readme.pdf for Psychiatry Genomics Consortium Distribution Files

PF Sullivan & Stephan Ripke, 04/2012

Introduction.

These are the results files from the Psychiatric Genomics Consortium mega-analyses (http://pgc.unc.edu).

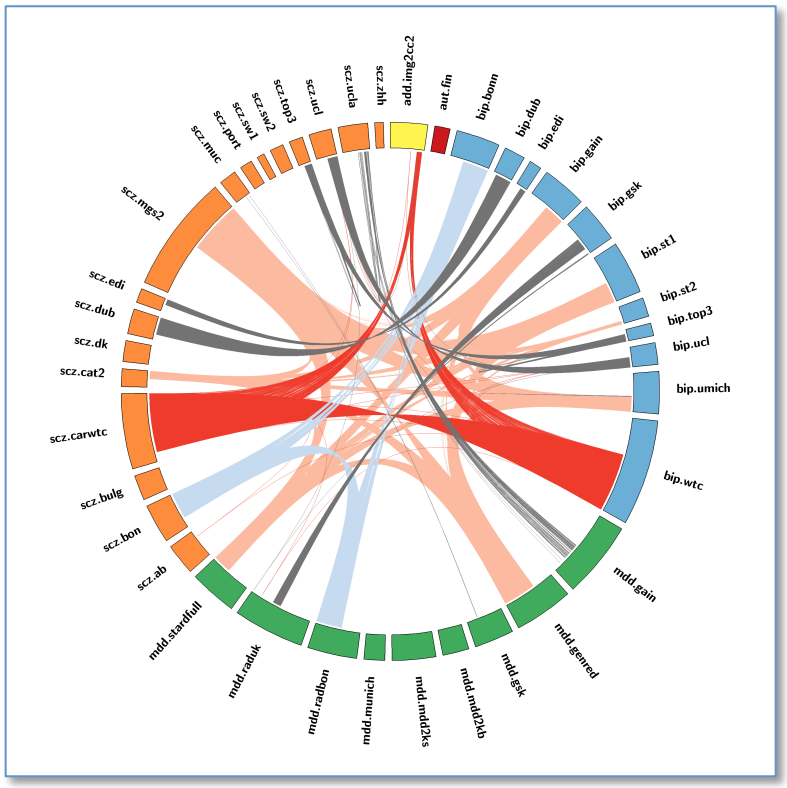
Disclaimer!

These data are provided "as is", and without warranty, for scientific and educational use only. If you download these data, you acknowledge that these data will be used only for non-commercial research purposes; that the investigator is in compliance with all applicable state, local, and federal laws or regulations and institutional policies regarding human subjects and genetics research; that secondary distribution of the data without registration by secondary parties is prohibited; and that the investigator will cite the appropriate PGC publication in any communications or publications arising directly or indirectly from these data.

Put more bluntly: there are a multitude of ways to screw up using these data. The onus is on you to use these data correctly. If you mess it up, it is completely on you.

CAUTION. Some controls were used by multiple studies (e.g., the SCZ and BIP control samples contain the exact same Gejman/MGS/GAIN control subjecgts). If you wish to compare results between these disorders, you must account for overlapping controls.

The circos plot below gives you an approximate idea of the degrees of control overlap. This is a non-trivial, non-ignorable issue.



Identifiability.

There have been extensive discussions in the human genetics community about SNP-level data release. Similarly, there have been extensive discussions within the PGC about data sharing.

Under some circumstances, it is possible to determine if a person was a case or a control in a GWAS. This is not generally the case, and requires near-ideal circumstances and an independent DNA source and genotyping results for a large number of markers (we note that this could require illegal behavior).

The data included in this public PGC distribution differ importantly from the idealized scenario. No individual data are being released. No case/control allele frequencies are included (in aggregate across samples or within each sample). Summary data per SNP were generated by imputing individual data onto a common backbone, analysis with PCA and study covariates, and then results from many different studies (usually 10 or more) were combined to yield summary results.

The distribution includes SNP information, odds ratio, standard error, p-value, and HapMap CEU allele frequency. The risk of identifiability was discussed at length with expert statistical geneticists (Mark Daly, Peter Visscher, Shaun Purcell, Bernie Devlin, and others). These experts agreed that the risk of identifiability, even in the strange case in which a DNA sample from a member of a PGC study was obtained and analyzed, is extremely small.

The PIs of the PGC approved this plan for release of results. Representatives of the NIMH also reviewed the plan, and confirmed it was consistent with NIMH policies.

It is possible to obtain individual data and more complete summary results via application to the NIMH repository (http://www.nimhgenetics.org), but these data are not available here.

Files.

See the primary papers (below) for full technical details. Again, appropriate use of these files is entirely your responsibility.

The distribution is a .zip file containing this .pdf, the full results file, and a "clumped" version that can be used for polygenetic risk profile analysis. DISORDER is adhd, bip, mdd, scz. DATE is the file preparation date (e.g., "2012-04") for April 2012.

pgc.DISORDER.full.DATE.txt

This file contains the full results based on the Stage 1 GWAS mega-analysis results. It is the results file on which the paper was based, a combined analysis of the imputed genotype dosages. Header row plus ~1.2 million SNPs (HapMap3, ADHD, AUT, MDD, SCZ) or ~2.4 million SNPs (HapMap2, BIP). It does not include any replication results.

pgc.DISORDER.clump.DATE.txt

This is the file that was used for common variant polygenic risk profile analyses. It is a subset of the full results file created using LD pruning. Clumping was done with plink using these steps: (a) drop SNPs with allele freq ≤ 0.02 or ≥ 0.98, drop imputation INFO < 0.9; and (b) perform p-value informed LD clumping (within 500kb window, r^2 = 0.25, and for SCZ include only one MHC SNP). Header row plus around 100K SNPs.

File contents.

Positions are UCSC hg18 / NCBI b36. Missing values are denoted by a period (“.”). For example, the first few rows of the SCZ files are:

==> pgc.scz.full.2012-04.txt <==

snpid hg18chr bp a1 a2 or se pval info ngt CEUaf

rs3131972	1	742584	A	G	1.0257	0.0835	0.761033	0.1613	0	0.16055
rs3131969	1	744045	A	G	1.0221	0.0801	0.784919	0.2225	0	0.133028
rs3131967	1	744197	T	C	1.0227	0.0858	0.79352	0.206	0	.

==> pgc.scz.clump.2012-04.txt <==

snpid hg18chr bp a1 a2 or se pval info ngt

rs10907175	1	1120590	A	C	1.0142	0.0354	0.69151	0.9922	13
rs2887286	1	1145994	T	C	0.9862	0.0278	0.617802	0.9989	17
rs11260562	1	1155173	A	G	0.9997	0.0466	0.995511	0.9281	11

snpid	SNP rs ID
hg18chr	hg18 chromosome (1-22)
bp	hg18 base position of SNP
a1	reference allele (not necessarily minor allele)
a2	alternate allele
or	odds ratio from logistic regression including PCA covariates (see papers)
se	standard error of the odds ratio
pval	asymptotic p-value
info	INFO score from imputation, ratio of variances, can exceed 1
ngt	number of studies in which this SNP directly genotyped (not imputed)
CEUaf	frequency of a1 in HapMap3 CEU (HapMap2 for BIP)

Citations.

If you use these data, you ***must*** cite the appropriate PGC publication.

Neale et al., Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry 49, 884 (Sep, 2010).

Psychiatric GWAS Consortium Bipolar Disorder Working Group, Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics 43, 977 (2011).

Major Depressive Disorder Working Group of the PGC, A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry. (In press).

Schizophrenia PGC, Genome-wide association study of schizophrenia identifies five novel loci. Nature Genetics 43, 969 (2011).

Autism and cross-disorder results will follow when in press.