PGC Major Depression 2 GWAS results public data release

Date: 22 February 2018; Updated March 16 2018
Authors: Maciej Trzaskowski & Naomi Wray (U Queensland, Brisbane, Australia)
m.trzaskowski@imb.uq.edu.au; naomi.wray@uq.edu.au

PGC-MDD2, Wray, Ripke, Mattheisen, Trzaskowski et al. (2018)
Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 2018.

Abstract:
Major depressive disorder (MDD) is a common illness accompanied by considerable morbidity, mortality, costs, and heightened risk of suicide. We conducted a genome-wide association (GWA) meta-analysis based in 135,458 cases and 344,901 control. We identified 44 independent and significant loci. The genetic findings were associated with clinical features of major depression, and implicated brain regions exhibiting anatomical differences in cases. Targets of antidepressant medications and genes involved in gene splicing were enriched for smaller association signal. We found important relations of genetic risk for major depression with educational attainment, body mass, and schizophrenia: lower educational attainment and higher body mass were putatively causal whereas major depression and schizophrenia reflected a partly shared biological etiology. All humans carry lesser or greater numbers of genetic risk factors for major depression. These findings help refine and define the basis of major depression and imply a continuous measure of risk underlies the clinical phenotype.

This meta-analysis included 75,607 cases and 231,747 controls from 23andMe. The 23andMe Data Transfer Agreement explicitly stipulates that we can only make 10K SNPs available for public download.

Data availability. The PGC's policy is to make genome-wide summary results public. Summary statistics for a combined meta-analysis of PGC29 with five of the six expanded samples (deCODE, Generation Scotland, GERA, iPSYCH, and UK Biobank) are available on the PGC web site (URLs). Results for 10,000 SNPs for all seven cohorts are also available on the PGC web site.

GWA summary statistics for the Hyde et al. cohort (23andMe, Inc.) must be obtained separately. These can be obtained by qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Contact David Hinds (dhinds@23andme.com) to apply for access to the data. Researchers who have the 23andMe summary statistics can readily recreate our results by meta-analyzing the six cohort results file with the Hyde et al. results file from 23andMe.28

Genome build. All genomic coordinates are given in NCBI Build 37/UCSC hg19.

Hence we make available two data sets on the PGC download website:
1) Genome-wide SNPs from meta-analysis excluding 23andMe
    daner_pgc_mdd_meta_no23andMe.gz
2) 10K SNPs from meta-analysis including 23andMe
    daner_pgc_mdd_fm_to10k_report_170228.gz

Data download
If a data file is empty, try using a different web browser.

File format
The files are standard ricopili daner format
https://sites.google.com/a/broadinstitute.org/ricopili/cvas#TOC-Output-Files
SE is the standard error of the log odds ratio

On Chr21 there are 61 SNPs that have only 12 instead of 19 fields. This is because only one cohort contributed these SNPs
use this syntax to remove the 61 SNPs from chromosome 21.
zcat daner_pgc_mdd_meta_no23andMe.gz | awk '(NF==19)' > daner_pgc_mdd_meta_19fields

to keep all SNPs and to make data readable by R and other software use:
zcat daner_pgc_mdd_meta_no23andMe.gz | cut -f1-12 > daner_pgc_mdd_meta_12fields

to get only the 61 SNPs from chromosome 21 use:
zcat daner_pgc_mdd_meta_no23andMe.gz | awk '(NF==12)' > daner_mdd_61SNPs

Strategy for selection of 10K SNPs for public release
We wished to provide the most informative 10K SNPs for use in profile scoring analyses.

- Only SNPs that are imputed well in our study (INFO > 0.9*), assuming that they will also impute well in other studies.
- MAF > 0.1*
- Clumped using r^2 threshold 0.1*
- Select SNPs with smallest p-value
- The 10K must include SNP results included in the manuscript (N= 245)

* Because we must include in the released 10K SNPs the 245 SNPs reported in the manuscript, the released SNP list includes some SNPs with MAF < 0.1, INFO < 0.9, r^2 with other SNPs of > 0.1.

We have reported 245 SNPs with their effect sizes in the main manuscript and the supplementary tables. These have to be included in the final file. Reported SNPs come from:

- Main Manuscript Table 1
- Supplementary Table 5 (b,c,d)
- Supplementary Table 9 (SMR)
- Supplementary Table 10 (TWAS/FUSION)

Generation of clumped GWAS association results:
Starting filename: daner_pgc_mdd_meta_align_170228.gz
1. Clump main meta-analysis file using the following:
   - info > 0.9
   - maf > 0.1
   - p1 and p2 = 1
   - - window = 500kb
   - LD Rsq> 0.1

Output filename: daner_pgc_mdd_meta_align_170228_gwclump.gz.clumped.xmhc.gz

2. Sort Clumped by p-value and select top 10k
3. Extract meta-results for the reported 245 SNPs
4. Add reported SNPs to the clumped top 10k
5. Remove duplicates
6. Reduce file to 10k making sure that none of the reported are removed

Final filename: daner_pgc_mdd_fm_top10k_report_170228.gz

NOTE: final file has a column 'report', which indicates whether the SNP was reported or not. The codes are:

- 0: SNP not reported in the manuscript (N=9755)
- 1: SNP reported and present in the clumped file (N=104)
- 2: SNP reported but NOT present in the clumped file (N=141)

Out-of-sample prediction
To benchmark the relative utility of the GWAS results released for public download, we provide out-of-sample polygenic risk score (PRS) using the Muenster sample (960 cases, 874 controls) as the target sample. The Muenster sample was not included in the meta-analysis, but did feature in out-of-sample prediction in the paper (Figure 2a).

For this sample, we show (next page) that the maximum Nagelkerke's R^2 / 10^{th} deciles odds ratio were

a) Full meta-analysis including 23andMe (as published clumped, MAF 0.1, INFO 0.9): 0.026/3.127 (Not available for public download)
b) As a) above but limited to top 9859 SNPs: 0.022/2.503
c) As b) above plus SNPs reported in manuscript 10K SNPs: 0.019/2.458

d) Full meta-analysis excluding 23andMe (as published clumped, MAF 0.1, INFO 0.9): 0.023/2.763

NB. Reporting maxima are subject to winner's curse.

Why do we explain less variance for c) over b)? In the absence of other data sets to evaluate out of sample prediction, we attribute this observation to chance. The same observation may not be seen with other target samples.

Full meta-analysis including 23andMe, reduced to MAF > 0.1, INFO >0.9, clumped at r2 < 0.1					
$P_{threshold}$	NSNP	NKR^2	P	AUC	OR
0.00000005	45	0.000	8.46E-01	0.497	1.022
0.000001	110	0.001	2.92E-01	0.516	1.185
0.0001	662	0.005	1.04E-02	0.533	1.206
0.001	2143	0.010	2.16E-04	0.541	1.856
0.01	7703	0.020	2.14E-07	0.565	2.503
0.05	19567	0.026	3.50E-09	0.570	3.103
0.1	28994	0.024	1.07E-08	0.569	3.127
0.2	42874	0.017	1.75E-06	0.560	2.283
0.5	67699	0.019	3.95E-07	0.565	2.215
1	85098	0.019	4.41E-07	0.565	2.761
Full meta-analysis including 23andMe, top 9859 SNPs*					
0.00000005	45	0.000	8.46E-01	0.497	1.022
0.000001	110	0.001	2.92E-01	0.516	1.185
0.0001	662	0.005	1.04E-02	0.533	1.206
0.001	2143	0.010	2.16E-04	0.541	1.856
0.01	7703	0.020	2.14E-07	0.565	2.503
0.05	9856	0.022	6.32E-08	0.568	2.365
Full meta-analysis including 23andMe, 10K SNPs released*					
0.00000005	86	0.000	8.24E-01	0.501	1.021
0.000001	197	0.000	6.00E-01	0.508	0.895
0.0001	770	0.002	8.12E-02	0.525	1.299
0.001	2256	0.007	2.43E-03	0.538	1.717
0.01	7820	0.017	2.27E-06	0.560	2.277
0.05	9975	0.019	5.63E-07	0.564	2.458
Full meta-analysis excluding 23andMe, reduced to MAF > 0.1, INFO >0.9, clumped at r2 < 0.1					
0.00000005	5	0.001	3.67E-01	0.514	0.702
0.000001	18	0.001	2.34E-01	0.518	0.987
0.0001	275	0.005	9.64E-03	0.533	1.776
0.001	1228	0.005	6.63E-03	0.531	1.719
0.01	5823	0.010	1.97E-04	0.542	1.820
0.05	16829	0.015	5.13E-06	0.554	2.347
0.1	25969	0.021	1.13E-07	0.563	2.345
0.2	39226	0.023	2.59E-08	0.568	2.311
0.5	62868	0.021	1.02E-07	0.563	2.763
1	78534	0.021	9.20E-08	0.563	2.309

$P_{threshold}$ = p-value threshold from discovery sample.
NSNP = number of SNPs in PRS, after matching with Muenster sample SNPs
NKR^2 = Naglekerke's R^2;
P = p-value for PRS in logistic regression.
AUC = area under the curve; probability of a case ranking higher than control.
OR= Odds ratio for being a case for those in 10^{th} decile of PRS compared to the first decile
*The NSNP is smaller because of matching with SNPs available in Muenster sample.