scz2.readme.pdf

2014 Psychiatry Genomics Consortium Schizophrenia Results

PF Sullivan & S Ripke, 07/2014

Introduction

These are the results files from the second PGC schizophrenia mega-analysis (http://pgc.unc.edu). Citation for all studies that use any of these data: Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014.

Disclaimer!

Methods. See paper for full details. Briefly:

…we report the results of a multi-stage schizophrenia genome-wide association study of up to 36,989 cases and 113,075 controls … We obtained genome-wide genotype data from which we constructed 49 ancestry matched, non-overlapping case-control samples (46 of European and three of East Asian ancestry, 34,241 cases and 45,604 controls) and 3 family-based samples of European ancestry (1,235 parent affected-offspring trios). These samples comprise the primary PGC GWAS meta-analysis. Genotype data from all studies were processed by the PGC using unified quality control procedures followed by imputation of SNPs and insertion-deletions using the 1000 Genomes Project reference panel. In each sample, association testing was conducted with PLINK using imputed marker dosages and principal components (PCs) to control for population stratification. The results were combined using an inverse-weighted fixed effects model. After quality control (imputation INFO score ≥ 0.6, MAF ≥ 0.01, and successfully imputed in ≥ 20 samples), we considered around 9.5 million variants … For the subset of LD-independent SNPs with $P < 1x10^{-6}$ in the meta-analysis, we next obtained data from deCODE genetics (1,513 cases and 66,236 controls of European ancestry). We define LD-independent SNPs as those with low LD ($r^2 < 0.1$) to a more significantly associated SNP within a 500 kb window. Given that high LD in the extended MHC region spans ~8 Mb, we conservatively include only a single MHC SNP. The deCODE data were then combined with those from the primary GWAS to give a dataset of 36,989 cases and 113,075 controls. In this final analysis, ***128*** *LD-independent SNPs surpassed genome-wide significance ($P \leq 5x10^{-8}$) … We defined an associated locus as the physical region containing all SNPs correlated with each of the 128 index SNPs at $r^2 > 0.6$. Associated loci within 250 kb of each other were merged. This process resulted in* ***108*** *physically distinct associated loci.*

Files for public distribution. Four files are available for download. All files are tab delimited text with a header row.

File	Description
scz2.rep.128.txt	Locations of 128 independent SNPs meeting genome-wide significance
scz2.anneal.108.txt	Locations of 108 genomic locations after annealing
scz2.prs.txt.gz	Polygenic risk profile training set based on 52 PGC SCZ2 samples (102K SNPs)
scz2.snp.results.txt.gz	Large file (171 MB)!! 1000 Genomes imputed SNP results (9.4M SNPs)

File descriptions

scz2.rep.128.txt *(header row and 128 entries)*

region.rank	significance rank, 1-128
snpid	rs ID of SNP or indel ("chr2_200825237_I") with minimum p-value
hg19chrc	hg19 chromosome as character string (chr1-chr22, chrX)
chr	hg19 chromosome as number (1-22, chrX=23)
bp	hg19 base pair position of SNP with minimum p-value
six1	start of LD association interval (defined by r2>0.6)
six2	end of LD association interval (defined by r2>0.6)
a1a2	alleles, a1=reference allele for frequency and OR
frqa	frequency of a1 in cases (Affected)
frqu	frequency of a1 in controls (Unaffected)
info	imputation quality score
p	p-value in PGC GWAS data
or	odds ratio in PGC GWAS data
se	standard error of ln(OR) in PGC GWAS data
p_rep	p-value in replication sample
or_rep	odds ratio in replication sample
se_rep	standard error of ln(OR) in replication sample
q_rep	heterogeneity Q-value in replication sample
p_comb	p-value in PGC+replication
or_comb	odds ratio in PGC+replication
se_comb	standard error of ln(OR) in PGC+replication
q_comb	heterogeneity Q value in PGC+replication

scz2.anneal.108.txt *(header row and 108 entries)*

anneal.rank	significance rank, 1-108
hg19chrc	hg19 chromosome as character string (chr1-chr22, chrX)
bestsnp	rs ID of SNP or indel ("chr2_200825237_I") with minimum p-value
pmin	minimum p-value in region in PGC+replication
anneal1	hg19, start of genomic region
anneal2	hg19, end of genomic region
spananneal	size of region in kb

scz2.prs.txt.gz *(header row and 102,636 SNPs)*

snpid	rs ID of SNP
a1	reference allele for OR (may not be minor allele)
a2	alternate allele
info	imputation quality score
or	odds ratio in PGC GWAS data
se	standard error of ln(OR) in PGC GWAS data
p	p-value in PGC GWAS data
ngt	number of samples in which SNP directly genotyped

scz2.snp.results.txt.gz *(header row and 9,444,230 SNPs)*

hg19chrc	hg19 chromosome as character string (chr1-chr22, chrX)
snpid	rs ID of SNP
a1	reference allele for OR (may not be minor allele)
a2	alternate allele
bp	hg19 base pair position of SNP
info	imputation quality score
or	odds ratio in PGC GWAS data
se	standard error of ln(OR) in PGC GWAS data
p	p-value in PGC GWAS data
ngt	number of samples in which SNP directly genotyped