liver, adipose tissue, heart, intestine, and skeletal muscle RNA (AMBION/Applied Biosystems, les Ulis, France). A total of 80 samples (40 paired samples from CID1 and

CID2) were randomly hybridized to Agilent 4x44K whole human genome microarrays, which are comprised of over 41,000 unique 60-mer oligonucleotide human sequences and transcripts. Sample preparation, hybridization, and microarray washing were performed according to manufacturer's recommendations (Agilent Technologies, Massy, France). Arrays were scanned using a GenePix 4000A Scanner (Axon Instruments-Molecular Devices, Sunnyvale, CA). The complete dataset is available in the NCBI Omnibus (http://www.ncbi.nlm.nih.gov/geo/) through the following series accession number: GSE24432.

Real time RT-PCR analysis: A subset of genes were validated by real-time reverse transcriptase PCR (RT–PCR) in 17 WM and 17 WR (sufficient RNA was not available for all 40 subjects). Reverse transcription was performed with 0.5 mg of total RNA and random hexamer primers, according to manufacturer's instructions (Promega, Charbonnieres-les-Bains, France). RT–PCR amplification was performed using an ABI 7300 (Applied Biosystems, Foster City, CA, USA) with the following thermal cycling conditions: 2 min at 50ºC, 10 min at 95ºC, followed by 40 cycles of 95ºC for 15 s and 60ºC for 1 min for denaturation, annealing and elongation. All samples were normalized to 18S gene expression (18S rRNA Control kit; Eurogentec, Seraing, Belgium). Differences in gene expression were assessed using a two-tailed, homoscedastic Student's t-test. Specific primers and probes were designed using Universal ProbeLibrary Assay Design Center by Roche Applied Science (https://www.roche-applied-science.com).

Statistical analyses: Changes in bio-clinical and anthropometric parameters between groups (WM versus WR) and between times (CID1 versus CID2) were analyzed using JMP Genomics Version 4.1 platform (SAS, Cary, NC, USA). An ANOVA model was generated and least square means were estimated for the differences between groups, times, and the interaction between group and time (group*time), taking into account repeated measurements for all subjects. A post-hoc student's T-test was used to determine significance in specific pair-wise analyses. The area under the curve (AUC) was calculated using the trapezoid rule for both glucose and insulin response to the 120 minute OGTT. All data is presented as mean ± SEM.

Microarray normalization was carried out by subtracting the median intensity background signal prior to intra-slide Loess normalization of log-transformed data (Goulphar Version 1.1.3 package (24)). All data was then uploaded into the JMP Genomics Version 4.1 platform (SAS, Cary, NC, USA) and further normalized using a quantile inter-slide intensity method. Multiple probes corresponding to the same gene were averaged to provide a single value for each GeneID per microarray. An ANOVA model was created using group (WM, WR), time (CID1, CID2) and the interaction between group and time (group*time) as fixed effects. Because each of the 40 individuals provided a biopsy at both time points, the model included subjects as a random effect. Least square means were estimated for the difference between groups, times, and group*time. A false discovery rate (FDR) of 0.01 was used to account for multiple testing.

FunNet analysis: The functional analysis of gene expression data was performed using FunNet (25). FunNet identifies Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathways overrepresented in gene expression data lists while accounting for tests of multiplicity (FDR=0.05). Four lists of significant GeneIDs were obtained from the ANOVA analysis and used for the functional analyses: 1) Weight Maintainers, which corresponded to genes up- and down-regulated by the LCD that were identified in WM subjects only, 2) Weight Regainers, which corresponded to genes up- and down-regulated by the LCD that were identified in WR subjects only, 3) directional concordance (DC), which corresponded to genes up- and down-regulated by the LCD that were identified in both WM and WR subjects, and 4) oppositely regulated (OPP), which corresponded to genes that were regulated in both WM and WR subjects during the LCD, but in an opposite manner. More specifically, up-regulated pathways correspond to those genes up-regulated in WR and down-regulated in WM, while down-regulated pathways correspond to those genes down-regulated in WR and up-regulated in WM.

Reporter Metabolite Analysis: The global predictive analysis of enzyme-induced transcriptional changes on metabolite concentrations was performed by Reporter Metabolite Analysis (26, 27) based on the Edinburgh human metabolic network (EHMN) reconstruction, which represents a high-confidence reconstructed network of metabolism (28). The metabolic reconstruction forms a bipartite network containing two kinds of nodes, enzymes and metabolites. A metabolite is connected to an enzyme if it is catalyzed (i.e. produced or consumed) by that particular enzyme. Therefore metabolites will only be linked to enzymes and never to each other, while enzymes will only be linked to metabolites and never to each other. The reporter metabolite algorithm relies on a rigorous statistical framework to identify metabolite nodes that are enriched in differentially expressed enzymatic genes among their connected enzyme nodes. We compared our results with those previously reported by Capel et al (20). Note that although the EHMN metabolic network was used in both studies, the number of enzymes detected in each network differed because the quality control measures applied to the gene expression data were not identical. The present analysis covered 1860 enzymes (80% of the EHMN metabolic network), 2166 metabolic reactions (76%), and 2225 metabolites in EHMN. See Supplementary Table 1 for the EHMN

coverage for the Capel et al analysis (20). The reporter metabolite figures were generated using Cytoscape 2.7.0 software (29).

Results

Anthropometric and bio-clinical data

Subjects were classified into WM or WR groups according to the percentage weight regain during the 6-month weight maintenance phase, where WM and WR corresponded to participants who regained between 0-10% and 50-100% of the weight lost after the LCD, respectively (**FIGURE 1**). Importantly, WM and WR groups were established by ensuring there was no difference at CID1 (baseline) or CID2 (after the LCD) for the anthropometric and bio-clinical parameters listed in **TABLE 1**. All subjects lost a minimum of 8% of their initial body weight during the LCD phase. Energy restriction led to significant decreases in body weight, BMI, cholesterol, glucose, and insulin in both WM and WR subjects (**TABLE 1**). HDL-cholesterol, adiponectin, CRP, fructosamin, and HOMA-IR were not significantly changed by the LCD in either group.

Circulating triglyceride levels were significantly reduced after the LCD in WM only (CID1 → 1.3 ± 0.1 mmol/L; CID2 → 1.0 ± 0.1 mmol/L, p=0.0007). This change appeared to be specific to WM subjects, as no significant difference was observed in WR subjects. This group-specific effect is further implied by the borderline significant differences (p=0.0627) in the group*time interaction analysis. Reductions in fasting insulin after the LCD appeared more significant in WM (WM: CID1 → 11.8 ± 1.1 μIU/mL; CID2 → 6.7 ± 0.7 μIU/mL, $P < 0.0001$) than WR subjects (CID1 → 8.9 ± 0.7 μIU/mL; CID2 → 6.2 ± 0.6 μIU/mL, $P = 0.0004$). Significant group-specific changes in fasting insulin were identified by the group*time interaction analysis (p=0.0459) (**TABLE 1**).

Energy intake during the protocol

Energy intake in all subjects was assessed at CID1, CID2, and CID3 using three day dietary records. Energy intake during the LCD was fixed at 3300 kJ/d for all study participants. While energy intake decreased significantly from CID1 to CID2 in both groups, there was no difference in energy intake between the groups at either CID1 or CID2 (data not shown). Furthermore, there were no difference in energy intake between the WM group (6226 ± 533 kJ/day) and WR (6788 ± 875 kJ/day) when measured at the end of the 6-month weight maintenance phase (CID3). This indicates that *ad libitum* consumption of low-fat diets during the 6-month weight maintenance phase did not influence weight regain or maintenance in study participants; thereby reinforcing that changes observed in gene expression and bio-clinical parameters were not related to differences in energy intake.

Changes in insulin secretion predicts 6-month weight maintenance

The area under the curve (AUC) for glucose and insulin response after an OGTT was calculated in all subjects before and after the LCD. No significant differences in glycemic response were detected between WM and WR at either CID1 or CID2. In contrast, the OGTT induced insulin secretion was markedly higher in WM compared to WR at CID1 (WM → AUC 8245 ± 881; WR → AUC 5674 ± 509; $P < 0.0001$); despite similar baseline fasting insulin levels. The LCD resulted in a significant decrease in insulin secretion in the WM group only. At CID2 there was no significant difference in insulin secretion between WM and WR. The group*time interaction analysis reinforced that changes in insulin secretion were specific to the WM group ($P = 0.0123$).

Gene Expression Analysis

Gene expression differences between WM and WR before and after the LCD

Adipose tissue gene expression in WM and WR was first examined at CID1 and CID2 independently. Although there were no significant differences in bio-clinical parameters at CID1 between WM and WR subjects, a gene expression analysis revealed that 1292 genes were differentially expressed between the two groups prior to commencing the LCD. Despite this large number of differentially expressed genes, a functional pathway analysis failed to detect any differences in KEGG biological pathways between the two groups. At CID2 the two groups of subjects appeared more similar with regards to their scAT gene expression profiles, with only 77 genes identified as differentially expressed between WM and WR. Again, no KEGG pathways were found to differ significantly between the groups.

The effects of LCD-induced weight loss on gene expression in WM and WR

The primary goal of the current study was to assess the differences in LCD-induced changes in gene expression in subjects classified as WM and WR. When comparing changes in gene expression between CID1 and CID2, 1291 and 1298 genes were differentially expressed in WM and WR, respectively. More specifically, 583 genes were up-regulated and 708 were down-regulated in the WM, while 628 genes were up-regulated and 670 were down-regulated in the WR. The most significant down-regulated gene in both WM and WR was stearoyl-CoA desaturase (*SCD1*; WM -3.4 fold; WR -2.5 fold), the rate-limiting enzyme responsible for the

conversion of saturated fatty acids into monounsaturated fatty acids (30). In WM, the most up-regulated gene was cell death-inducing DFFA-like effector a (*CIDEA*; +2.2 fold); which was not differentially regulated in WR. This gene plays an important role in adipose tissue energy expenditure and lipid accumulation, in particular increasing fat oxidation (31). In WR, the up-regulated gene was vimentin (*VIM*; +2.0 fold). This gene, expressed in fibroblasts and preadipocytes, was unique to WR and is thought to play an important role in the cellular remodelling that occurs during adipocyte differentiation (32). Expression changes in *SCD1* and *CIDEA* were confirmed by real-time RT-PCR (data not shown); however expression changes for *VIM* were not significant ($P = 0.17$).

The two gene lists for WM and WR were further dissected in order to better explore the shared and unique gene expression responses to the LCD between the two groups. 1027 and 1034 genes were uniquely regulated in WM and WR, respectively (**FIGURE 2**). Although there were a large number of differentially expressed genes unique to the two groups, the functional analysis revealed that these genes tended to belong to similar functional pathways (**FIGURE 2**). The LCD caused an increase in ribosomal genes and decreases in *oxidative phosphorylation* and metabolism pathways in both groups. The genes associated with *oxidative phosphorylation* are also found in other pathways, which is why pathways related to Alzheimer's, Huntington's, and Parkinson's disease appear in **FIGURE 2**; however, it is the *oxidative phosphorylation* pathway that is most relevant when considering adipose tissue gene expression. The LCD caused a decrease in *valine, isoleucine, and leucine degradation* pathway (related to 10 genes: *ABAT*, *ACAA2*, *ALDH6A1*, *AOX1*, *BCKDHB*, *DLD*, *HIBADH*, *HMGCS1*, *HSD17B10*, and *MCCC1*) in WM subjects, while the LCD caused a decrease in the *fructose and mannose metabolism* pathway (related to 6 genes: *ALDOA*, *ALDOB*, *KHK*, *MPI*, *PFKM*, and *PFKP*) in WR subjects.

After removing genes uniquely regulated in the two groups, 264 genes were differentially expressed in both WM and WR; however, directional concordance was not always maintained. As depicted in **FIGURE 2**, 170 genes were in directional concordance, meaning that the LCD had a similar effect on gene expression in both WM and WR groups. In contrast, 94 genes were regulated oppositely, meaning that the LCD had a different effect on gene expression in each group.

Those genes in directional concordance (DC) suggest the LCD caused a decrease in the *biosynthesis of unsaturated fatty acids* and *alpha-linoleic metabolism* pathways in both WM and WR. These pathways include such genes as fatty acid desaturase 1 (*FADS1*), fatty acid desaturase 2 (*FADS2*), acyl-CoA oxidase 1 (*ACOX*), and stearoyl-CoA desaturase (*SCD1*). Both *SCD1* and *FADS1* changes in expression were confirmed by real-time RT-PCR (data not shown). In addition, a subset of genes related to ribosomal pathways was up-regulated by the LCD in WM and WR.

We also examined pathways that were oppositely regulated by the LCD (OPP in **FIGURE 2**) in WM and WR groups. Several genes related to *focal adhesion* functions were up-regulated in WR and down-regulated in WM following the LCD: catenin beta 1 (*CTNNB1*), fibronectin 1 (*FN1*), mitogen-activated protein kinase 1 (*MAPK1*), PTK2 protein tyrosine kinase 2 (*PTK2*), β-actin (*ACTB*), and caveolin 1 (*CAV1*). These genes play important roles in the coordination of the extra-cellular matrix, and mediate processes such as cell growth and differentiation, and intracellular signalling; suggesting that a LCD had different effects on extracellular matrix remodelling in the two groups. Interestingly, the LCD resulted in the increased expression of genes related to *apoptosis* and the *p53 signalling pathway* in WM subjects and not WR subjects. More specifically, caspase 3 (*CASP3*) and caspase 8 (*CASP8*) are up-regulated in WM, while these genes are down-regulated in WR. *CASP8* gene expression values were validated by real-time RT-PCR. Both *CASP3* and *CASP8* play crucial roles initiating programmed cell death; suggesting that greater cell death in scAT during a LCD may underlie successful short-term weight maintenance.

Reporter Metabolite Analysis

To assess how the transcriptional differences in WM and WR most likely affected downstream metabolism, we overlaid our gene expression data with the EHMN reconstruction (28) and used Reporter Metabolite Analysis (26) to identify metabolites that may represent biomarkers for successful weight maintenance. It is apparent by the network structure that the LCD induces a more highly coordinated response in WM subjects compared to WR subjects, as represented by the dense and highly inter-connected network (**FIGURES 3A and B)**; however, there are some shared and distinct features within these two networks that are noteworthy.

Metabolites identified as significantly down-regulated by the LCD in both WM and WR were (2R, 4S)-2, 4-diaminopentanoate and 2-amino-4-oxopentanoic acid from the D-arginine and D-ornithine pathway ($P = 4.5 \times 10^{-4}$, $P = 2.0 \times 10^{-6}$, respectively). Their significance was driven by the *GAPDH* gene, which was down-regulated in both groups during caloric restriction (WM -1.16 fold; WR -1.17 fold).

Interestingly, we observed a global pattern in the metabolite network that was unique to WM and related to a large number of differentially expressed enzymes. The LCD resulted in a marked coordinated down-regulation in enzymes associated with fatty acid metabolism, citric acid cycle, and oxidative phosphorylation in

WM subjects; a signature that is absent in WR. More specifically, several metabolites displayed differential expression in their associated enzymes: NADPH, NADP+, NADH, H+, CoA, acetyl-CoA, acyl-CoA, stearoyl-CoA, oleoyl-CoA, palmitoyl-CoA, and palmitoleoyl-CoA (**TABLE 2**). The majority of these metabolites (9 out of 11) were previously identified in the Capel et al study (20), where the authors also reported a marked down-regulation in the same enzymes during energy restriction. Most of these metabolites were related to a large number of enzymes (reflected by the numerous connections to several enzymes in the metabolic network), which reinforced the differences in network connectivity observed between WM and WR, as these metabolites were not significantly regulated in the WR group.

Discussion

Considerable inter-individual variability in weight maintenance following caloric restriction has been observed. The present study was designed to assess whether changes in scAT gene expression profiles during the weight loss phase of a dietary intervention protocol could be used to predict changes in body weight during a subsequent 6-month weight maintenance phase. This analysis revealed that an 8-week low calorie diet (LCD) triggered distinct changes in scAT gene expression in subjects classified as weight maintainers (WM; 0-10% weight regain) compared to weight regainers (WR; 50-100% weight regain). Furthermore, only the WM group experienced changes in plasma triglyceride levels and insulin secretion during the LCD.

Fasting triglyceride levels were significantly decreased by the LCD in the WM group only. Schwab et al previously reported decreases in triglycerides enriched in saturated and short-chain fatty acids following energy restriction, which were associated with improved insulin sensitivity (33). A larger follow-up study failed to find significant decreases in plasma triglycerides (15), suggesting variability in response to an energy restricted diet. Weight maintenance was not assessed in either study; therefore it is unclear whether a relationship exists between the changes in triglycerides and successful 6-month weight maintenance.

A significant reduction in insulin secretion during the LCD was observed in WM, but not WR, subjects. Although the fasting insulin levels in WM and WR groups were not different at CID1, it is important to note that data regarding insulin secretion in response to an OGTT is not routinely used for bio-clinical matching of different subject groups. Therefore the significantly higher CID1 insulin secretion measured in WM compared to WR may serve as a novel predictor for successful weight maintenance following a LCD phase. A number of studies have attempted to determine whether insulin secretion plays a role in long term body weight regulation (34, 35). Previous research has tended to examine whether insulin secretion affects weight gain. Baseline insulin resistance was shown to not predict weight loss in healthy obese women who consumed a hypocaloric diet (36); our data suggests that baseline insulin resistance does not predict successful weight maintenance either. Recently, Crujeiras et al reported that baseline fasting plasma insulin levels do not predict weight regain (37); our results confirm this result as well. Schwartz et al demonstrated that reduced insulin secretion was a significant predictor for weight gain (34) and Chaput et al showed that 30 minute insulin levels during an OGTT were positively associated with 6-year weight gain (38). It is difficult to directly compare the outcomes of our study and these other studies because of different experimental designs; however, our data suggests that insulin secretion may indeed be a useful predictor of changes in body weight.

The present study provides a novel contribution to the existing literature by analyzing whether LCD-induced changes in scAT gene expression can be used to predict successful short-term weight maintenance. The functional analysis of gene expression data showed that *focal adhesion, apoptosis* and *p53 signalling* pathways were differentially regulated during a LCD in WM and WR. In the WM group, subjects experienced a decrease in the *focal adhesion* pathway; which consists of extracellular matrix genes associated with diverse functions such as inflammation, and cell growth and differentiation. Because the present study used a hypothesis generating approach and the genes related to focal adhesion have wide-ranging roles in various signalling pathways, it is difficult to predict whether extracellular remodelling is higher or lower in each group. Rather, we report here that a LCD has different effects on the extracellular matrix in WM and WR subjects.

The LCD caused an increase in caspase gene expression in WM subjects (i.e. *apoptosis* pathway), suggesting that these subjects may be experiencing an increase in scAT apoptosis. In addition, CIDEA, the most up-regulated gene in WM has also been shown to regulate apoptosis in different cell types, including adipocytes (39). Little previous work has examined the impact of diet-induced weight loss on adipose tissue apoptosis. Aubin et al studied obese subjects and found that an inhibitor of cellular adipose apoptosis was increased in the stroma-vascular fraction of scAT following weight loss (40). This work aligns with that of Alkhouri et al who recently showed that caspase-3 was up-regulated (i.e. increased apoptosis) in diet-induced obese mice (41). Our results suggest the opposite, where a LCD increased *CASP3* and *CASP8* expression only in individuals experiencing successful short-term weight maintenance. Although our study and that of Alkouri et al appear to conflict, there are several noteworthy differences. Firstly, different fat depots were used in these studies, suggesting that omental and subcutaneous fat depots may regulate apoptosis pathways differently following changes in body weight. Secondly, Alkouri et al compared morbidly obese and lean individuals, while we recruited only moderately obese participants. Despite these differences, both studies demonstrate that changes in body weight may influence adipose tissue apoptosis. Because of the variable response observed between individuals following caloric restriction, it appears likely that stratifying our population into WM and WR groups has better highlighted subtle differences in scAT apoptosis. The notion that greater scAT apoptosis during a LCD may predispose individuals to successful weight maintenance is intriguing; however, future studies are required in order to confirm this finding, identify the specific adipose tissue cell-type in which the apoptosis pathway is increased, and determine the physiological outcome for this increase.

The FunNet analysis revealed that a LCD regulated oxidative phosphorylation and lipid metabolism pathways similarly in both WM and WR subjects. However, the *valine, isoleucine, and leucine degradation* pathway was detected in WM subjects only and the *fructose and mannose metabolism* pathway was detected in WR only. The experimental design used in the present study is unique; therefore identifying pathways regulated by a LCD and associating this with successful and/or unsuccessful weight maintenance represents a novel finding that requires further examination.

Interestingly, the metabolic network analysis was able to pick up several metabolites related to fatty acid metabolism, the citric acid cycle, and oxidative phosphorylation that were specifically regulated by the LCD in WM and associated with a large number of down-regulated enzymes. It is most likely that this analysis was able to detect these differences because it incorporates metabolic network topology, a feature which is often lacking in classical bioinformatic functional analyses such as FunNet. The marked differential expression and major down-regulation in enzymes catalyzing fatty acid metabolism, the citric acid cycle, and oxidative phosphorylation observed during the LCD in WM suggests that individuals predisposed for successful weight maintenance may be able to decrease fat accumulation by coordinating a better overall metabolic response.

In conclusion, the current study demonstrates that LCD-induced changes in bio-clinical parameters and scAT gene expression may foreshadow weight maintenance and weight regain. More specifically, the LCD led to significant decreases in plasma triglyceride levels and insulin secretion only in subjects who subsequently experience successful short-term weight maintenance. Global gene expression profiling in scAT revealed that a LCD up-regulated pathways related to apoptosis in WM compared to WR. Moreover, metabolic network analyses revealed that genes related to fatty acid metabolism, the citric acid cycle, and oxidative phosphorylation are significantly down-regulated during the LCD in these same subjects. While it remains unclear to what extent LCD-induced changes in gene expression can be used to confidently predict short-term weight maintenance, our study reinforces the continued need to explore the relevance of genetic and metabolic factors for predicting changes in body weight.

ACKNOWLEDGEMENTS

The authors' responsibilities were as follows: DMM, MRT, VP, WHMS, JDZ, and KC determined the study design; JAM, DB, MAvB, THD, CGW, and AA were responsible for conducting the clinical investigation; CH was responsible for data integration; AMQ, NV, and DL were responsible for the RNA bank; DMM and VP performed the microarray work; VP performed the real-time RT-PCR validation; DMM and THP were responsible for statistical and bioinformatic analyses; DMM, THP, and KC prepared the first and final version of the manuscript; and all authors read and provided feedback on the different versions of the manuscript. None of the authors had a conflict of interest.

Table 1: LCD response in each group, as reflected in commonly measured bio-clinical parameters[1].

Bio-clinical Parameter	Weight Regainers (WR) (n = 20)		Main effect TIME	Weight Maintainers (WM) (n = 20)		Main effect TIME	GROUP*TIME interaction
	CID1	CID2	*P*-value[2]	CID1	CID2	*P*-value[2]	*P*-value[3]
Weight (kg)	91.9 ± 2.8	83.2 ± 2.6	<0.0001**	91.8 ± 2.7	82.1 ± 2.6	<0.0001**	0.1568
Body mass index (BMI)	33.5 ± 0.9	30.3 ± 0.9	<0.0001**	33.5 ± 0.9	29.9 ± 0.8	<0.0001**	0.1053
Fasting Cholesterol (mmol/l)	4.9 ± 0.2	4.1 ± 0.2	<0.0001**	5.1 ± 0.2	4.2 ± 0.2	<0.0001**	0.5989
Fasting Triglycerides (mmol/l)	1.1 ± 0.1	1.0 ± 0.1	0.3522	1.3 ± 0.1	1.0 ± 0.1	0.0007**	0.0627
Fasting HDL (mmol/l)	1.3 ± 0.1	1.1 ± 0.1	0.0039**	1.2 ± 0.1	1.1 ± 0.1	0.2918	0.1552
Fasting Fructosamin (µmol/l)	202 ± 4	200 ± 5	0.6280	206 ± 5	207 ± 4	0.8761	0.6489
Fasting Adiponectin (µg/ml)	9.8 ± 1.0	10.4 ± 1.0	0.4430	9.2 ± 1.0	10.4 ± 1.0	0.1276	0.5832
Fasting CRP (mg/l)	3.9 ± 0.6	4.8 ± 1.3	0.3666	4.6 ± 0.8	3.4 ± 0.9	0.1941	0.1238
Fasting glucose (mmol/l)	5.0 ± 0.1	4.7 ± 0.1	0.0032**	5.0 ± 0.1	4.7 ± 0.1	0.0007**	0.7126
Fasting insulin (µIU/ml)	8.9 ± 0.7	6.2 ± 0.6	0.0004**	11.8 ± 1.1	6.7 ± 0.7	<0.0001**	0.0479*
HOMA-IR	2.3 ± 0.2	1.5 ± 0.2	0.2496	2.9 ± 0.3	2.2 ± 0.8	0.3491	0.8704

[1]All values are means ± SEM. CID1 refers to the time point before caloric restriction; CID2 refers to the time point after 8 weeks of caloric restriction; WR, weight regainers; WM, weight maintainers; HOMA-IR, homeostasis model assessment of insulin resistance; CRP, C-reactive protein. The table includes bio-clinical variables at CID1 and CID2 in 20 women classified as weight regainers (WR) and 20 women classified as weight maintainers (WM). There was no GROUP effect at either CID1 or CID2, reinforcing that groups were well matched.

[2]$P < 0.05$ for TIME effects determined using an ANOVA and a post-hoc student's T-test to identify significant differences.

[3]$P < 0.05$ for significant interactions between GROUP (WM and WR) and TIME (CID1 and CID2). Subjects were paired at the two time points (i.e. CID1 and CID2) and bio-clinical parameters analyzed using an ANOVA model. An ANOVA model was generated and least square means were estimated for the differences between groups, times, and the interaction between group and time (GROUP*TIME), taking into account repeated measurements for all subjects.

Table 2. Reporter metabolites identified in WM and WR, and their validation with an independent study[1].

	Weight Maintainers (WM)			Weight Regainers (WR)			Capel et al (Ref #20)		
Reporter metabolite	p-value[2]	↑ reactions	↓ reactions	p-value[2]	↑ reactions	↓ reactions	p-value[2]	↑ reactions	↓ reactions
NADPH	0.008	57	88	0.217	75	70	0.026	6	16
Oleoyl-CoA	0.009	4	4	0.111	2	6	0.020	0	2
CoA	0.009	42	67	0.427	45	64	0.002	6	15
Acetyl-CoA	0.012	24	46	0.768	26	44	0.035	2	10
Stearoyl-CoA	0.013	5	8	0.164	3	10	0.020	0	2
NADP+	0.019	58	90	0.237	76	72	0.020	6	16
H+	0.031	128	204	0.117	149	183	0.037	9	38
NADH	0.033	50	74	0.217	75	70	0.085	3	21
Palmitoyl-CoA	0.039	10	17	0.105	7	20	0.054	1	2
Acyl-CoA	0.044	18	18	0.202	13	23	0.000	0	6
Palmitoleoyl-CoA	0.050	7	8	0.154	6	9	0.005	0	2

[1] To assess how the transcriptional differences in WM and WR most likely affected downstream metabolism, we overlaid our gene expression data with the Edinburgh human metabolic network (EHMN) reconstruction and used the reporter metabolite algorithm to identify metabolites that vary between the two groups. We compared our results with those of Capel et al (20). Note that even though EHMN reconstruction was used in both studies, the number of enzymes detected in each network differed because the different algorithms used to identify differentially regulated genes.

[2]The reporter metabolites' unadjusted *P*-values denote their significance of being metabolic "hot-spots". In other words, these metabolites are connected to more differentially regulated enzymes (between CID1 and CID2) than expected by chance. The ↑ and ↓ arrows indicate the number of metabolic reactions in which the metabolite is catalyzed by an enzyme that is either up- or down-regulated during the LCD. $P < 0.05$ for significant changes in reporter metabolites.

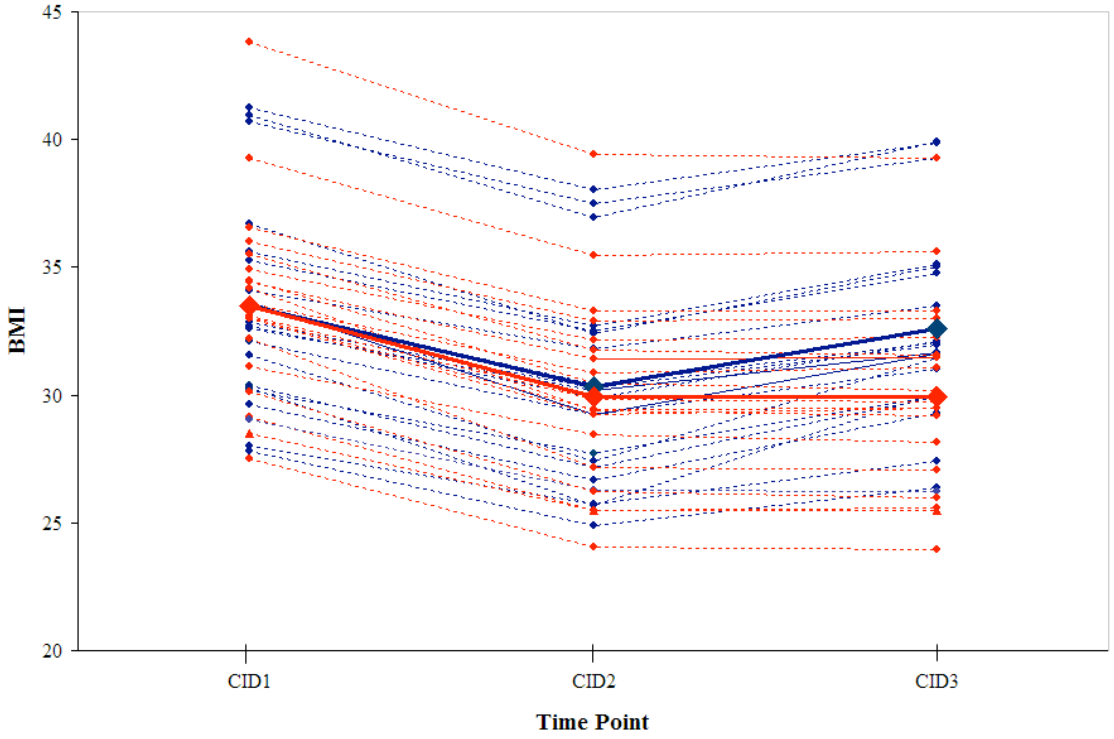


Figure 1: BMI evolution over the course of the intervention period. Each dotted line depicts the weight curve of an individual, where red and blue lines are for WM and WR, respectively. The solid red and blue lines depict the group average for WM and WR, respectively. CID1 refers to the time point before caloric restriction; CID2 refers to the time point after 8 weeks of caloric restriction; CID3 refers to the time point after the 6-month weight maintenance phase.

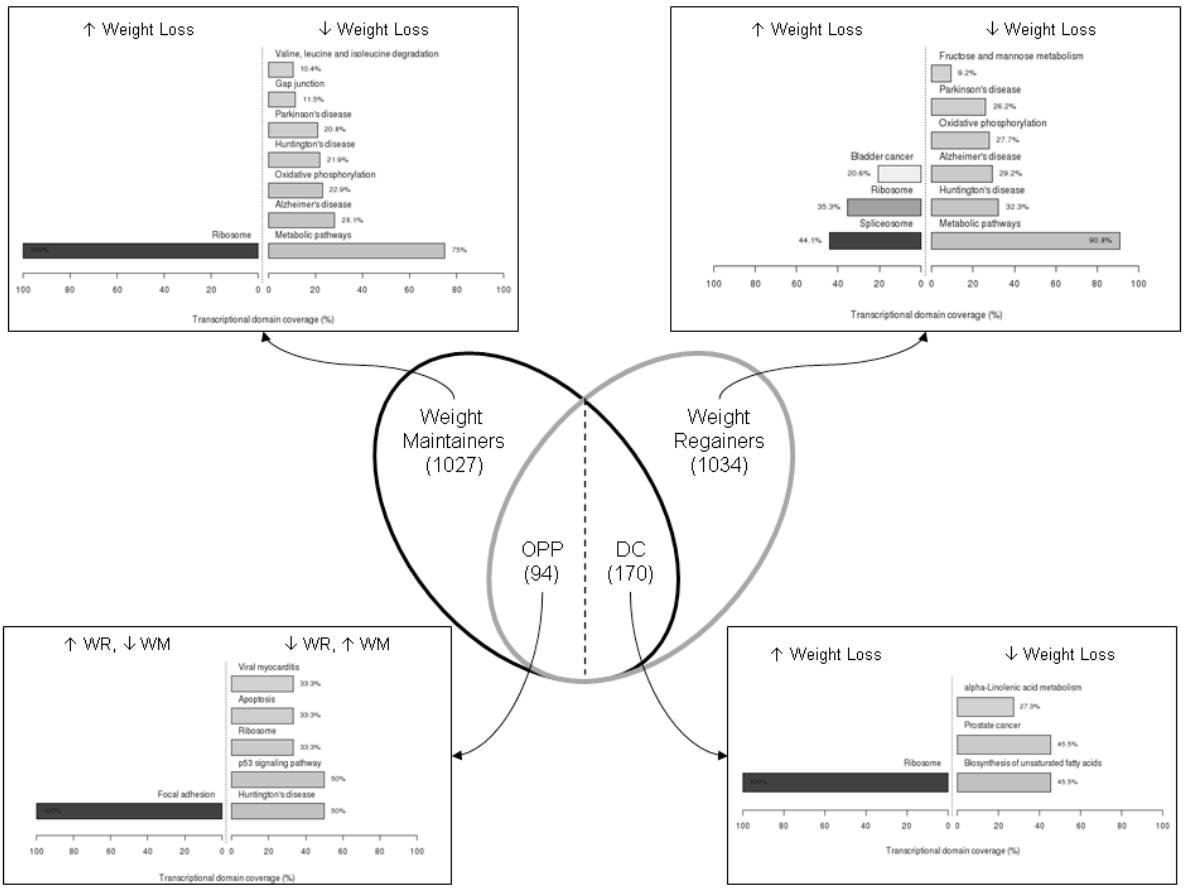
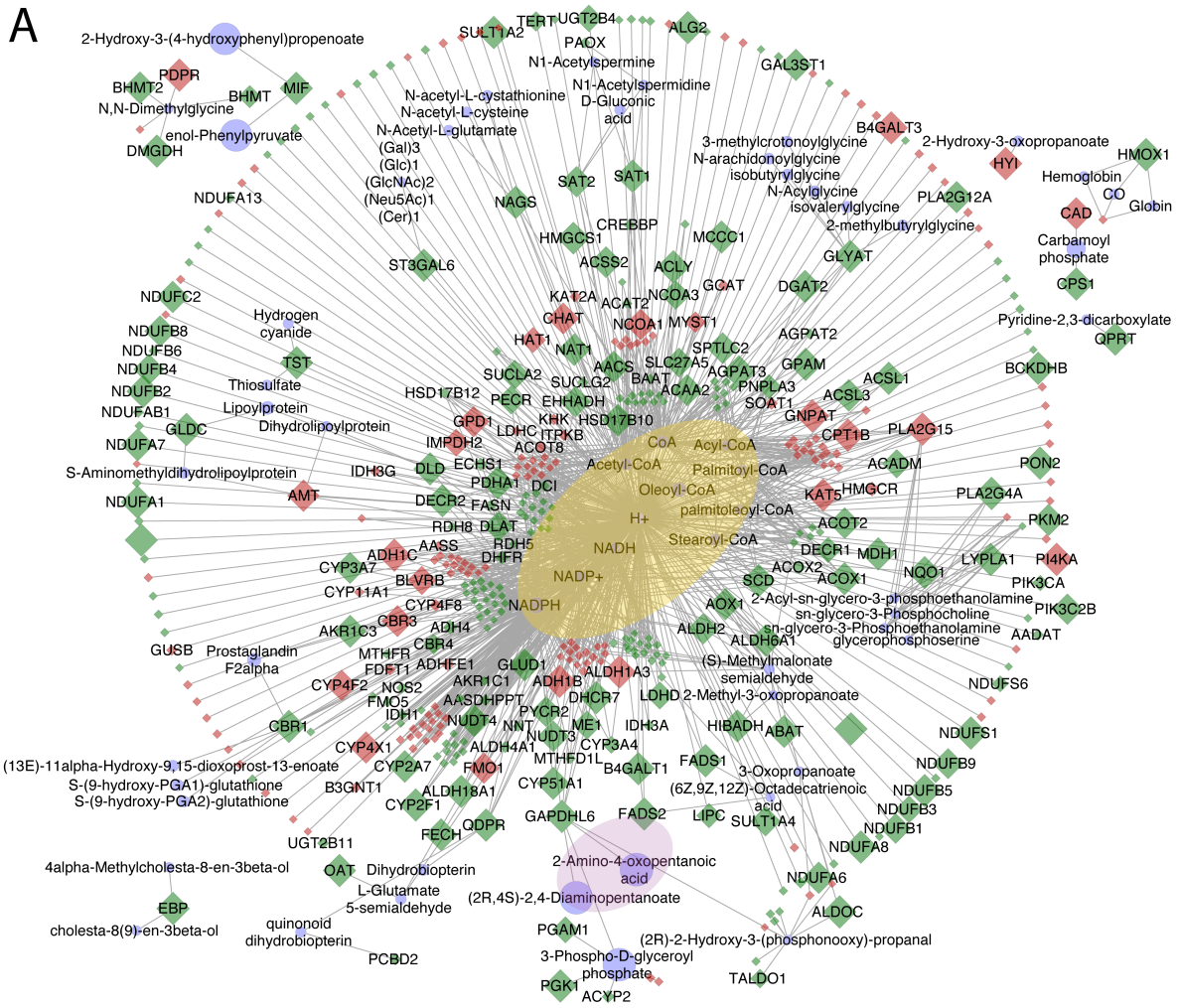


Figure 2: Venn diagram depicting overlap in differentially expressed genes following the LCD in WM and WR. Functional analyses revealed that although the genes regulated by the LCD in WM and WR differ they are related to similar functional processes. The LCD caused a significant decrease in the *valine, leucine, and isoleucine degradation* pathway in WM subjects. The LCD caused a significant decrease in the *fructose and mannose metabolism* pathway in WR subjects. Genes that are differentially expressed by the LCD in both WM and WR were not always directionally concordant. DC indicates genes in directional concordance, i.e. up or down in WM and WR subjects. *Oxidative phosphorylation* and *biosynthesis of unsaturated fatty acids* pathways were decreased by the LCD in both WM and WR. OPP indicates genes that are not directionally concordant, i.e. up in WM and down in WR, or vice versa. The *apoptosis* pathway was up-regulated by the LCD in WM compared to WR. ↑ and ↓ indicate an increase and decrease, respectively, during the LCD.

A


2-Hydroxy-3-(4-hydroxyphenyl)propenoate
PDPR
BHMT2
BHMT
MIF
N,N-Dimethylglycine
enol-Phenylpyruvate
DMGDH
NDUFA13
TERT
UGT2B4
SULT1A2
ALG2
PAOX
N1-Acetylspermine
N1-Acetylspermidine
D-Gluconic acid
GAL3ST1
N-acetyl-L-cystathionine
N-acetyl-L-cysteine
N-Acetyl-L-glutamate
(Gal)3
(Glc)1
(GlcNAc)2
(Neu5Ac)1
(Cer)1
NAGS
SAT2
SAT1
B4GALT3
3-methylcrotonoylglycine
N-arachidonoylglycine
isobutyrylglycine
N-Acylglycine
isovalerylglycine
2-Hydroxy-3-oxopropanoate
HYI
HMOX1
Hemoglobin
CO
Globin
PLA2G12A
CAD
2-methylbutyrylglycine
CREBBP
HMGCS1
MCCC1
GLYAT
Carbamoyl phosphate
CPS1
ST3GAL6
ACSS2
ACLY
GCAT
DGAT2
NDUFC2
KAT2A
ACAT2
NCOA3
Pyridine-2,3-dicarboxylate
QPRT
Hydrogen cyanide
CHAT
NCOA1
MYST1
AGPAT2
NDUFB8
NDUFB6
HAT1
NAT1
SPTLC2
NDUFB4
TST
SUCLA2
AACS
SLC27A5
AGPAT3
GPAM
BCKDHB
NDUFB2
Thiosulfate
SUCLG2
BAAT
ACAA2
PNPLA3
SOAT1
ACSL1
ACSL3
HSD17B12
PECR
EHHADH
NDUFAB1
Lipoylprotein
Dihydrolipoylprotein
GPD1
KHK
HSD17B10
GNPAT
CPT1B
PLA2G15
GLDC
LDHC
ITPKB
IMPDH2
ACOT8
CoA
Acyl-CoA
NDUFA7
ACADM
PON2
S-Aminomethyldihydrolipoylprotein
IDH3G
DLD
ECHS1
Acetyl-CoA
Palmitoyl-CoA
PDHA1
DCI
Oleoyl-CoA
HMGCR
PLA2G4A
NDUFA1
AMT
DECR2
FASN
KAT5
H+
palmitoleoyl-CoA
ACOT2
PKM2
RDH8
DLAT
Stearoyl-CoA
AASS
RDH5
NADH
DECR1
MDH1
LYPLA1
PI4KA
ADH1C
DHFR
NADP+
ACOX2
NQO1
CYP3A7
SCD
ACOX1
PIK3CA
BLVRB
CYP11A1
2-Acyl-sn-glycero-3-phosphoethanolamine
CYP4F8
NADPH
AOX1
sn-glycero-3-Phosphocholine
PIK3C2B
CBR3
AKR1C3
ADH4
ALDH2
sn-glycero-3-Phosphoethanolamine
glycerophosphoserine
AADAT
GUSB
Prostaglandin F2alpha
CBR4
MTHFR
ALDH6A1
FDFT1
ADHFE1
GLUD1
(S)-Methylmalonate semialdehyde
ALDH1A3
CYP4F2
NOS2
AKR1C1
ADH1B
NDUFS6
FMO5
AASDHPPT
DHCR7
LDHD
2-Methyl-3-oxopropanoate
IDH1
NUDT4
PYCR2
CBR1
NNT
ME1
IDH3A
HIBADH
ABAT
NDUFS1
NUDT3
CYP3A4
CYP4X1
ALDH4A1
MTHFD1L
FADS1
NDUFB9
(13E)-11alpha-Hydroxy-9,15-dioxoprost-13-enoate
CYP2A7
FMO1
B4GALT1
3-Oxopropanoate
S-(9-hydroxy-PGA1)-glutathione
B3GNT1
CYP51A1
(6Z,9Z,12Z)-Octadecatrienoic acid
NDUFB5
S-(9-hydroxy-PGA2)-glutathione
ALDH18A1
CYP2F1
GAPDHL6
FADS2
LIPC
SULT1A4
NDUFB3
QDPR
FECH
NDUFB1
UGT2B11
NDUFA8
4alpha-Methylcholesta-8-en-3beta-ol
OAT
Dihydrobiopterin
2-Amino-4-oxopentanoic acid
NDUFA6
L-Glutamate
5-semialdehyde
(2R,4S)-2,4-Diaminopentanoate
EBP
ALDOC
cholesta-8(9)-en-3beta-ol
quinonoid dihydrobiopterin
PGAM1
(2R)-2-Hydroxy-3-(phosphonooxy)-propanal
PCBD2
3-Phospho-D-glyceroyl phosphate
TALDO1
PGK1
ACYP2

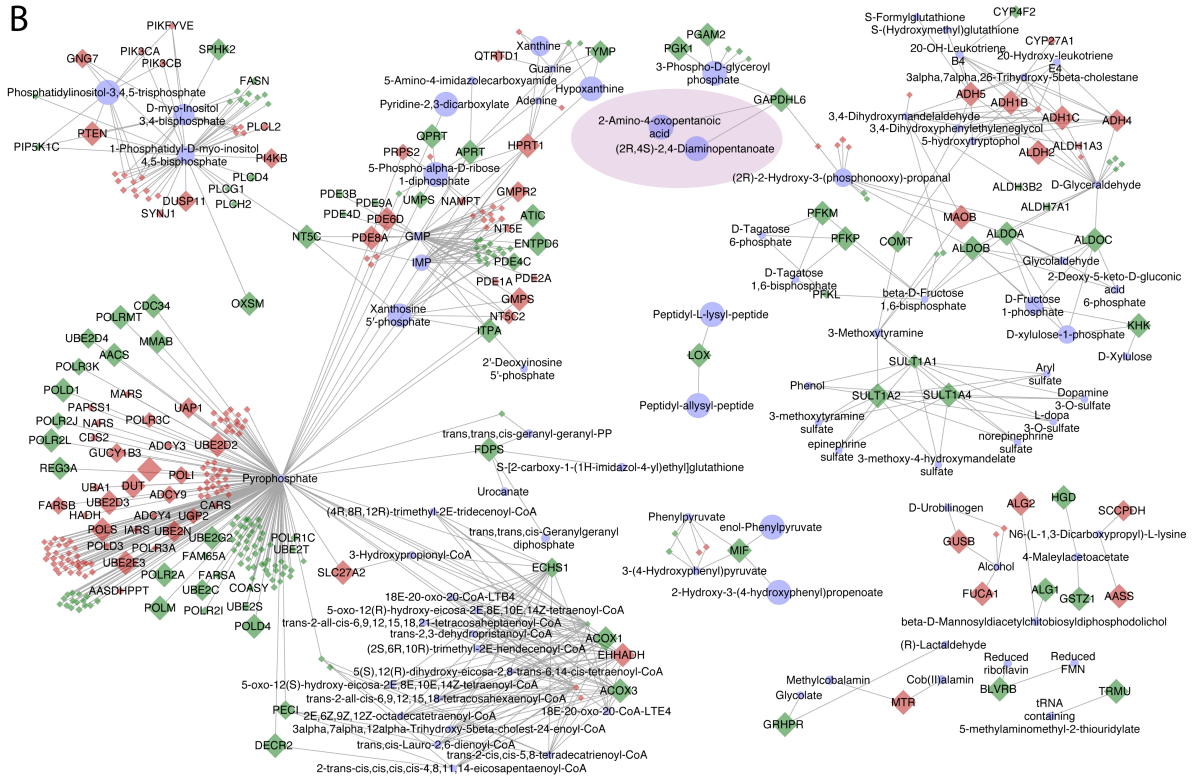


Figure 3. Metabolic reconstruction network analysis corresponding to LCD-induced changes in gene expression in WM (A) and WR (B). We used a high-confidence metabolic network reconstruction to search for metabolites that are catalyzed by enzymes that exhibit coordinated changes in gene expression levels during the LCD. The metabolic networks consist of metabolites (circles) that are connected to enzymes (diamonds) that catalyze the metabolites. Only metabolites linked to enzymes for which the underlying genes are differentially expressed during the LCD are shown. Metabolite circles are scaled according to their significance (i.e. larger circles reflect smaller *P*-values), where the *P*-value denotes the significance of being a metabolic "hot-spot". A metabolite that is connected to a large number of differentially regulated enzymes (between CID1 and CID2) will be depicted by a larger circle. The ranges of *P*-values, which correspond to circle size, are (A) 2.0×10^{-5} to 0.05 and (B) 8.9×10^{-10} to 0.05. For enzymes, red indicates an up-regulation and green a down-regulation in gene expression. The impact of the LCD on scAT gene expression is more highly coordinated in WM than WR, as represented by the dense and highly inter-connected network. In WM (A) the LCD elicits a marked coordinated down-regulation of genes coding for enzymes associated with fatty acid metabolism, citric acid cycle, and oxidative phosphorylation (shaded in yellow). That signature is absent in the WR analysis (B) in which alternate sites in metabolism are active. Highly significant metabolites common to both WM and WR are shaded in purple.

References

1. Bessesen DH. Update on obesity. J Clin Endocrinol Metab 2008;93:2027-34.
2. Wu T, Gao X, Chen M, van Dam RM. Long-term effectiveness of diet-plus-exercise interventions vs. diet-only interventions for weight loss: a meta-analysis. Obes Rev 2009;10:313-23.
3. Barte JC, Ter Bogt NC, Bogers RP, et al. Maintenance of weight loss after lifestyle interventions for overweight and obesity, a systematic review. Obes Rev 2010.
4. Shaw KA, O'Rourke P, Del Mar C, Kenardy J. Psychological interventions for overweight or obesity. Cochrane Database of Systematic Reviews 2005:Art. No.: CD003818. DOI: 10.1002/14651858.CD003818.pub2.
5. Elfhag K, Rossner S. Who succeeds in maintaining weight loss? A conceptual review of factors associated with weight loss maintenance and weight regain. Obes Rev 2005;6:67-85.
6. Wing RR, Phelan S. Long-term weight loss maintenance. Am J Clin Nutr 2005;82:222S-225S.
7. Weiss EC, Galuska DA, Kettel Khan L, Gillespie C, Serdula MK. Weight regain in U.S. adults who experienced substantial weight loss, 1999-2002. Am J Prev Med 2007;33:34-40.
8. Sacks FM, Bray GA, Carey VJ, et al. Comparison of weight-loss diets with different compositions of fat, protein, and carbohydrates. N Engl J Med 2009;360:859-73.
9. Delbridge EA, Prendergast LA, Pritchard JE, Proietto J. One-year weight maintenance after significant weight loss in healthy overweight and obese subjects: does diet composition matter? Am J Clin Nutr 2009;90:1203-14.
10. Capel F, Viguerie N, Vega N, et al. Contribution of energy restriction and macronutrient composition to changes in adipose tissue gene expression during dietary weight-loss programs in obese women. J Clin Endocrinol Metab 2008;93:4315-22.
11. Marquez-Quinones A, Mutch DM, Debard C, et al. Adipose tissue transcriptome reflects variations between subjects with continued weight loss and subjects regaining weight 6 mo after caloric restriction independent of energy intake. Am J Clin Nutr 2010.
12. Marti A, Martinez-Gonzalez MA, Martinez JA. Interaction between genes and lifestyle factors on obesity. Proc Nutr Soc 2008;67:1-8.
13. Clement K, Viguerie N, Poitou C, et al. Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. FASEB J. 2004;18:1657-1669.
14. Dahlman I, Linder K, Arvidsson NE, et al. Changes in adipose tissue gene expression with energy-restricted diets in obese women. Am J Clin Nutr 2005;81:1275-1285.
15. Kolehmainen M, Salopuro T, Schwab US, et al. Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. Int J Obes (Lond) 2008;32:292-303.
16. Viguerie N, Poitou C, Cancello R, Stich V, Clement K, Langin D. Transcriptomics applied to obesity and caloric restriction. Biochimie. 2005;87:117-123.
17. Henegar C, Tordjman J, Achard V, et al. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. Genome Biol 2008;9:R14.
18. Bouchard L, Rabasa-Lhoret R, Faraj M, et al. Differential epigenomic and transcriptomic responses in subcutaneous adipose tissue between low and high responders to caloric restriction. Am J Clin Nutr 2010;91:309-20.
19. Mutch DM, Temanni MR, Henegar C, et al. Adipose gene expression prior to weight loss can differentiate and weakly predict dietary responders. PLoS ONE. 2007;2:e1344.
20. Capel F, Klimcakova E, Viguerie N, et al. Macrophages and adipocytes in human obesity: adipose tissue gene expression and insulin sensitivity during calorie restriction and weight stabilization. Diabetes 2009;58:1558-67.
21. Larsen TM, Dalskov S, van Baak M, et al. The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries - a comprehensive design for long-term intervention. Obes Rev 2009.
22. Moore CS, Lindroos AK, Kreutzer M, et al. Dietary strategy to manipulate ad libitum macronutrient intake, and glycaemic index, across eight European countries in the Diogenes Study. Obes Rev 2009.
23. Larsen TM, Dalskov SM, van Baak M, et al. Diets with high or low protein content and glycemic index for weight-loss maintenance. N Engl J Med 2010;363:2102-13.
24. Lemoine S, Combes F, Servant N, Le CS. Goulphar: rapid access and expertise for standard two-color microarray normalization methods. BMC Bioinformatics 2006;7:467.
25. Prifti E, Zucker JD, Clement K, Henegar C. FunNet: an integrative tool for exploring transcriptional interactions. Bioinformatics 2008.
26. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 2005;102:2685-9.

27. Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR. Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. PLoS Comput Biol 2010;6:e1000729.
28. Ma H, Sorokin A, Mazein A, et al. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 2007;3:135.
29. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498-504.
30. Merino DM, Ma DW, Mutch DM. Genetic variation in lipid desaturases and its impact on the development of human disease. Lipids Health Dis 2010;9:63.
31. Puri V, Ranjit S, Konda S, et al. Cidea is associated with lipid droplets and insulin sensitivity in humans. Proc Natl Acad Sci U S A 2008;105:7833-8.
32. Teichert-Kuliszewska K, Hamilton BS, Roncari DA, et al. Increasing vimentin expression associated with differentiation of human and rat preadipocytes. Int J Obes Relat Metab Disord 1996;20 Suppl 3:S108-13.
33. Schwab U, Seppanen-Laakso T, Yetukuri L, et al. Triacylglycerol fatty acid composition in diet-induced weight loss in subjects with abnormal glucose metabolism--the GENOBIN study. PLoS ONE. 2008;3:e2630.
34. Schwartz MW, Boyko EJ, Kahn SE, Ravussin E, Bogardus C. Reduced insulin secretion: an independent predictor of body weight gain. J Clin Endocrinol Metab 1995;80:1571-6.
35. Weyer C, Hanson K, Bogardus C, Pratley RE. Long-term changes in insulin action and insulin secretion associated with gain, loss, regain and maintenance of body weight. Diabetologia 2000;43:36-46.
36. McLaughlin T, Abbasi F, Carantoni M, Schaaf P, Reaven G. Differences in insulin resistance do not predict weight loss in response to hypocaloric diets in healthy obese women. J Clin Endocrinol Metab 1999;84:578-81.
37. Crujeiras AB, Goyenechea E, Abete I, et al. Weight Regain after a Diet-Induced Loss Is Predicted by Higher Baseline Leptin and Lower Ghrelin Plasma Levels. J Clin Endocrinol Metab 2010.
38. Chaput JP, Tremblay A, Rimm EB, Bouchard C, Ludwig DS. A novel interaction between dietary composition and insulin secretion: effects on weight gain in the Quebec Family Study. Am J Clin Nutr 2008;87:303-9.
39. Ito M, Nagasawa M, Hara T, Ide T, Murakami K. Differential roles of CIDEA and CIDEC in insulin-induced anti-apoptosis and lipid droplet formation in human adipocytes. J Lipid Res 2010;51:1676-84.
40. Aubin D, Gagnon A, Grunder L, Dent R, Allen M, Sorisky A. Adipogenic and antiapoptotic protein levels in human adipose stromal cells after weight loss. Obes Res 2004;12:1231-4.
41. Alkhouri N, Gornicka A, Berk MP, et al. Adipocyte apoptosis, a link between obesity, insulin resistance, and hepatic steatosis. J Biol Chem 2010;285:3428-38.

[Supplementary Files and Tables can be found in online supplement of the paper]

Concluding remarks

6

In this Thesis I have presented and discussed methodologies that go beyond approaches that solely proceed within a single data type, such as traditional GWA analysis. Limitations in the latter type of approaches are becoming increasingly apparent, and integrative approaches provide a promising alternative.

One of the major challenges for integrative approaches is that complex traits most often are characterized by substantial genetic heterogeneity that is distributed across processes, which are still poorly represented in interaction databases, and incompletely captured by existing high-throughput technologies. However, absence of evidence does not necessarily imply evidence of absence. In the remainder of the Thesis, I will outline how data integration can be used to prioritize rare variants for targeted follow-up studies, and conclude with some final remarks.

The major theme in this Thesis has been that GWA studies have proven successful in identifying novel etiologic loci, but that these single-data type-based analyses comprise several inherent weaknesses as well. The three most pronounced being:

i) Despite estimations that genetic factors account for at least one third of the variation in most complex traits, findings from GWA studies currently explain less than 10% of the genetic variation for most of them. Let me use obesity as an example; despite estimations that this risk-phenotype is 40-70% heritable, the established 32 body-mass index-associated SNPs identified by the GIANT Consortium (a collaboration on GWA studies with focus on anthropometric measures) [Speliotes et al., 2010] account for less than 2% of the genetic variation in the trait.

ii) For most loci from GWA studies it is not clear what gene is the relevant one (several genes overlap most GWA loci), and in situations where the relevant gene could be mapped, it is often unclear what its function is.

iii) For most heterogenic traits (including obesity) genetic interactions between risk variants remain elusive. In addition, no specific etiologic pathways have been assigned to the majority of the loci identified in GWA studies.

In the post GWA analysis era, the genetics research community is turning towards sequencing to detect rare variants that my explain larger parts of the genetic variation in complex traits (and in some cases identify risk enhancing or risk decreasing genes unambiguously – at least for coding variants). Examples on such efforts are (1) imputation of GWA study data with variation data from the 1000 Genomes Project [1000

Genomes Project Consortium et al., 2010], (2) re-sequencing of loci that have been associated with complex traits [Hardy and Singleton, 2009], (3) new microarrays that capture rare variants,[1] (4) exome sequencing, and (5) whole-genome sequencing. Especially, the latter two analytical avenues have resulted in the discovery of causal variants for several Mendelian disorders [Ng et al., 2010], and are producing a wealth of rare variant data.

Arguably, the massive amount of data may soon become a 'burden' rather than valuable new information, as the identification of rare causal variants in heterogenic traits is believed to become challenging. Specifically, smaller sample sizes (due to relatively high cost of sequencing), and difficulties in distinguishing background variation from causal variation, will limit statistical power to detect etiologic variants.

A key hypothesis underlying the work presented in this Thesis, was that a given risk-phenotype is rarely the consequence of a single polymorphic gene, but rather caused by a complex interplay of various risk variants acting upon networks of genes. I have presented three different methodologies that address points (i) - (iii) by using data integration as a means to augment analyses of genetic variation underlying complex traits. In **Paper I**, my co-workers and I presented a method that identifies associations with individually moderate effects, but in aggregate significant effects on the phenotype. In **Paper II**, we showed that integration of several complementary data types identifies associations that were missed in the original studies. In **Paper III**, we reported that information on known associations' enrichment in specific protein complexes, may incriminate other, hitherto uncharacterized, susceptibility genes. In **Papers IV-V**, we showed that coordinated changes in gene expression levels may foreshadow downstream phenotypic changes. These approaches may lead to a better understanding of genetic variation in complex traits, by (1) identifying novel moderate effect size risk factors, (2) placing genes in etiological contexts, and (3) pinpointing pathobiological pathways.

Oftentimes less is more, but when it comes to deciphering biology *more* can make the difference.

[1] For instance the new Metabochip, a custom-designed Illumina microarray that holds 185,000 medium rare SNPs within regions that previous GWA studies have been associated with metabolic and cardiovascular traits [Ingelsson, 2010].

Bibliography

[1000 Genomes Project Consortium et al., 2010] 1000 Genomes Project Consortium, Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061--1073.

[Allen et al., 2010] Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J.-H. H., Yang, J., Gudbjartsson, D., Heard-Costa, N. L., Randall, J. C., Qi, L., Smith, A. V., Mägi, R., Pastinen, T., Liang, L., Heid, I. M., Luan, J., Thorleifsson, G., Winkler, T. W., Goddard, M. E., Lo, K. S., Palmer, C., Workalemahu, T., Aulchenko, Y. S., Johansson, A., Zillikens, M. C., Feitosa, M. F., Esko, T. o., Johnson, T., Ketkar, S., Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N. L., Hayward, C., Hottenga, J.-J. J., Jacobs, K. B., Knowles, J. W., Kutalik, Z., Monda, K. L., Polasek, O., Preuss, M., Rayner, N. W., Robertson, N. R., Steinthorsdottir, V., Tyrer, J. P., Voight, B. F., Wiklund, F., Xu, J., Zhao, J. H., Nyholt, D. R., Pellikka, N., Perola, M., Perry, J. R., Surakka, I., Tammesoo, M.-L. L., Altmaier, E. L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D. I., Chen, C., Coin, L., Cooper, M. N., Dixon, A. L., Gibson, Q., Grundberg, E., Hao, K., Junttila, M. J., Kaplan, L. M., Kettunen, J., König, I. R., Kwan, T., Lawrence, R. W., Levinson, D. F., Lorentzon, M., McKnight, B., Morris, A. P., Müller, M., Ngwa, J. S., Purcell, S., Rafelt, S., Salem, R. M., Salvi, E., Sanna, S., Shi, J., Sovio, U., Thompson, J. R., Turchin, M. C., Vandenput, L., Verlaan, D. J., Vitart, V., White, C. C., Ziegler, A., Almgren, P., Balmforth, A. J., Campbell, H., Citterio, L., De Grandi, A., Dominiczak, A., Duan, J., Elliott, P., Elosua, R., Eriksson, J. G., Freimer, N. B., Geus, E. J., Glorioso, N., Haiqing, S., Hartikainen, A.-L. L., Havulinna, A. S., Hicks, A. A., Hui, J., Igl, W., Illig, T., Jula, A., Kajantie, E., Kilpeläinen, T. O., Koiranen, M., Kolcic, I., Koskinen, S., Kovacs, P., Laitinen, J., Liu, J., Lokki, M.-L. L., Marusic, A., Maschio, A., Meitinger, T., Mulas, A., Paré, G., Parker, A. N., Peden, J. F., Petersmann, A., Pichler, I., Pietiläinen, K. H., Pouta, A., Ridderstråle, M., Rotter, J. I., Sambrook, J. G., Sanders, A. R., Schmidt, C. O., Sinisalo, J., Smit, J. H., Stringham, H. M., Walters, G. B., Widen, E., Wild, S. H., Willemsen, G., Zagato, L., Zgaga, L., Zitting, P., Alavere, H., Farrall, M., McArdle, W. L., Nelis, M., Peters, M. J., Ripatti, S., van Meurs, J. B., Aben, K. K., Ardlie, K. G., Beckmann, J. S., Beilby, J. P., Bergman, R. N., Bergmann, S., Collins, F. S., Cusi, D., den Heijer, M., Eiriksdottir, G., Gejman, P. V., Hall, A. S., Hamsten, A., Huikuri, H. V., Iribarren, C., Kähönen, M., Kaprio, J., Kathiresan, S., Kiemeney, L., Kocher, T., Launer, L. J., Lehtimäki,

T., Melander, O., Mosley, T. H., Musk, A. W., Nieminen, M. S., O'Donnell, C. J., Ohlsson, C., Oostra, B., Palmer, L. J., Raitakari, O., Ridker, P. M., Rioux, J. D., Rissanen, A., Rivolta, C., Schunkert, H., Shuldiner, A. R., Siscovick, D. S., Stumvoll, M., Tönjes, A., Tuomilehto, J., van Ommen, G.-J. J., Viikari, J., Heath, A. C., Martin, N. G., Montgomery, G. W., Province, M. A., Kayser, M., Arnold, A. M., Atwood, L. D., Boerwinkle, E., Chanock, S. J., Deloukas, P., Gieger, C., Grönberg, H., Hall, P., Hattersley, A. T., Hengstenberg, C., Hoffman, W., Lathrop, G. M., Salomaa, V., Schreiber, S., Uda, M., Waterworth, D., Wright, A. F., Assimes, T. L., Barroso, I., Hofman, A., Mohlke, K. L., Boomsma, D. I., Caulfield, M. J., Cupples, L. A., Erdmann, J., Fox, C. S., Gudnason, V., Gyllensten, U., Harris, T. B., Hayes, R. B., Jarvelin, M.-R. R., Mooser, V., Munroe, P. B., Ouwehand, W. H., Penninx, B. W., Pramstaller, P. P., Quertermous, T., Rudan, I., Samani, N. J., Spector, T. D., Völzke, H., Watkins, H., Wilson, J. F., Groop, L. C., Haritunians, T., Hu, F. B., Kaplan, R. C., Metspalu, A., North, K. E., Schlessinger, D., Wareham, N. J., Hunter, D. J., O'Connell, J. R., Strachan, D. P., Wichmann, H.-E. E., Borecki, I. B., van Duijn, C. M., Schadt, E. E., Thorsteinsdottir, U., Peltonen, L., Uitterlinden, A. G., Visscher, P. M., Chatterjee, N., Loos, R. J., Boehnke, M., McCarthy, M. I., Ingelsson, E., Lindgren, C. M., Abecasis, G. R., Stefansson, K., Frayling, T. M., and Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, advance online publication.

[Altshuler et al., 2010] Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Marie Muzny, D., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarroll, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Cristina Manca, M., Marshall, P. A., Matsuda, I., Ngare, D., Ota Wang, V., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52--58.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25--29.

[Ashe et al., 1997] Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P., and Proudfoot, N. J. (1997). Intergenic transcription and transinduction of the human beta-globin locus. *Genes & development*, 11(19):2494--2509.

[Bader et al., 2001] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T., and Hogue, C. W. V. (2001). BIND--The Biomolecular Interaction Network Database. *Nucl. Acids Res.*, 29(1):242--245.

[Barabási et al., 2011] Barabási, A.-L. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56--68.

[Baranzini et al., 2009] Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B. M., Kappos, L., Polman, C. H., Matthews, P. M., Hauser, S. L., Gibson, R. A., Oksenberg, J. R., and Barnes, M. R. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet*, 18(11):2078--90.

[Barrett et al., 2008] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.-P., de Vos, M., Vermeire, S., Louis, E., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, 40(8):955--962.

[Becker et al., 2004] Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The Genetic Association Database. *Nature Genetics*, 36(5):431--432.

[Bochukova et al., 2009] Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., Hurles, M. E., and Farooqi, I. S. (2009). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666--670.

[Boissel et al., 2009] Boissel, S., Reish, O., Proulx, K., Kawagoe-Takaki, H., Sedgwick, B., Yeo, G., Meyre, D., Golzio, C., Molinari, F., Kadhom, N., Etchevers, H., Saudek, V., Farooqi, I., Froguel, P., Lindahl, T., O'Rahilly, S., Munnich, A., and Colleaux, L. (2009). Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *American Journal of Human Genetics*, 85(1):106--111.

[Brown and Jurisica, 2005] Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics (Oxford, England)*, 21(9):2076--2082.

[Burgard et al., 2004] Burgard, A. P., Nikolaev, E. V., Schilling, C. H., and Maranas, C. D. (2004). Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions. *Genome Research*, 14(2):301--312.

[Bush et al., 2009] Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 368--379.

[Butland et al., 2005] Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli . *Nature*, 433(7025):531--537.

[Cantor et al., 2010] Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American journal of human genetics*, 86(1):6--22.

[Capel et al., 2009] Capel, F., Klimčáková, E., Viguerie, N., Roussel, B., Vítková, M., Kováčiková, M., Polák, J., Kováčová, Z., Galitzky, J., Maoret, J.-J., Hanáček, J., Pers, T. H., Bouloumié, A., Štich, V., and Langin, D. (2009). Macrophages and Adipocytes in Human Obesity. *Diabetes*, 58(7):1558--1567.

[Chatr-aryamontri et al., 2007] Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35(suppl 1):D572--D574.

[Cheng et al., 2005] Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science (New York, N.Y.)*, 308(5725):1149--1154.

[Chuang et al., 2007] Chuang, H.-Y. Y., Lee, E., Liu, Y.-T. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3.

[Church et al., 2009] Church, C., Lee, S., Bagg, E. A. L., McTaggart, J. S., Deacon, R., Gerken, T., Lee, A., Moir, L., Mecinović, J., Quwailid, M. M., Schofield, C. J., Ashcroft, F. M., and Cox, R. D. (2009). A Mouse Model for the Metabolic Effects of the Human Fat Mass and Obesity Associated FTO Gene. *PLoS Genet*, 5(8):e1000599+.

[Church et al., 2010] Church, C., Moir, L., McMurray, F., Girard, C., Banks, G. T., Teboul, L., Wells, S., Brüning, J. C., Nolan, P. M., Ashcroft, F. M., and Cox, R. D. (2010). Overexpression of Fto leads to increased food intake and results in obesity. *Nature Genetics*, 42(12):1086--1092.

[Cirulli and Goldstein, 2010] Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415--425.

[Cordell, 2009] Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392--404.

[Croft et al., 2011] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691--D697.

[Dalgaard et al., 2011] Dalgaard, M., Weinhold, N., Edsgrd, D., Silver, J., Pers, T., Jørgensen, N., Juul, A., Gerds, T., Giwercman, A., Giwercman, Y., Cedermark, G., Virtanen, H., Toppari, J., Daugaard, G., Jensen, T., Brunak, S., Rajpert-De Meyts, E., Skakkebæk, N., Leffers, H., and Gupta, R. (2011). A genome-wide association study of men with symptoms of testicular dysgenesis syndrome and its network biology interpretation. *Submitted*.

[de Bakker et al., 2008] de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, 17(R2):R122--R128.

[Dina et al., 2007] Dina, C., Meyre, D., Gallina, S., Durand, E., Korner, A., Jacobson, P., Carlsson, L. M. S., Kiess, W., Vatin, V., Lecoeur, C., Delplanque, J., Vaillant, E., Pattou, F., Ruiz, J., Weill, J., Levy-Marchal, C., Horber, F., Potoczna, N., Hercberg, S., Le Stunff, C., Bougneres, P., Kovacs, P., Marre, M., Balkau, B., Cauchi, S., Chevre, J.-C., and Froguel, P. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics*, 39(6):724--726.

[Duarte et al., 2007] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777--1782.

[Dupuis et al., 2010] Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., Lindgren, C. M., Mägi, R., Morris, A. P., Randall, J., Johnson, T., Elliott, P., Rybin, D., Thorleifsson, G., Steinthorsdottir, V., Henneman, P., Grallert, H., Dehghan, A., Jan, J., Franklin, C. S., Navarro, P., Song, K., Goel, A., Perry, J. R., Egan, J. M., Lajunen, T., Grarup, N., Sparsø, T., Doney, A., Voight, B. F., Stringham, H. M., Li, M., Kanoni, S., Shrader, P., Cavalcanti-Proença, C., Kumari, M., Qi, L., Timpson, N. J., Gieger, C., Zabena, C., Rocheleau, G., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S. A., Payne, F., Maria, R., Pattou, F., Sethupathy, P., Ardlie, K., Ariyurek, Y., Balkau, B., Barter, P., Beilby, J. P., Ben-Shlomo, Y., Benediktsson, R., Bennett, A. J., Bergmann, S., Bochud, M., Boerwinkle, E., Bonnefond, A., Bonnycastle, L. L., Borch-Johnsen, K., Böttcher, Y., Brunner, E., Bumpstead, S. J., Charpentier, G., Ida, Y.-D., Chines, P., Clarke, R., Coin, L. J., Cooper, M. N., Cornelis, M., Crawford, G., Crisponi, L., Day, I. N., de Geus, E. J., Delplanque, J., Dina, C., Erdos, M. R., Fedson, A. C., Fischer-Rosinsky, A., Forouhi, N. G., Fox, C. S., Frants, R., Grazia, M., Galan, P., Goodarzi, M. O., Graessler, J., Groves, C. J., Grundy, S., Gwilliam, R., Gyllensten, U., Hadjadj, S., Hallmans, G., Hammond, N., Han, X., Hartikainen, A.-L. L., Hassanali, N., Hayward, C., Heath, S. C., Hercberg, S., Herder, C., Hicks, A. A., Hillman, D. R., Hingorani, A. D., Hofman, A., Hui, J., Hung, J., Isomaa, B., Johnson, P. R., Jørgensen, T., Jula, A., Kaakinen, M., Kaprio, J., Kesaniemi, A. A., Kivimaki, M., Knight, B., Koskinen, S., Kovacs, P., Ohm, K., Lathrop, M. M., Lawlor, D. A., Le Bacquer, O., Lecoeur, C., Li, Y., Lyssenko, V., Mahley, R., Mangino, M., Manning, A. K., Teresa, M., McAteer, J. B., McCulloch, L. J., McPherson, R., Meisinger, C., Melzer, D., Meyre, D., Mitchell, B. D., Morken, M. A., Mukherjee, S., Naitza, S., Narisu, N., Neville, M. J., Oostra, B. A., Orrù, M., Pakyz, R., Palmer, C. N., Paolisso, G., Pattaro, C., Pearson, D., Peden, J. F., Pedersen, N. L., Perola, M., Pfeiffer, A. F., Pichler, I., Polasek, O., Posthuma, D., Potter, S. C., Pouta,

A., Province, M. A., Psaty, B. M., Rathmann, W., Rayner, N. W., Rice, K., Ripatti, S., Rivadeneira, F., Roden, M., Rolandsson, O., Sandbaek, A., Sandhu, M., Sanna, S., Aihie, A., Scheet, P., Scott, L. J., Seedorf, U., Sharp, S. J., Shields, B., Sigurethsson, G., Sijbrands, E. J., Silveira, A., Simpson, L., Singleton, A., Smith, N. L., Sovio, U., Swift, A., Syddall, H., Syvänen, A.-C. C., Tanaka, T., Thorand, B., Tichet, J., Tönjes, A., Tuomi, T., Uitterlinden, A. G., Willems, K., van Hoek, M., Varma, D., Visvikis-Siest, S., Vitart, V., Vogelzangs, N., Waeber, G., Wagner, P. J., Walley, A., Walters, B. B., Ward, K. L., Watkins, H., Weedon, M. N., Wild, S. H., Willemsen, G., Witteman, J. C., Yarnell, J. W., Zeggini, E., Zelenika, D., Zethelius, B., Zhai, G., Hua, J., Zillikens, C. C., DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, Borecki, I. B., Loos, R. J., Meneton, P., Magnusson, P. K., Nathan, D. M., Williams, G. H., Hattersley, A. T., Silander, K., Salomaa, V., Davey, G., Bornstein, S. R., Schwarz, P., Spranger, J., Karpe, F., Shuldiner, A. R., Cooper, C., Dedoussis, G. V., Serrano-Ríos, M., Morris, A. D., Lind, L., Palmer, L. J., Hu, F. B., Franks, P. W., Ebrahim, S., Marmot, M., Kao, L. H., Pankow, J. S., Sampson, M. J., Kuusisto, J., Laakso, M., Hansen, T., Pedersen, O., Paul, P., Wichmann, E. E., Illig, T., Rudan, I., Wright, A. F., Stumvoll, M., Campbell, H., Wilson, J. F., Anders Hamsten on behalf of Procardis Consortium, MAGIC investigators, Bergman, R. N., Buchanan, T. A., Collins, F. S., Mohlke, K. L., Tuomilehto, J., Valle, T. T., Altshuler, D., Rotter, J. I., Siscovick, D. S., Penninx, B. W., Boomsma, D. I., Deloukas, P., Spector, T. D., Frayling, T. M., Ferrucci, L., Kong, A., Thorsteinsdottir, U., Stefansson, K., van Duijn, C. M., Aulchenko, Y. S., Cao, A., Scuteri, A., Schlessinger, D., Uda, M., Ruokonen, A., Jarvelin, M.-R. R., Waterworth, D. M., Vollenweider, P., Peltonen, L., Mooser, V., Abecasis, G. R., Wareham, N. J., Sladek, R., Froguel, P., Watanabe, R. M., Meigs, J. B., Groop, L., Boehnke, M., McCarthy, M. I., Florez, J. C., and Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105--116.

[Editorial, 2010] Editorial, N. G. (2010). Primary research on existing data. *Nature Genetics*, 42(6):467.

[Elbers et al., 2009] Elbers, C. C., van Eijk, K. R., Franke, L., Mulder, F., van der Schouw, Y. T., Wijmenga, C., and Onland-Moret, N. C. (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.*, 33(5):419--431.

[Emily et al., 2009] Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231--1240.

[Engström et al., 2006] Engström, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S. L., Yang, L., Kunarso, G., Ng, E. L., Batalov, S., Wahlestedt, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Wells, C., Bajic, V. B., Orlando, V., Reid, J. F., Lenhard, B., and Lipovich, L. (2006). Complex Loci in Human and Mouse Genomes. *PLoS Genet*, 2(4):e47+.

[Fawcett and Barroso, 2010] Fawcett, K. A. and Barroso, I. (2010). The genetics of obesity: FTO leads the way. *Trends in genetics : TIG*, 26(6):266--274.

[Finn et al., 2008] Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J. J., Hotz, H.-R. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A.

(2008). The Pfam protein families database. *Nucleic acids research*, 36(Database issue):D281--D288.

[Fischer et al., 2009] Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T., Bruning, J. C., and Ruther, U. (2009). Inactivation of the Fto gene protects from obesity. *Nature*, 458(7240):894--898.

[Frayling et al., 2007] Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M.-R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D., The, Hattersley, A. T., and Mccarthy, M. I. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, 316(5826):889--894.

[Frazer et al., 2009] Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241--251.

[Galwey, 2009] Galwey, N. W. (2009). A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic epidemiology*, 33(7):559--568.

[Garrod, 2002] Garrod, A. (2002). The incidence of alkaptonuria: a study in chemical individuality. *Yale Journal of Biological Medicine*, 75(5):221--231.

[Gavin et al., 2006] Gavin, A.-C. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631--636.

[Gerken et al., 2007] Gerken, T., Girard, C. A., Tung, Y.-C. L., Webby, C. J., Saudek, V., Hewitson, K. S., Yeo, G. S. H., McDonough, M. A., Cunliffe, S., McNeill, L. A., Galvanovskis, J., Rorsman, P., Robins, P., Prieur, X., Coll, A. P., Ma, M., Jovanovic, Z., Farooqi, I. S., Sedgwick, B., Barroso, I., Lindahl, T., Ponting, C. P., Ashcroft, F. M., O'Rahilly, S., and Schofield, C. J. (2007). The Obesity-Associated FTO Gene Encodes a 2-Oxoglutarate-Dependent Nucleic Acid Demethylase. *Science*, 318(5855):1469--1472.

[Gille et al., 2010] Gille, C., Bolling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hubner, K., Karlstadt, A., Ganeshan, R., Konig, M., Rother, K., Weidlich, M., Behre, J., and Holzhutter, H.-G. (2010). HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular Systems Biology*, 6.

[Goh et al., 2007] Goh, K.-I. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685--8690.

[Goto et al., 2002] Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*, 30(1):402--404.

[Güldener et al., 2006] Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.-W., and Stümpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Research*, 34(suppl 1):D436--D441.

[Hao et al., 2010] Hao, T., Ma, H. W., Zhao, X. M., and Goryanin, I. (2010). Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics*, 11(1):393+.

[Hardy and Singleton, 2009] Hardy, J. and Singleton, A. (2009). Genomewide Association Studies and Human Disease. *N Engl J Med*, 360(17):1759--1768.

[Hartwell et al., 1999] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47--C52.

[Heard et al., 2010] Heard, E., Tishkoff, S., Todd, J. A., Vidal, M., Wagner, G. A. P., Wang, J., Weigel, D., and Young, R. (2010). Ten years of genetics and genomics: what have we achieved and where are we heading? *Nature Reviews Genetics*, 11(10):723--733.

[Heid et al., 2010] Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, C. C., Speliotes, E. K., Mägi, R., Workalemahu, T., White, C. C., Bouatia-Naji, N., Harris, T. B., Berndt, S. I., Ingelsson, E., Willer, C. J., Weedon, M. N., Luan, J., Vedantam, S., Esko, T. o., Kilpeläinen, T. O., Kutalik, Z., Li, S., Monda, K. L., Dixon, A. L., Holmes, C. C., Kaplan, L. M., Liang, L., Min, J. L., Moffatt, M. F., Molony, C., Nicholson, G., Schadt, E. E., Zondervan, K. T., Feitosa, M. F., Ferreira, T., Lango, H., Weyant, R. J., Wheeler, E., Wood, A. R., MAGIC, Estrada, K., Goddard, M. E., Lettre, G., Mangino, M., Nyholt, D. R., Purcell, S., Vernon, A., Visscher, P. M., Yang, J., McCarroll, S. A., Nemesh, J., Voight, B. F., Absher, D., Amin, N., Aspelund, T., Coin, L., Glazer, N. L., Hayward, C., Heard-Costa, N. L., Hottenga, J.-J. J., Johansson, A., Johnson, T., Kaakinen, M., Kapur, K., Ketkar, S., Knowles, J. W., Kraft, P., Kraja, A. T., Lamina, C., Leitzmann, M. F., McKnight, B., Morris, A. P., Ong, K. K., Perry, J. R., Peters, M. J., Polasek, O., Prokopenko, I., Rayner, N. W., Ripatti, S., Rivadeneira, F., Robertson, N. R., Sanna, S., Sovio, U., Surakka, I., Teumer, A., van Wingerden, S., Vitart, V., Hua, J., Cavalcanti-Proença, C., Chines, P. S., Fisher, E., Kulzer, J. R., Lecoeur, C., Narisu, N., Sandholt, C., Scott, L. J., Silander, K., Stark, K., Tammesoo, M.-L. L., Teslovich, T. M., John, N., Watanabe, R. M., Welch, R., Chasman, D. I., Cooper, M. N., Jansson, J.-O. O., Kettunen, J., Lawrence, R. W., Pellikka, N., Perola, M., Vandenput, L., Alavere, H., Almgren, P., Atwood, L. D., Bennett, A. J., Biffar, R., Bonnycastle, L. L., Bornstein, S. R., Buchanan, T. A., Campbell, H., Day, I. N., Dei, M., Dörr, M., Elliott, P., Erdos, M. R., Eriksson, J. G., Freimer, N. B., Fu, M., Gaget, S., Geus, E. J., Gjesing, A. P., Grallert, H., Grässler, J., Groves, C. J., Guiducci, C., Hartikainen, A.-L. L., Hassanali, N., Havulinna, A. S., Herzig, K.-H. H., Hicks, A. A., Hui, J., Igl, W., Jousilahti, P., Jula, A., Kajantie, E., Kinnunen, L., Kolcic, I., Koskinen, S., Kovacs, P., Kroemer, H. K., Krzelj, V., Kuusisto, J., Kvaloy, K., Laitinen, J., Lantieri, O., Lathrop, M. M., Lokki, M.-L. L., Luben, R. N., Ludwig, B., McArdle, W. L., McCarthy, A., Morken, M. A., Nelis, M., Neville, M. J., Paré, G., Parker, A. N., Peden, J. F., Pichler,

I., Pietiläinen, K. H., Platou, C. G., Pouta, A., Ridderstråle, M., Samani, N. J., Saramies, J., Sinisalo, J., Smit, J. H., Strawbridge, R. J., Stringham, H. M., Swift, A. J., Teder-Laving, M., Thomson, B., Usala, G., van Meurs, J. B., van Ommen, G.-J. J., Vatin, V., Volpato, C. B., Wallaschofski, H., Walters, B. B., Widen, E., Wild, S. H., Willemsen, G., Witte, D. R., Zgaga, L., Zitting, P., Beilby, J. P., James, A. L., Kähönen, M., Lehtimäki, T., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Raitakari, O., Ridker, P. M., Stumvoll, M., Tönjes, A., Viikari, J., Balkau, B., Ben-Shlomo, Y., Bergman, R. N., Boeing, H., Davey, G., Ebrahim, S., Froguel, P., Hansen, T., Hengstenberg, C., Hveem, K., Isomaa, B., Jørgensen, T., Karpe, F., Khaw, K.-T. T., Laakso, M., Lawlor, D. A., Marre, M., Meitinger, T., Metspalu, A., Midthjell, K., Pedersen, O., Salomaa, V., Schwarz, P. E., Tuomi, T., Tuomilehto, J., Valle, T. T., Wareham, N. J., Arnold, A. M., Beckmann, J. S., Bergmann, S., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Collins, F. S., Eiriksdottir, G., Gudnason, V., Gyllensten, U., Hamsten, A., Hattersley, A. T., Hofman, A., Hu, F. B., Illig, T., Iribarren, C., Jarvelin, M.-R. R., Kao, L. H., Kaprio, J., Launer, L. J., Munroe, P. B., Oostra, B., Penninx, B. W., Pramstaller, P. P., Psaty, B. M., Quertermous, T., Rissanen, A., Rudan, I., Shuldiner, A. R., Soranzo, N., Spector, T. D., Syvanen, A.-C. C., Uda, M., Uitterlinden, A., Völzke, H., Vollenweider, P., Wilson, J. F., Witteman, J. C., Wright, A. F., Abecasis, G. R., Boehnke, M., Borecki, I. B., Deloukas, P., Frayling, T. M., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R. C., North, K. E., O'Connell, J. R., Peltonen, L., Schlessinger, D., Strachan, D. P., Hirschhorn, J. N., Assimes, T. L., Wichmann, H.-E. E., Thorsteinsdottir, U., van Duijn, C. M., Stefansson, K., Cupples, A. A., Loos, R. J., Barroso, I., McCarthy, M. I., Fox, C. S., Mohlke, K. L., and Lindgren, C. M. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics*, 42(11):949--960.

[Hermjakob et al., 2004] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic acids research*, 32(Database issue).

[Herold et al., 2009] Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics (Oxford, England)*, 25(24):3275--3281.

[Hindorff et al., 2009] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362--9367.

[Holden et al., 2008] Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, 24(23):2784--5.

[Holmans et al., 2009] Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A. R., Purcell, S. M., Sklar, P., Owen, M. J., O'Donovan, M. C., and Craddock, N. (2009). Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *The American Journal of Human Genetics*, 85(1):13--24.

[Hutz et al., 2008] Hutz, J. E., Kraja, A. T., McLeod, H. L., and Province, M. A. (2008). CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.*, 32(8):779--790.

[Ideker et al., 2002] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)*, 18 Suppl 1(suppl 1):S233--S240.

[Ideker and Sharan, 2008] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome research*, 18(4):644--652.

[Illig et al., 2010] Illig, T., Gieger, C., Zhai, G., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmuller, G., Kato, B. S., Mewes, H.-W., Meitinger, T., de Angelis, M. H., Kronenberg, F., Soranzo, N., Wichmann, H.-E., Spector, T. D., Adamski, J., and Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, 42(2):137--141.

[Ingelsson, 2010] Ingelsson, E. (2010). Large-scale genome-wide association studies consortia: blessing, burden, or necessity? *Circulation Cardiovascular Genetics*, 3(10):475--483.

[International HapMap Consortium, 2003] International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789--796.

[International HapMap Consortium, 2005] International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299--1320.

[Ioannidis et al., 2009] Ioannidis, J. P. A., Thomas, G., and Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, 10(5):318--329.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569--4574.

[Iwai and Naraba, 2005] Iwai, N. and Naraba, H. (2005). Polymorphisms in human pre-miRNAs. *Biochemical and biophysical research communications*, 331(4):1439--1444.

[Jia et al., 2011] Jia, P., Zheng, S., Long, J., Zheng, W., and Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, 27(1):95--102.

[Johansson et al., 2010] Johansson, A., Marroni, F., Hayward, C., Franklin, C., Kirichenko, A., Jonasson, I., Hicks, A., Vitart, V., Isaacs, A., Axenovich, T., Campbell, S., Floyd, J., Hastie, N., Knott, S., Lauc, G., Pichler, I., Rotim, K., Wild, S., Zorkoltseva, I., Wilson, J., Rudan, I., Campbell, H., Pattaro, C., Pramstaller, P., Oostra, B., Wright, A., van Duijn, C., Aulchenko, Y., Gyllensten, and Consortium., E. (2010). Linkage and genome-wide association analysis of obesity-related phenotypes: association of weight with the MGAT1 gene. *Obesity*, 18(4):803--808.

[Kamburov et al., 2011] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(suppl 1):D712--D717.

[Kandasamy et al., 2010] Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., Telikicherla, D., Navarro, J. D., Mathivanan, S., Pecquet, C., Gollapudi, S. K. K., Tattikota, S. G. G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H. K., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R., Ramachandra, Y. L., Rahiman, B. A., Prasad, K. S., Lin, J.-X. X., Houtman, J. C., Desiderio, S., Renauld, J.-C. C., Constantinescu, S. N., Ohara, O., Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G. D., Sander, C., Leonard, W. J., and Pandey, A. (2010). NetPath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):R3+.

[Kasowski et al., 2010] Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., and Snyder, M. (2010). Variation in transcription factor binding among humans. *Science (New York, N.Y.)*, 328(5975):232--235.

[Kathiresan et al., 2009] Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R., Zelenika, D., Kubalanza, K. A., Morken, M. A., Scott, L. J., Stringham, H. M., Galan, P., Swift, A. J., Kuusisto, J., Bergman, R. N., Sundvall, J., Laakso, M., Ferrucci, L., Scheet, P., Sanna, S., Uda, M., Yang, Q., Lunetta, K. L., Dupuis, J., de Bakker, P. I. W., O'Donnell, C. J., Chambers, J. C., Kooner, J. S., Hercberg, S., Meneton, P., Lakatta, E. G., Scuteri, A., Schlessinger, D., Tuomilehto, J., Collins, F. S., Groop, L., Altshuler, D., Collins, R., Lathrop, G. M., Melander, O., Salomaa, V., Peltonen, L., Orho-Melander, M., Ordovas, J. M., Boehnke, M., Abecasis, G. R., Mohlke, K. L., and Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*, 41(1):56--65.

[Kawaji et al., 2006] Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucl. Acids Res.*, 34(suppl_1):D632--636.

[Kelly et al., 2008] Kelly, T., Yang, W., Chen, C. S., Reynolds, K., and He, J. (2008). Global burden of obesity in 2005 and projections to 2030. *International Journal of Obesity*, aop(current).

[Klein et al., 2005] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):385--389.

[Klein et al., 2001] Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D. E., Rubin, D. L., Shafa, F., Stuart, J. M., and Altman, R. B. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *The pharmacogenomics journal*, 1(3):167--170.

[Kong et al., 2009] Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K. T., Jonasdottir,

A., Frigge, M. L., Gylfason, A., Olason, P. I., Gudjonsson, S. A., Sverrisson, S., Stacey, S. N., Sigurgeirsson, B., Benediktsdottir, K. R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J. H., Johannsson, O. T., Hreidarsson, A. B., Sigurdsson, G., Ferguson-Smith, A. C., Gudbjartsson, D. F., Thorsteinsdottir, U., and Stefansson, K. (2009). Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275):868--874.

[Kopelman, 2007] Kopelman, P. (2007). Health risks associated with overweight and obesity. *Obesity Reviews*, 8(8):13--17.

[Kraft and Raychaudhuri, 2009] Kraft, P. and Raychaudhuri, S. (2009). Complex diseases, complex genes: keeping pathways on the right track. *Epidemiology (Cambridge, Mass.)*, 20(4):508--511.

[Ku et al., 2010] Ku, C. S. S., Loy, E. Y. Y., Pawitan, Y., and Chia, K. S. S. (2010). The pursuit of genome-wide association studies: where are we now? *Journal of human genetics*, 55(4):195--206.

[Lage et al., 2007] Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309--316.

[Lao et al., 2008] Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., and Comas, D. (2008). Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241--1248.

[Levy et al., 2007] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254+.

[Lindgren et al., 2009] Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., Speliotes, E. K., Thorleifsson, G., Willer, C. J., Herrera, B. M., Jackson, A. U., Lim, N., Scheet, P., Soranzo, N., Amin, N., Aulchenko, Y. S., Chambers, J. C., Drong, A., Luan, J., Lyon, H. N., Rivadeneira, F., Sanna, S., Timpson, N. J., Zillikens, C. C., Hua, J., Almgren, P., Bandinelli, S., Bennett, A. J., Bergman, R. N., Bonnycastle, L. L., Bumpstead, S. J., Chanock, S. J., Cherkas, L., Chines, P., Coin, L., Cooper, C., Crawford, G., Doering, A., Dominiczak, A., Doney, A. S., Ebrahim, S., Elliott, P., Erdos, M. R., Estrada, K., Ferrucci, L., Fischer, G., Forouhi, N. G., Gieger, C., Grallert, H., Groves, C. J., Grundy, S., Guiducci, C., Hadley, D., Hamsten, A., Havulinna, A. S., Hofman, A., Holle, R., Holloway, J. W., Illig, T., Isomaa, B., Jacobs, L. C., Jameson, K., Jousilahti, P., Karpe, F., Kuusisto, J., Laitinen, J., Lathrop, M. M., Lawlor, D. A., Mangino, M., McArdle, W. L., Meitinger, T., Morken, M. A., Morris, A. P., Munroe, P., Narisu, N., Nordström, A., Nordström, P., Oostra, B. A., Palmer, C. N., Payne, F., Peden, J. F., Prokopenko, I., Renström, F., Ruokonen, A., Salomaa, V., Sandhu, M. S., Scott, L. J., Scuteri, A., Silander, K., Song, K., Yuan, X., Stringham, H. M., Swift, A. J., Tuomi, T., Uda, M.,

Vollenweider, P., Waeber, G., Wallace, C., Walters, B. B., Weedon, M. N., Wellcome Trust Case Control Consortium, Witteman, J. C., Zhang, C., Zhang, W., Caulfield, M. J., Collins, F. S., Smith, G. D., Day, I. N., Franks, P. W., Hattersley, A. T., Hu, F. B., Jarvelin, M.-R. R., Kong, A., Kooner, J. S., Laakso, M., Lakatta, E., Mooser, V., Morris, A. D., Peltonen, L., Samani, N. J., Spector, T. D., Strachan, D. P., Tanaka, T., Tuomilehto, J., Uitterlinden, A. G., van Duijn, C. M., Wareham, N. J., Watkins, H., Procardis Consortia, Waterworth, D. M., Boehnke, M., Deloukas, P., Groop, L., Hunter, D. J., Thorsteinsdottir, U., Schlessinger, D., Wichmann, H.-E. E., Frayling, T. M., Abecasis, G. R., Hirschhorn, J. N., Loos, R. J., Stefansson, K., Mohlke, K. L., Barroso, I., McCarthy, M. I., and Giant Consortium (2009). Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS genetics*, 5(6).

[Liu et al., 2010] Liu, Y., Guo, Y., Zhang, L., Pei, Y., Yu, N., Yu, P., Papasian, C., and Deng, H. (2010). Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. *Obesity*, 18(12):2339--2346.

[Ma and Goryanin, 2008] Ma, H. and Goryanin, I. (2008). Human metabolic network reconstruction and its impact on drug discovery and development. *Drug Discovery Today*, 13(9-10):402--408.

[Ma et al., 2007] Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology*, 3.

[Maher, 2008] Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18--21.

[Mailman et al., 2007] Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., and Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181--1186.

[Manolio, 2010] Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine*, 363(2):166--176.

[Manolio et al., 2008] Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation*, 118(5):1590--1605.

[Manolio et al., 2009] 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., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747--753.

[Mattick, 2004] Mattick, J. S. (2004). RNA regulation: a new genetics? *Nature Review Genetics*, 5(4):316--323.

[Mattick and Makunin, 2006] Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, 15(suppl 1):R17--R29.

[McCarthy, 2010] McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *The New England journal of medicine*, 363(24):2339--2350.

[McCarthy et al., 2009] McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., Perkins, D. O., Dickel, D. E., Kusenda, M., Krastoshevsky, O., Krause, V., Kumar, R. A., Grozeva, D., Malhotra, D., Walsh, T., Zackai, E. H., Kaplan, P., Ganesh, J., Krantz, I. D., Spinner, N. B., Roccanova, P., Bhandari, A., Pavon, K., Lakshmi, B., Leotta, A., Kendall, J., Lee, Y.-h., Vacic, V., Gary, S., Iakoucheva, L. M., Crow, T. J., Christian, S. L., Lieberman, J. A., Stroup, T. S., Lehtimaki, T., Puura, K., Haldeman-Englert, C., Pearl, J., Goodell, M., Willour, V. L., DeRosse, P., Steele, J., Kassem, L., Wolff, J., Chitkara, N., McMahon, F. J., Malhotra, A. K., Potash, J. B., Schulze, T. G., Nothen, M. M., Cichon, S., Rietschel, M., Leibenluft, E., Kustanovich, V., Lajonchere, C. M., Sutcliffe, J. S., Skuse, D., Gill, M., Gallagher, L., Mendell, N. R., Craddock, N., Owen, M. J., O'Donovan, M. C., Shaikh, T. H., Susser, E., DeLisi, L. E., Sullivan, P. F., Deutsch, C. K., Rapoport, J., Levy, D. L., King, M.-C., and Sebat, J. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics*, 41(11):1223--1227.

[Mcpherson et al., 2007] Mcpherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2007). A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science*, 316(5830):1488--1491.

[Medina et al., 2009] Medina, I., Montaner, D., Bonifaci, N., Pujana, M. A., Carbonell, J., Tarraga, J., Al-Shahrour, F., and Dopazo, J. (2009). Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res*, 37(Web Server issue):W340--4.

[Meyre et al., 2009] Meyre, D., Delplanque, J., Chèvre, J.-C. C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F., Proença, C., Gaget, S., Körner, A., Kovacs, P., Kiess, W., Tichet, J., Marre, M., Hartikainen, A.-L. L., Horber, F., Potoczna, N., Hercberg, S., Levy-Marchal, C., Pattou, F., Heude, B., Tauber, M., McCarthy, M. I., Blakemore, A. I., Montpetit, A., Polychronakos, C., Weill, J., Coin, L. J., Asher, J., Elliott, P., Järvelin, M.-R. R., Visvikis-Siest, S., Balkau, B., Sladek, R., Balding, D., Walley, A., Dina, C., and Froguel, P. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature genetics*, 41(2):157--159.

[Moffatt et al., 2007] Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S. A., Wong, K. C., Illig, T., Vogelberg, C., Weiland, S. K., von Mutius, E., Abecasis, G. R., Farrall, M., Gut, I. G., Lathrop, G. M., and Cookson, W. O. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, 448(7152):470--473.

[Mohlke et al., 2008] Mohlke, K., Boehnke, M., and Abecasis, G. (2008). Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Human Molecular Genetics*, 15(15):102--108.

[Mootha and Hirschhorn, 2010] Mootha, V. K. and Hirschhorn, J. N. (2010). Inborn variation in metabolism. *Nature Genetics*, 42(2):97--98.

[Ng et al., 2010] Ng, S. B., Nickerson, D. A., Bamshad, M. J., and Shendure, J. (2010). Massively parallel sequencing and rare disease. *Human Molecular Genetics*, 19(R2):R119--R124.

[Ng et al., 2009] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272--276.

[Nica et al., 2010] Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genet*, 6(4):e1000895+.

[Nicolae et al., 2010] Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet*, 6(4):e1000888+.

[Novembre et al., 2008] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98--101.

[O'Dushlaine et al., 2009] O'Dushlaine, C., Kenny, E., Heron, E. A., Segurado, R., Gill, M., Morris, D. W., and Corvin, A. (2009). The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, 25(20):2762--2763.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 27(1):29--34.

[Pagel et al., 2005] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W. W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, 21(6):832--834.

[Patil and Nielsen, 2005] Patil, K. R. and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685--2689.

[Pedersen et al., 2006] Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol*, 2(4):e33+.

[Peri et al., 2003] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R. R., Suresh, S., Ghosh,

N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10):2363--2371.

[Pers et al., 2011] Pers, T., Hansen, N. T., Lage, K., Koefoed, P., Dworzynski, P., Miller, M., Flint, T., Mellerup, E., Dam, H., Andreassen, O., Djurovic, S., Melle, I., Børglum, A., Werge, T., Purcell, S., Ferreira, M., Kouskoumvekaki, I., Workman, C., Hansen, T., Mors, O., and Brunak, S. (2011). Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genetic Epidemiology*.

[Petronis, 2010] Petronis, A. (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, 465(7299):721--727.

[Pfaffl et al., 2004] Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnology letters*, 26(6):509--515.

[Phizicky and Fields, 1995] Phizicky, E. M. and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94--123.

[Pomerantz et al., 2009] Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M., Yao, K., Kehoe, S. M., Lenz, H.-J., Haiman, C. A., Yan, C., Henderson, B. E., Frenkel, B., Barretina, J., Bass, A., Tabernero, J., Baselga, J., Regan, M. M., Manak, J. R., Shivdasani, R., Coetzee, G. A., and Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature Genetics*, 41(8):882--884.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904--909.

[Qin et al., 2010] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59--65.

[Quigley and Balmain, 2009] Quigley, D. and Balmain, A. (2009). Systems genetics analysis of cancer susceptibility: from mouse models to humans. *Nature reviews. Genetics*, 10(9):651--657.

[Ramensky et al., 2002] Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic acids research*, 30(17):3894--3900.

[Rapley et al., 2009] Rapley, E. A., Turnbull, C., Al Olama, A. A. A., Dermitzakis, E. T., Linger, R., Huddart, R. A., Renwick, A., Hughes, D., Hines, S., Seal, S., Morrison, J., Nsengimana, J., Deloukas, P., UK Testicular Cancer Collaboration, Rahman, N., Bishop, D. T., Easton, D. F., and Stratton, M. R. (2009). A genome-wide association study of testicular germ cell tumor. *Nature genetics*, 41(7):807--810.

[Raychaudhuri et al., 2009] Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C., International Schizophrenia Consortium, Purcell, S. M., Sklar, P., Scolnick, E. M., Xavier, R. J., Altshuler, D., and Daly, M. J. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS genetics*, 5(6):e1000534+.

[Reich and Lander, 2001] Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in genetics : TIG*, 17(9):502--510.

[Robinson, 2010] Robinson, R. (2010). Common disease, multiple rare (and distant) variants. *PLoS biology*, 8(1):e1000293+.

[Romero et al., 2005] Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6(1):R2+.

[Ruepp et al., 2008] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N. N., Stümpflen, V., and Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(Database issue):D646--650.

[Saccone et al., 2010] Saccone, S. F., Bolze, R., Thomas, P., Quan, J., Mehta, G., Deelman, E., Tischfield, J. A., and Rice, J. P. (2010). SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic acids research*.

[Safran et al., 2010] Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation*, 2010(0):baq020+.

[Salwinski et al., 2004] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic acids research*, 32(Database issue):D449--D451.

[Sanchez-Pulido and Andrade-Navarro, 2007] Sanchez-Pulido, L. and Andrade-Navarro, M. A. (2007). The FTO (fat mass and obesity associated) gene codes for a novel member of the non-heme dioxygenase superfamily. *BMC Biochemistry*, 8:23+.

[Saunders et al., 2007] Saunders, M. A., Liang, H., and Li, W. H. (2007). Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A*, 104(9):3300--3305.

[Saxena et al., 2010] Saxena, R., Hivert, M.-F. F., Langenberg, C., Tanaka, T., Pankow, J. S., Vollenweider, P., Lyssenko, V., Bouatia-Naji, N., Dupuis, J., Jackson, A. U., Kao, L. H., Li, M., Glazer, N. L., Manning, A. K., Luan, J., Stringham, H. M., Prokopenko, I., Johnson, T., Grarup, N., Boesgaard, T. W., Lecoeur, C., Shrader, P., O'Connell, J., Ingelsson, E., Couper, D. J., Rice, K., Song, K., Andreasen, C. H., Dina, C., Köttgen, A., Le Bacquer, O., Pattou, F., Taneera, J., Steinthorsdottir, V., Rybin, D., Ardlie, K., Sampson, M., Qi, L., van Hoek, M., Weedon, M. N., Aulchenko, Y. S., Voight, B. F., Grallert, H., Balkau, B., Bergman, R. N., Bielinski, S. J., Bonnefond, A., Bonnycastle, L. L., Borch-Johnsen, K., Böttcher, Y., Brunner, E., Buchanan, T. A., Bumpstead, S. J., Cavalcanti-Proença, C., Charpentier, G., Chen, Y.-D. I. D., Chines, P. S., Collins, F. S., Cornelis, M., J Crawford, G., Delplanque, J., Doney, A., Egan, J. M., Erdos, M. R., Firmann, M., Forouhi, N. G., Fox, C. S., Goodarzi, M. O., Graessler, J., Hingorani, A., Isomaa, B., Jørgensen, T., Kivimaki, M., Kovacs, P., Krohn, K., Kumari, M., Lauritzen, T., Lévy-Marchal, C., Mayor, V., McAteer, J. B., Meyre, D., Mitchell, B. D., Mohlke, K. L., Morken, M. A., Narisu, N., Palmer, C. N., Pakyz, R., Pascoe, L., Payne, F., Pearson, D., Rathmann, W., Sandbaek, A., Sayer, A. A. A., Scott, L. J., Sharp, S. J., Sijbrands, E., Singleton, A., Siscovick, D. S., Smith, N. L., Sparsø, T., Swift, A. J., Syddall, H., Thorleifsson, G., Tönjes, A., Tuomi, T., Tuomilehto, J., Valle, T. T., Waeber, G., Walley, A., Waterworth, D. M., Zeggini, E., Zhao, J. H. H., GIANT consortium, MAGIC investigators, Illig, T., Wichmann, H. E., Wilson, J. F., van Duijn, C., Hu, F. B., Morris, A. D., Frayling, T. M., Hattersley, A. T., Thorsteinsdottir, U., Stefansson, K., Nilsson, P., Syvänen, A.-C. C., Shuldiner, A. R., Walker, M., Bornstein, S. R., Schwarz, P., Williams, G. H., Nathan, D. M., Kuusisto, J., Laakso, M., Cooper, C., Marmot, M., Ferrucci, L., Mooser, V., Stumvoll, M., Loos, R. J., Altshuler, D., Psaty, B. M., Rotter, J. I., Boerwinkle, E., Hansen, T., Pedersen, O., Florez, J. C., McCarthy, M. I., Boehnke, M., Barroso, I., Sladek, R., Froguel, P., Meigs, J. B., Groop, L., Wareham, N. J., and Watanabe, R. M. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nature genetics*, 42(2):142--148.

[Scott et al., 2007] Scott, L. J. J., Mohlke, K. L. L., Bonnycastle, L. L. L., Willer, C. J. J., Li, Y., Duren, W. L. L., Erdos, M. R. R., Stringham, H. M. M., Chines, P. S. S., Jackson, A. U. U., Prokunina-Olsson, L., Ding, C.-J. J., Swift, A. J. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.-Y. Y., Conneely, K. N. N., Riebow, N. L. L., Sprau, A. G. G., Tong, M., White, P. P. P., Hetrick, K. N. N., Barnhart, M. W. W., Bark, C. W. W., Goldstein, J. L. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A. A., Watanabe, R. M. M., Valle, T. T. T., Kinnunen, L., Abecasis, G. R. R., Pugh, E. W. W., Doheny, K. F. F., Bergman, R. N. N., Tuomilehto, J., Collins, F. S. S., and Boehnke, M. (2007). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*.

[Scuteri et al., 2007] Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G. R. (2007). Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genetics*, 3(7):e115+.

[Segrè et al., 2010] Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., Altshuler, D., Consortium, D., and Investigators, M. (2010). Common Inherited Variation in Mi-

tochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genet*, 6(8):e1001058+.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498--2504.

[Shlomi et al., 2008] Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, 26(9):1003--1010.

[Siepel et al., 2005] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034--1050.

[Sørensen et al., 2010] Sørensen, T. I. A., Virtue, S., and Vidal-Puig, A. (2010). Obesity as a clinical and public health problem: Is there a need for a new definition based on lipotoxicity effects? *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*.

[Speliotes et al., 2010] 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., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., Wheeler, E., Wood, A. R., Ferreira, T., Weyant, R. J., Segre, A. V., Estrada, K., Liang, L., Nemesh, J., Park, J.-H., Gustafsson, S., Kilpelainen, T. O., Yang, J., Bouatia-Naji, N., Esko, T., Feitosa, M. F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A. V., Welch, R., Zhao, J. H., Aben, K. K., Absher, D. M., Amin, N., Dixon, A. L., Fisher, E., Glazer, N. L., Goddard, M. E., Heard-Costa, N. L., Hoesel, V., Hottenga, J.-J., Johansson, A., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M. F., Myers, R. H., Narisu, N., Perry, J. R. B., Peters, M. J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L. J., Timpson, N. J., Tyrer, J. P., van Wingerden, S., Watanabe, R. M., White, C. C., Wiklund, F., Barlassina, C., Chasman, D. I., Cooper, M. N., Jansson, J.-O., Lawrence, R. W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M. T. S., Almgren, P., Arnold, A. M., Aspelund, T., Atwood, L. D., Balkau, B., Balmforth, A. J., Bennett, A. J., Ben-Shlomo, Y., Bergman, R. N., Bergmann, S., Biebermann, H., Blakemore, A. I. F., Boes, T., Bonnycastle, L. L., Bornstein, S. R., Brown, M. J., Buchanan, T. A., Busonero, F., Campbell, H., Cappuccio, F. P., Cavalcanti-Proenca, C., Chen, Y.-D. I., Chen, C.-M., Chines, P. S., Clarke, R., Coin, L., Connell, J., Day, I. N. M., Heijer, M., Duan, J., Ebrahim, S., Elliott, P., Elosua, R., Eiriksdottir, G., Erdos, M. R., Eriksson, J. G., Facheris, M. F., Felix, S. B., Fischer-Posovszky, P., Folsom, A. R., Friedrich, N., Freimer, N. B., Fu, M., Gaget, S., Gejman, P. V., Geus, E. J. C., Gieger, C., Gjesing, A. P., Goel, A., Goyette, P., Grallert, H., Graszler, J., Greenawalt, D. M., Groves, C. J., Gudnason, V., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Hall, A. S., Havulinna, A. S., Hayward, C., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K. B., Jarick, I., Jewell, E., John, U., Jorgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L. M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J. W., Kolcic, I., Konig, I. R.,

Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaloy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L. J., Lecoeur, C., Lehtimaki, T., Lettre, G., Liu, J., Lokki, M.-L., Lorentzon, M., Luben, R. N., Ludwig, B., Manunta, P., Marek, D., Marre, M., Martin, N. G., McArdle, W. L., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G. W., Morken, M. A., Morris, A. P., Mulic, R., Ngwa, J. S., Nelis, M., Neville, M. J., Nyholt, D. R., O'Donnell, C. J., O'Rahilly, S., Ong, K. K., Oostra, B., Pare, G., Parker, A. N., Perola, M., Pichler, I., Pietilainen, K. H., Platou, C. G. P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N. W., Ridderstrale, M., Rief, W., Ruokonen, A., Robertson, N. R., Rzehak, P., Salomaa, V., Sanders, A. R., Sandhu, M. S., Sanna, S., Saramies, J., Savolainen, M. J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D. S., Smit, J. H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A. J., Tammesoo, M.-L., Tardif, J.-C., Teder-Laving, M., Teslovich, T. M., Thompson, J. R., Thomson, B., Tonjes, A., Tuomi, T., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C. I. G., Voight, B. F., Waite, L. L., Wallaschofski, H., Walters, G. B., Widen, E., Wiegand, S., Wild, S. H., Willemsen, G., Witte, D. R., Witteman, J. C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J. P., Farooqi, I. S., Hebebrand, J., Huikuri, H. V., James, A. L., Kahonen, M., Levinson, D. F., Macciardi, F., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Ridker, P. M., Stumvoll, M., Beckmann, J. S., Boeing, H., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Chanock, S. J., Collins, F. S., Cupples, L. A., Smith, G. D., Erdmann, J., Froguel, P., Gronberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T. B., Hattersley, A. T., Hayes, R. B., Heinrich, J., Hu, F. B., Hveem, K., Illig, T., Jarvelin, M.-R., Kaprio, J., Karpe, F., Khaw, K.-T., Kiemeney, L. A., Krude, H., Laakso, M., Lawlor, D. A., Metspalu, A., Munroe, P. B., Ouwehand, W. H., Pedersen, O., Penninx, B. W., Peters, A., Pramstaller, P. P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N. J., Schwarz, P. E. H., Shuldiner, A. R., Spector, T. D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T. T., Wabitsch, M., Waeber, G., Wareham, N. J., Watkins, H., Wilson, J. F., Wright, A. F., Zillikens, M. C., Chatterjee, N., McCarroll, S. A., Purcell, S., Schadt, E. E., Visscher (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11):937--948.

[Stark et al., 2006] Stark, C., Breitkreutz, B.-J. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535--D539.

[Stefansson et al., 2008] Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, A., Mattiasdottir, S., Blondal, T., Haraldsson, M., Magnusdottir, B. B., Giegling, I., Möller, H.-J. J., Hartmann, A., Shianna, K. V., Ge, D., Need, A. C., Crombie, C., Fraser, G., Walker, N., Lonnqvist, J., Suvisaari, J., Tuulio-Henriksson, A., Paunio, T., Toulopoulou, T., Bramon, E., Di Forti, M., Murray, R., Ruggeri, M., Vassos, E., Tosato, S., Walshe, M., Li, T., Vasilescu, C., Mühleisen, T. W., Wang, A. G., Ullum, H., Djurovic, S., Melle, I., Olesen, J., Kiemeney, L. A., Franke, B., GROUP, Sabatti, C., Freimer, N. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Andreassen, O. A., Ophoff, R. A., Georgi, A., Rietschel, M., Werge, T., Petursson, H., Goldstein, D. B., Nöthen, M. M., Peltonen,

L., Collier, D. A., St Clair, D., and Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210):232--236.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545--15550.

[Sulem et al., 2007] Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature genetics*, 39(12):1443--1452.

[Szklarczyk et al., 2011] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561--D568.

[Sørensen et al., 2006] Sørensen, T., Boutin, P., Taylor, M., Larsen, L., Verdich, C., Petersen, L., Holst, C., Echwald, S., Dina, C., Toubro, S., Petersen, M., Polak, J., Clement, K., Marinez, J., Langin, D., Oppert, J., Stich, V., Ia, Arner, P., Saris, W., Pedersen, O., Astrup, A., Froguel, P., and Consortium, N. (2006). Genetic polymorphisms and weight loss in obesity: a randomised trial of hypo-energetic high versus low-fat diets. *PLoS Clinical Trials*.

[Taylor et al., 2009] Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199--204.

[Teslovich et al., 2010] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Cho, Y. S., Go, M. J., Kim, Y. J., Lee, J.-Y. Y., Park, T., Kim, K., Sim, X., Ong, R. T., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Zhao, J. H., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y., Wright, A. F., Witteman, J. C., Wilson, J. F., Willemsen, G., Wichmann, H.-E. E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I.,

McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K., Lucas, G., Luben, R., Loos, R. J., Lokki, M.-L. L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., König, I. R., Khaw, K.-T. T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M.-R. R., Janssens, C. C., Ingelsson, E., Igl, W., Hovingh, G. K., Hottenga, J.-J. J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Döring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J., de Faire, U., Crawford, G., Collins, F. S., Chen, Y.-d. D., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, M. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E.-S. S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, H. A., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, A. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., and Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707--713.

[Thomas, 2010] Thomas, D. (2010). Gene-environment-wide association studies: emerging approaches. *Nature reviews. Genetics*, 11(4):259--272.

[Thomas et al., 2003] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9):2129--2141.

[Thorleifsson et al., 2008] Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., Jonsdottir, T., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Jonsson, F., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Lauritzen, T., Aben, K. K., Verbeek, A. L. M., Roeleveld, N., Kampman, E., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Becker, D. M., Gulcher, J., Kiemeney, L. A., Pedersen, O., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1):18--24.

[Torkamani et al., 2008] Torkamani, A., Topol, E. J., and Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265--272.

[Tung and Yeo, 2011] Tung, Y.-C. L. and Yeo, G. S. H. (2011). From GWAS to biology: lessons from FTO: From GWAS to biology: lessons from FTO. *Annals of the New York Academy of Sciences*, 1220(1):162--171.

[Turner et al., 2010] Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation*, 2010.

[Uhlén et al., 2005] Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K. A., Amini, B., Andersen, E., Andersson, A.-C. C., Angelidou, P., Asplund, A., Asplund, C., Berglund, L., Bergström, K., Brumer, H., Cerjan, D., Ekström, M., Elobeid, A., Eriksson, C., Fagerberg, L., Falk, R., Fall, J., Forsberg, M., Björklund, M. G. G., Gumbel, K., Halimi, A., Hallin, I., Hamsten, C., Hansson, M., Hedhammar, M., Hercules, G., Kampf, C., Larsson, K., Lindskog, M., Lodewyckx, W., Lund, J., Lundeberg, J., Magnusson, K., Malm, E., Nilsson, P., Odling, J., Oksvold, P., Olsson, I., Oster, E., Ottosson, J., Paavilainen, L., Persson, A., Rimini, R., Rockberg, J., Runeson, M., Sivertsson, A., Sköllermo, A., Steen, J., Stenvall, M., Sterky, F., Strömberg, S., Sundberg, M., Tegel, H., Tourle, S., Wahlund, E., Waldén, A., Wan, J., Wernérus, H., Westberg, J., Wester, K., Wrethagen, U., Xu, L. L. L., Hober, S., and Pontén, F. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics : MCP*, 4(12):1920--1932.

[van der Hoeven et al., 1994] van der Hoeven, F., Schimmang, T., Volkmann, A., Mattei, M., Kyewski, B., and Rther, U. (1994). Programmed cell death is affected in the novel mouse mutant Fused toes (Ft). *Development*, 120(9):2601--2607.

[Vandesompele et al., 2002] Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7).

[Walters et al., 2010] Walters, R. G., Jacquemont, S., Valsesia, A., de Smith, A. J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., Delobel, B., Stutzmann, F., El-Sayed Moustafa, J. S., Chevre, J. C., Lecoeur, C., Vatin, V., Bouquillon, S., Buxton, J. L., Boute, O., Holder-Espinasse, M., Cuisset, J. M., Lemaitre, M. P., Ambresin, A. E., Brioschi, A., Gaillard, M., Giusti, V., Fellmann, F., Ferrarini, A., Hadjikhani, N., Campion, D., Guilmatre, A., Goldenberg, A., Calmels, N., Mandel, J. L., Le Caignec, C., David, A., Isidor, B., Cordier, M. P., Dupuis-Girod, S., Labalme, A., Sanlaville, D., Beri-Dexheimer, M., Jonveaux, P., Leheup, B., Ounap, K., Bochukova, E. G., Henning, E., Keogh, J., Ellis, R. J., MacDermot, K. D., van Haelst, M. M., Vincent-Delorme, C., Plessis, G., Touraine, R., Philippe, A., Malan, V., Mathieu-Dramard, M., Chiesa, J., Blaumeiser, B., Kooy, R. F., Caiazzo, R., Pigeyre, M., Balkau, B., Sladek, R., Bergmann, S., Mooser, V., Waterworth, D., Reymond, A., Vollenweider, P., Waeber, G., Kurg, A., Palta, P., Esko, T., Metspalu, A., Nelis, M., Elliott, P., Hartikainen, A. L., McCarthy, M. I., Peltonen, L., Carlsson, L., Jacobson, P., Sjostrom, L., Huang, N., Hurles, M. E., O/'Rahilly, S., Farooqi, I. S., Mannik, K., Jarvelin, M. R., Pattou, F., Meyre, D., Walley, A. J., Coin, L. J. M., Blakemore, A. I. F., Froguel, P., and Beckmann, J. S. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, 463(7281):671--675.

[Wang et al., 2007] Wang, K., Li, M., and Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics*, 81(6):1278--1283.

[Wang et al., 2009] Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J. P., Russell, R. K., Sleiman, P. M., Imielinski, M., Glessner, J., and Hou, C. (2009). Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. *The American Journal of Human Genetics*.

[Wang et al., 2005] Wang, W. Y. S., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature reviews. Genetics*, 6(2):109--118.

[Washietl et al., 2005a] Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A., and Stadler, P. F. (2005a). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnology*, 23(11):1383--1390.

[Washietl et al., 2005b] Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005b). Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454--2459.

[Wellcome Trust Case Control Consortium, 2007] Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661--678.

[Wellcome Trust Case Control Consortium et al., 2010] Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking, L., Howard, E., Howard, P., Howson, J. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D. C., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R., Phillips, A., Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. A., Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St Clair, D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. A., Su, Z., Symmons, D. P., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. A., Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C., Hall, A. S., Hattersley, A. T., Hill, A. V., Mathew, C. G.,

Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. A., Samani, N. J., and Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713--720.

[Willer et al., 2008a] Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W.-M. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G. R. (2008a). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*, 40(2):161--169.

[Willer et al., 2008b] Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., Sanna, S., Scheet, P., Weedon, M. N., Wheeler, E., Zhao, J. H., Jacobs, L. C., Prokopenko, I., Soranzo, N., Tanaka, T., Timpson, N. J., Almgren, P., Bennett, A., Bergman, R. N., Bingham, S. A., Bonnycastle, L. L., Brown, M., Burtt, N. P., Chines, P., Coin, L., Collins, F. S., Connell, J. M., Cooper, C., Smith, G. D., Dennison, E. M., Deodhar, P., Elliott, P., Erdos, M. R., Estrada, K., Evans, D. M., Gianniny, L., Gieger, C., Gillson, C. J., Guiducci, C., Hackett, R., Hadley, D., Hall, A. S., Havulinna, A. S., Hebebrand, J., Hofman, A., Isomaa, B., Jacobs, K. B., Johnson, T., Jousilahti, P., Jovanovic, Z., Khaw, K.-T., Kraft, P., Kuokkanen, M., Kuusisto, J., Laitinen, J., Lakatta, E. G., Luan, J., Luben, R. N., Mangino, M., McArdle, W. L., Meitinger, T., Mulas, A., Munroe, P. B., Narisu, N., Ness, A. R., Northstone, K., O'Rahilly, S., Purmann, C., Rees, M. G., Ridderstråle, M., Ring, S. M., Rivadeneira, F., Ruokonen, A., Sandhu, M. S., Saramies, J., Scott, L. J., Scuteri, A., Silander, K., Sims, M. A., Song, K., Stephens, J., Stevens, S., Stringham, H. M., Tung, Y. C. L., Valle, T. T., Van Duijn, C. M., Vimaleswaran, K. S., Vollenweider, P., Waeber, G., Wallace, C., Watanabe, R. M., Waterworth, D. M., Watkins, N., Witteman, J. C. M., Zeggini, E., Zhai, G., Zillikens, M. C., Altshuler, D., Caulfield, M. J., Chanock, S. J., Farooqi, I. S., Ferrucci, L., Guralnik, J. M., Hattersley, A. T., Hu, F. B., Jarvelin, M.-R., Laakso, M., Mooser, V., Ong, K. K., Ouwehand, W. H., Salomaa, V., Samani, N. J., Spector, T. D., Tuomi, T., Tuomilehto, J., Uda, M., Uitterlinden, A. G., Wareham, N. J., Deloukas, P., Frayling, T. M., Groop, L. C., Hayes, R. B., Hunter, D. J., Mohlke, K. L., Peltonen, L., Schlessinger, D., Strachan, D. P., Wichmann, H.-E., McCarthy, M. I., Boehnke, M., Barroso, I., Abecasis, G. R., and Hirschhorn, J. N. (2008b). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, 41(1):25--34.

[Wu et al., 2008] Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T. P., and Hautaniemi, S. (2008). Integrated network analysis platform for protein-protein interactions. *Nature Methods*, 6(1):75--77.

[Yanovski and Yanovski, 2011] Yanovski, S. and Yanovski, J. (2011). Obesity prevalence in the United States--up, down, or sideways? *New England Journal of Medicine*, 11(364):987--989.

[Yon Rhee et al., 2008] Yon Rhee, S., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509--515.

[Yu et al., 2008] Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M. J. (2008). A navigator for human genome epidemiology. *Nature genetics*, 40(2):124--125.

[Zamar et al., 2009] Zamar, D., Tripp, B., Ellis, G., and Daley, D. (2009). Path: a tool to facilitate pathway-based genetic association analysis. *Bioinformatics (Oxford, England)*, 25(18):2444--2446.

[Zimmermann et al., 2009] Zimmermann, E., Kring, S., Berentzen, T., Holst, C., Pers, T., Hansen, T., Pedersen, O., Sørensen, T., and Jess, T. (2009). Fatness-associated FTO gene variant increases mortality independent of fatness--in cohorts of Danish men. *PLoS ONE*, 118(4).

Appendices

FTO analysis: Methods and materials used for the expression analysis

cDNA synthesis

RNA from the large mouse tissue panel was prepared by Zyagen (San Diego, CA, USA). RNA for the large human tissue panel was purchased from Clontech, Mountain View, Ca. The small human brain tissue panel was purchased from BioChain Institute, Hayward, CA. Aliquots of all human RNAs was treated with DNAse I and using DNAse treated RNA and three primer sets specific for 3 untranscribed chromosomal regions (chromosomes 7, 13 and 20, human RNA samples used in this study were tested negative for the presence of genomic DNA. Aliquots of 2.5 μg RNA were used as template for cDNA synthesis with an RNAseH-deficient reverse transcriptase derived from MoMLV (SuperScript) and a poly-dT primer. Aliquots of a subset of the RNAs were reverse transcribed both with a poly-dT and a random hexamere primer. cDNA from all samples of each cDNA panel was synthesized at the same time using the same mastermix to avoid technical variations.

Quantitative real-time PCR

All primers were designed using Oligo 6.0 software (Molecular Biology Insights, Cascade, CO) and ordered from TAG Copenhagen A/S, Denmark (Supplementary Material). Where possible, primers were designed to be intron spanning amplifying 150 to 400 bps. optimal annealing temperature for each primer set was determined by gradient RT-PCR (PTC-225, Bio-Rad, Hercules, CA) with cDNA prepared from Universal Human Reference RNA (Stratagene, La Jolla, CA) as template. Gel electrophoresis and melting curve analysis were used to verify that a single PCR product of the predicted size was generated. The product was subsequently isolated using the GENECLEAN II Kit (Qbiogene Inc., Irvine, CA) and serially diluted and used for generation of a standard curve. Using approximately 20 ng of each cDNA sample as template, Q-PCR was done in duplicates in an Opticon-2 thermocycler (Bio-Rad, Hercules, CA), using LightCycler-FastStart DNA Master SYBR Green I kit (Roche, Germany). All amplifications were performed in a total volume of 10 μl containing 3 mM MgCl2, 12% sucrose, and 1x reaction buffer included in the LightCycler kit. The PCR cycling profile consisted of a 10-min pre-denaturation step at 98∞C followed by 35 three-step cycles; at 98∞C for 10 s, at the optimized annealing temperature indicated in Table 1 and 2 for the specific gene for 20 s, and finally at 72∞C for 20 s. For quantitation, a Bestkeeper [Pfaffl et al., 2004] normalization factor was calculated using the most stable housekeeping genes. Normalized relative amount of RNA was calculated as described [Vandesompele et al., 2002].

In situ hybridization

In situ hybridization of the non-coding RNA was performed on 10-μm frozen tissue sections from adult (BalB/C) mouse brain. Sections were fixed in 4% paraformaldehyde and acetylated in acetic anhydride/triethanolamine, each followed by washes in PBS. Sections were then prehybridized in hybridization solution (50% formamide, 5x SSC, 0.5 mg/mL yeast tRNA, 1x Denhardt's solution) at 37∞C for 2 hours. 38 to 40 bps long oligonucleotide probes have been designed complementary to the putative non-coding RNA both on the sense and the antisense strand since the direction of transcription was not clear. A set of control probes has also been designed outside the sequence of the

candidate RNA since the borders of the gene was not known. The probes were purchased FITC-labelled at both ends (TAGC, Denmark). The probes on the same strand are either used in combination or separately. Ten-picomole probes were hybridized at 44∞C overnight. After post-hybridization washes in 50% formamide at 50∞C ,0.1x SSC at 54∞C and 0.5 %SSC at 54∞C, the *in situ* hybridization signals were detected using the tyramide signal amplification system (Perkin Elmer) according to the manufacturer's instructions. Slides were mounted in Prolong Gold containing DAPI (Invitrogen) and analyzed with an Olympus MVX10 microscope equipped with a CCD camera and Olympus CellP software.

Supplementary figures

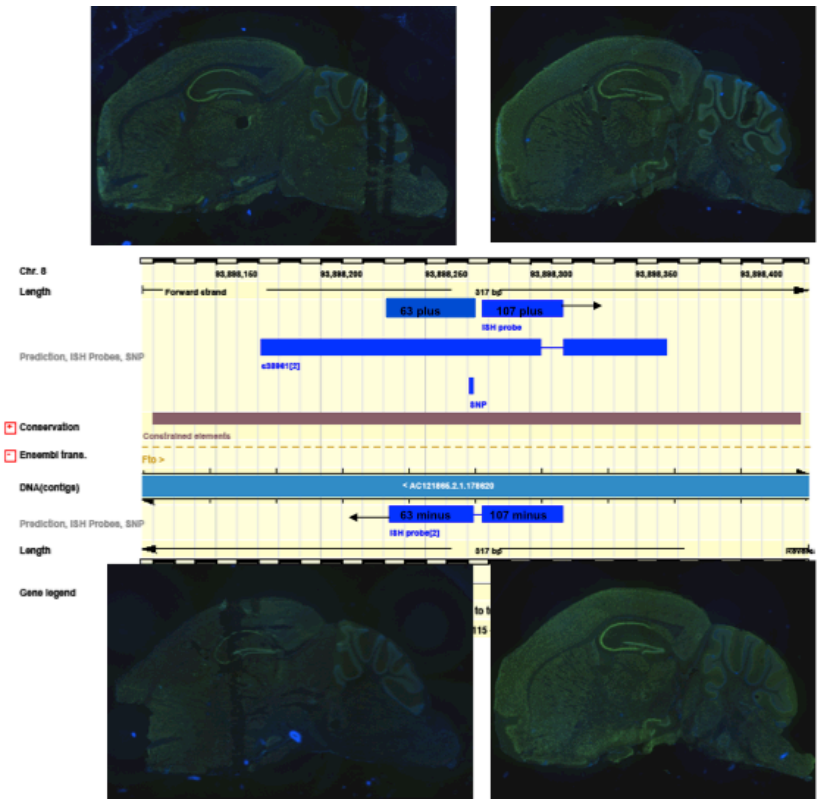


Figure 1: *In situ* analysis results of coronal mouse brain sections.

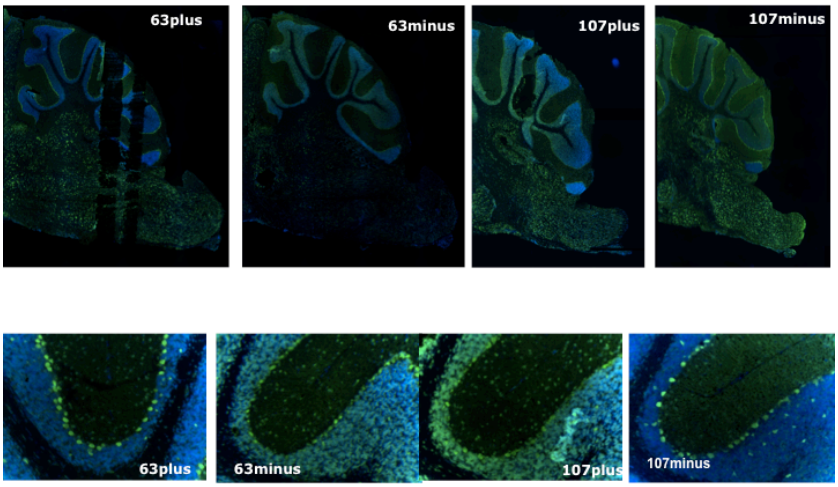


Figure 2: *In situ* analysis results of from the 4 different primers (2 for each strand).