Eating Disorders Working Group of the Psych Genomics Consortium
Anorexia Nervosa GWAS Summary Statistics
11 July 2019

Prepared by: Hunna Watson, Zeynep Yilmaz, and Patrick Sullivan

Primary genome-wide association study (GWAS) publication: papers that use these data must cite:

Watson et al. (2019). Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nature Genetics*.

Cohort description

PGC-ED is a collaboration representing researchers and clinicians from around the world, founded with the goal of identifying the genetic risk factors involved in the etiology of anorexia nervosa (AN) and other eating disorders. The Freeze 2 AN sample comprises 16,992 individuals with AN and 55,525 controls from 33 cohorts.[1] Cases met DSM-IV criteria for either lifetime AN (restricting or binge-purge subtype) or lifetime eating disorders 'not otherwise specified' AN-subtype (i.e., exhibiting the core features of AN).[2] Detailed information on recruitment and case ascertainment can be found elsewhere.[1,3,4,5,6] Out of the 33 cohorts, the majority of samples came from the Anorexia Nervosa Genetics Initiative study, a multi-site recruitment effort in the USA, Sweden, Denmark, and Australia, with assistance from New Zealand. These samples (12,537 cases post-QC) are included for the first time in an AN GWAS in the publication above. A cohort from the UK Biobank was included, and these samples also appear for the first time in an AN GWAS in this publication. The remaining cohorts were from the Genetic Consortium for Anorexia Nervosa (GCAN)/Wellcome Trust Case Control Consortium 3 (WTCCC3) and Children's Hospital of Philadelphia (CHOP)/Price Foundation Collaborative Group (PFCG). Most of the GCAN/WTCCC3 samples and CHOP/PFCG samples were in previous AN GWAS.[3,4,7]

QC overview

We performed uniform QC, followed by relatedness testing, genotype imputation, case-control analysis within each dataset, and meta-analysis across samples. QC was performed on the 33 datasets using the PGC RICOPILI pipeline. All cohorts were QC'd with default thresholds. First, a pre-filter 95% variant call rate was applied, which is helpful for merges of cases and controls from different studies. This was followed by sample filtering to retain only high-quality samples with a call rate ≥ 98% and inbreeding coefficient (Fhet) < |0.2|. Then, variant filtering was applied to retain only high-quality variants with MAF ≥ 0.01, < 2% missingness, differences in call rates between cases and controls ≤ 2%, and passing Hardy-Weinberg Equilibrium (HWE) checks, 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. Some cohorts required additional QC detailed here.[1] Principal components analysis (PCA) was conducted in the RICOPILI pipeline for each dataset using EIGENSTRAT and SmartPCA in fastmode.[8,9,10] In the first round, the 1000 Genomes Phase 3 populations were used for detecting ancestry outliers.[11] PCA was repeated on each cohort without the reference samples, and outliers were removed as necessary. This process was repeated until cases and controls appeared evenly interspersed across pairs of PCs. Duplicates and related individuals (PiHat > 0.2) were removed. The LD score regression intercept[12] for each cohort after QC ranged from 0.98 (SE 0.01) to 1.03 (SE 0.01). Imputation to the 1000 Genomes Phase 3 reference[11] was performed with Eagle for pre-phasing[13] and Minimac3[14] for imputation.

Association analysis

GWAS for each dataset was performed on imputed variant dosages using an additive model in the RICOPILI pipeline, which uses PLINK.[15] The first 5 ancestry principal components and principal components that differed nominally significantly ($p < 0.5$) between cases and controls were included as covariates. Summary statistics results were next filtered for SNPs with MAF ≥ 0.01 and imputation quality (INFO) scores ≥ 0.7. A fixed-effects meta-analysis across the 33 data sets using inverse-variance weighting was conducted with METAL[16] in the RICOPILI pipeline. The meta-analysis LD intercept was 1.02 (s.e. = 0.01), consistent with minimal population stratification or other systematic biases.

Note. rs73088112 (chr3:49413352) was initially identified as the 9th significant variant. However, it was not an independent association and was merged into a locus with rs9821797 (chr3:48718253).

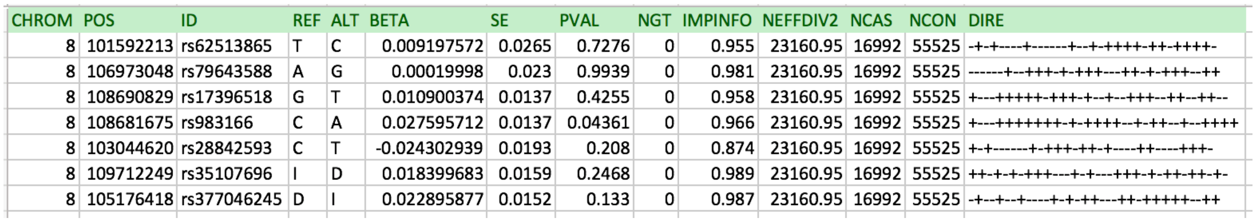
File description

Filename: **pgcAN2.2019-07.vcf.tsv.gz** (size: 319.2 MB)

PGC sumstats files now have a VCF-like header where all fields are explicitly defined. The key elements that led to the file are included (genome build, imputation reference, methods section, Ncase, Ncontrol, Ntrio, etc). Header lines begin with "##". A few of the header lines are below:

```
##genomeReference="GRCh37"
##imputationReference="1000GenomesPhase3"
##shortName="PGC-AN2"
##dependentVariable="anorexia nervosa"
##dependentVariableType="discrete"
##model="dependentVariable=SNPi+PC1-PC5"
##nCase="16992"
##nControl="55525"
##nTrio="0"
##nVariants=8219102
```

After the header, the rest of the file is standard tab-delimited text with a column definition row. A few lines are below:



CHROM	POS	ID	REF	ALT	BETA	SE	PVAL	NGT	IMPINFO	NEFFDIV2	NCAS	NCON	DIRE
8	101592213	rs62513865	T	C	0.009197572	0.0265	0.7276	0	0.955	23160.95	16992	55525	-+-+----+------+--+-++++-++-++++-
8	106973048	rs79643588	A	G	0.00019998	0.023	0.9939	0	0.981	23160.95	16992	55525	------+--+++-+-+++---++-+-+++--++
8	108690829	rs17396518	G	T	0.010900374	0.0137	0.4255	0	0.958	23160.95	16992	55525	+---+++++-+++-+--+--+++--++--++--
8	108681675	rs983166	C	A	0.027595712	0.0137	0.04361	0	0.966	23160.95	16992	55525	+---+++++++-+-++++--+-++--+--++++
8	103044620	rs28842593	C	T	-0.024302939	0.0193	0.208	0	0.874	23160.95	16992	55525	+-+------+-+++-++-+----++----+++-
8	109712249	rs35107696	I	D	0.018399683	0.0159	0.2468	0	0.989	23160.95	16992	55525	++-+-+-+++---+-+---+++-+-++-++-+-
8	105176418	rs377046245	D	I	0.022895877	0.0152	0.133	0	0.987	23160.95	16992	55525	-+--+--+----+-+-++---++-+++++--++

Note that REF (A2) and ALT (A1) are per the VCF file definitions. The header explicitly defines the columns. For example, the effect size column:

```
##INFO=<ID=BETA,Number=1,Type=Float,Description="beta or ln(OR) of ALT">
```

To read this file, skip to the row beginning with "CHROM" and then input as usual.

```
#=== R code to read in the TSV version of the VCF
library(data.table)
x <- fread(file="pgcAN2.2019-07.vcf.tsv.gz",
           skip="CHROM\tPOS",
           stringsAsFactors=FALSE, data.table=FALSE)
```

Additional notes:

- For long insertion/deletion variants, ALT/REF (A1/A2) alleles are truncated to the first 13 bases with a specification of the remaining length
- For multiallelic variants, “m” is appended to the marker name for different alternative alleles in order to ensure that the marker name is unique
- Allele frequencies and case/control counts per variant are currently omitted from the public release for data privacy. For inquiries about accessing the data, please contact the Eating Disorder Data Access Committee (DAC) representative (pgc.dac.ano@gmail.com).

Data use agreement:

The PGC has made the full results from all published PGC studies available for download. If you download these data, you and your immediate collaborators (“investigators”) acknowledge and agree to all of the following conditions:

1. These data are provided on an "AS-IS" basis, without warranty of any type, expressed or implied, including but not limited to any warranty as to their performance, merchantability, or fitness for any particular purpose;
2. Investigators will use these results for scientific research and educational use only;
3. Downloaded PGC results can be shared among collaborators but the reposting or public distribution of PGC results files is prohibited;
4. Investigators certify that they are in compliance with all applicable local, state, and federal laws or regulations and institutional policies regarding human subjects and genetics research;
5. Investigators will cite the appropriate PGC publication(s) in any communications or publications arising directly or indirectly from these data; and
6. Investigators will never attempt to identify any participant.

References:

[1] Watson HJ, Yilmaz Z, Thornton LM, Hübel C, Coleman JRI, . . ., Bulik CM. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat. Genet. 2019.

[2] American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 4th edition, text revision. 2000. Washington, DC: American Psychiatric Association.

[3] Wang K, Zhang H, Bloss CS, Duvvuri V, Kaye W,. . ., Price Foundation Collaborative Group. A genome-wide association study on common SNPs and rare CNVs in anorexia nervosa. Mol. Psychiatry. 2011;16:949-59.

[4] Boraska V, Franklin CS, Floyd JA, Thornton LM, Huckins LM,. . ., Bulik CM. A genome-wide association study of anorexia nervosa. Mol. Psychiatry. 2014;19:1085-94.

[5] Thornton LM, Munn-Chernoff MA, Baker JH, Juréus A, Parker R, …, Bulik CM. The Anorexia Nervosa Genetics Initiative (ANGI): overview and methods. Contemp. Clin. Trials. 2018;74;61-69.

[6] Kirk KM, Martin FC, Mao O, Parker R, Maguire S, . . ., Martin NG. The Anorexia Nervosa Genetics Initiative: study description and sample characteristics of the Australian and New Zealand arm. Aust. N. Z. J. Psychiatry. 2017;51;583-94.

[7] Duncan L, Yilmaz Z, Gaspar H, Walters R, Goldstein J, …, Bulik CM. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. Am. J. Psychiatry. 2017;174:850-58.

[8] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, & Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 2006;38;904-09.

[9] Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, . . ., Price AL. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. 2016; 98; 456-72.

[10] Galinsky KJ, Loh PR, Mallick S, Patterson NJ, & Price AL. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. Am. J. Hum. Genet. 2016;99;1130-39.

[11] 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature;536;68-74.

[12] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, . . ., Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 2015;47;291-5.

[13] Loh PR, Palamara PF, & Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 2016;48; 811-16.

[14] Das S, Forer L, Schönherr S, Sidore C, Locke AE, . . ., Fuchberger C. Next-generation genotype imputation service and methods. Nat. Genet. 2016;48;1284-87.

[15] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, . . ., Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007;81;559-75.

[16] Willer C, Li Y, & Abecasis G. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190-2191.