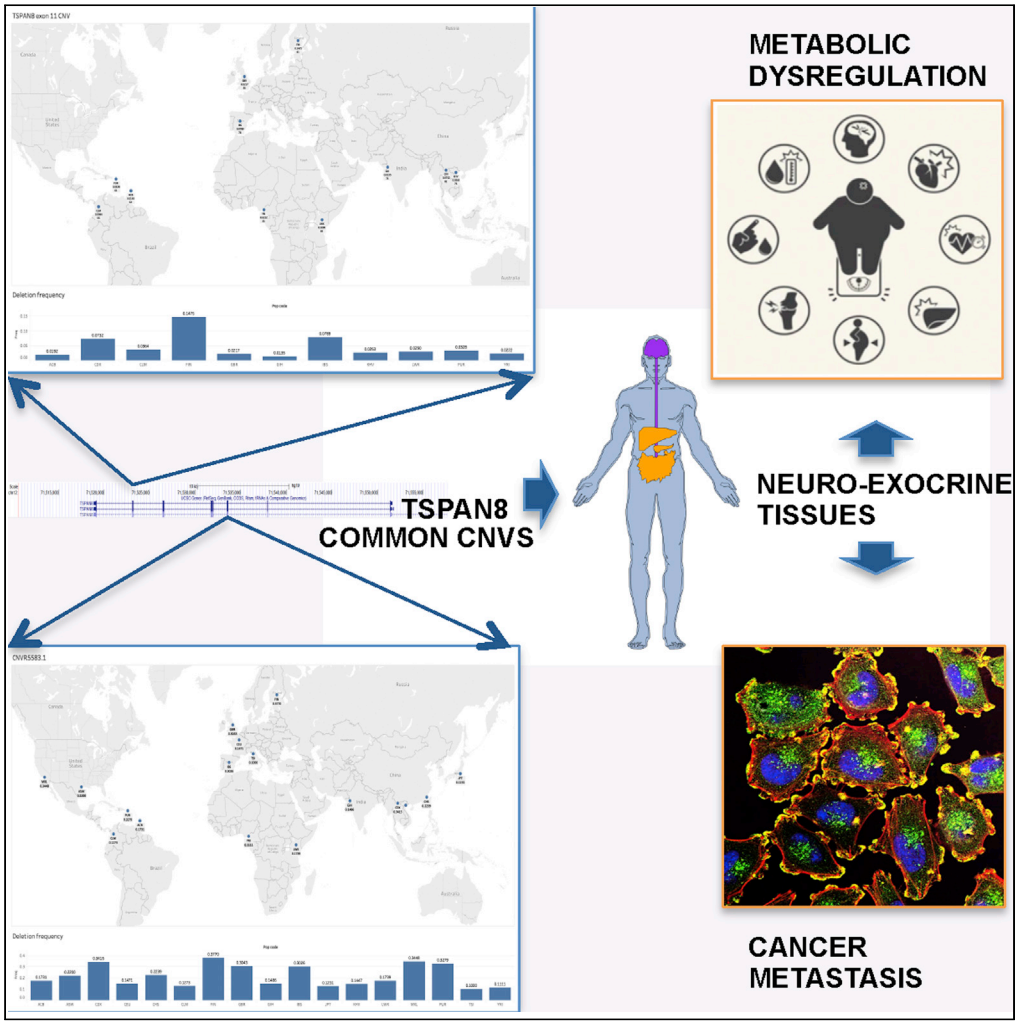
iScience

CellPress
OPEN ACCESS

Article

Signatures of TSPAN8 variants associated with human metabolic regulation and diseases



Tisham De, Angela Goncalves, Doug Speed, ..., Michael R. Johnson, Marjo-Riitta Jarvelin, Lachlan JM. Coin

tisham.de08@imperial.ac.uk

Highlights

We demonstrate neuro-exocrine axis for type 2 diabetes and metabolic regulation

Human induced pluripotent stem cells was successfully applied for disease modeling

We note germline CNV deletions are reversed to somatic amplifications

We characterizedgene variants associated with obesity with greater odds ratio than FTO SNPs

De et al., iScience 24, 102893
August 20, 2021
https://doi.org/10.1016/j.isci.2021.102893

iScience

CellPress
OPEN ACCESS

Article

Signatures of TSPAN8 variants associated with human metabolic regulation and diseases

Tisham De,[1,2,3,17,*] Angela Goncalves,[4,5] Doug Speed,[6,7,8] Philippe Froguel,[2,3,9] NFBC consortium, Daniel J. Gaffney,[4] Michael R. Johnson,[10,16] Marjo-Riitta Jarvelin,[11,12,13,14,16] and Lachlan JM. Coin[1,2,15,16]

SUMMARY

Here, with the example of common copy number variation (CNV) in the *TSPAN8* gene, we present an important piece of work in the field of CNV detection, that is, CNV association with complex human traits such as ^{1}H NMR metabolomic phenotypes and an example of functional characterization of CNVs among human induced pluripotent stem cells (HipSci). We report *TSPAN8* exon 11 (ENSE00003720745) as a pleiotropic locus associated with metabolomic regulation and show that its biology is associated with several metabolic diseases such as type 2 diabetes (T2D) and cancer. Our results further demonstrate the power of multivariate association models over univariate methods and define metabolomic signatures for variants in *TSPAN8*.

INTRODUCTION

In human genetics, the concept of common genetic variation in common diseases has been the central tenet of research for more than two decades. Landmark studies such as the Wellcome Trust Case Control Consortium (WTCCC) analysis of eight common diseases first reported a common CNV (CNVR5583.1, *TSPAN8* exon 7 deletion, ENSE00000871916) associated with type 2 diabetes (T2D) (Wellcome Trust Case Control Consortium et al., 2010). CNVR5583.1 was validated by polymerase chain reaction (PCR) and was found to have an allele frequency of 36% and 40% for cases and controls, respectively. One of the best tagging single-nucleotide polymorphisms (SNPs) for CNVR5583.1 was reported to be rs1705261 with r^2 = 0.998 with highest linkage disequilibrium (LD) among all SNPs. CNVR5583.1, a common exonic variant for controls (minor allele frequency [MAF] = 40%; highest CNV frequency among all WTCCC disease and control cohorts), has not been reported or rediscovered in any of the recent large-scale CNV discovery projects. These include the thousand genomes project (1KG) (n = 2,504), the gnomAD project (n = 141,456), and more recently the CNV analysis from UK biobank (UKBB) (Aguirre et al., 2019) (n = 472,228). Furthermore, well-established longitudinal studies such as the Northern Finland Birth Cohorts (Rantakallio, 1988; Järvelin et al., 1997) (NFBC) and UKBB (Bycroft et al., 2018) are powerful resources for uncovering the effect of common genetic variants on quantitative traits and lifestyle phenotypes such as socioeconomic status, medication, and diet.

Building on the theme of common genetic variants and their role in common diseases and by integrating insights from current important landmark human genetic resources, our study here exemplifies that common human genetic variation, in particular common CNVs in the *TSPAN8* gene, can play an important and common role in the pathogenesis of diabetes and cancer. Furthermore, these manifestations are most likely caused through metabolic dysregulation. Through in-depth gene expression analysis including from the human induced pluripotent stem cells project (HipSci) and PheWAS results for *TSPAN8*, *METTL7B* (a *trans* CNV-QTL for *TSPAN8*), and *NKX2-2* (a common transcription factor), we suggest that *TSPAN8*, *METTL7B* and *NKX2-2* are expressed in tandem in different tissues of the body in humans and in other species and are likely to be linked through molecular functions.

RESULTS

TSPAN8 CNVs

Here, we have reported the rediscovery of CNVR5583.1 in the 1KG next-generation sequence (NGS) data for multiple human populations including Finnish (FIN) and British (GBR) populations. Using cnvHitSeq (see STAR Methods), we report the CNV deletion frequency for CNVR5583.1 in FIN and GBR as 37.7% and 30%,

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

[2]Department of Genomics of Common Diseases, Imperial College London, London, UK

[3]Department of Metabolism, Imperial College London, London, UK

[4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

[5]German Cancer Research Center (DKFZ) and DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany

[6]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark

[7]Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Aarhus, Denmark

[8]Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

[9]Inserm UMR1283, CNRS UMR8199, European Genomic Institute for Diabetes (EGID), Université de Lille, Institut Pasteur de Lille, Lille University Hospital, Lille, France

[10]Department of Brain Sciences, Imperial College London, London, UK

[11]Centre for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland

[12]Unit of Primary Health Care and Medical Research Center, Oulu University Hospital, Oulu, Finland

[13]Department of Epidemiology and Biostatistics, Medical Research Council–Public Health England Centre for Environment and Health, Imperial College London, London, UK

[14]Biocenter Oulu, University of Oulu, Oulu, Finland

Continued

respectively (Figure 1 and Table S1A). Next, guided by the NGS derived CNV breakpoint information for common exonic CNVs in 1KG data and SNP tagging CNV information (LD) from the WTCCC 16K CNV study, we identified CNVs in *TSPAN8* exon 10 and TSPAN8 exon 11 in two Northern Finland population cohorts – NFBC 1986 (n = 4,060) and NFBC 1966 (n = 5,240). These two CNV regions lie close to the most significant SNP associated with T2D in *TSPAN8* (Figure 2). LD information from the 1KG data when juxtaposed on the UKBB PheWAS results (57 million TOPMed-imputed variants in 400,000 British white individuals) indicate high population specificity (Figure 3). This LD structure in *TSPAN8* was more pronounced than that in *PCSK9*. These results indicate that additional evolutionary, migratory, or human adaptation factors are likely to be involved at these genomic loci.

In NFBC 1986, genotyped on Illumina Cardio-Metabochip platform (Voight et al., 2012), we rediscovered CNVR5583.1 with an allele frequency of ~8% (Table S1B) tagged by rs1705261 with $r^2 = 0.942$ (Table S1C). In addition, a common CNV (MAF ~5% in NFBC, 1986 and 1KG FIN) overlapped with exon 11 in *TSPAN8* which was found to be in weak LD with CNVR5583.1. The LD results were SNP-CNV $r^2 = 0.623$ and CNV-CNV $r^2 = 0.68$ (Figures S1A and S1B). PennCNV results for NFBC and other population cohorts seem to indicate undercalling of CNVs in *TSPAN8* (Figure S2 and Table S1D).

We highlight that in the public release of CNV data from gnomAD consortium, three common CNV deletions with MAF >5% (MAFs 51%, 26% and 9%) were reported in *TSPAN8* but none of these were exonic or overlapped with CNVR5583.1 or *TSPAN8* exon 11 (Table S1E). However, we find that there are marked visual differences in sequencing depth coverage across *TSPAN8* exon 11 and exon 7 (CNVR5583.1), indicating the presence of structural variation in these regions (Table S1F). The 1KG CNV release reported no common CNVs within the *TSPAN8* gene (Table S1G). In the Memorial Sloan Kettering Cancer Center (MSKCC, url : https://www.mskcc.org/about.) portal for pan-cancer data (The Cancer Genome Atlas (TCGA) project data included), consisting of ~87,000 samples across 287 different cancer types, we observed that *TSPAN8* common germline deletions, including CNVR5583.1, are almost completely depleted (deletion allele frequency <0.01%). In contrast, in most cancer types where *TSPAN8* was found to be altered ~2% of the 87,823 patients (91,339 samples from 287 studies), most patient genomes had amplifications with allele frequency >5% (Figure S3). CNV analysis of the HipSci patient germline genomes and the donor-derived cell lines data indicated a similar pattern. We found *TSPAN8* CNV deletions with MAF ~5% in germline genomes (Figure S4), and this was reduced to allele frequency of <0.01% in the patient-derived induced pluripotent stem (iPS) cell lines.

Metabolomic signatures of TSPAN8 variants

Metabolomic signatures were obtained by applying univariate and multivariate approaches (Multiphen, see STAR Methods) using cnvHap-derived CNV genotypes. Across *TSPAN8* and within a window of one megabase around *TSPAN8*, the strongest CNV-metabolome association signal was discovered within *TSPAN8* exon 11 (chr12:71523134), closely followed by exon 10 and other nearby probes (Figures 4A and 4B). At chr12:71523134, on meta-analysis (inverse variance fixed effects) of 228 metabolic phenotypes in NFBC 1986 and NFBC 1966 (n = 9,190), we found the top metabolic phenotype to be **HDL_TG** (Triglycerides in high-density lipoprotein [HDL], p value = 0.00102, Table S2A iii). In our multivariate signature analysis, a signature consisting of several subclasses of HDL was found to be associated with multivariate joint signature with a p value of 0.00368, located at 12:71526593 (Figures 4C and S2A vi). Genome-wide univariate inflation factors for CNV **HDL_TG** associations were found to be 1.004 and 1.157 in NFBC 1986 and NFBC 1966, respectively (Table S2B). Using genotyping platform intensity measurement log-r ratio (LRR)–based association model (association independent of cnvHap genotypes or CNV calling), **HDL_TG** replicated in the meta-analysis of NFBC 1986 and NFBC 1966 with a p value of 0.0873 (Table S2C iii). In a separate British replication cohort (Whitehall), the strongest lipid association signal in the *TSPAN8* gene was observed for HDL lipid at 12:71526064, near exon 10 with univariate LRR p value of 5.02×10^{-6} (Table S2D i, Figure S5). Across all cohorts and association approaches, the strongest signal in *TSPAN8* was found at chr12:71523134 (exon 11) with a p value = 7.33×10^{-233} (Table S2C vi) with a metabolomic signature consisting of 27 metabolites. Influence of sex on association results is reported for Whitehall cohort in Table S2D. In NFBC cohorts, this is reported in Table S2H iii and in Figures S18–S23.

Furthermore, CNV at chr12:71523134 (MAF ~2%) exhibited strong pleiotropy with >50% (115/228) of the metabolites having a significant p value < 0.05 (Figure 5). In contrast, significant SNP association results at the same position (MAF = 43%) showed pleiotropy of only 2%. In addition, individuals with CNV deletion

[15]Department of Microbiology and Immunology, University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

[16]These authors contributed equally

[17]Lead contact

*Correspondence: tisham.de08@imperial.ac.uk

https://doi.org/10.1016/j.isci.2021.102893

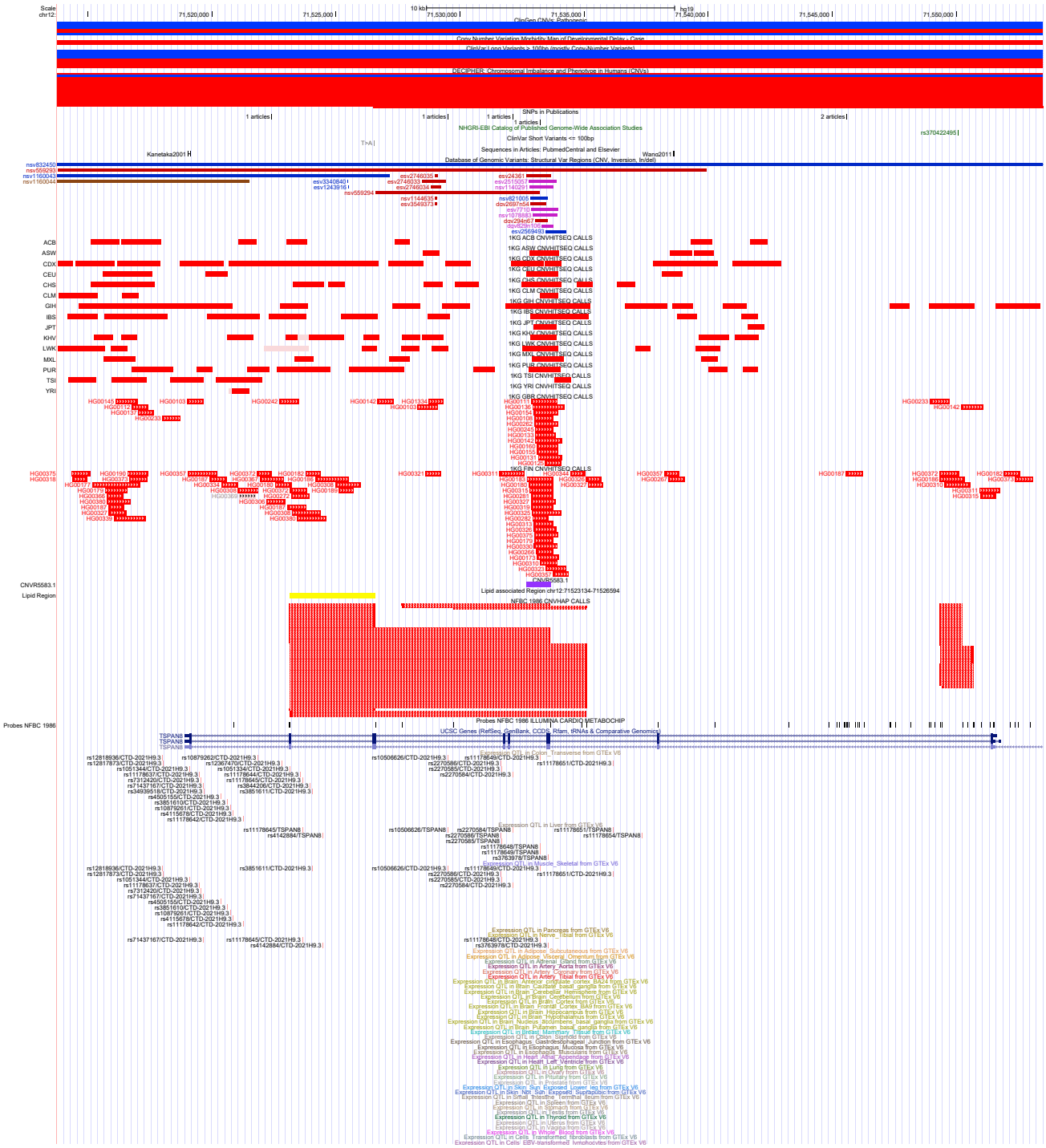


Figure 1. UCSC genome browser plot for CNV breakpoints in *TSPAN8* (hg19)

CNV breakpoints determined by cnvHitSeq for 1KG NGS data. Additional annotation for cnvHap-derived breakpoints in NFBC 1986, Illumina Cardio-Metabochip probe locations, and other publically available published CNVs breakpoints are also marked. The bottom section shows significant eQTL results from the GTeX project. Of note, significant eQTLs are tissue specific and are concentrated in the downstream regions near *TSPAN8* exon 11 and CNVR5583.1.

at chr12:71523134 had significantly higher levels of metabolite levels, particularly for low-density lipoprotein (LDL) and its subcategories (Table S2E). We found 61% (22/36) of LDL and its subclasses had a significant p value < 0.05 for higher metabolite levels.

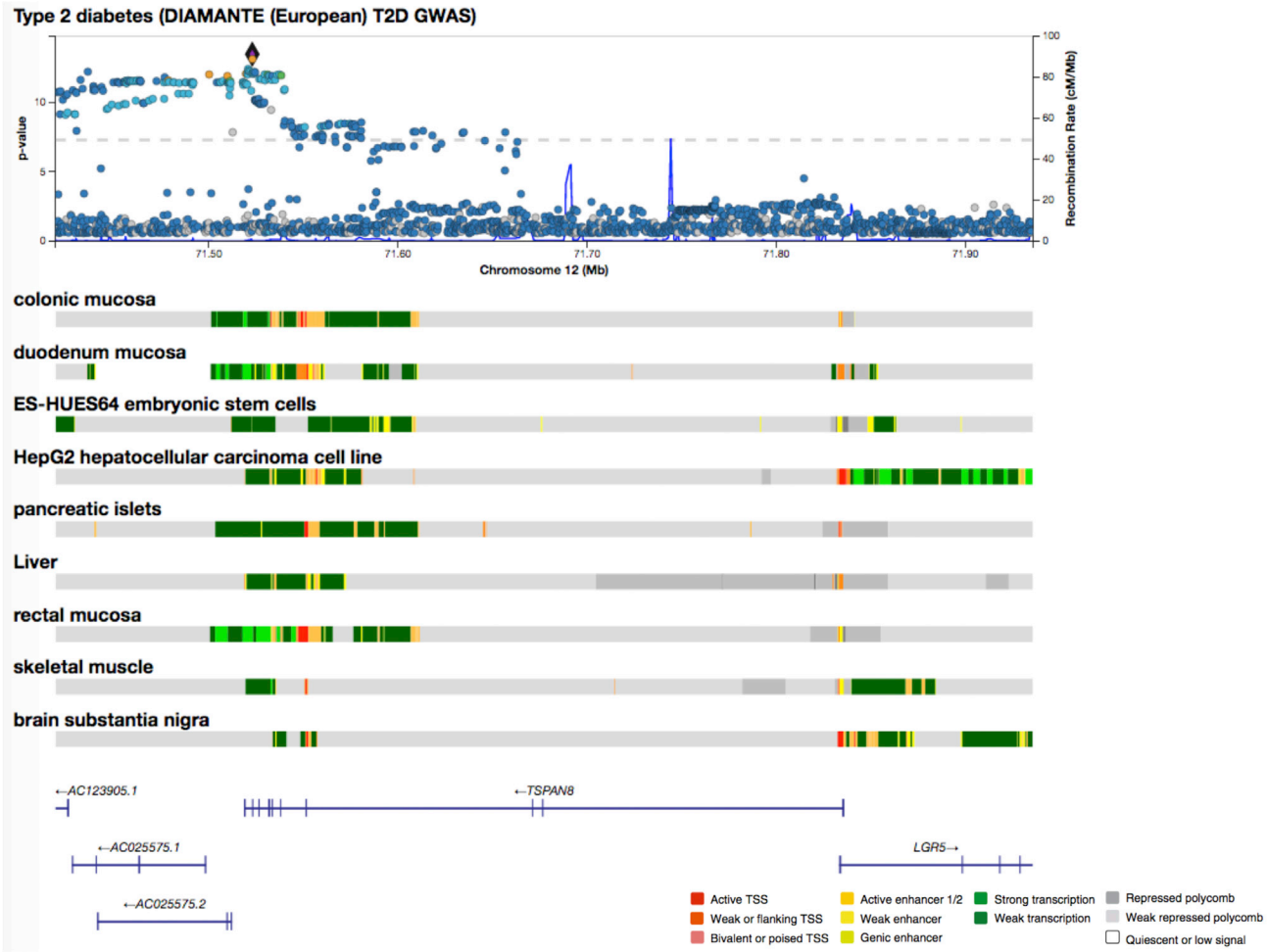


Figure 2. Regional Manhattan plot

SNP association results for *TSPAN8* in T2D DIAMANTE GWAS analysis (N ~1 million) obtained from the T2D knowledge portal (url: http://www.type2diabetesgenetics.org/). Schematic includes epigenetic annotations for transcriptional activity within *TSPAN8*. Of note, within a window of one megabase, the most significant SNP association locus for T2D outcome lies near *TSPAN8* exon 11 (rs1796330, chr12:71522953, p value = 3.20×10^{-14}).

We report all univariate and multivariate CNV genotype associations and validation results for all common and rare CNVs in the *TSPAN8* gene with 228 metabolomic phenotypes in NFBC 1986 and NFBC 1966 characterized by nuclear magnetic resonance (NMR) in Tables S2A, S2C, and S2D and their subsections. We highlight two metabolites of interest from previous genome-wide association study (GWAS) of human metabolome (Suhre et al., 2011; Shin et al., 2014) which lie near *TSPAN8* exon 10, namely **1)** ratio of 7-methylguanine to mannose (chr12:71524858, p value = 6.58×10^{-7}) and **2)** 3,4-dihydroxybutyrate (chr12:71526064, p value = 7.36×10^{-5} synonym: 3,4-dihydroxybutyric acid) (Table S2F).

Functional characterization and biology of *TSPAN8*

To understand the function of *TSPAN8* CNVs further, we carried out genome-wide *TSPAN8* CNV-QTL (germline CNVs association with iPS cell line gene expression) analysis in human induced pluripotent stem cells from the HipSci project (Kilpinen et al., 2017). One of the top hits included *METTL7B* (genome-wide rank 3, p value = 0.000195, Q value = 0.865, Table S3A). We observe that in addition to these results, it might be possible to assign a priori assumption for functional relationship between *TSPAN8* and *METTL7B* based on current knowledge of common transcriptional factors, gene and protein co-expression, and PheWAS analysis.

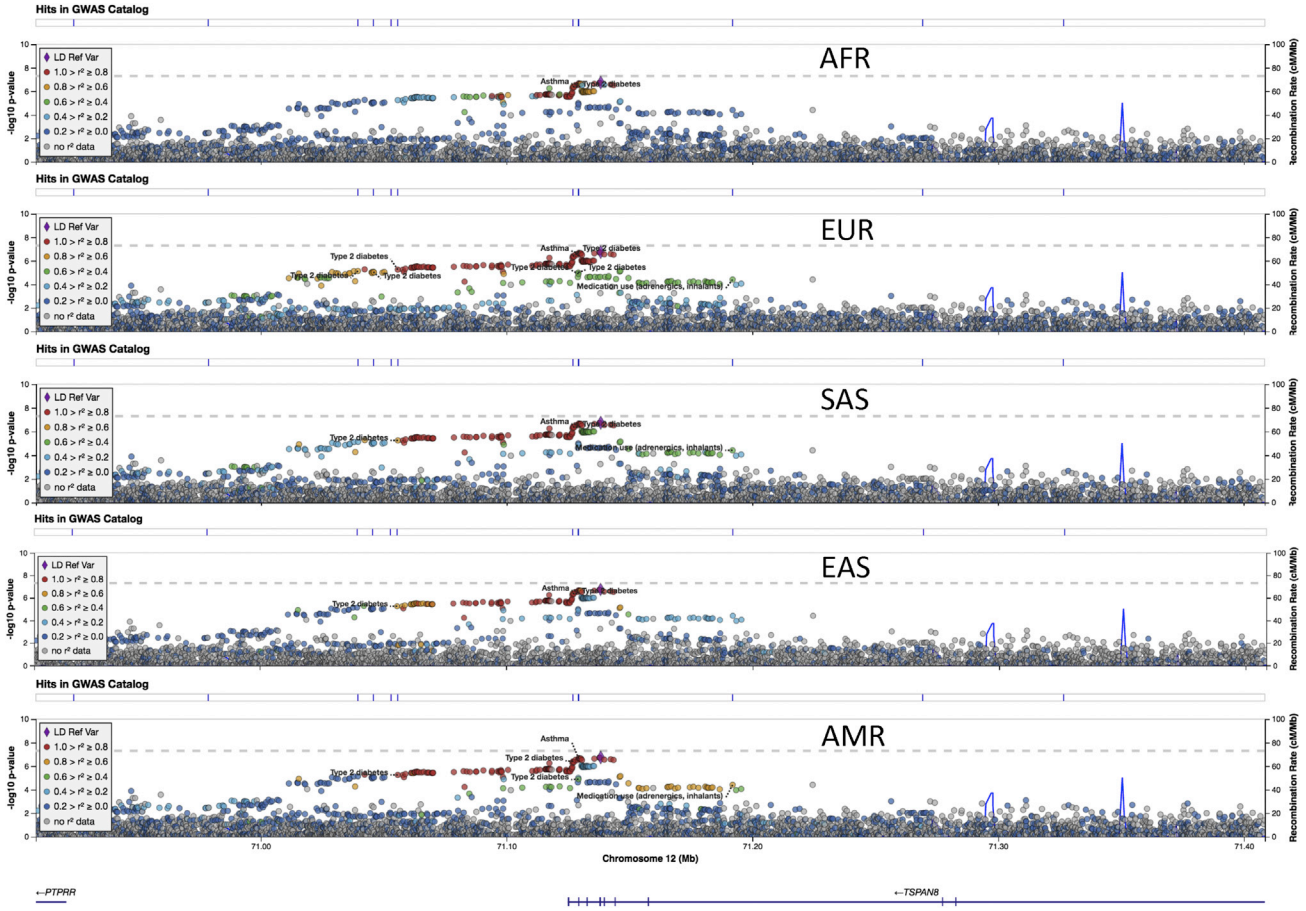


Figure 3. LD results from 1KG

Regional Manhattan plot for association of results of SNPs with **1400** EHR-derived broad PheWAS codes from the UK BioBank (n = 400,000 url: https://pheweb.org/). Results are next stratified by LD results for different population ancestries from the 1KG project. Of note, the LD structure seems to be correlated with human migration routes from Africa.

NKX2-2 is a common epigenetic regulator for *TSPAN8* and *METTL7B* active in adult human pancreatic islets (Figures 3D and S6, Table S3B). Main evidence of tissue-specific expression for these three genes included GTEx and Novartis whole-body gene expression maps (Boos et al., 2013) (Figures S7 and S8), significant tissue-specific SNP-eQTLs (GTEx, Table S3C), eQTL colocalization, and causality results reported by the T2D knowledge portal (Table S3D), single-cell gene expression databases (Figures 6A and 6B and Tables S3E, S3F, and S3G), and whole-body gene expression results in mouse (Tabula Muris, Figure S9), *Papio anubis*, *Ovis aries*, and *Xenopus laevis* (Table S3H). In developing *Xenopus laevis*, *NKX2-2* and *TSPAN8* expression seemed to have a positive correlation from Nieuwkoop and Faber (Nieuwkoop and Faber, 1994) (NF) stage 12 to 35-36, but from NF stage 35/36, they become negatively correlated, suggesting additional transcriptional repression factors in play. Such factors are unknown at the moment and warrant further investigation (Figure S10).

Thus, combining evidence from multiple databases and publications, we have demonstrated strong evidence of in-tandem RNA and protein co-expression for *NKX2-2*, *TSPAN8*, and *METTL7B*. We further hypothesize that these three genes together are likely to be functionally linked in energy homeostasis and glucose metabolism in the body through their coordinated action in tissues and organs related to insulin, hormones, other signaling molecules – **1)** production: pancreas, **2)** processing: liver and gut, **3)** regulation: brain and central nervous system, and 4) uptake: muscles/other organs.

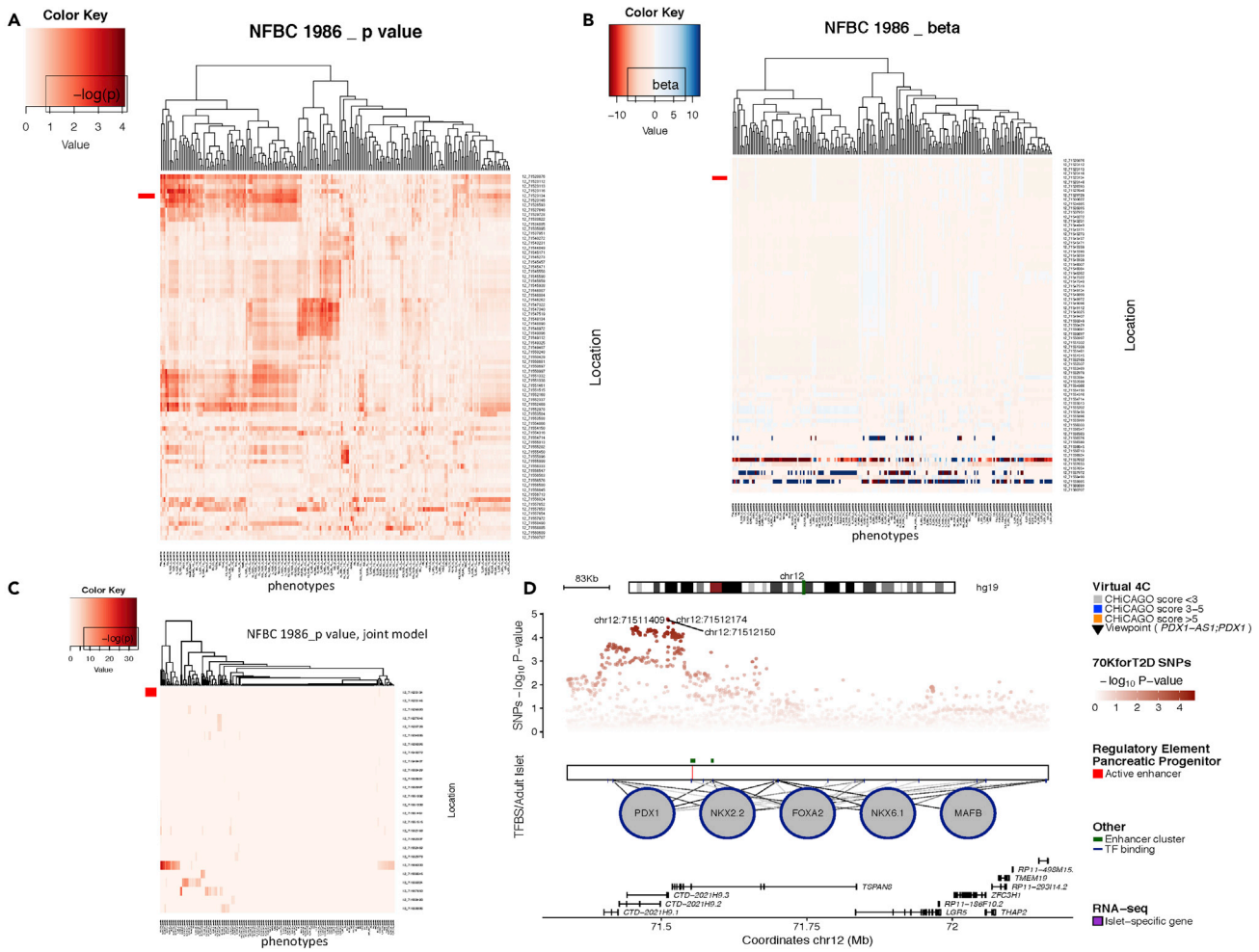


Figure 4. Metabolomic signatures in NFBC and epigenetic annotations

(A) Metabolomic signatures in NFBC 1986: Heatmap showing association results for *TSPAN8* CNV genotypes with metabolomic measurements in NFBC 1986. *TSPAN8* exon 11 deletion showing high degree of pleiotropy is marked in red.

(B) Same as 4A, but showing beta values from the association results from A.

(C) Same as 4A) but showing results for MultiPhen reverse regression with variable selection or multivariate metabolomic signature

(D) Epigenetic annotation and T2D trait association: Transcription-factor-binding sites (TFBS) in *TSPAN8* determined in the adult human pancreatic islets. Schematic includes 70KforT2D SNP association results for type 2 diabetes along with regulatory annotations for *TSPAN8* (negative HiC results). This schematic was generated by the Islet regulome browser, url: http://pasqualilab.upf.edu/app/isletregulome with following settings: *Chromatin maps*: Pancreatic progenitors (Cebola L, et al., 2015), Enhancer clustering algorithm: *Enhancer clusters (Pasquali L, et al 2014)*, transcription factors: *Adult islets – Tissue specific (Pasquali L, et al 2014)* and SNP association results**: *70KforT2D***.

The protein structure of *TSPAN8* remains unknown at the moment. However, the protein structure of *TSPAN28* which has strong sequence similarity with *TSPAN8* (e-value = 2×10^{-38}) has been determined and further shown to bind with cholesterol (Zimmerman et al., 2016) (Figure 6C). Here, we show that *TSPAN8* exon 11 deletion when mapped to *TSPAN28* might lie close to the closest cholesterol-binding pocket, thus opening up the possibility that *TSPAN8* exon 11 deletion might have an effect on cellular cholesterol transport, binding, and metabolism.

Disease, lifestyle, and exposome analysis of *TSPAN8*, *NKX2-2*, and *METTL7B*

Here, we report CNV association with disease outcomes for *TSPAN8*, *NKX2-2*, and *METTL7B* in previously characterized cohorts, namely DESIR for T2D (Vaxillaire et al., 2008), child obesity cohort (Walters et al., 2010), and adult obesity cohort (Walters et al., 2010) (Table S4A). LRR-based association p values for these

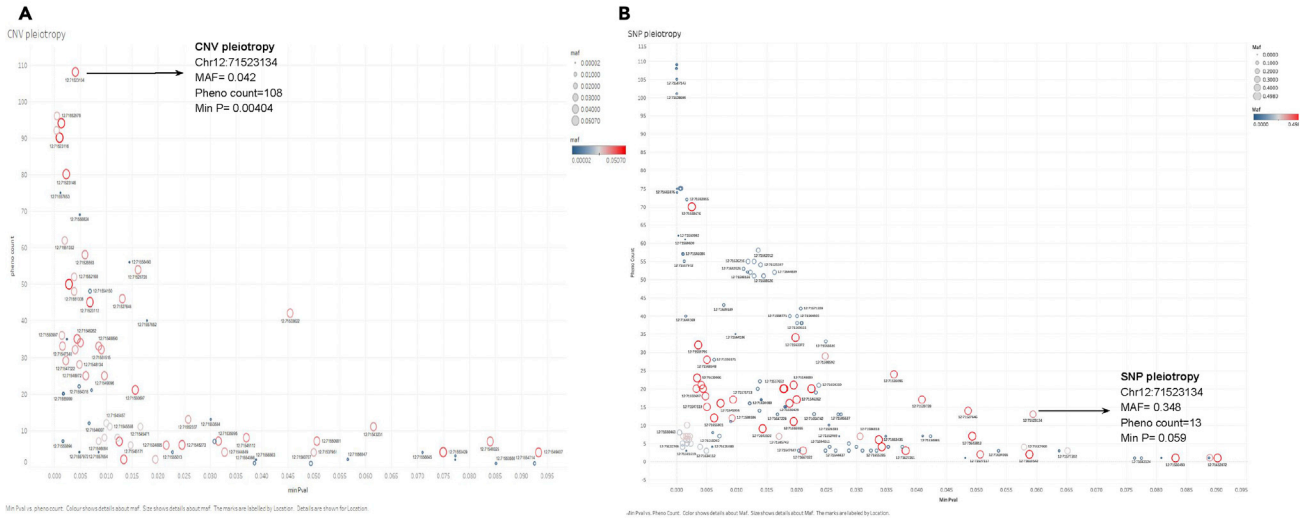


Figure 5. Pleiotropic nature of *TSPAN8* CNVs

Schematic showing pleiotropic nature of *TSPAN8* variants (A) CNVs and (B) SNPs in NFBC 1986. Every probe location in the *TSPAN8* gene is denoted by a circle and represents univariate association results for cnvHap genotypes with metabolomic measurements. Size and color of each circle correspond to CNV allele frequency. Higher allele frequency is denoted by both color (blue to red) and size of the circle. Y axis denotes phenotype count, i.e. the number of metabolomic phenotypes found associated with given CNV genotype at a p value threshold of ~0.05. X axis denotes the minimum p value observed at a given probe location.

cohorts for probes in *TSPAN8* exon 11 (chr12:7152314) were 0.0651, 0.00305, and 2.57×10^{-12}, respectively. These results included several genomic loci with significant CNV disease signals.

In NFBC 1986 and NFBC 1966, the top PheWAS traits associated with *TSPAN8*, *METTL7B*, and *NKX2-2* included insulin medication, glycemic traits, and smoking (Tables S4B and S4C), while common phenotypes from the FINNGEN project included T2D, diabetes with coma (both type 1 and 2), neurological complications, and several categories of glycemic traits (Table S4E). Some common lifestyle and exposome phenotypes from several public databases and GWAS catalogs included death at home, medication use, body mass index (BMI), hip circumference, waist hip ratio in females, and balding pattern in males (Figure S11; Tables S4F–S4K). A common metabolomic signature for CNVs in *TSPAN8*, *NKX2-2*, and *METTL7B* included **XXL_VLDL_L** (total lipids in chylomicrons and extremely large very-low-density lipoprotein [VLDL]) which was recently reported to be associated with increased all-cause mortality rate in humans (Deelen et al., 2019). In cancer biology, *TSPAN8* has been well characterized and is mainly implicated in cancer hallmarks related to metastasis and angiogenesis. By comparing and contrasting mutations including CNVs, single-nucleotide variants (SNVs), and gene expression, with a well-known classic tumor suppressor gene such as *PTEN* (Figure S3C), we propose that *TSPAN8* is likely to be an oncogene, involved with cancer metabolism through CNV amplifications and overexpression. Thus, since *TSPAN8* SNVs are quite sparse, *TSPAN8* CNVs are more likely to be cancer driver events. Overall survival estimates for patients with overexpression in TSPAN8 in many cancer types were also found to be significantly lower (Table S4L).

DISCUSSION

Finnish populations are known to be enriched for deleterious variants and hence are likely to be of added value for understanding molecular mechanisms of common disease such as T2D and metabolic disorders. Here, we have reported in-depth association analyses of CNVs using univariate and multivariate approaches in the *TSPAN8* gene with 228 circulating plasma metabolites in more than 9,300 Finnish individuals. In our analysis, we have highlighted some important aspects related to CNV detection and association approaches for cohorts with large sample sizes, commonly characterized through microarrays and NGS platforms. Some salient points included successful application of 'population aware' methods for CNV detection, application of probabilistic measures for CNV genotypes for improved CNV-phenotype associations, and leveraging intensity-based approaches for independent validation of CNV-phenotype associations. We demonstrate that CNVs are prevalent in germline, somatic, and iPS cell line genomes; however, their characterization, especially determining correct breakpoints and allele frequency, remains

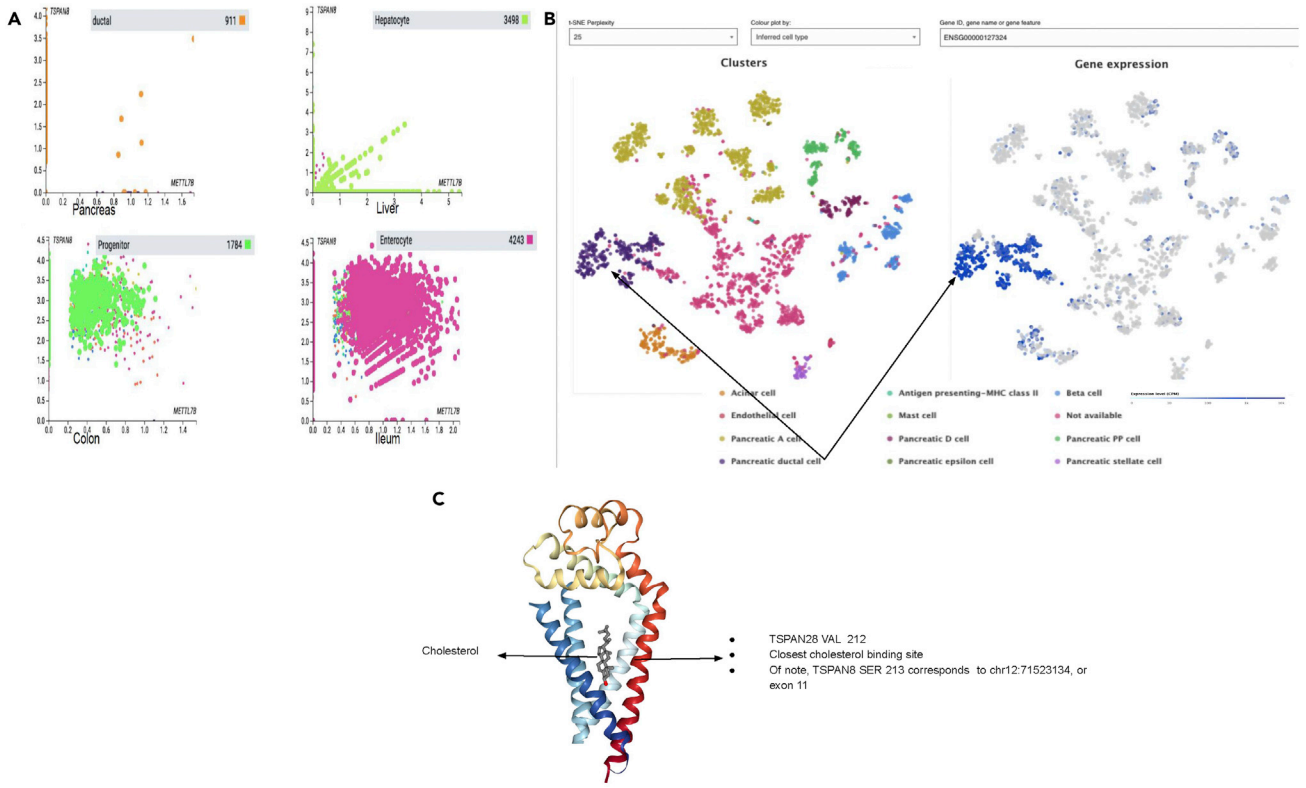


Figure 6. Functional characterization

(A) Single-cell gene expression results from the Human Cell Atlas. Correlation of single-cell gene expression data for *TSPAN8* and *METTL7B* in specific cell types, namely, ductal cells in the pancreas, hepatocytes in the liver, progenitor cells in the colon, and enterocytes in the ileum. Single-cell gene expression was obtained from the Human Cell Atlas project and analyzed and visualized through the cellxgene software (https://data.humancellatlas.org/analyze/portals/cellxgene).

(B) Single-cell gene expression data in the human pancreas. Published single-cell gene expression data showing tissue specificity of *TSPAN8* in the ductal cells of the human pancreas (Single Cell Gene Expression Atlas, Segerstolpe Å, Palasantza A et al. (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes.).

(C) 3-dimensional protein structure of ***TSPAN28***: 3D protein structure from protein databank showing a cholesterol-binding pocket in *TSPAN28* (PDB: 5TCX). The closest cholesterol binding site in TSPAN28 is located at VAL 212 (at a distance of 5.4 ångström (Å) from cholesterol). In TSPAN8, amino acid SER 213 maps to exon 11 or chr12:71523134.

challenging and underexplored. Importantly, approaches for delineating functional impact of bystander CNVs from real disease-causing pathogenic variants remain limited at the moment. New technologies based on CRISPR-based genome engineering, long read sequencing, and sequence-guided reanalysis of published GWAS microarray data sets are some promising leads to address these challenges.

Using a modest sample size of ~9,100, our multivariate approach of using all 228 metabolomic phenotypes in a single model allowed us to pinpoint the most significant and also perhaps the functionally important region in *TSPAN8* located within or near exon 11. In contrast, the multiethnic DIAMANTE meta-analysis for T2D (Mahajan et al., 2018) reported the most significant SNP in *TSPAN8* near exon 10 at 12:71522953 with a p value = 3.2×10^{-14} using a sample size of ~1 million (74,124 T2D cases and 824,006 controls). This result highlights the power of multivariate metabolomics analysis for genomics and highlights its relevance for rare variant analysis which usually requires extremely large sample sizes.

In the HipSci data, we rediscovered *TSPAN8* CNV deletions in iPS donor genomes with MAF ~5% and subsequent CNV-QTL analysis led to the discovery of *METTL7B*, as a potential new *trans* CNV-QTL for *TSPAN8*. Furthermore, *NKX2-2* was found to be a common transcription factor for these two genes active in pancreatic islets. The initial evidence from the iPS cell lines analysis is suggestive but weak owing to

nonsignificant Q value; however, several additional results from epigenetic and single-cell RNA-Seq data reinforced our hypothesis that *TSPAN8*, *METTL7B*, and *NKX2-2* are likely to be functionally linked in a very tissue-specific manner in humans and other species. This evidence enabled us to build a robust a priori hypothesis for *TSPAN8* and *METTL7B* and gives more weight to the HipSci results. An additional important result we would like to highlight is the possible involvement of *TSPAN8*, *METTL7B*, and *NKX2-2* in the PPAR pathway. To elucidate further, we note that *NKX2-2* has been experimentally shown to regulate *TSPAN8* and *PPARG* (Tables S3B and S2). In addition to this, the fact that *METTL7B* (Synonym: *ALDI*- Associated With Lipid Droplets 1 (Turró et al., 2006)) physically co-localizes with *PLIN1* on peroxisomes in the cell cytosol, suggests that the PPAR pathway might indeed be a common denominator (Figure S12).

Another observation we make here is that in addition to strong tissue-specific gene expression in humans and other species, *TSPAN8*, *NKX2-2*, and *METTL7B* further tend to be expressed in pairs but never together, i.e. all three genes being expressed in the same tissue is rarely seen. This phenomenon of pair exclusivity of gene expression was also indirectly reflected through Kaplan-Meier survival curve estimates for many cancer types (Table S4L). Some highlights of such patterns included pancreatic ductal carcinoma and kidney cancer, both of which have strong germline tissue expression. One exception to this pattern was cervical cancer where all three genes were found to be overexpressed. Cervical cancer has links to human papillomavirus (HPV), and thus, it might be a genuine outlier. However, we caution that these observations are preliminary and require further experimental investigation before any definitive conclusions can be made.

TSPAN8 and *METTL7B* both seem to have strong evidence of being involved with obesity. *TSPAN8*'s role in obesity is strongly indicated by knockout experiment in mice leading resistance to weight gain (−15.6%) and also corroborated by our novel association results for child obesity, where we found deletions in *TSPAN8* are protective against obesity with an odds ratio of 24.59 (p value = 1.268×10^{-6}) (Table SiL i). *METTL7B*'s role in obesity is a relatively new observation. Of importance is a recent GWAS analysis of childhood onset obesity (Riveros-McKay et al., 2019) where the authors reported *rs540249707* near *METTL7B* to have an odds ratio of 3.6 (95% confidence interval [CI] = 2.13–6.08, p value = 1.77×10^{-6}) which was higher than that of the FTO variant rs9928094 with an odds ratio of 1.44 (95% CI = 1.33–1.57, p value = 1.42×10^{-18}). Furthermore, *METTL7B* variants have also been reported as one of the top hits in GWAS of amphetamine response (Table S4H ii). Although discontinued, amphetamines are known to be prescribed as antiobesity medication (Ricca et al., 2009) with side effects related to increased alertness. Whether association of *TSPAN8* and *METTL7B* with obesity, central nervous system, or other traits is driven by independent molecular mechanisms or through common molecular pathways is left unvalidated at the moment.

Findings from single-cell data indicate that *TSPAN8* is mainly expressed in pancreatic ductal and acinar cells, thus highlighting its involvement of the neuro-exocrine axis for energy homeostasis and metabolism. Furthermore, *NKX2-2* and *TSPAN8* seem to be strongly coexpressed in similar regions of the human brain, in particular in the midbrain region around the hypothalamus, neural stem, and spinal cord. *METTL7B* on the other hand is overexpressed in glioblastoma (Figure S3B). Observations of several fold high expressions for *TSPAN8* and *NKX2-2* are replicated in the UK Brain Expression Study (Ramasamy et al., 2014) (Figure S13) and were also reflected through results from MetaXcan, eCAVIAR, and COLOC analysis for *TSPAN8* (Table S3D). These observations for *TSPAN8* and *NKX2-2* suggest genetic links in the neuro-exocrine axis for energy and metabolic homeostasis in humans. Our neurological observations are further intriguing owing to an earlier reported association of *TSPAN8* SNPs in exon 10 with 3,4-dihydroxybutyrate (synonym: 3,4-dihydroxybutyric acid). Butyrate has hormone-like properties and can induce enhanced secretion of glucagon and insulin (Gao et al., 2009) in the pancreas and has known beneficial effects on intestinal homeostasis for energy metabolism via the gut-brain axis (Li et al., 2018). Importantly, 3,4-dihydroxybutyric acid is known to be linked to satiety (Shimizu et al., 1984; Minami et al., 1988) and with ultra-rare succinic semialdehyde dehydrogenase deficiency (SSADH).

Using principles similar to reverse genetics, through PheWAS and phenotypic trait analysis, we further strengthen our metabolomic and gene expression findings. One such example is a common metabolomic signature for *TSPAN8*, *METTL7B*, and *NKX2-2* CNVs, **XXL_VLDL_L**, which was recently found to be associated with all-cause mortality (Deelen et al., 2019). The mortality risk factor is further corroborated by strong PheWAS signal for traits related to death at home in UK Biobank results which were common for all three

genes and had the most significant p value = 1.87×10^{-33} for *TSPAN8* (Table S4i). Some of the other phenotypic traits of interest included high medication use, diabetes with neurological complications, several categories of glycemic traits, BMI, hip circumference, and fat mass.

Our observations from current scientific literature indicate that all common germline CNV deletions (at least 5 CNVs with MAF >5%) in *TSPAN8* are nearly depleted in almost all somatic cancer genomes. The fact that they are also depleted in iPS cell line genomes postulates that *TSPAN8* CNVs are likely to be under unknown somatic evolutionary forces. In contrast, genes such as *GSTM1* or *RHD* which also harbor common germline CNV deletions with MAF >30% seem to retain CNV deletions during their somatic evolution (data not presented). This phenomenon indicates that human germline genomes might have inbuilt safety mechanisms or harbor tumor-suppressive variants, in order to provide inherent protection against uncontrolled cell proliferation or cancer. Similar to *TSPAN8* CNV deletions, one might expect such tumor-suppressive events to be present as 'common variants' in various human populations.

Of note, germline metabolomic signatures of *TSPAN8* and its associated genes can shed light on cancer metabolism, which can be exploited for diagnostic, therapeutic, or palliative interventions. One such possibility which warrants further investigation is our observation that *TSPAN8* and *METTL7B* (active but with weak expression) are expressed with high specificity in triple-negative breast cancer (Figure S14).

To conclude, our results robustly demonstrate the strong pleiotropic effects of *TSPAN8*, *METTL7B*, and *NKX2-2* on a wide range of human phenotypes, suggesting common molecular mechanisms and biological pathways, which opens up possibilities for diagnostic and therapeutic approaches for metabolic diseases.

Limitations of the study

There are several limitations in our study. First, we were restricted to 228 metabolites (mostly lipids, lipoproteins, and fatty acids) measured and provided by Nightingale Healthcare Limited (https://nightingalehealth.com/). In reality, the human metabolome is quite large and complex (N >>228). Hence, the real effect of *TSPAN8* variants on other categories of metabolite classes remains unknown. In addition, the genotyping platforms used to detect CNVs in our study are not sufficiently dense to map CNVs to high resolution. The fact that the cnvHap algorithm leverages probe-by-probe reclustering of LRR and BAF values to estimate CNV genotypes alleviates this problem to some extent.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENT MODEL AND SUBJECT DETAILS
  - Study cohorts
  - Metabolomic measurements in NFBC 1986 and NFBC 1966
  - Lipid measurements in WH-II
- METHODS DETAILS
  - CNV analysis
  - cnvHap: Normalization and quality control
  - CNV predictions using cnvHap
  - CNV segmentation
  - Expected CNV genotypes
  - Association analysis
  - Intensity-based validation of CNV association
  - Multiple testing
  - Linkage disequilibrium
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102893.

ACKNOWLEDGMENTS

We would like to thank Luise Cederkvist Kristiansen for her preliminary work and analysis of NFBC cohorts. We would like to thank Mr Matthias Heger and Mrs Pat Murphy for administraive help and guidance. Dr Tisham De was supported by Imperial College London School of Public Health PhD scholarship from 2011 to 2014.

AUTHOR CONTRIBUTIONS

T.D., L.J.M.C., M.R.J., and M-.R.J. were involved in study design, performed analysis, and wrote the manuscript. **A.G. and D.G.** carried out analysis of HipSci data and helped in writing the manuscript. **D.S.** advised on statistical inference and helped in writing the manuscript. **P.F.** provided access and advised on obesity and type 2 diabetes cohorts.

DECLARATION OF INTEREST

No external or financial interests to be declared.

Received: January 8, 2021
Revised: June 18, 2021
Accepted: July 20, 2021
Published: August 20, 2021

REFERENCES

Aguirre, M., Rivas, M.A., and Priest, J. (2019). 'Phenome-wide burden of copy-number variation in the UK biobank'. Am. J. Hum. Genet. *105*, 373–383.

Bellos, E., Johnson, M.R., and Coin, L.J.M. (2012). 'cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data'. Genome Biol. *13*, R120.

Boos, J.A., Kirk, D.W., Piccolotto, M.-L., Zuercher, W., Gfeller, S., Neuner, P., Dattler, A., Wishart, W.L., Von Arx, F., Beverly, M., et al. (2013). Whole-body scanning PCR; a highly sensitive method to study the biodistribution of mRNAs, noncoding RNAs and therapeutic oligonucleotides. Nucleic Acids Res. *41*, e145.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Coin, L.J.M., Asher, J.E., Walters, R.G., Moustafa, J., De Smith, A.J., Sladek, R., Balding, D.J., Froguel, F., and Blakemore, A.I.F. (2010). cnvHap: an integrative population and haplotype–based multiplatform model of SNPs and CNVs. Nature methods *7*, 541–546.

Deelen, J., Kettunen, J., Fischer, K., van der Spek, A., Trompet, S., Kastenmüller, G., Boyd, A., Zierer, J., van den Akker, E.B., Ala-Korpela, M., et al. (2019). A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. Nat. Commun. *10*, 3346.

Eleftherohorinou, H., Andersson-Assarsson, J.C., Walters, R.G., El-Sayed Moustafa, J.S., Coin, L., Jacobson, P., Carlsson, L.M.S., Blakemore, A.I.F., Froguel, P., Walley, A.J., et al. (2011). famCNV: copy number variant association for quantitative traits in families. Bioinformatics *27*, 1873–1875.

Gao, Z., Yin, J., Zhang, J., Ward, R.E., Martin, R.J., Lefevre, M., Cefalu, W.T., and Ye, J. (2009). Butyrate improves insulin sensitivity and increases energy expenditure in mice. Diabetes *58*, 1509–1517.

Järvelin, M.R., Elliott, P., Kleinschmidt, I., Martuzzi, M., Grundy, C., Hartikainen, A.L., and Rantakallio, P. (1997). Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986. Paediatric perinatal Epidemiol. *11*, 298–312.

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P., Culley, O.J., et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. Nature *546*, 370–375.

Li, Z., Yi, C.-X., Katiraei, S., Kooijman, S., Zhou, E., Chung, C.K., Gao, Y., van den Heuvel, J.K., Meijer, O.C., Berbée, J.F.P., et al. (2018). Butyrate reduces appetite and activates brown adipose tissue via the gut-brain neural circuit. Gut *67*, 1269–1279.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.

Minami, T., Oomura, Y., Nabekura, J., and Fukuda, A. (1988). Direct effects of 3,4-dihydroxybutanoic acid gamma-lactone and 2,4,5-trihydroxypentanoic acid gamma-lactone on lateral and ventromedial hypothalamic neurons. Brain Res. *462*, 258–264.

Nieuwkoop, P.D., and Faber, J. (1994). Normal table of Xenopus laevis (Daudin) garland publishing. New York, 252.

Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am. J. Hum. Genet. *74*, 765–769.

O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C.F., Elliott, P., Jarvelin, M.-R., and Coin, L.J.M. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS one *7*, e34861.

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., UK Brain Expression Consortium, North American Brain Expression Consortium, Coin, L., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. *17*, 1418–1428.

Rantakallio, P. (1988). The longitudinal study of the northern Finland birth cohort of 1966. Paediatric Perinatal Epidemiol. *2*, 59–88.

Ricca, V., Castellini, G., Mannucci, E., Monami, M., Ravaldi, C., Gorini Amedei, S., Lo Sauro, C., Rotella, C.M., and Faravelli, C. (2009). Amphetamine derivatives and obesity. Appetite *52*, 405–409.

Riveros-McKay, F., Mistry, V., Bounds, R., Hendricks, A., Keogh, J.M., Thomas, H., Henning, E., Corbin, L.J., Understanding Society Scientific Group, O'Rahilly, S., et al. (2019). 'Genetic architecture of human thinness compared to severe obesity'. PLoS Genet. *15*, e1007603.

Shimizu, N., Oomura, Y., and Sakata, T. (1984). Modulation of feeding by endogenous sugar acids acting as hunger or satiety factors. Am. J. Physiol. *246*, R542–R550.

Shin, S.-Y., Fauman, E.B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. Nat. Genet. *46*, 543–550.

Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., CARDIoGRAM, Deloukas, P., Erdmann, J., et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54–60.

Turró, S., Ingelmo-Torres, M., Estanyol, J.M., Tebar, F., Fernández, M.A., Albor, C.V., Gaus, K., Grewal, T., Enrich, C., Pol, A., et al. (2006). Identification and characterization of associated with lipid droplet protein 1: a novel membrane-associated protein that resides on hepatic lipid droplets. Traffic *7*, 1254–1269.

Vaxillaire, M., Veslot, J., Dina, C., Proença, C., Cauchi, C., Charpentier, G., Tichet, J., Fumeron, F., Marre, M., Meyre, D., et al. (2008). Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. Diabetes *57*, 244–254.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. *8*, e1002793.

Walters, R.G., Jacquemont, S., Valsesia, A., de Smith, A.J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., et al. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature *463*, 671–675.

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., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.

Zimmerman, B., Kelly, B., McMillan, B.J., Seegar, T.C.M., Dror, R.O., Kruse, A.C., and Blacklow, S.C. (2016). Crystal structure of a full-length human tetraspanin reveals a cholesterol-binding pocket. Cell *167*, 1041–1051.e11.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Human induced pluripotent stem cells (HipSci)	https://www.phe-culturecollections.org.uk/	Catalog no 77650042
Software and algorithms		
R	https://www.r-project.org/	v3.6.3
Java	https://www.oracle.com/uk/java/	JDK 7
cnvHap	https://www.imperial.ac.uk/people/l.coin	NA
cnvHitSeq	https://sourceforge.net/projects/cnvhitseq/	NA
MultiPhen	https://github.com/lachlancoin/MultiPhen	NA

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr Tisham De (tisham.de08@imperial.ac.uk or de.tisham@gmail.com).

Materials availability

NFBC material can be requested from the consortium website https://www.oulu.fi/nfbc/materialrequest.

Data and code availability

- Data

NFBC data are available with appropriate access permissions. Further details are available here https://www.oulu.fi/nfbc/materialrequest.

Data related to the WH-II study and their phenotypes are available at the following website https://www.ucl.ac.uk/whitehallII/

- Code

All codes used to process and analyze data are published, and the source code is currently available at.

Multiphen: https://github.com/lachlancoin/MultiPhen

cnvHap: https://www.imperial.ac.uk/people/l.coin

cnvHitSeq: https://sourceforge.net/projects/cnvhitseq/

- Additional information

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

This study did not generate new reagents or software code.

EXPERIMENT MODEL AND SUBJECT DETAILS

Study cohorts

All cohorts reported in this study including data from 1KG project, NFBC 1986, NFBC 1966, Whitehall II study (WH-II), DESIR, Child Obesity cohort, Adult Obesity cohort, 1958 British Birth Cohort 1958 (BC1958), National Blood Survey (NBS), Helsinki Birth cohort (HBCS), and HipSci samples have prior ethical approval and consent from all study subjects involved. Further details including aims and methods have

been reported earlier. In our analysis, we refer to NFBC 1986 as the primary discovery cohort for CNVs and metabolomic signatures and NFBC 1966 and WH-II as replication cohorts. BC1958, NBS, and HBCS were used as control cohorts for ascertaining CNV allele frequencies. Child obesity, Adult obesity, and DESIR cohorts were used for replicating disease outcomes. 1KG NGS data were used for CNV breakpoint and frequency calculations. HipSci data were used for functional characterization of CNVs.

Metabolomic measurements in NFBC 1986 and NFBC 1966

Metabolomic measurements for NFBC 1986 (n=228) and NFBC 1966 (n=228) cohorts were carried out using high-throughput ^{1}H nuclear magnetic resonance (NMR) technology developed by Nightingale Healthcare Limited (https://nightingalehealth.com/).

Further details of aims and methods for characterization of various lipoprotein species, ratios, and size along with other metabolites have been described earlier. A complete list of metabolomic phenotypes used in our analysis, their names, and categories is listed in Table S5. Clinical characteristics including age and gender for NFBC cohorts are reported in Table S6.

Lipid measurements in WH-II

After obtaining relevant permissions, we had access to the following lipid measurements for our analysis – apoprotein A1 (Apo A1), apoprotein B (Apo B), cholesterol total (Bchol), cholesterol HDL (HDL), intermediate-density lipoprotein (IDL), triglycerides (Trig), lipoprotein A (LPA), and cholesterol LDL (LDL).

METHODS DETAILS

CNV analysis

NGS-based CNV identification. First, CNV calls were generated using the cnvHiTSeq algorithm (Bellos et al., 2012) in *TSPAN8* genic region using NGS low-coverage data from 1KG project for 17 different populations. cnvHiTSeq uses a Hidden Markov Model (HMM)–based probabilistic model for genotyping and discovering CNVs from NGS platforms. It incorporates various signatures from sequencing data such as read depth, read pair, and allele frequency information and then integrates them into a single HMM model to provide improved sensitivity for CNV detection. Normalization of the sequence data prior to CNV analysis using cnvHitSeq was performed in the following manner: sequencing files in binary alignment format (bam) for the different populations were first downloaded from the 1KG website. For each population, samples were normalized in a sliding window of 25 base pairs and were corrected for wave effects and GC content. Next, cnvHiTSeq was run with a combination of read depth and split read information, with an initial transition probability of 0.15 and 15 expectation maximization (EM) training iterations.

cnvHap: Normalization and quality control

Cohorts genotyped on the Illumina platform were processed through the Illumina Beadstudio (now called Genomestudio2.0) software. LRR, B-Allele frequency (BAF), and sample SNP genotypes were exported from the Beadstudio software as 'final reports' for subsequent CNV analysis. Prior to CNV calling, data normalization was done in a genotyping-plate-specific manner in order to correct for batch effects. For every genotyping plate, data were adjusted for LRR median correction and LRR variance. Genomic wave effects were accounted for by fitting a localized loess function with a 500-kbp window. Next, the processed LRR and BAF values with relevant covariates such as genotyping plate, BAF, LRR variance were used as input by the cnvHap software for CNV calling.

CNV predictions using cnvHap

CNVs in the TSPAN8 gene were called in various cohorts using the cnvHap algorithm (Coin et al., 2010). This algorithm uses a haplotype HMM for simultaneously discovering and genotyping CNVs from various high-throughput SNP genotyping platforms such as Illumina Cardio-Metabochip and Agilent aCGH arrays. The haplotype HMM of cnvHap uses combined information of CNVs (LRR) and SNP (BAF) data in population aware mode for CNV predictions. cnvHap has specific emission parameters for different genotyping platforms. In our analysis, we used Illumina-platform-specific emission parameters in all cases. cnvHap was used in its population aware mode where all samples were simultaneously used to train the model. In contrast, CNV detection methods such as PennCNV trains the HMM one sample at a time and does not leverage population-level information for CNV prediction.

CNV segmentation

cnvHap calculates the most probable linear sequence of copy number states (hidden state of the HMM model) for each sample by using dynamic programming and outputs this sequence as CNV breakpoints.

Additionally, cnvHap also calculates probabilistic CNV genotypes or expected CNV genotypes (described later) based on posterior probabilities. Of note, CNV allele frequency based on breakpoints information might differ from frequency calculated using posterior probabilities of CNV genotypes.

Expected CNV genotypes

The haplotype HMM of cnvHap calculates the probability of deletion and duplication for each sample at a given probe which we refer to as the expected CNV genotypes. For example, at a particular probe, if a sample has CNV genotype assigned as 1 (heterozygous deletion) with probability of 0.8, then the expected CNV genotype is calculated as

$$1*0.8 + 2*0.2 = 1.2$$

The expected CNV genotypes were calculated separately for deletions and duplications for every sample and at every probe location. We have used expected CNV genotypes for all our association analyses and results.

Association analysis

Next, using the MultiPhen software (O'Reilly et al., 2012), we carried out both univariate and multivariate approaches for associating expected CNV genotypes with metabolomic phenotypes in all cohorts. In the univariate analysis, for every probe location, P values for association were calculated using expected CNV genotypes as predictors and metabolomic phenotypes as the outcome. For common genomic probe locations, meta-analysis of NFBC 1986 and NFBC 1966 was performed using the inverse variance fixed-effect model.

For multivariate analysis, we used the MultiPhen software which implements a reverse regression model where phenotypes are used as predictors and CNV genotypes are used as outcome. We refer to this model as a multivariate joint model. The multivariate joint model uses ordinal probit regression to associate CNV genotypes (outcome) with multiple metabolomic phenotypes (predictors) simultaneously and provides a single joint p value capturing the effect of all phenotypes together. In addition, we have further implemented a variable selection method into this model by using a custom backward-selection algorithm. This backward-selection method reduces the correlation structure in the phenotypic space through an iterative process and in the end provides a set of uncorrelated variables. This uncorrelated set of variables is next used in the standard MultiPhen multivariate joint model to obtain a single P value and effect size for all phenotypes. In our analysis, we refer to this subset of phenotypes as metabolomic signatures. In all univariate and multivariate regression analyses, phenotypes were transformed using quantile normalization and 50 LRR principal components (PCs), LRR variance, and sex were used as covariates.

Intensity-based validation of CNV association

There have been several reports regarding the use of direct raw signal data from various technology platforms without using intermediate processing or software as an input for bioinformatics methods. Such approaches have previously been applied for CNV-phenotype association studies where LRR intensity measurements from genotyping platforms were used (Eleftherohorinou et al., 2011). Here, we have leveraged a similar approach by using LRR-phenotype association results as an alternate method to validate CNV genotype-phenotype association results. Similar to CNV association analysis, we applied univariate and multivariate approaches from MultiPhen for the LRR data and used for 50 LRR PCs, LRR variance, and sex as covariates in the model.

Multiple testing

Previous studies have reported the presence of a high degree of correlation in metabolomic and lipid phenotypes. In order to adjust for multiple testing thresholds in the presence of such correlation structure, several alternate methods to Bonferroni correction such as the Sidak-Nyholt correction have been

proposed (Nyholt, 2004). Briefly in this method, for calculating the net number of effective tests M_{eff} in the presence of correlation structure in the phenotypes, the following formula can be used.

$$M_{eff} = 1 + (M - 1)\left(1 - \frac{\mathrm{Var}(\lambda_{obs})}{M}\right)$$

Here $lambda_{obs}$ is the eigen decomposition of the correlation matrix of metabolomic phenotypes. The net effective number of tests M_{eff} obtained can then be applied to the Sidak formula or the Bonferroni correction, in order to determine the correct p value threshold. On applying this correction to Sidak-Nyholt and the Bonferroni method, the adjusted multiple testing the p value thresholds obtained were 8.05e-4 and 7.8e-4, respectively.

Linkage disequilibrium

In NFBC 1986 and other cohorts, LD calculation was done using Pearson correlation coefficient for LRR and CNV genotype data from genotyping arrays and sequence data. In addition, a linear regression model was also used to calculate LD between CNV genotype and the number of B-alleles. In NFBC 1966, no probes were found to be in LD (r^2>0.5) with *TSPAN8* exon 11 or CNVR5583.1 and hence not reported.

QUANTIFICATION AND STATISTICAL ANALYSIS

All quantitative and statistical analyses are described in detail in the methods section of STAR Methods.

ADDITIONAL RESOURCES

Further details regarding the Northern Finland Birth Cohorts longitudinal study are available here- https://www.oulu.fi/nfbc/nfbc1966_1986.