**Data Supplement for for Duncan et al., Genome-Wide Association Study Reveals First Locus for Anorexia Nervosa and Metabolic Correlations. Am J Psychiatry (doi: 10.1176/appi.ajp.2017.16121402)**

## Contents

**Supplementary Text**

**Note: Tables S3–S6** are in separate Excel files, and **Figures S1–S8** are in a separate PDF file.

## Materials and Methods for the Genome-Wide Association Analysis

Additional information on controls. Controls included in this report fall into three categories. Children’s Hospital of Philadelphia/Price Foundation Collaborative Group (CHOP/PFCG) controls were collected per Wang et al. (1), and ‘Italy–South’ and ‘Greece’ were collected per the Genetic Consortium for Anorexia Nervosa/Wellcome Trust Case Control Consortium 3 (GCAN/WTCCC3) Boraska et al. (2) For the remaining nine datasets (single asterisks in Table S1), controls in the present report were controls in other genomic studies of psychiatric disorders. It is possible that some of these controls had AN or genetically correlated psychiatric or other diagnoses. Screening for AN in all controls would be ideal, but retrospective contact to screen for AN was not allowed or feasible. The consequence of this – if any – is likely to be attenuated signal in the GWAS because of misclassification; thus, we do not expect spurious results because of this aspect of our study design. Informed consent was obtained from all participants and local institutional review boards (IRBs) approved all individual studies. IRBs at the University of

North Carolina, Massachusetts General Hospital, and Stanford University approved analytical work for this investigation.

Individual dataset information. Ascertainment characteristics relevant to all samples are given in the main text. For each individual dataset (named by country of sample ascertainment), we provide relevant publications for additional information about samples. For all but the CHOP/PGCG samples, all cases are GCAN/WTCCC3, with this corresponding publication: Boraska V, Franklin CS, Floyd JA, Thornton LM, Huckins LM, Southam L, et al. A genome-wide association study of anorexia nervosa. Mol Psychiatry. 2014;19:1085-94.
**CHOP/PFCG** – Cases and controls described here:
Wang K, Zhang H, Bloss CS, Duvvuri V, Kaye W, Schork NJ, et al. A genome-wide association study on common SNPs and rare CNVs in anorexia nervosa. Mol Psychiatry. 2011;16:949-59.
**Greece & Italy-South** – Both cases and controls for these two datasets are from (GCAN/WTCCC3).
**Czech Republic** – Cases are from GCAN/WTCCC3. Controls described here:
Nelis, M. et al. Genetic structure of Europeans: a view from the North-East. PloS One 4, e5472 (2009).
**Finland** – Cases are from GCAN/WTCCC3. No publications on control samples are available.
**France** – Cases were from GCAN/WTCCC3. Genotyping of controls was provided by the Centre National de Génotypage (Institut de Génomique, Commissariat à l'énergie atomique et aux énergies alternatives, Evry, France).
**Germany** Cases are from GCAN/WTCCC3. Controls were described in the publication listed below:
Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421-7. (see ‘scz_boco_eur’ sample).

> “These German samples were collected by separate groups within the MooDS Consortium in Mannheim, Bonn, Munich and Jena. For the PGC analyses, the samples were combined by chip and ancestry. In Bonn/Mannheim, cases were ascertained as previously described. 17 Controls were drawn from three population-based epidemiological studies (PopGen), the Cooperative Health Research in the Region of

> Augsburg (KORA) study, and the Heinz Nixdorf Recall (HNR) study. All participants gave written informed consent and the local ethics committees approved the human subjects protocols. Additional controls were randomly selected from a Munich-based community sample and screened for the presence of anxiety and affective disorders using the Composite International Diagnostic Screener. Only individuals negative for the above mentioned disorders were included in the sample."

**Netherlands** – Cases are from GCAN/WTCCC3. Two publications describe controls:
Demirkan et al.: Psychol Med. 2016 Jun;46(8):1613-23. Somatic, positive and negative domains of the Center for Epidemiological Studies Depression (CES-D) scale: a meta-analysis of genome-wide association studies.
Hofman A, Brusselle GG, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, Ikram MA, Klaver CC, Nijsten TE, Peeters RP, Stricker BH, Tiemeier HW, Uitterlinden AG, Vernooij MW. Eur J Epidemiol. 2015 Aug;30(8):661-708. The Rotterdam Study: 2016 objectives and design update.
**Norway** – Cases are from GCAN/WTCCC3. No publications on control samples are available.
**Spain** – Cases are from GCAN/WTCCC3. Controls described here:
Nelis, M. et al. Genetic structure of Europeans: a view from the North-East. PloS One 4, e5472 (2009).
**UK** – Cases are from GCAN/WTCCC3. Controls are from WTCCC2, as described in the European Genome-phenome Archive: https://www.ebi.ac.uk/ega/studies/EGAS00000000028
**United States / Canada** – Cases are from GCAN/WTCCC3. Controls described here:
Scott, L. J. et al. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc. Natl. Acad. Sci. 106, 7501–7506 (2009).

QC and analytical overview. We performed uniform quality control (QC), followed by relatedness testing, imputation to phase 3 of 1000 Genomes data (3), case-control analysis within each dataset, meta-analysis across samples, and then heritability estimation and genetic correlation analysis in the combined dataset.

QC procedures were performed on each of the 12 individual datasets. First, single nucleotide polymorphisms (SNPs) with missingness rate >0.05 and those that were invariant (minor allele frequency: MAF=0) were removed, then individuals with missingness >0.02 were removed. Second, individuals with heterozygosity ($F_{het}$) > |0.2| and individuals failing sex checks were removed. Third, after these initial filters to remove poorly performing SNPs and individuals with the worst global QC metrics, we again employed SNP missingness filters, but with a more stringent threshold (>0.02). Fourth, we excluded SNPs with differential missingness between cases and controls >0.02. Finally, SNPs failing Hardy-Weinberg equilibrium (HWE) checks were excluded, using a more stringent filter in controls ($p<1x10^{-6}$) than in cases ($p<1x10^{-10}$), given that associated alleles could be out of HWE in cases. All analyses were performed using second generation PLINK (4).

Principal components analysis (PCA) was performed within each dataset and then also across all datasets using FastPCA (5). PCA was conducted on high quality SNPs with low linkage disequilibrium (LD) that passed the following filters: (1) SNP directly genotyped in all datasets; (2) minor allele frequency (MAF) >0.05; (3) HWE $p > 1x10^{-4}$; (4) no strand ambiguous (AT or GC) SNPs; (5) no SNPs in known high linkage disequilibrium (LD) regions (the extended major histocompatbility complex (MHC) chr6:25-35Mb and chromosome 8 inversion chr8:7-13Mb); and (6) $r^2$ between SNPs <0.2 (i.e., the PLINK option: '--indep-pairwise 200 100 0.2' which was applied twice). Within each dataset, scatterplots of principal components were visually examined for outliers, which were then removed. Following this step, PCA was re-run, plots were re-inspected, newly identified outliers were removed as necessary, and this process was repeated until cases and controls appeared evenly interspersed across pairs of PCs. Imputation to the 1000 Genomes phase 3 reference (3) was performed within the Psychiatric Genomics Consortium (PGC) pipeline, per previously reported procedures (6) using SHAPEIT for phasing (7) and IMPUTE2 for imputation (8).

Following imputation, samples were combined for relatedness testing and calculation of ancestry principal component covariates. The same filters as above were employed, and individuals with proportion identical by descent (IBD) values >0.2 were removed, preferentially retaining cases when a pair of related individuals contained a case and a control.

## Methods for Heritability, Genetic Correlation, Gene-based, and Pathway Analyses

SNP-chip heritability ($h^2_{\text{SNP}}$) and genetic correlation estimation with Linkage Disequilibrium Score regression (LDSC). LDSC (9, 10) was used on the SNP level summary statistics from the meta-analysis across twelve studies. General instructions are provided here: https://github.com/bulik/ldsc. For $h^2_{\text{SNP}}$ estimates in this paper, we used the *constrained* option, thus constraining the LDSC regression intercept to be 1 (i.e., the expected chi-square for a single SNP not influenced by population stratification). This approach was appropriate because—using individual-level genotype data and sample-level summary statistics—we were able to rule out the presence of related individuals and population stratification. The population prevalence used was 0.9% for liability scale $h^2_{\text{SNP}}$ estimation.

LD score regression was applied to our results in two additional ways. First, for genetic correlation analyses performed with LDSC, we used the *unconstrained* analysis because we were unable to rule out sample overlap between our AN cohort and individuals in the studies of 159 phenotypes we tested for genetic correlation. Sample overlap across pairs of traits tested for genetic correlation with LDSC does not pose a problem; however, if the precise amount of sample overlap is not known, it precludes use of *constrained* genetic correlation analysis. Datasets and phenotypes used for genetic correlation analysis come from the PGC data repository (https://www.med.unc.edu/pgc/results-and-downloads) and LD-Hub (11).

Second, we evaluated whether genomic regions associated with anorexia nervosa in our GWAS analysis tended to include functional features in the human genome (while accounting for the impact of LD). We used LD score regression with a "baseline model" including 52 annotation categories. The categories are described elsewhere (12); they included conserved region (13), University of California Santa Cruz (UCSC) gene models [exons, introns, promoters, untranslated regions (UTRs)], and functional genomic annotations constructed using data from ENCODE (14) and the Roadmap Epigenomics Consortium (15), grouped into cell type specific annotations. The 10 cell type/tissue annotation groups are adrenal/pancreas, cardiovascular, central nervous system, connective tissue/bone, gastrointestinal, immune, kidney, liver, skeletal muscle, and other.

To assess significance of heritability enrichment, we ran LD score regression ten times, each time adding one of these ten groups to the baseline model. Here, we wanted to control for the 52 annotation categories in the baseline model when identifying disease-relevant cell type groups, and so we report regression coefficients (i.e., the contribution of a cell-type grouping to per-SNP heritability controlling for all categories in the baseline model) instead of the proportion of heritability divided by the proportion of SNPs as in the analysis of the 52 annotation categories. See Table S3 for the results.

Gene-based association and pathway analysis. Our approach was guided by rigorous method comparisons of type I error rates of different algorithms (16, 17). *P*-values quantifying the degree of association of genes and gene-sets were generated using MAGMA (v1.03) (17). Gene analysis in MAGMA uses a multiple regression approach to incorporate LD between markers and to detect multi-marker effects, using F-tests to compute gene *P*-values. This model first projects the SNP matrix for a gene onto its principal components (PC), pruning away PCs with very small eigenvalues, and then uses the remaining PCs as predictors for the phenotype in a linear regression model. This improves power by removing redundant parameters, and guarantees that the model is identifiable in the presence of highly collinear SNPs. By default, only 0.1% of the variance in the SNP data matrix is pruned away. We used ENSEMBL gene models for 20,011 genes giving a Bonferroni corrected *P*-value threshold of $2.5 \times 10^{-6}$. Other than multiple genes in the genome-wide significant region on chromosome 12, no genes reached significance.

Pathway (gene-set) *P*-values were obtained using a competitive analysis, which tests if genes in a particular gene-set are more strongly associated with the phenotype than other gene-sets. We used European-ancestry subjects from the 1,000 Genomes Project (Phase 3 v5a, imputation INFO > 0.6, MAF ≥ 0.01) (3) for the LD reference. The gene window used was 35 kb upstream and 10 kb downstream to include regulatory elements. Gene-sets were extracted from MSigDB v5.1 (18), which includes canonical pathways (CP) and Gene Ontology (GO) gene sets. CP were curated from BioCarta, KEGG, Matrisome, Pathway Interaction Database, Reactome, SigmaAldrich, Signaling Gateway, Signal Transduction KE, and SuperArray. Pathways containing 10-1000 genes were included for a total of 2,737 pathways (1309 CP, 1428 GO). Principal components analysis of gene-set membership indicated that there were 1,900 independent pathways (i.e., the number of principal components explaining >99.5% of the variance), yielding 2.63 x $10^{-5}$ (0.05/1900) as the corrected 5% level of significance for pathway testing. No pathway reached significance. See Table S3 for full gene-based and pathway analyses results.

## Exploratory Mouse Tissue Analysis

In an exploratory series of experiments in order to assess which genes in our GWAS significant region might be preferentially involved in the responses to fasting we assessed, via quantitative polymerase chain reaction (qPCR), the gene expression of the genes in mouse hypothalamus.

Animals and diet. Unless stated otherwise, male C57BL/6J mice were fed *ad libitum* with either a standard chow diet (Harlan Teklad LM-485; 5.6% kcal fat) or a high-fat diet (D12331; Research Diets, New Brunswick, NJ, USA; 58% kcal fat). The mice had free access to water and were maintained under constant ambient conditions (22±1°C, constant humidity, 12h/12h light/dark cycle). All animal studies were performed in Cincinnati, OH, USA and approved by the Animal Ethics Committee of Cincinnati, OH, USA.

Mouse gene expression analysis. To assess effects on fasting and re-feeding (mimicking some clinical treatments for AN), hypothalamic gene expressions were profiled in 27/28-week-old male C57BL/6J mice which either: (1) were fed ad libitum with a regular chow diet; (2) had been fasting for 12, 24 or 36 h; (3) had been fasting for 36 h and then re-fed for 6 h using a fat-free diet (FFD); or (4) had been fasting for 36 h and then re-fed for 6 h using a high-fat diet (HFD). There were 6-8 mice in each group. Hypothalamic expression of target genes was further assessed in age-matched male C57BL/6J mice fed either a regular chow diet (body weight 32.69 ± 0.45 g) or a high-fat diet (body weight 54.72±1.25g; N=7–8 mice per group).

Target genes in the genome-wide significant region (*IKZF4, RPS26, ERBB3, PA2G4, ZC3H10, ESYT1, SUOX*, *RAB5B, CDK2, PMEL, DGKA)* were amplified using the ViiA 7 real-time PCR system (Life Technologies, Darmstadt, Germany), and results were normalized to the housekeeping genes *hypoxanthine guanine phosphoribosyltransferase 1* (*HPRT*) or *peptidylprolyl isomerase B* (*PPIB*).

The used primer sequences were as follows: *IKZF4-F: 3’*-GAGGAGCACAAGGAGAGGTG-5’, *IKZF4-R: 3’*-AATGAAAGTTGGCCGTTCAG-5’; *RPS26-F: 3’*-CCAAGGATAAGGCCATCAAG-*5’; RPS26-R: 3’*-CGGGATCGATTCCTAACAAC-*5’; ERBB3-F: 3’*-TACTGGTGGCCATGAATGAA-*5’; ERBB3-R: 3’*-CTCAATGTAAACGCCCCCTA-*5’; PA2G4-F: 3’*-GGTCAAACCTGGAAACCAGA-*5’; PA2G4-R: 3’*-TCATGCACCTCAAATTCTGC-*5’; ZC3H10-F: 3’*-CTGGCCACCAATGAGGTACT-*5’; ZC3H10-R: 3’*-TGGCTGCTCAGAGTGGTATG-*5’; ESYT1-F: 3’*-GGGTGAAAAGCCATTACGAA-*5’; ESYT1-R: 3’*-GTCGGGCGTTTGATAAAGAA-*5’; SUOX-F: 3’*-CTTCCACAGGCCATCAGAGT-*5’; SUOX -R: 3’*-TGCTCATGGTAGACCAGCAC-*5’; RAB5B-F: 3’*-GAGAGTCTGCAGTGGGGAAG-*5’; RAB5B -R: 3’*-CAGCAGTGTCCCAGATCTCA-*5’; CDK2-F: 3’*-GCCCTATTCCCTGGAGATTC-*5’; CDK2-R: 3’*-GGGGTCATAGTGCAGCATTT-*5’; PMEL-F: 3’*-CACCGACACCATAATGCTTG-*5’; PMEL -R: 3’*-GCAGGACACAGTCAGCTCAA-*5’; DGKA-F: 3’*-GGAGGTTCCCCATCACCTAT-*5’; DGKA -R: 3’*-TTTCCACTTCCGTGCTATCC-*5’*

Figure S7 shows the results from this analysis.

## References for Supplementary Text

1. Wang K, Zhang H, Bloss CS, Duvvuri V, Kaye W, Schork NJ, et al. A genome-wide association study on common SNPs and rare CNVs in anorexia nervosa. Mol Psychiatry. 2011;16:949-59.
2. Boraska V, Franklin CS, Floyd JA, Thornton LM, Huckins LM, Southam L, et al. A genome-wide association study of anorexia nervosa. Mol Psychiatry. 2014;19:1085-94.
3. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The International Genome Sample Resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 2016.
4. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
5. Galinsky K, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson N, et al. Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. bioRxiv. 2015:018143.
6. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421-7.
7. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9:179-81.

8. Howie B, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
9. Bulik-Sullivan B, Finucane H, Anttila V, Gusev A, Day F, ReproGen Consortium, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236-41.
10. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291-5.
11. Zheng J, Erzurumluoglu M, Elsworth B, Howe L, Haycock P, Hemani G, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. bioRxiv. 2016:051094.
12. Finucane H, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228-35.
13. Lindblad-Toh K, Garber M, Zuk O, Lin M, Parker B, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478:476-82.
14. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57-74.
15. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317-30.
16. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nat Neurosci. 2015;18:199-209.
17. de Leeuw C, Mooij J, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.
18. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417-25.

**TABLE S1**. **Descriptive information for 12 contributing studies**. Case, control, and SNP numbers reflect final QC. Other than CHOP/PFCG, all cases, plus cases and controls from Greece and Italy-South were from the Genetic Consortium for Anorexia Nervosa/Wellcome Trust Case Control Consortium 3 (GCAN/WTCCC3). Multiple rounds of QC were necessary for 9 of 12 studies given that controls were sourced from diverse repositories.

| | After QC | | | | | Genotyping Array | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Cases** | **Controls** | **Total** | **Lambda** | **N SNPs** | **Platform Cases** | **Platform Controls** |
| CHOP/PFCG** | 1,031 | 3,627 | 4,658 | 1.021 | 9,663,045 | Ill. Hum.Hap610 | Ill. Hum.Hap610 |
| Czech Republic* | 72 | 41 | 113 | 1.062 | 9,161,728 | Ill. Hum.660W | Ill. 311K |
| Finland* | 131 | 524 | 655 | 1.011 | 9,691,342 | Ill. Hum.660W | Ill. 550K |
| France* | 293 | 979 | 1,272 | 1.034 | 9,345,816 | Ill. Hum.660W | Ill. 311K |
| Germany* | 556 | 2,164 | 2,720 | 1.027 | 9,515,296 | Ill. Hum.660W | Ill. 550K |
| Greece | 70 | 79 | 149 | 1.03 | 9,302,402 | Ill. Hum.660W | Ill. Hum.660W |
| Italy-South | 75 | 50 | 125 | 1.05 | 9,506,517 | Ill. Hum.660W | Ill. Hum.660W |
| Netherlands* | 348 | 1,362 | 1,710 | 1.049 | 9,538,824 | Ill. Hum.660W | Ill. 550K |
| Norway* | 82 | 315 | 397 | 1.052 | 9,379,255 | Ill. Hum.660W | Ill. 317K |
| Spain* | 186 | 117 | 303 | 1.046 | 9,329,196 | Ill. Hum.660W | Ill. 311K |
| UK* | 237 | 964 | 1,201 | 1.012 | 9,554,456 | Ill. Hum.660W | Ill. 1M |
| US/Canada* | 414 | 760 | 1,174 | 1.018 | 9,537,730 | Ill. Hum.660W | Ill. 550K |
| Total | 3,495 | 10,982 | 14,477 | 1.045 | 10,641,224 | | |

**Both cases and controls are new (i.e. not previously published in AN meta-analyses). *Controls only are new; not previously used in published AN meta-analyses. SNPs=single nucleotide polymorphisms, CHOP=Children's Hospital of Philadelphia, PFCG=Price Foundation Collaborative Group, Ill.Hum.660W=Illumina Human 660W-Quad, Ill.=Illumina.

## TABLE S2. Top six loci (represented by sentinel variants after clumping of results)

| CHR | VARIANT | BP | A1 | A2 | FRQ Case | FRQ Control | INFO | OR | SE | P |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | rs4622308 | 56469185 | T | C | 0.480 | 0.442 | 0.90 | 1.20 | 0.03 | 4.252E-09 |
| 5 | rs200312312 | 104002346 | T | C | 0.706 | 0.677 | 0.90 | 1.20 | 0.03 | 6.73E-08 |
| 12 | rs117957029 | 127685233 | T | C | 0.969 | 0.977 | 0.71 | 0.58 | 0.10 | 1.62E-07 |
| 12 | rs11174202 | 62252257 | A | G | 0.582 | 0.547 | 0.97 | 1.17 | 0.03 | 3.11E-07 |
| 12 | chr12:69435103 | 69435103 | GTATATACATA | G | 0.830 | 0.807 | 0.77 | 1.24 | 0.04 | 7.22E-07 |
| 4 | rs13125782 | 7428266 | T | C | 0.239 | 0.215 | 0.93 | 1.19 | 0.04 | 9.20E-07 |

CHR=chromosome, BP=base position, A1=allele1, A2=allele2, FRQ=frequency, INFO=imputation quality score, OR=odds ratio, SE=standard error, P=p-value