



**NORMENT**

Norwegian Centre for  
Mental Disorders Research

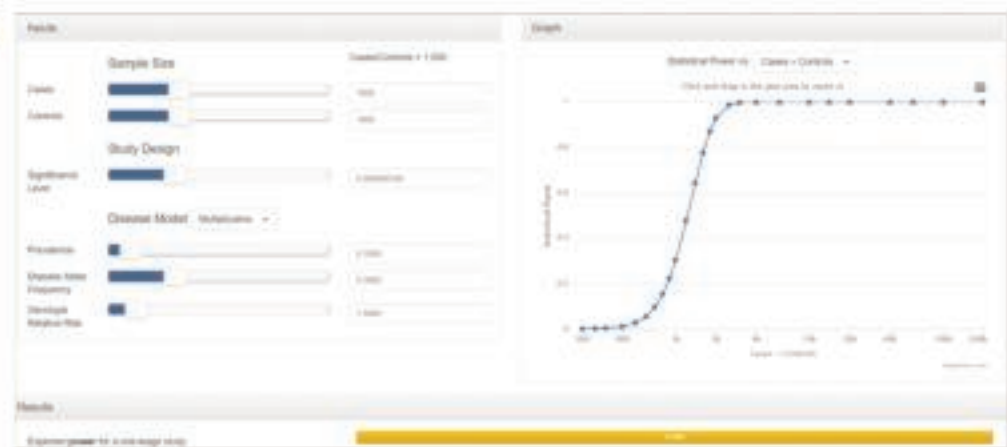
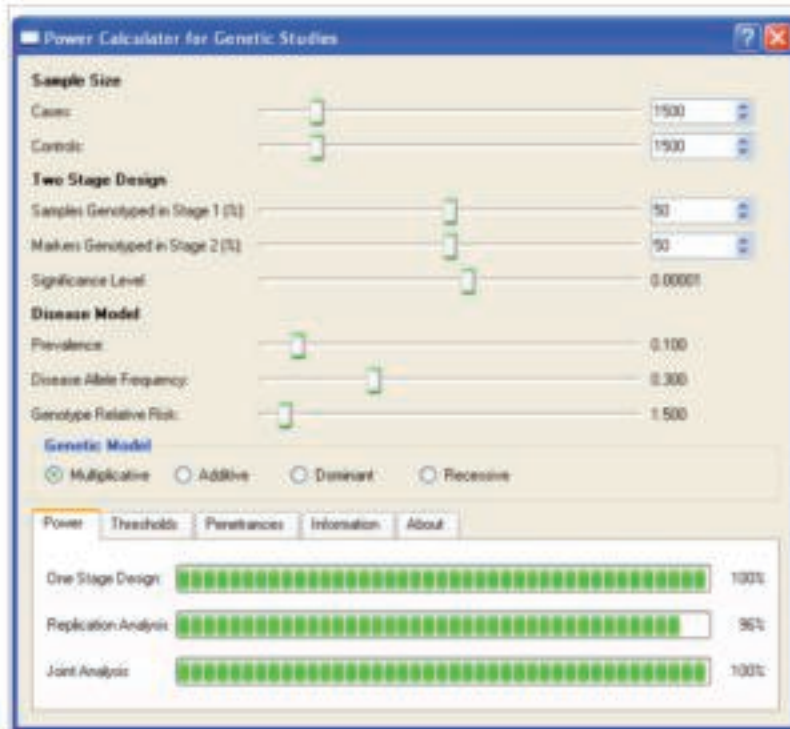
# How to Run a GWAS?

Kevin O'Connell

# Why?, before How

- The aim of genome-wide association studies (GWAS) is to identify single nucleotide polymorphisms (SNPs) of which the allele frequencies vary systematically as a function of phenotypic trait values
  - between cases with schizophrenia and healthy controls
  - between individuals with high vs. low scores on cognition
- Identification of trait-associated SNPs may subsequently reveal new insights into the biological mechanisms underlying these phenotypes.
- These analyses require the execution of several quality checks and careful conductance of statistical analyses to avoid spurious associations due to several potential sources of confounding (e.g., population stratification). In addition, knowledge of genetic power calculation is necessary to avoid performing underpowered studies.

# Power Calculations



GAS

[http://csg.sph.umich.edu/abecasis/gas\\_power\\_calculator/index.html](http://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html)

CaTS

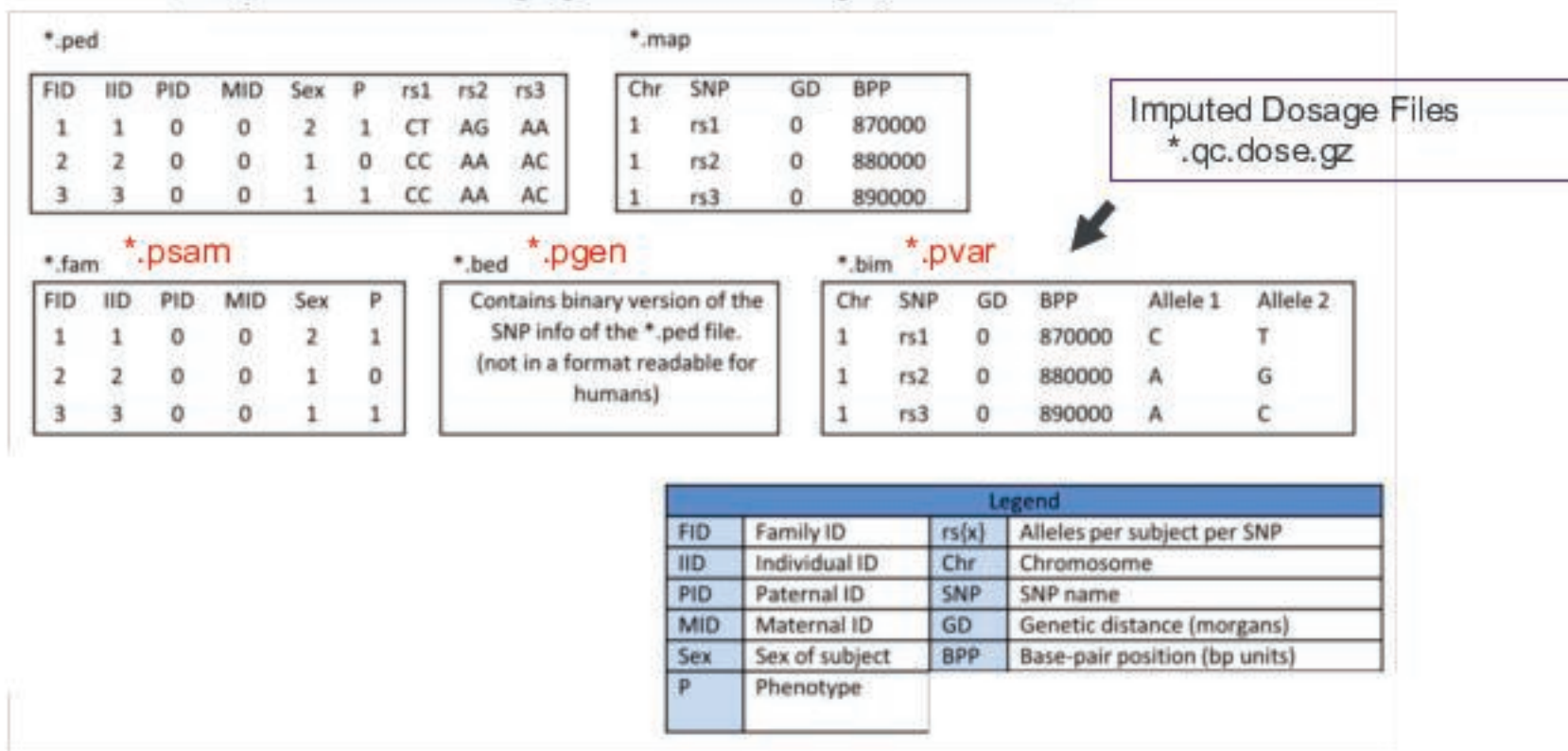
<http://csg.sph.umich.edu/abecasis/CaTS/index.html>

# Software and Data Format



## Plink

- <https://www.cog-genomics.org/plink/1.9/>
- <https://www.cog-genomics.org/plink/2.0/>





# Data manipulation

Convert imputation dosage files:

```
plink2  
--import-dosage  
/tsd/p33/data/durable/vault/genetics/imputation/TOP2017/chunk*.qc.dose.gz  
--fam top2017.fam  
--make-pgen  
--out out_directory/chunk
```

Switch file formats

```
--make-pgen (plink2: pgen/psam/pvar)  
--make-bed (bed/bim/fam)  
--recode (ped/map)
```

Merge files

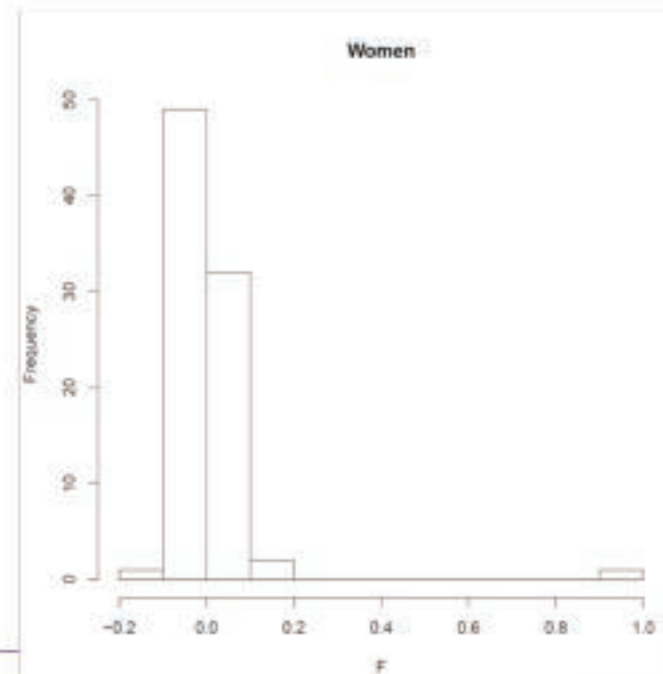
```
--pfile X --merge Y --out  
--pfile X --merge-list list.txt --out
```

# QC Steps

Step	Command	Function	Thresholds and explanation
1: Missingness of SNPs and individuals	--geno  --mind	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.  Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).  Note. SNP filtering should be performed before individual filtering.
2: Sex discrepancy	--check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.

FID	IID	PEDSEX	SNPSEX	STATUS	F
1328	NA06989	2	2	OK	-0.01184
1377	NA11891	1	1	OK	1
1349	NA11843	1	1	OK	1
1330	NA12341	2	2	OK	-0.01252
1444	NA12739	1	1	OK	1
1344	NA10850	2	2	OK	0.01496
1328	NA06984	1	1	OK	1
1463	NA12877	1	1	OK	1
1418	NA12275	2	2	OK	-0.1028
1346	NA12043	1	1	OK	1
1375	NA12264	1	1	OK	1
1349	NA10854	2	1	PROBLEM	0.99
1459	NA12865	2	2	OK	0.0302

plink --bfile X --impute-sex --make-bed --out Y



# QC Steps

Step	Command	Function	Thresholds and explanation
1: Missingness of SNPs and individuals	--geno  --mind	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed.  Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02).  Note, SNP filtering should be performed before individual filtering.
2: Sex discrepancy	--check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.
3: Minor allele frequency (MAF)	--maf	Includes only SNPs above the set MAF threshold.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds. Respectively, for large ( $N = 100,000$ ) vs. moderate samples ( $N = 10,000$ ), 0.01 and 0.05 are commonly used as MAF threshold.
4: Hardy-Weinberg equilibrium (HWE)	--hwe	Excludes markers which deviate from Hardy-Weinberg equilibrium.	Common indicator of genotyping error, may also indicate evolutionary selection. For binary traits we suggest to exclude: HWE $p$ value < $1e-10$ in cases and < $1e-6$ in controls. Less strict case threshold avoids discarding disease-associated SNPs under selection (see online tutorial at <a href="https://github.com/MareesAT/GWA_tutorial/">https://github.com/MareesAT/GWA_tutorial/</a> ). For quantitative traits, we recommend HWE $p$ value < $1e-6$ .
5: Heterozygosity		Excludes individuals with high or low heterozygosity rates	Deviations can indicate sample contamination, inbreeding. We suggest removing individuals who deviate $\pm 3$ SD from the samples' heterozygosity rate mean.

# QC Steps

6: Relatedness	--genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs ( <b>pruning</b> ) for this analysis and limit it to autosomal chromosomes only. Cryptic relatedness can interfere with the association analysis. If you have a family-based sample (e.g., parent-offspring), you do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population based sample we suggest to use a pi-hat threshold of 0.2, which is in line with the literature (Anderson et al., 2010; Guo et al., 2014).
	--min	Sets threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, pi-hat >0.2 (i.e., second degree relatives) can be detected.	
7: Population stratification	--genome	Calculates identity by descent (IBD) of all sample pairs.	Use independent SNPs ( <b>pruning</b> ) for this analysis and limit it to autosomal chromosomes only.
	--cluster --mds-plot k	Produces a k-dimensional representation of any substructure in the data, based on IBS.	
	--pca	Produces values for the first 20 principal components. You can change the number by passing a numeric parameter	



# Population Stratification

- An important source of systematic bias in GWAS is population stratification.
- It has been shown that even subtle degrees of population stratification within a single ethnic population can exist (Abdellaoui et al., 2013; Francioli et al., 2014).
- Testing and controlling for the presence of population stratification is essential.

# Testing and Controlling for Stratification

- After imputation, identify autosomal SNPs with very high imputation quality (INFO >0.8) and low missingness (<1%).

`--geno 0.01`

- Further identify SNPs after linkage disequilibrium pruning ( $r^2 > 0.02$ ) and frequency filtering (MAF > 0.05).

`--indep-pairwise 50 2 0.02`      `--maf 0.05`

- The derived SNP set should then be used for robust relatedness testing and population structure analysis.
- Relatedness testing is done with PLINK and pairs of subjects with  $\pi_{\text{hat}} > 0.2$  identified, with one member of each pair removed at random after preferentially retaining cases over controls.

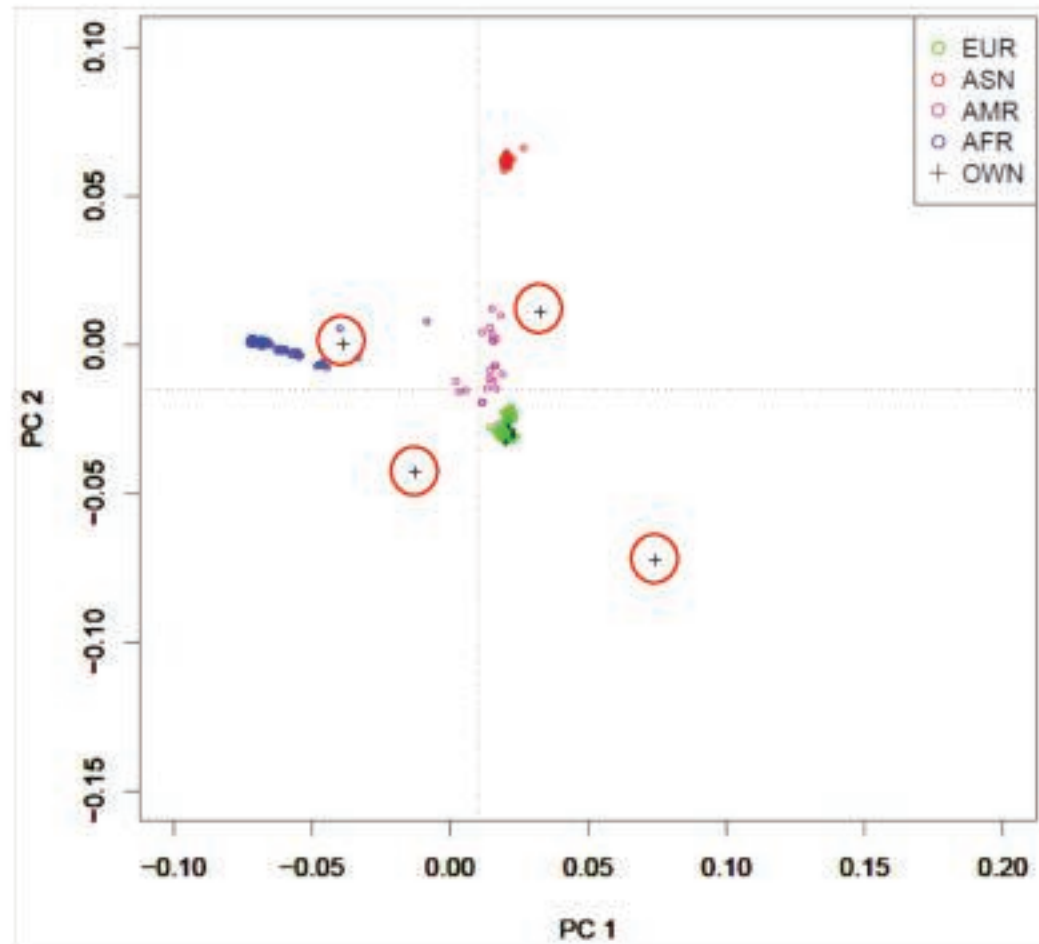
`--genome --min 0.2`

- Principal component estimation should be performed with the same set of autosomal SNPs.

# Testing and Controlling for Stratification



--pca



# Covariates for GWAS

- Test the first 20 principal components for phenotype association (with study indicator variables included as covariates).
- Include significantly associated principal components in the association analyses.

```
$ head covar_pca_ect.txt | column -t
FID    IID      Age  Gender  C1      C2      C3      C4      C7
1328   NA06989  61   1       0.0160249 -0.0527081 0.0532834 -0.00151572 0.019497
1377   NA11891  23   1       0.00880326 -0.0302948 -0.0051995 0.0268125 -0.000492102
1349   NA11843  65   2       -0.000948607 0.0140868 -0.00435938 -0.0145398 -0.00908916
1330   NA12341  48   2       -0.0133513 -0.0112818 0.00311679 0.00684165 -0.0295584
1328   NA06984  34   2       0.00775499 0.0128441 0.0194965 -0.0222438 -0.0126287
1418   NA12275  38   1       -0.0119731 -0.0060443 -0.00943407 0.0183443 0.0304177
13291  NA06986  35   1       -0.000383232 0.0113762 0.0048583 -0.0238018 -0.00481855
1418   NA12272  39   2       -0.0103709 -0.00907079 -0.010815 -0.0168651 -0.0181808
13292  NA07051  32   2       -0.00338742 -0.00631925 -0.00832884 0.0436293 -0.0209961
```



# Run the GWAS

plink2

--bfile *X*

--covar *covarfile.txt*

--logistic / --linear

--hide-covar (removes covariate-specific lines from the main report)

--adjust (reports the genomic inflation estimate  $\lambda$  in the log file)

--out *results*

```
$ head logistic_results.assoc.logistic
```

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs3131972	742584	A	ADD	109	1.957	1.552	0.1207
1	rs3131969	744045	A	ADD	109	2.322	1.8	0.07179
1	rs1048488	750775	C	ADD	108	1.986	1.591	0.1117
1	rs12562034	758311	A	ADD	109	0.3292	-1.912	0.05586
1	rs12124819	766409	G	ADD	109	0.9143	-0.2787	0.7805
1	rs4040617	769185	G	ADD	109	2.059	1.549	0.1214
1	rs4970383	828418	A	ADD	109	1.825	1.41	0.1587
1	rs4475691	836671	T	ADD	109	1.669	1.116	0.2643
1	rs1806509	843817	C	ADD	109	1.817	1.854	0.06371

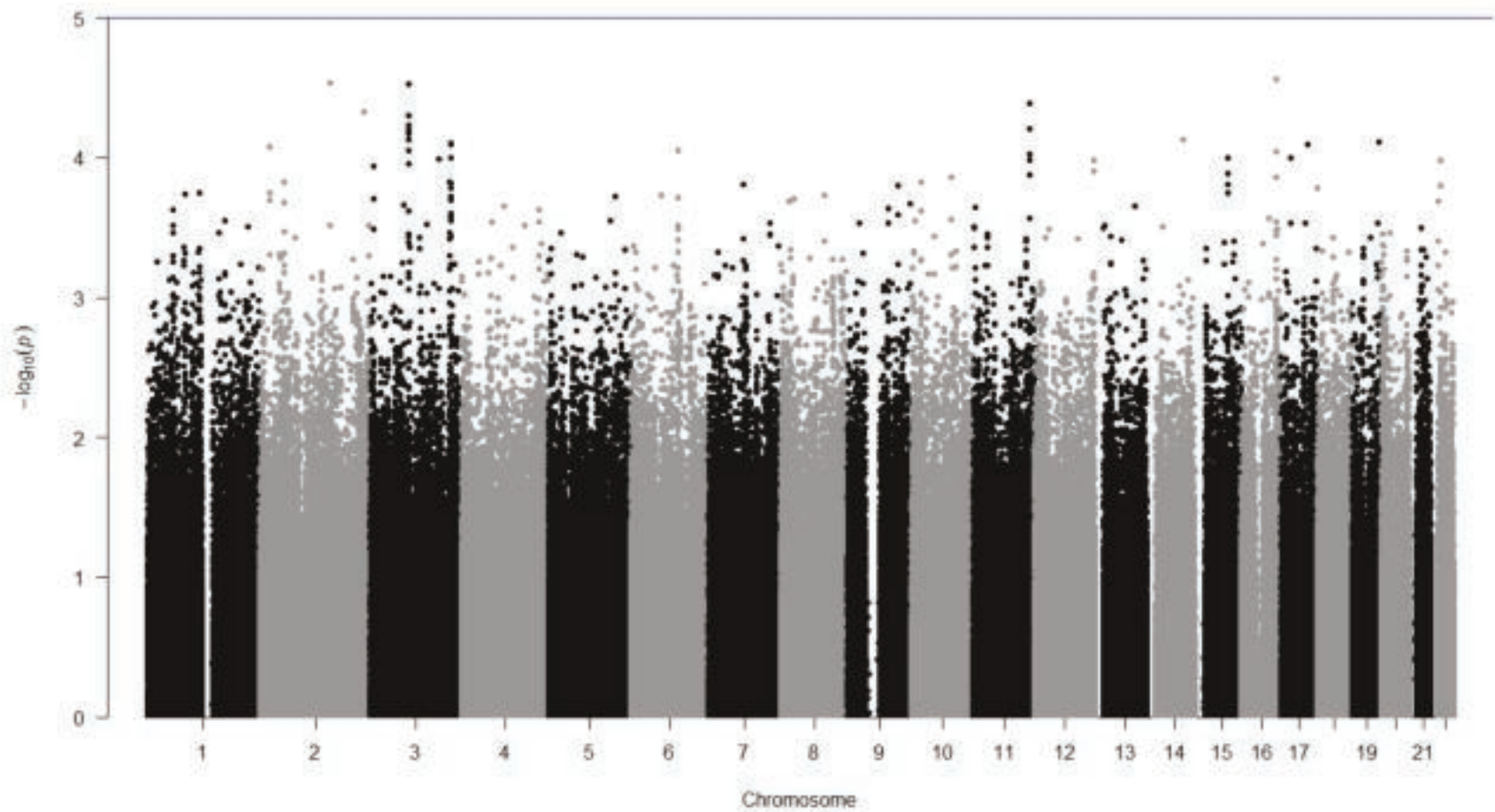
# Generate Manhattan Plot



```
install.packages("qqman",repos="http://cran.cnr.berkeley.edu/",lib=~" ")  
library("qqman",lib.loc=~")
```

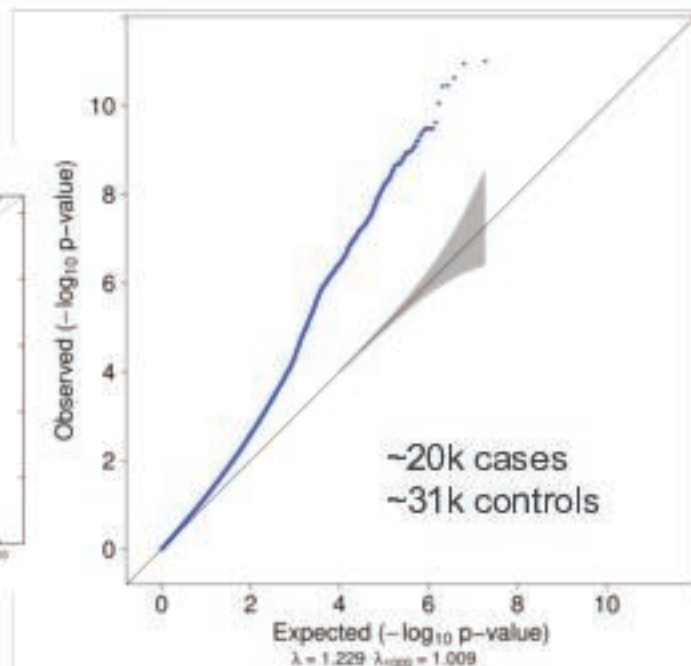
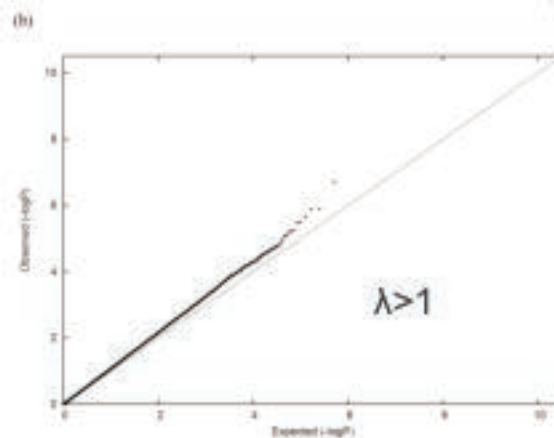
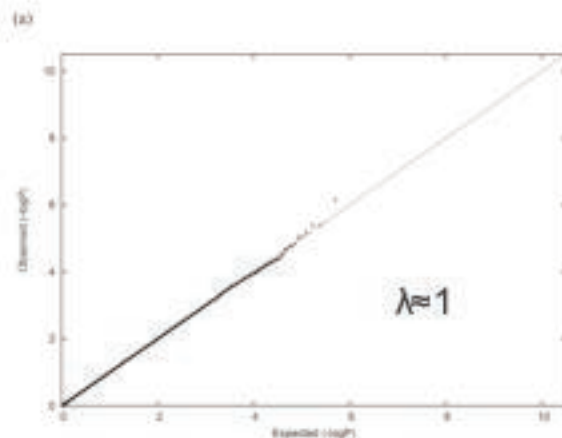
```
results_log <- read.table("new_logistic_results.assoc_2.logistic", head=T)  
jpeg("Logistic_manhattan.jpeg")  
manhattan(results_log,chr="CHR",bp="BP",p="P",snp="SNP")  
dev.off()
```

```
results_log <- read.table("logistic_results.assoc_2.logistic", head=T)  
jpeg("QQ-Plot_logistic.jpeg")  
qq(results_log$P, main = "QQ plot of GWAS p-values")  
dev.off()
```



# Determining Stratification

- $\lambda \approx 1$  indicates no stratification
- $\lambda > 1$  indicates potential stratification or other confounders such as family structure, cryptic relatedness or polygenicity
- Also consider  $\lambda$  is proportional to sample size





# Define GWS loci and lead SNPs

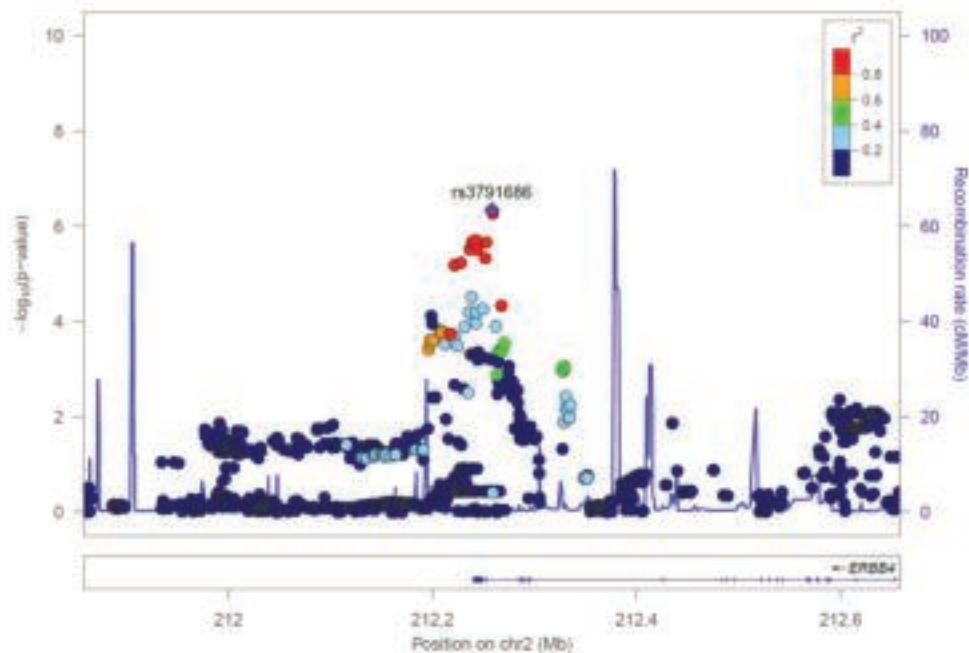
- GWAS findings implicate genomic regions “loci” containing multiple significant SNPs
- “Clumping” is used to convert associated SNPs to associated regions by identifying an index SNP with the smallest P-value in a genomic window and other SNPs in high LD with that index SNP
- Can use PLINK
  - Retain SNPs with association  $P < 0.0001$  and  $r^2 < 0.1$  within 3 Mb windows
  - `--clump-p1 1e-4 --clump-p2 1e-4 --clump-r2 0.1 --clump-kb 3000`
  - Bedtools (<https://bedtools.readthedocs.io>) is used to combine partially or wholly overlapping clumps within 50 kb.
- Can use FUMA SNP2GENE option (<http://fuma.ctglab.nl/>)
- Regional plots should be reviewed to identify and remove singleton associations

# Fuma Clumping Options

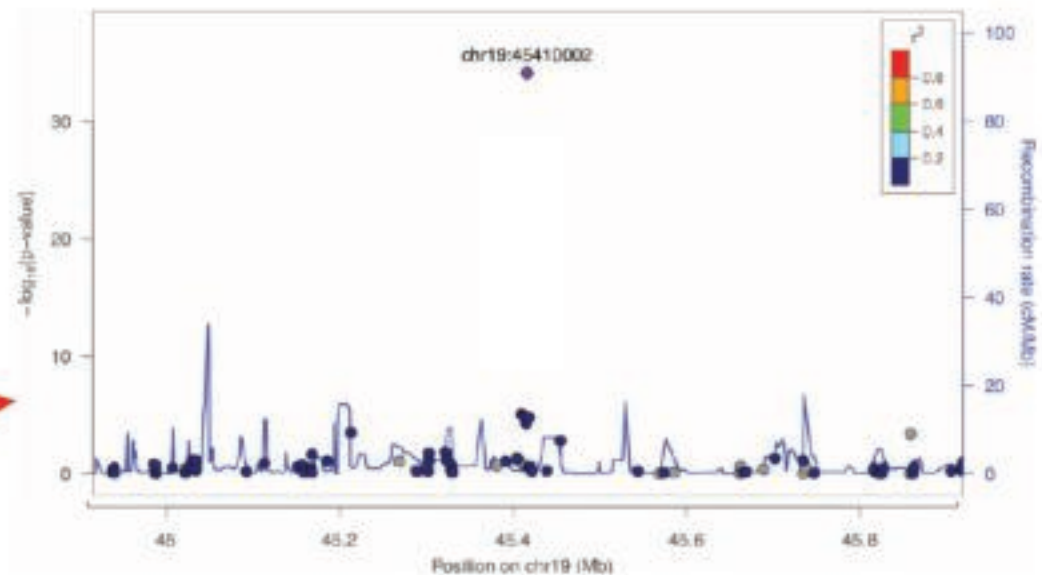
<http://fuma.ctglab.nl/>

## 2. Parameters for lead SNPs and candidate SNPs identification

Sample size (N) ?	Total sample size (integer): <input type="text"/>	
	OR Column name for N per SNP (text): <input type="text"/>	✓ OK. The defined column will be used for sample size per SNP.
Minimum P-value of lead SNPs (<)	<input type="text" value="5e-8"/>	✓ OK.
$r^2$ threshold to define LD structure of lead SNPs ( $\geq$ )	<input type="text" value="0.6"/>	✓ OK.
Maximum P-value cutoff (<) ?	<input type="text" value="0.05"/>	✓ OK.
Reference panel population	<input type="text" value="1000G Phase3 EUR"/> *	✓ OK.
Include variants in reference panel (non-GWAS tagged SNPs in LD) ?	<input type="text" value="Yes"/> *	✓ OK.
Minimum Minor Allele Frequency ( $\geq$ ) ?	<input type="text" value="0.01"/>	✓ OK.
Maximum distance between LD blocks to merge into a locus (< kb) ?	<input type="text" value="250"/> kb	✓ OK.



Plots generated using LocusZoom  
<http://locuszoom.sph.umich.edu/>



# Post-GWAS?

- Meta-analysis
  - PGC GWAS are all meta-analyses of numerous cohorts
  - METAL  
([https://genome.sph.umich.edu/wiki/METAL\\_Documentation](https://genome.sph.umich.edu/wiki/METAL_Documentation))
  - Forest plots should be reviewed to confirm that association signals arose from the majority of the cohorts

```
# === DESCRIBE AND PROCESS THE FIRST INPUT FILE ===
MARKER SNP
ALLELE REF_ALLELE OTHER_ALLELE
EFFECT BETA
PVALUE PVALUE
WEIGHT N
PROCESS inputfile1.txt

# === THE SECOND INPUT FILE HAS THE SAME FORMAT AND CAN BE PROCESSED IMMEDIATELY ===
PROCESS inputfile2.txt

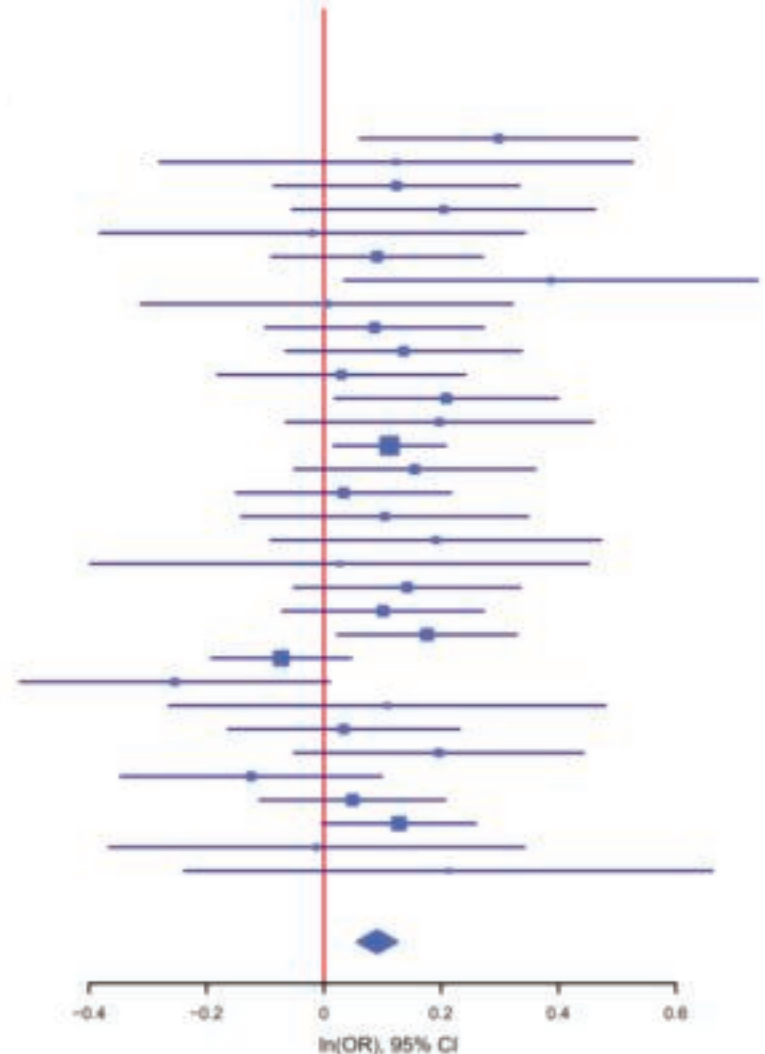
# === DESCRIBE AND PROCESS THE THIRD INPUT FILE ===
MARKER SNP
ALLELE A_REF OTHER_ALLELE
EFFECT BETA
PVALUE pvalue
WEIGHT N
PROCESS inputfile3.txt
```



# Plots using the forestplot package in R

<https://cran.r-project.org/web/packages/forestplot/vignettes/forestplot.html>

rs7544145	T/C			1:150138699			
+++++				het_P:	0.409	het_I:	-2.6
	ngt	info	p_value	f_ca(n)	f_co(n)	ln(OR)	STDerr
bip_bmau_eur	0	0.97	0.0142	0.853(329)	0.812(1810)	0.298	0.121
bip_bmg2_eur	0	1.01	0.551	0.847(181)	0.810(479)	0.123	0.206
bip_bmg3_eur	0	0.93	0.243	0.814(490)	0.798(880)	0.124	0.107
bip_bmpo_eur	0	0.96	0.122	0.860(410)	0.839(689)	0.204	0.132
bip_bmsp_eur	0	0.83	0.914	0.825(253)	0.836(291)	-0.0199	0.185
bip_bonn_eur	0	0.94	0.326	0.822(674)	0.808(1245)	0.0907	0.0922
bip_dub1_eur	0	0.98	0.0312	0.853(150)	0.802(796)	0.387	0.18
bip_edi1_eur	0	0.96	0.975	0.815(280)	0.817(275)	0.00509	0.162
bip_fat2_eur	0	0.92	0.363	0.832(728)	0.820(1106)	0.0865	0.0953
bip_fran_eur	0	0.95	0.188	0.830(451)	0.810(1630)	0.136	0.103
bip_gain_eur	0	0.91	0.781	0.825(703)	0.820(603)	0.03	0.108
bip_gsk1_eur	0	0.95	0.0316	0.846(741)	0.819(901)	0.209	0.0974
bip_hai2_eur	0	0.90	0.142	0.834(410)	0.813(485)	0.197	0.134
bip_icuk_eur	0	0.92	0.0223	0.830(2524)	0.814(4106)	0.112	0.0491
bip_jst5_eur	0	0.96	0.139	0.826(644)	0.802(624)	0.155	0.105
bip_may1_eur	0	0.91	0.719	0.810(934)	0.806(759)	0.0338	0.0938
bip_mich_eur	0	0.96	0.406	0.820(486)	0.811(436)	0.104	0.125
bip_pfl1_eur	0	0.92	0.185	0.813(378)	0.795(339)	0.191	0.144
bip_rom3_eur	0	0.87	0.901	0.862(233)	0.861(198)	0.0269	0.217
bip_st2c_eur	0	0.92	0.15	0.837(634)	0.819(1216)	0.142	0.0987
bip_stp1_eur	0	0.92	0.25	0.820(920)	0.807(999)	0.101	0.0878
bip_swa2_eur	0	0.91	0.0244	0.824(864)	0.800(2257)	0.176	0.0783
bip_swei_eur	0	0.92	0.235	0.803(1283)	0.812(3657)	-0.0725	0.061
bip_top7_eur	0	0.91	0.0599	0.791(449)	0.823(373)	-0.254	0.135
bip_top8_eur	0	0.94	0.589	0.815(149)	0.799(293)	0.108	0.19
bip_uci2_eur	0	0.88	0.737	0.806(726)	0.800(693)	0.0341	0.101
bip_uclo_eur	0	0.92	0.12	0.828(441)	0.802(495)	0.196	0.126
bip_ume4_eur	0	0.93	0.276	0.814(568)	0.830(569)	-0.124	0.114
bip_usc2_eur	0	0.91	0.547	0.829(1297)	0.825(1156)	0.0488	0.081
bip_wtcc_eur	0	0.94	0.0555	0.829(1636)	0.811(1632)	0.128	0.067
ms.bip_butr_eur	0	0.92	0.946	0.831(236)	0.833(236)	-0.0122	0.181
ms.bip_uktr_eur	0	1.02	0.355	0.840(130)	0.808(130)	0.212	0.23
<b>PGC_BIP32b_mds7a</b>	<b>0</b>	<b>0.93</b>	<b>4.83e-07</b>	<b>0.825(20352)</b>	<b>0.812(31358)</b>	<b>0.0909</b>	<b>0.0181</b>



# Post-GWAS?

- Meta-analysis
  - METAL  
([https://genome.sph.umich.edu/wiki/METAL\\_Documentation](https://genome.sph.umich.edu/wiki/METAL_Documentation))
  - Forest plots should be reviewed to confirm that association signals arose from the majority of the cohorts
- FUMA
  - Additional analyses and annotation of results

# Additional FUMA analyses

<http://fuma.ctglab.nl/>

3-1. Gene Mapping (positional mapping) 

3-2. Gene Mapping (eQTL mapping) 

3-3. Gene Mapping (3D Chromatin Interaction mapping) 

4. Gene types 

5. MHC region 

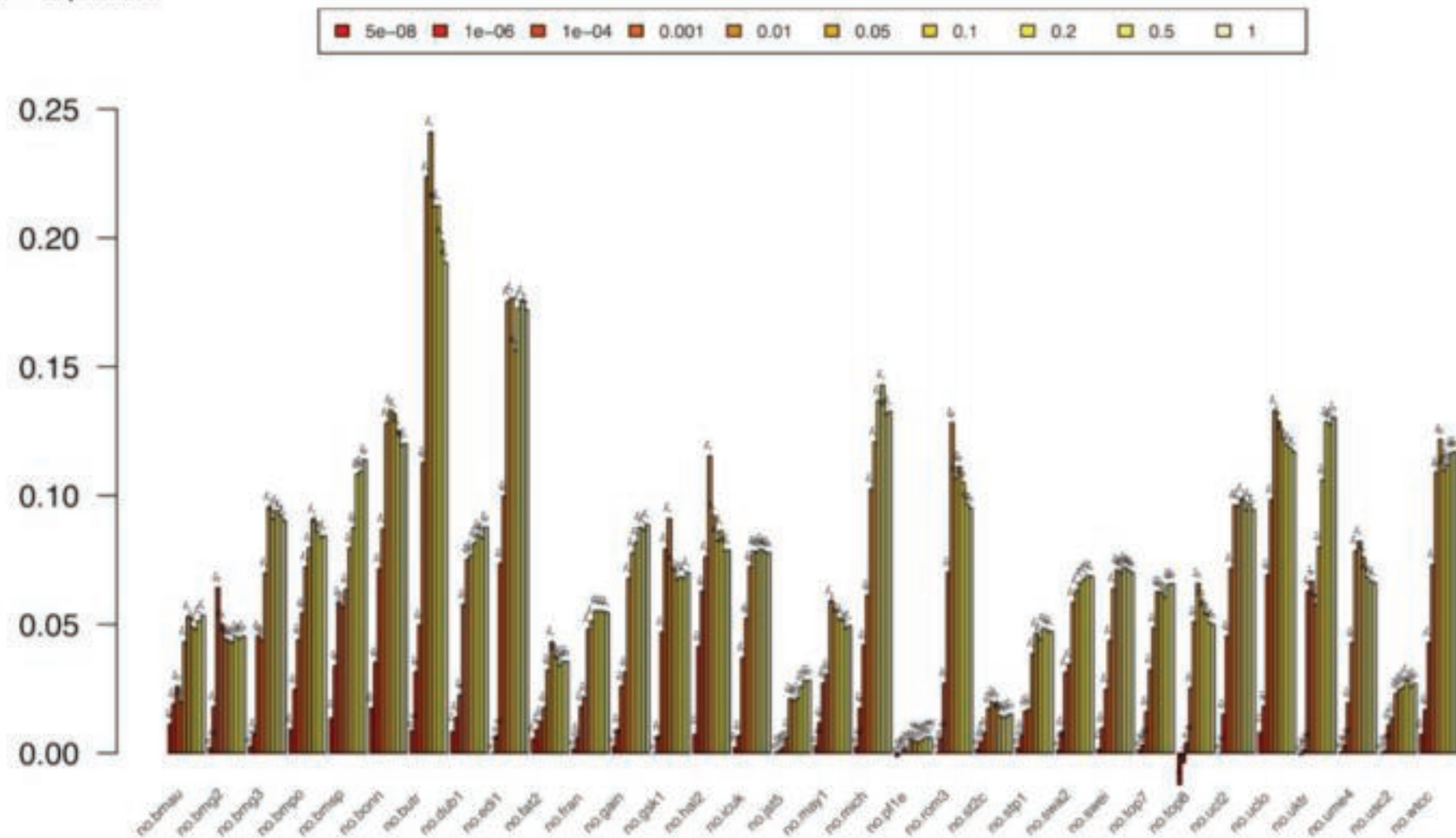
6. MAGMA analysis 

# Post-GWAS?

- Meta-analysis
  - METAL  
([https://genome.sph.umich.edu/wiki/METAL\\_Documentation](https://genome.sph.umich.edu/wiki/METAL_Documentation))
  - Forest plots should be reviewed to confirm that association signals arose from the majority of the cohorts
- FUMA
  - Additional analyses and annotation of results
- PRS
  - Leave-one-out analysis
  - More details on calculating and using PRS after Lunch



## R – squared



# Family-Based GWAS

- PLINK
  - --tdt (logistic)
  - It is possible to separately consider transmissions from heterozygous fathers versus heterozygous mothers to affected offspring, --poo (parent-of-origin)
  - --qfam (linear)
  - More detail at [https://www.cog-genomics.org/plink/1.9/fam\\_assoc](https://www.cog-genomics.org/plink/1.9/fam_assoc)




Specific software for family-based analyses listed


# Homework



INTERNATIONAL JOURNAL OF METHODS IN  
**PSYCHIATRIC RESEARCH**

ORIGINAL ARTICLE |  Open Access   

## A tutorial on conducting genome-wide association studies: Quality control and statistical analysis

Andries T. Marees , Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis,  
Cynthia Marie-Claire, Eske M. Derks

First published: 27 February 2018 | <https://doi.org/10.1002/mpr.1608> | Cited by: 1

<https://www.ncbi.nlm.nih.gov/pubmed/29484742>



**NORMENT**

Norwegian Centre for  
Mental Disorders Research

# Thank You

Oslo universitetssykehus HF  
Klinikk psykisk helse og avhengighet  
Seksjon for psykoseforskning/TOP  
Ullevål sykehus, bygg 49  
P.O. Box 4956 Nydalen  
NO-0424 Oslo  
Norway