GWAS of SSNS

Data management, QC and exploratory analyses

03/03/2020

GWAS SSNS

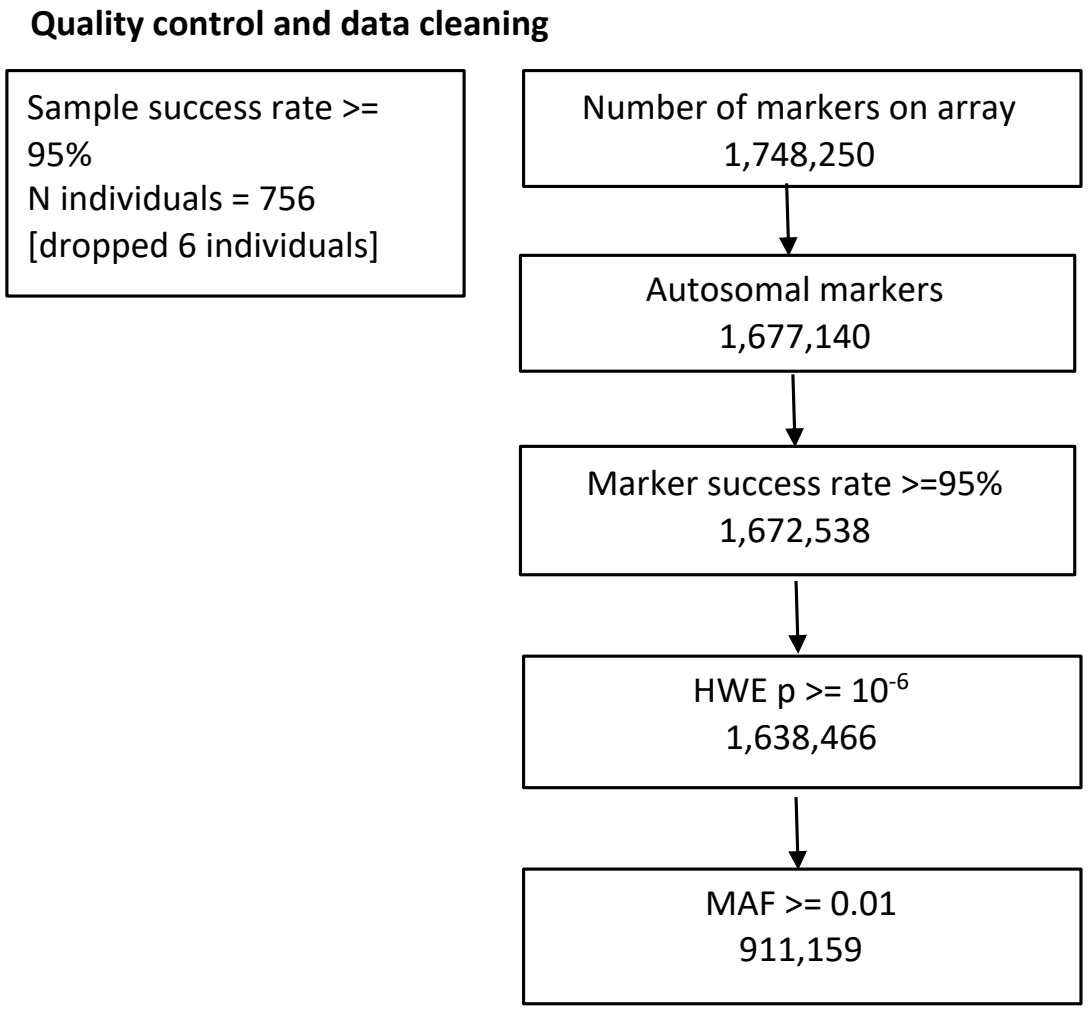
Cases

- Starting n = 762 multi-ethnic SSNS samples
- Ancestry/ethnicity – Asian 386 (50.6%), Black 161 (21.1%) , Caucasian 149 (19.6%), Hispanic/”Mixed” (8.7%)
- Genotyped on Illumina MEGA Global SNP chip

Controls

- Initial controls: 1000 Genomes Phase 3 version 5

SSNS samples



Data management 1

- QC’ed markers from SSNS samples used to extract matching markers by chromosome and position
- Both datasets aligned and alleles checked against hg37 fasta
- Markers passing QC reannotated against dbSNP151
- Initial check with bcftools +fixref plugin

SSNS dataset checked against hg37

- Pre

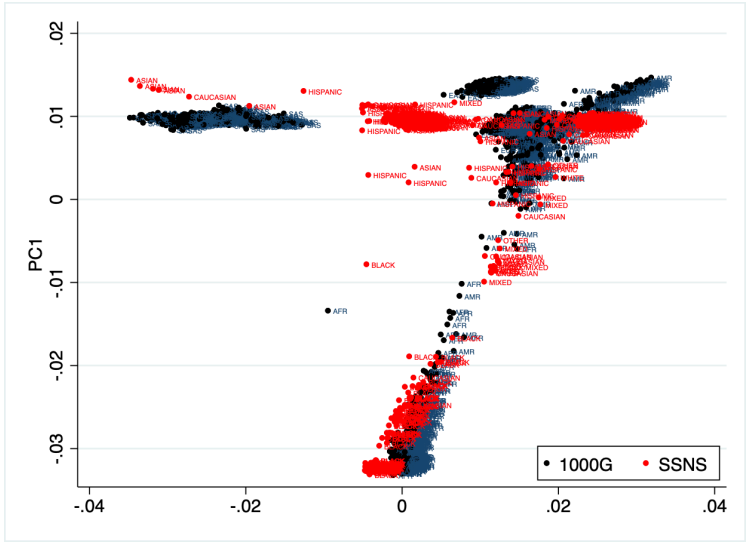
```
# SC, guessed strand convention
SC      TOP-compatible      0
SC      BOT-compatible      0
# ST, substitution types
ST      A>C     38099   4.2%
ST      A>G     143217          15.7%
ST      A>T     24618   2.7%
ST      C>A     42466   4.7%
ST      C>G     26512   2.9%
ST      C>T     179742          19.8%
ST      G>A     178918          19.7%
ST      G>C     26423   2.9%
ST      G>T     42696   4.7%
ST      T>A     24789   2.7%
ST      T>C     144037          15.8%
ST      T>G     38195   4.2%
# NS, Number of sites:
NS      total           911159
NS      ref match       708525          77.9%
NS      ref mismatch    201187          22.1%
NS      skipped         1447
NS      non-ACGT        1447
NS      non-SNP         0
NS      non-biallelic   0
```

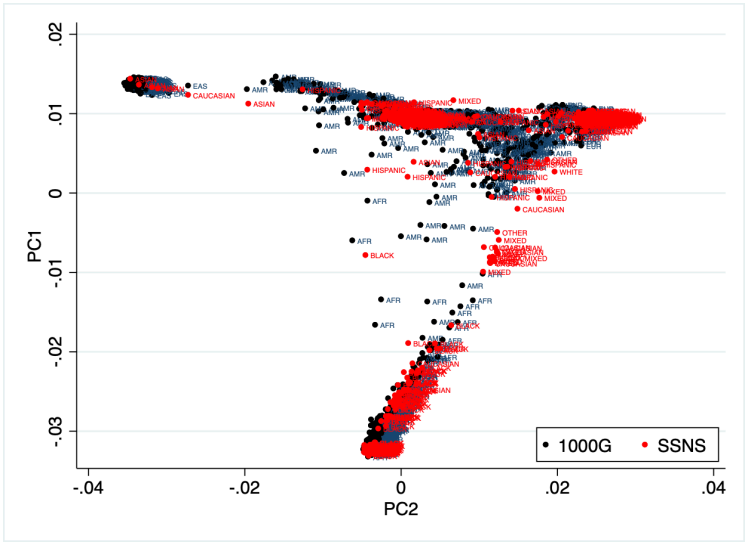
- Post

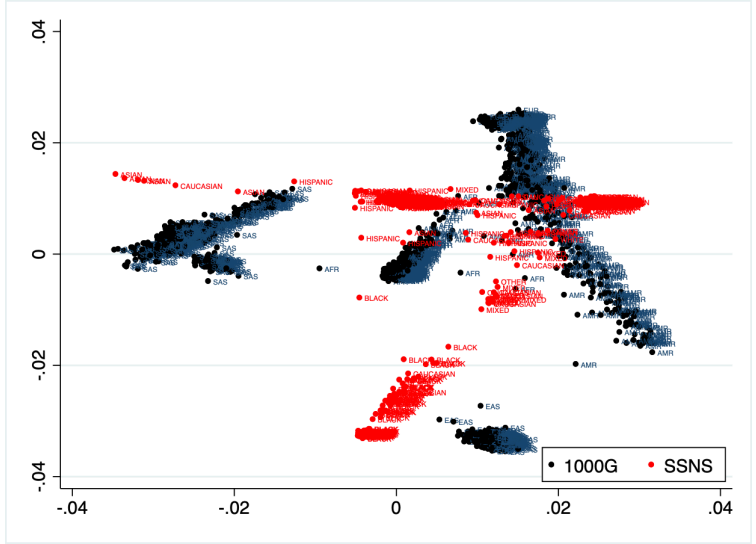
```
# SC, guessed strand convention
SC      TOP-compatible      0
SC      BOT-compatible      0
# ST, substitution types
ST      A>C     38394   4.2%
ST      A>G     147130          16.2%
ST      A>T     24703   2.7%
ST      C>A     42226   4.6%
ST      C>G     26512   2.9%
ST      C>T     175776          19.3%
ST      G>A     175072          19.2%
ST      G>C     26423   2.9%
ST      G>T     42417   4.7%
ST      T>A     24704   2.7%
ST      T>C     147936          16.3%
ST      T>G     38419   4.2%
# NS, Number of sites:
NS      total           911159
NS      ref match       901584          99.1%
NS      ref mismatch    8128    0.9%
NS      skipped         1447
NS      non-ACGT        1447
NS      non-SNP         0
NS      non-biallelic   0
```

1000G dataset checked against hg37

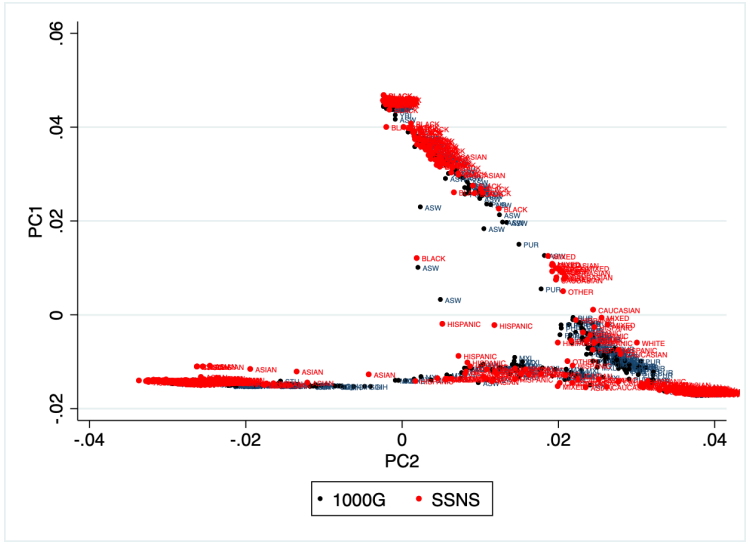
```
# SC, guessed strand convention
SC      TOP-compatible      0
SC      BOT-compatible      0
# ST, substitution types
ST      A>C     36095  4.1%
ST      A>G     142581         16.3%
ST      A>T     22265  2.5%
ST      C>A     39565  4.5%
ST      C>G     25512  2.9%
ST      C>T     171019         19.6%
ST      G>A     170458         19.5%
ST      G>C     25469  2.9%
ST      G>T     39531  4.5%
ST      T>A     22181  2.5%
ST      T>C     143301         16.4%
ST      T>G     35908  4.1%
# NS, Number of sites:
NS      total           878142
NS      ref match       873524          100.0%
NS      ref mismatch    361     0.0%
NS      skipped         4257
NS      non-ACGT        0
NS      non-SNP         4257
NS      non-biallelic   0
```

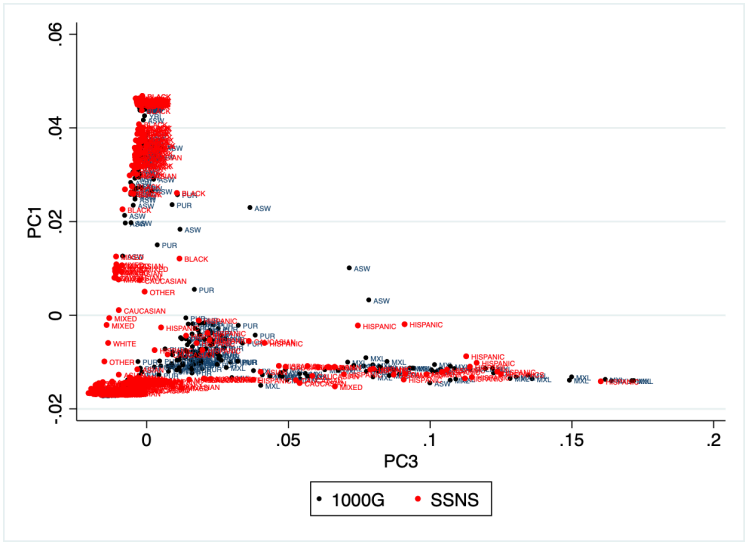


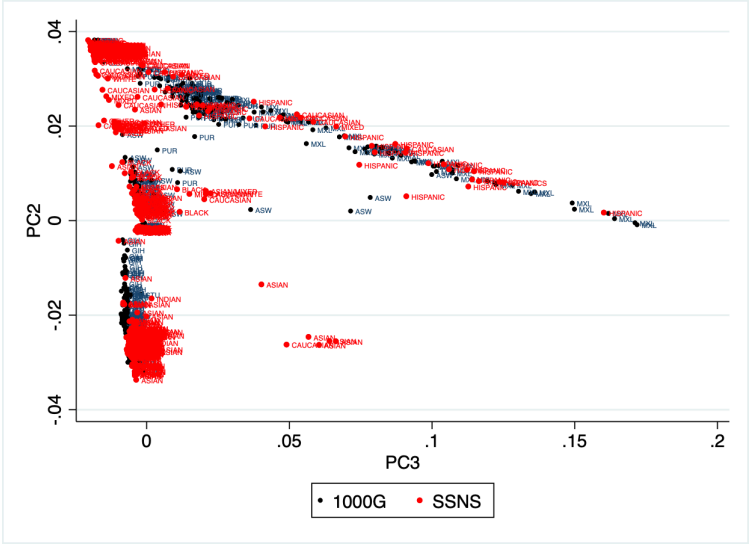




PCA check of SSNS samples vs 1000G pre-selection of controls







PCA check of SSNS samples vs 1000G: post-selection of controls

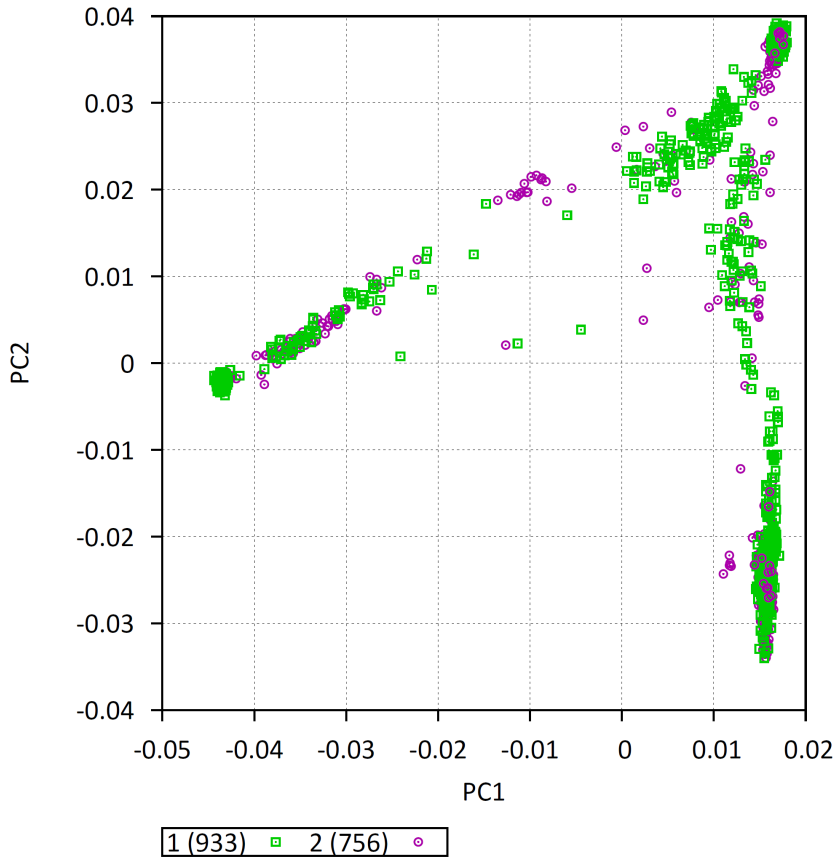
Post-selection of controls after PCA check

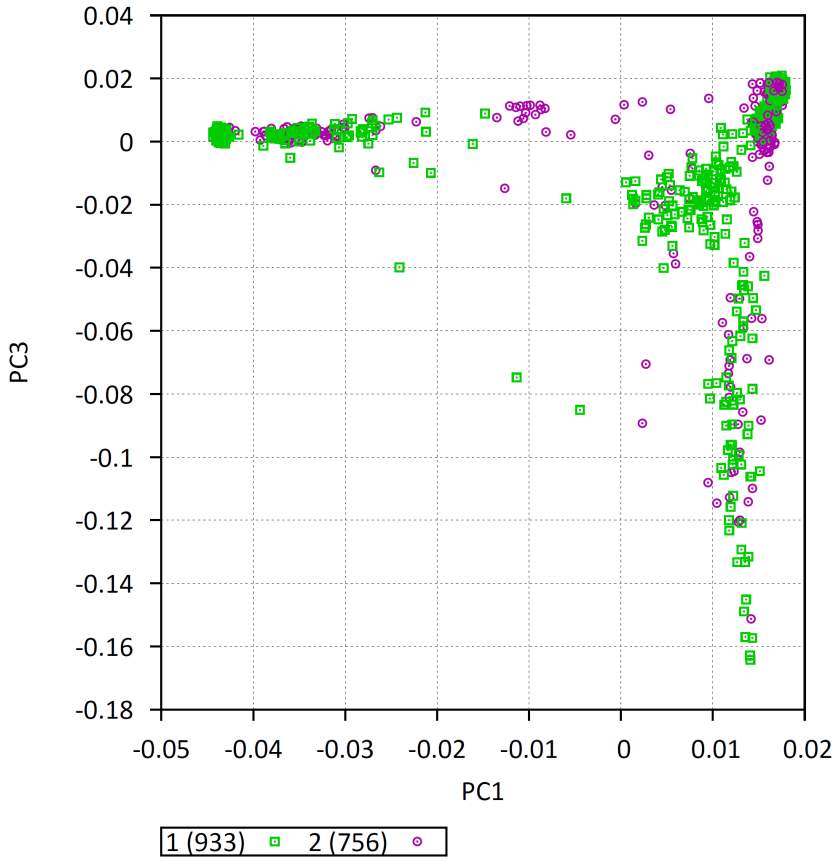
Group	Cases	Controls	Comments
Asian	383	307	Controls are SAS [STU, ITU, GIH]
European	162	190	Controls are EUR [CEU, GBR]
African	154	268	Controls are AFR [YRI, ESN, ASW]
Admixed	57	168	Controls are AMR [MXL, PUR]
All	756	933	

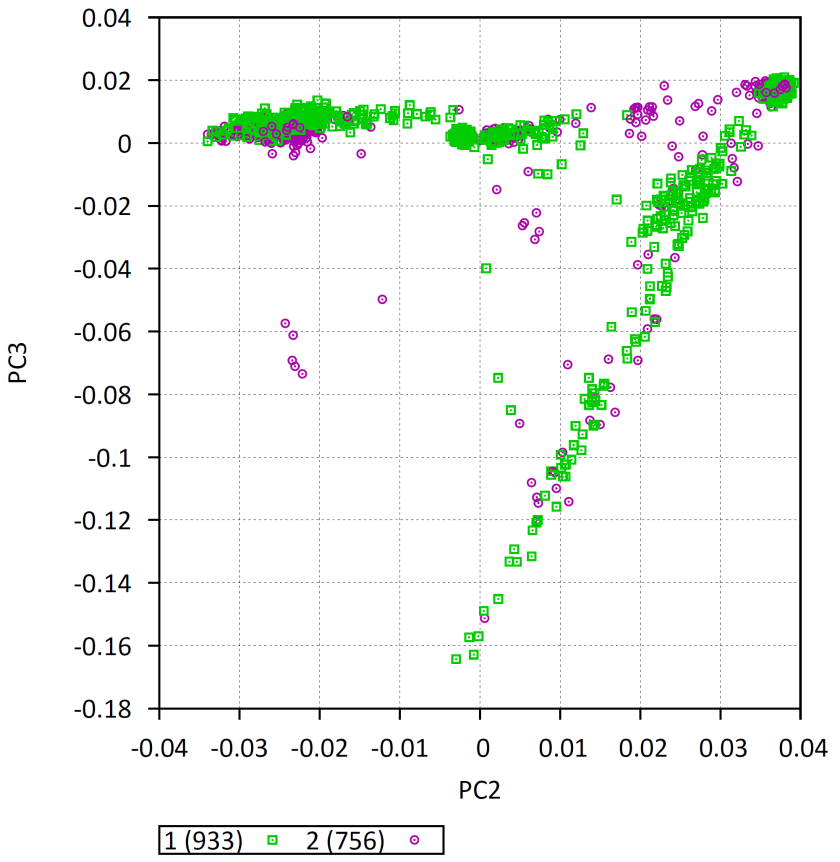
SSNS GWAS: Exploratory data analysis

- Pre-imputation QC with Rayner's HRC-1000G imputation preparation and checking toolbox and appropriate ancestry specific reference
- Genotype data imputation using Michigan Imputation Server
- 1000 Genomes reference imputation panel
- Variants imputed with high quality - R^2 >= 0.3 - analyzed

PCA plot: all cases and controls



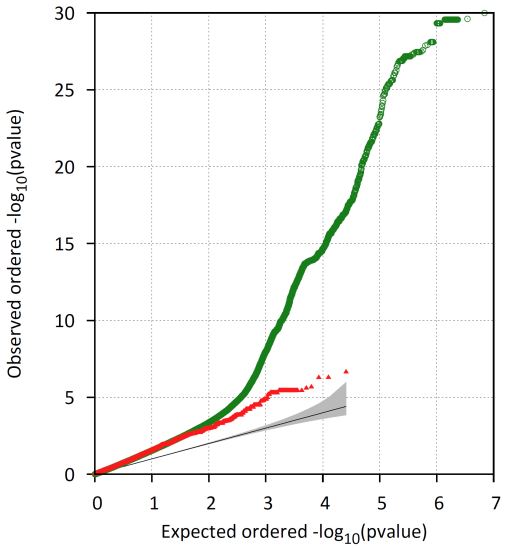




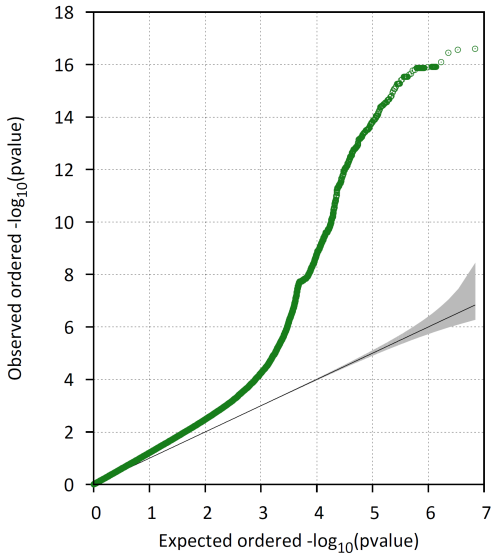
Association analysis by ancestry group

Association analysis per pop

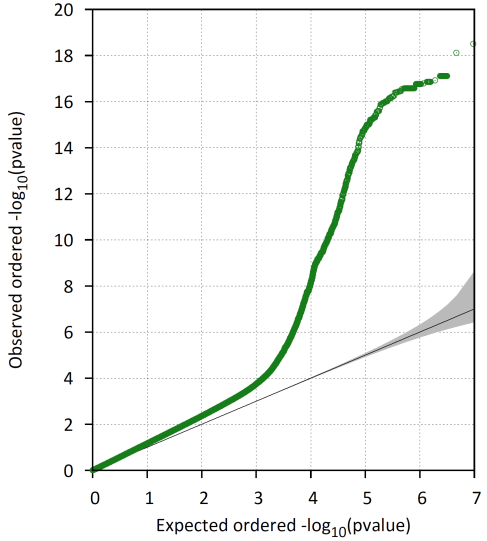
[MAF >= 0.05, no covariates, logistic model, Wald test]



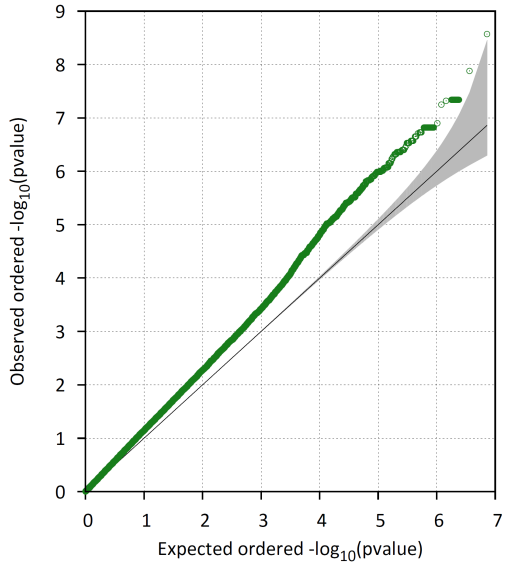
Asian



European



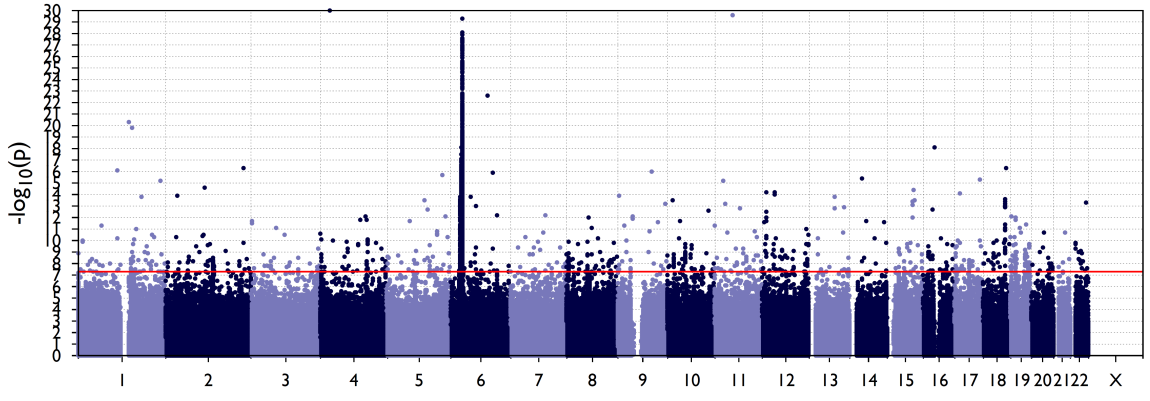
African



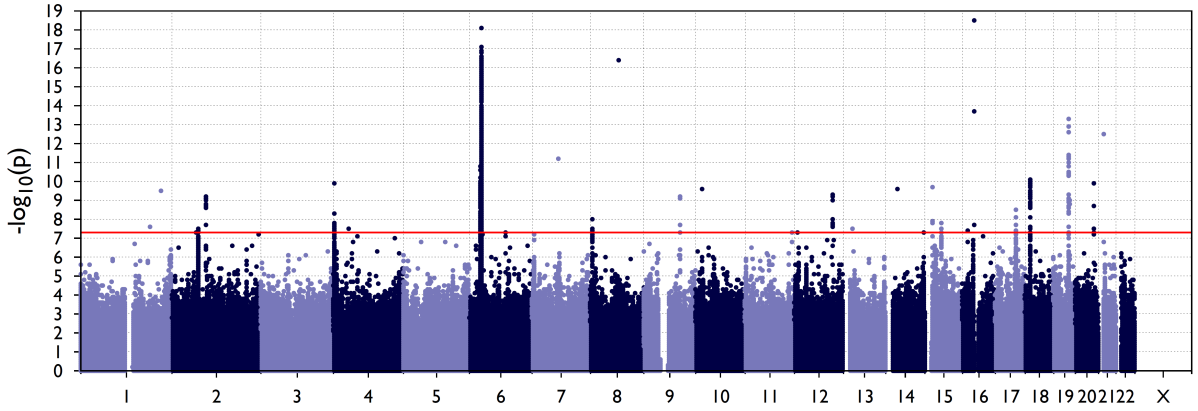
Mixed

Association analysis per pop

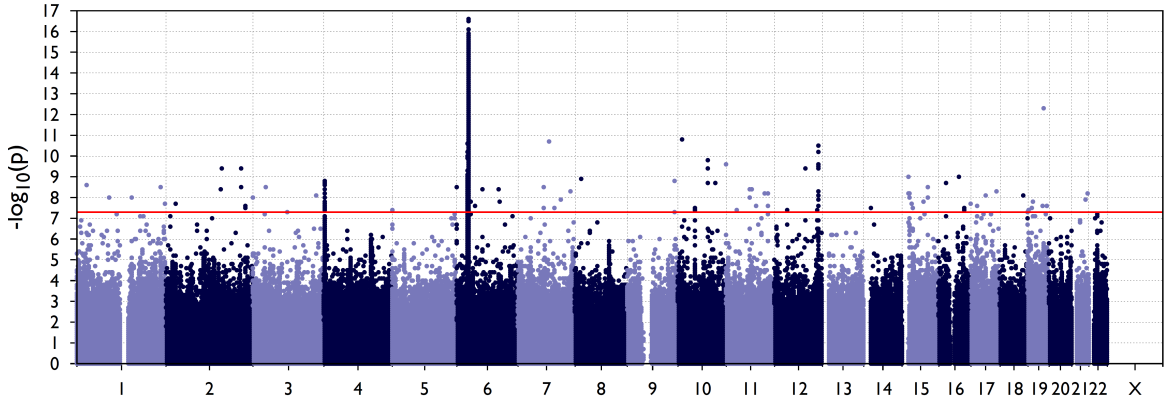
[MAF >= 0.05, no covariates, logistic model, Wald test]



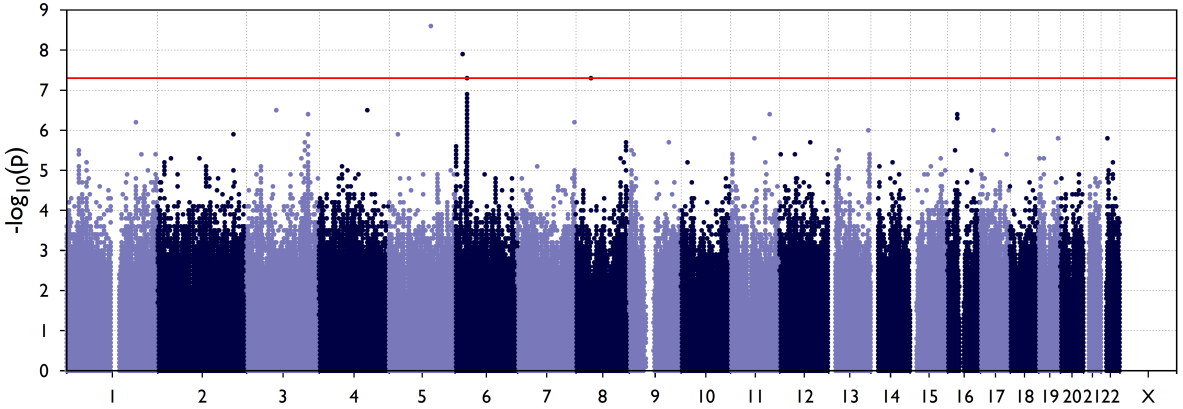
Asian



African



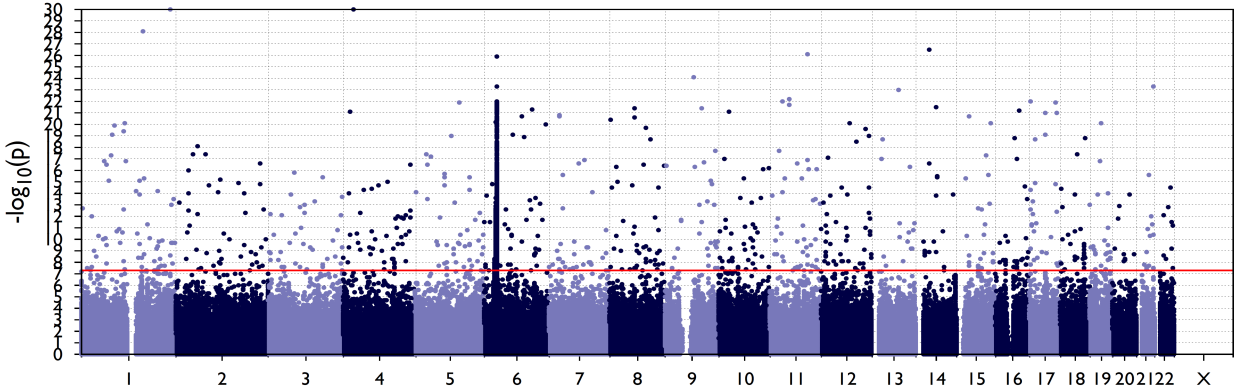
European

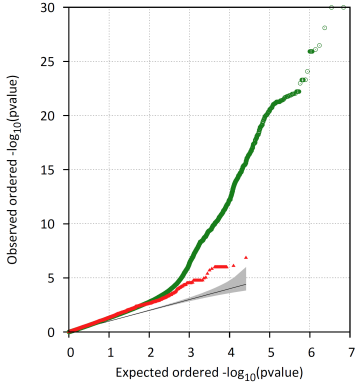


Mixed

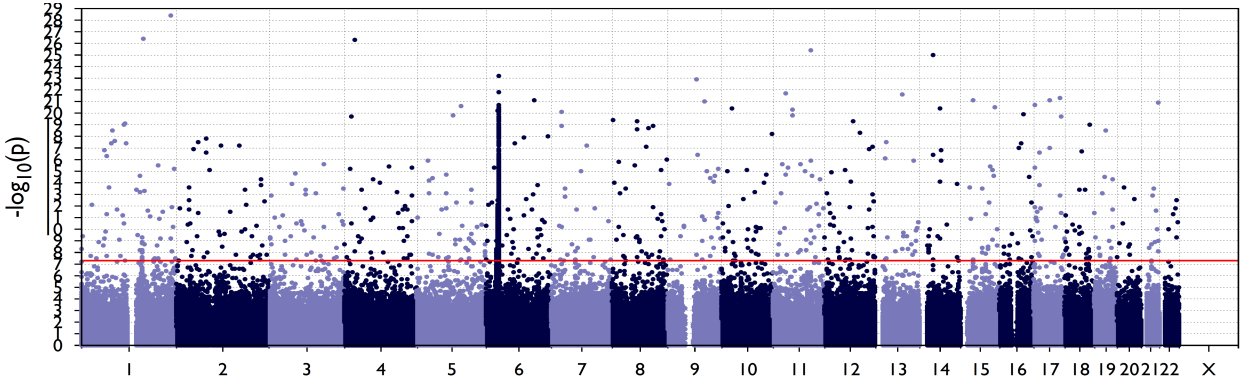
Asian ancestry: Association [Wald test]

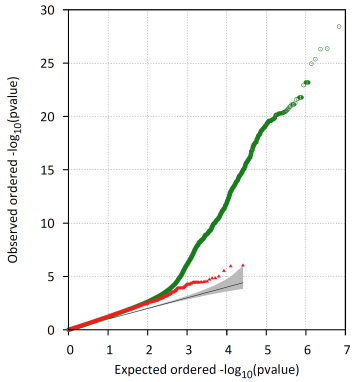
Adjusted for 1 PC



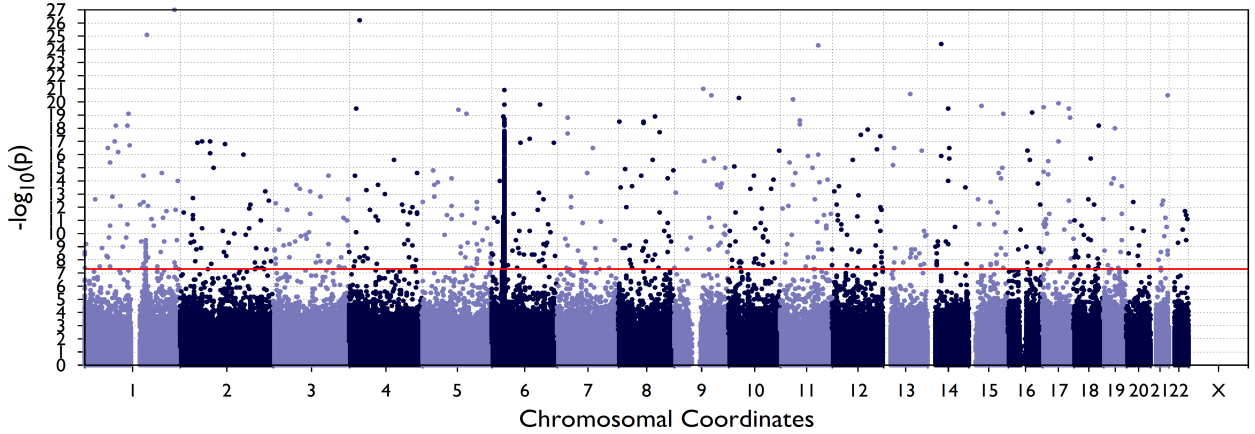


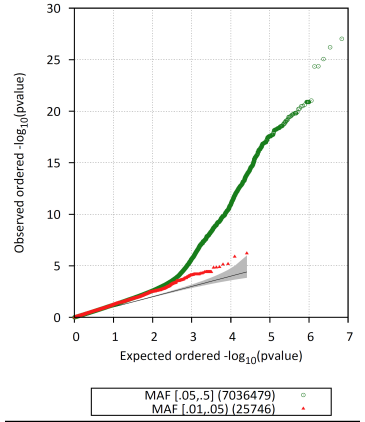
3 PCs





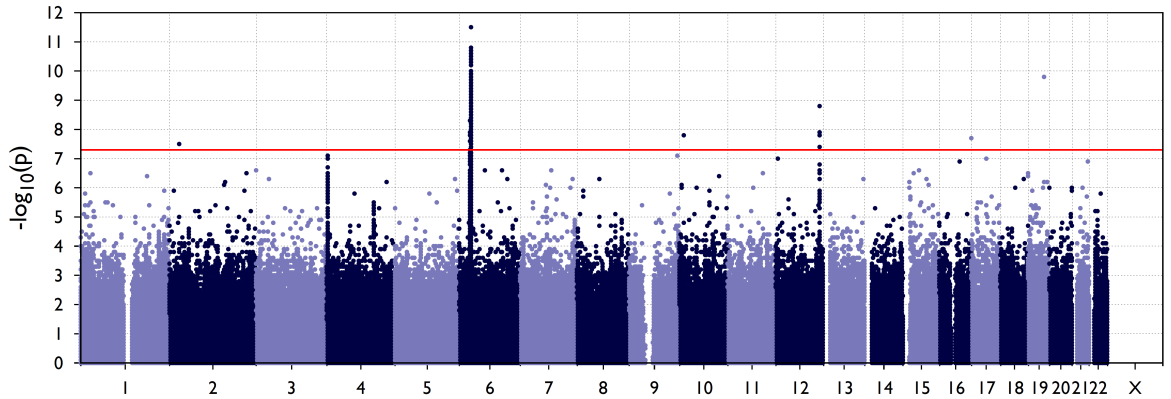
5 PCs

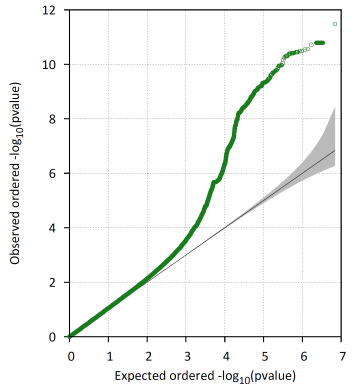




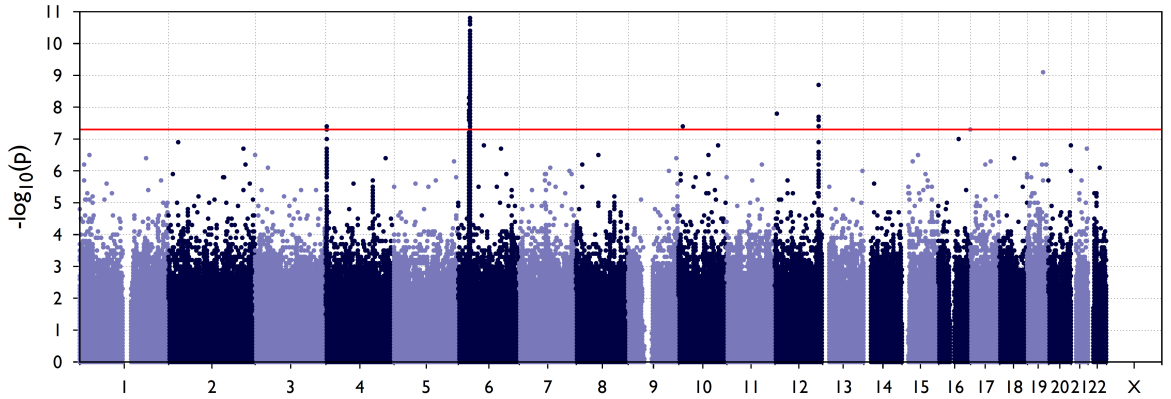
European ancestry: Association [Wald test]

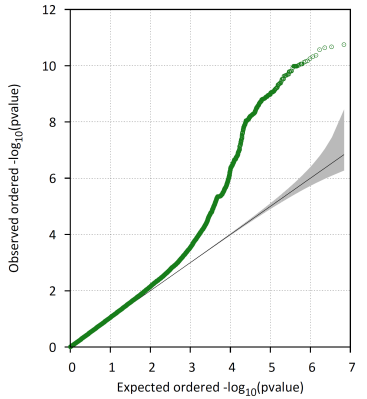
Adjusted for 1 PC



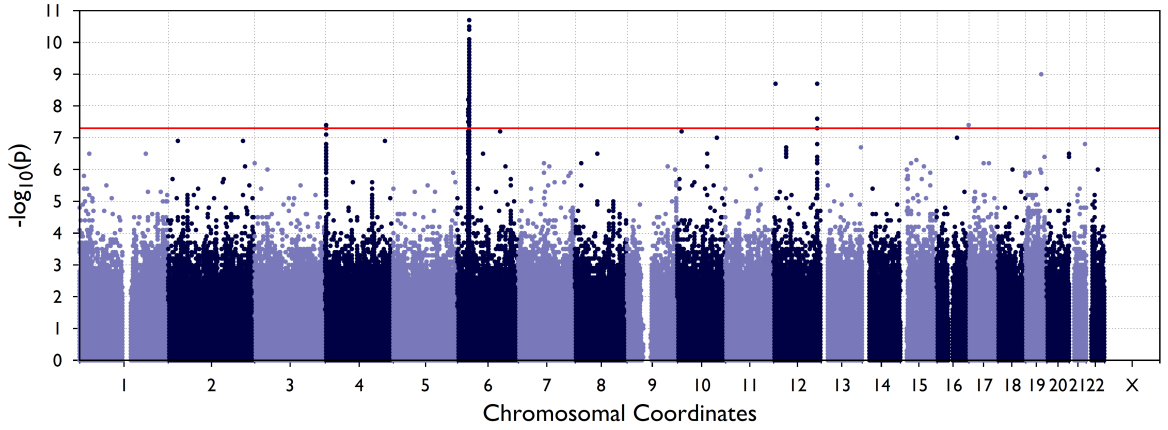


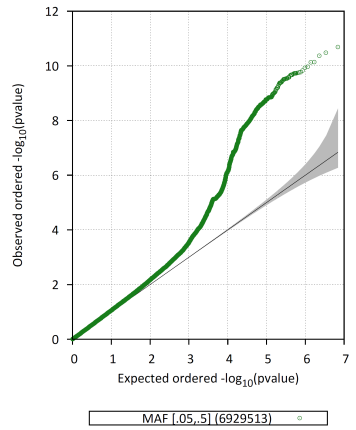
3 PCs





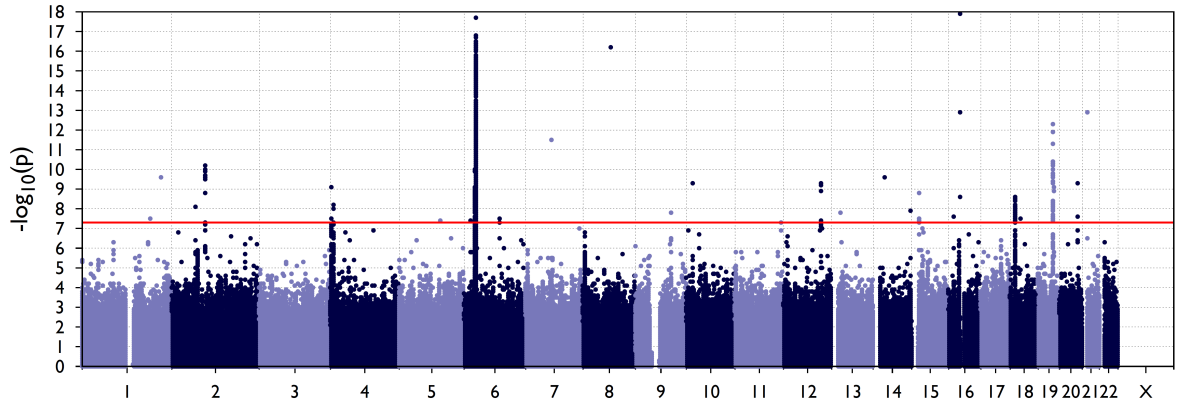
5 PCs

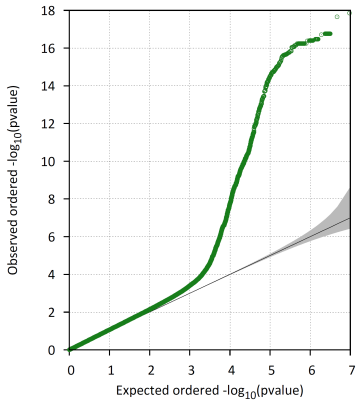




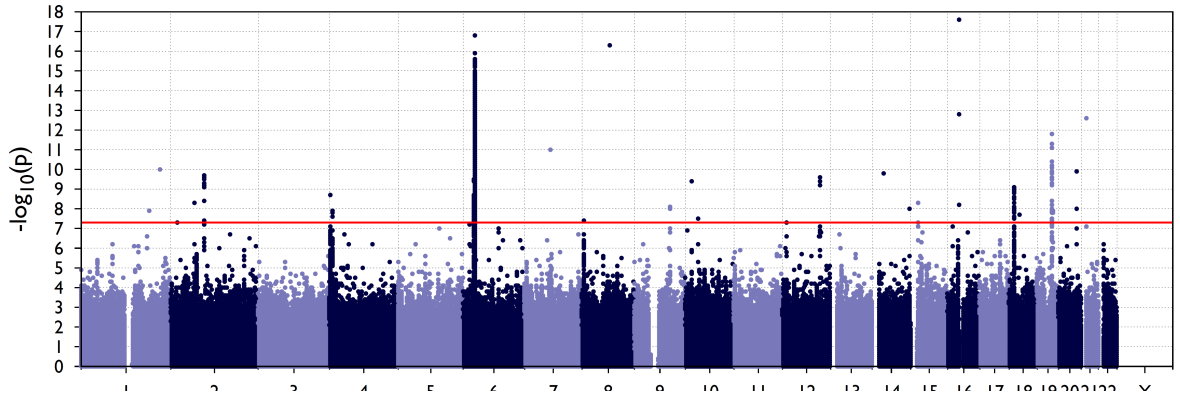
African ancestry: Association [Wald test]

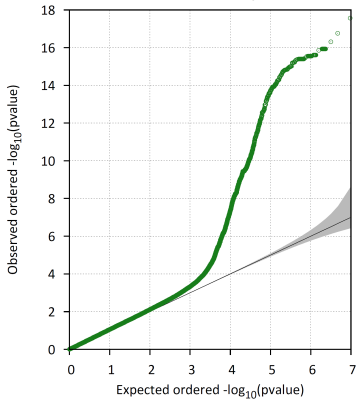
Adjusted for 1 PC



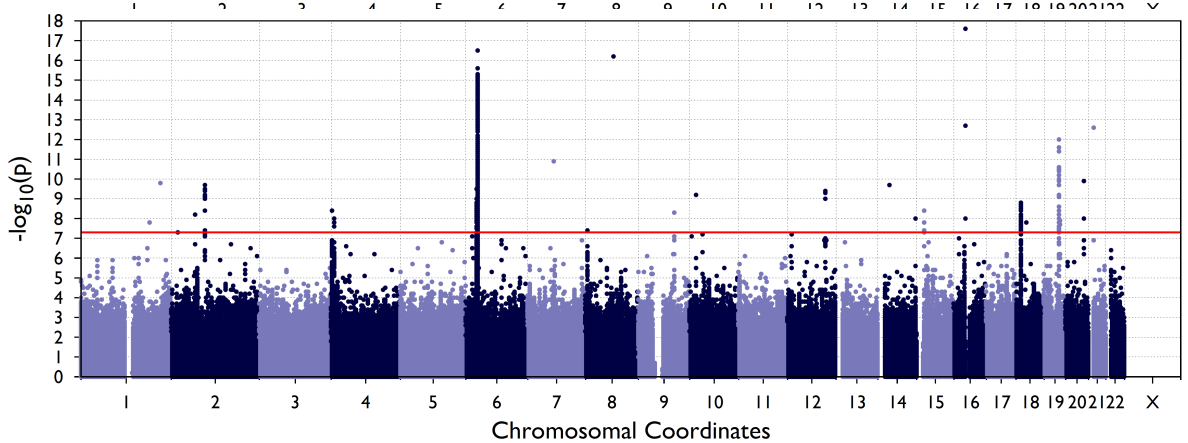


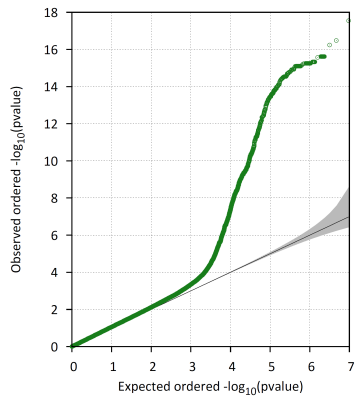
3 PCs





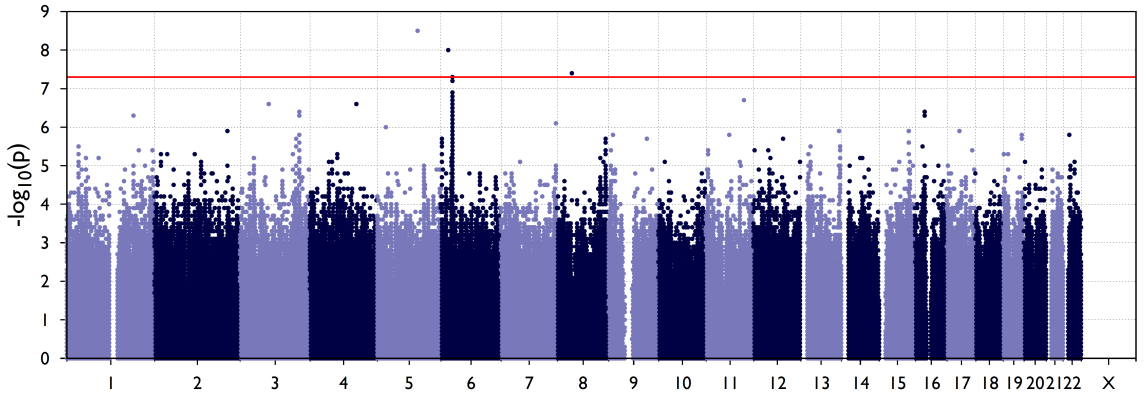
5 PCs

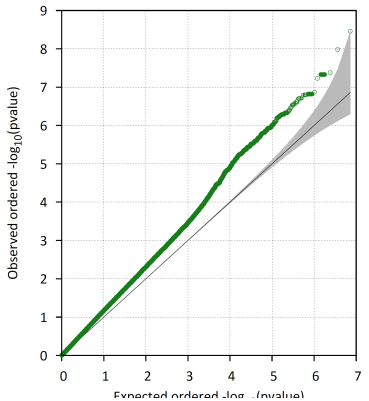




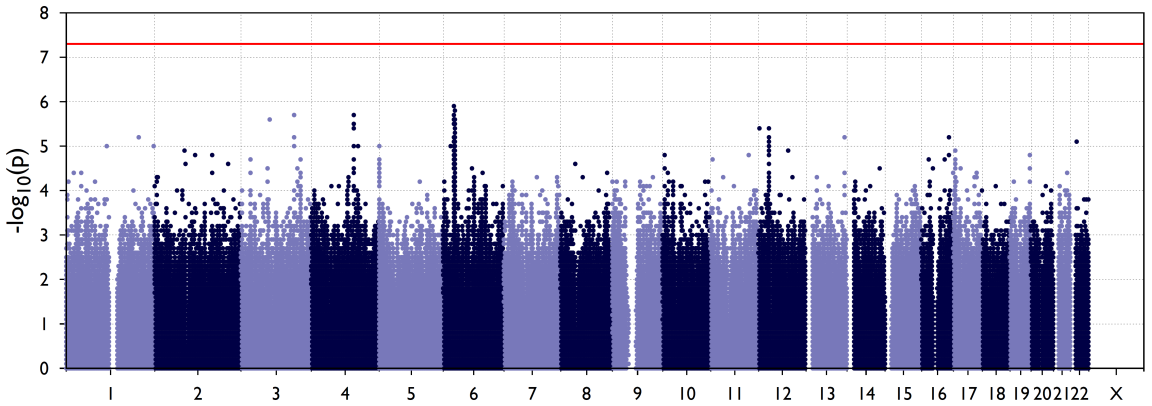
Mixed ancestry: Association [Wald test]

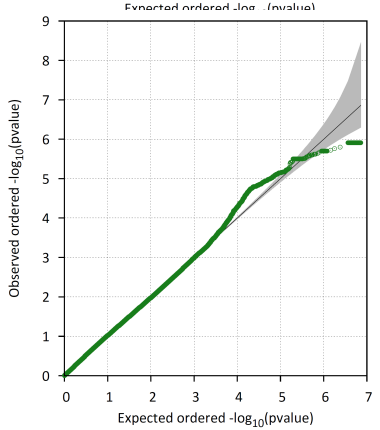
Adjusted for 1 PC



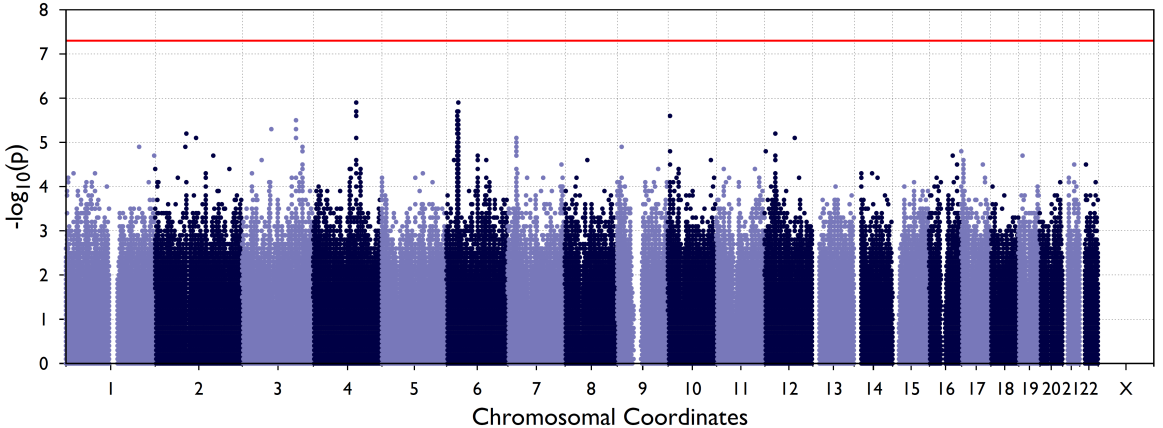


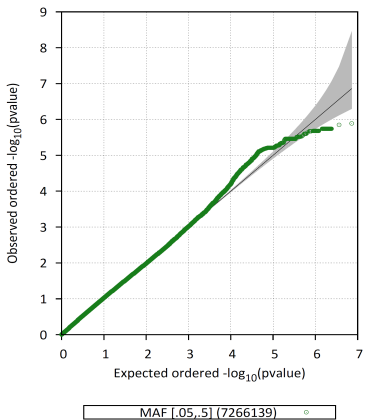
3 PCs





5 PCs

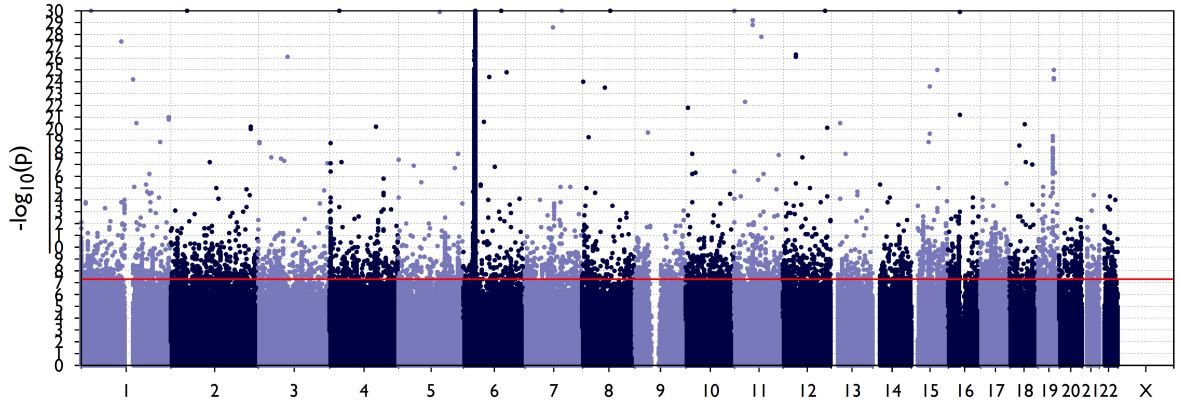


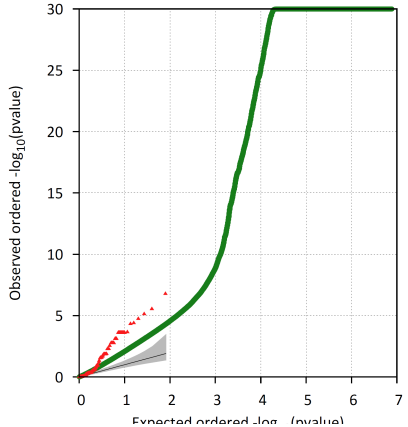


Joint analysis

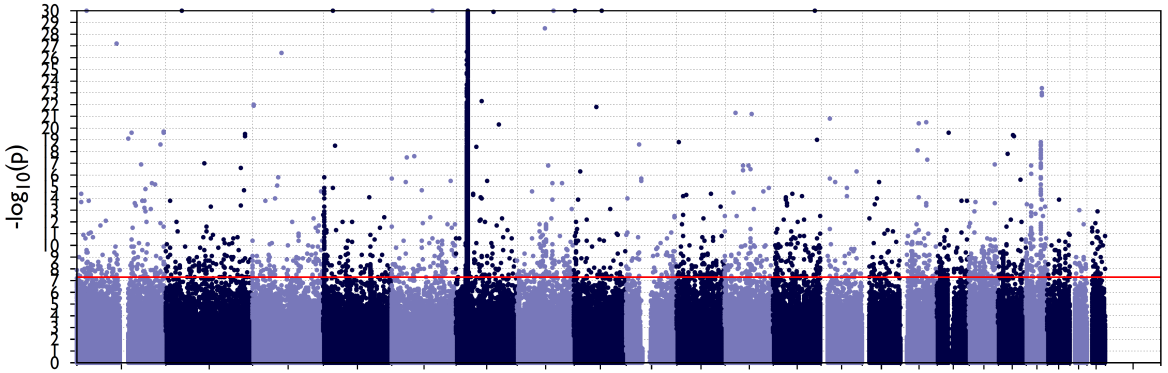
Exploratory joint analysis controlling for PCs

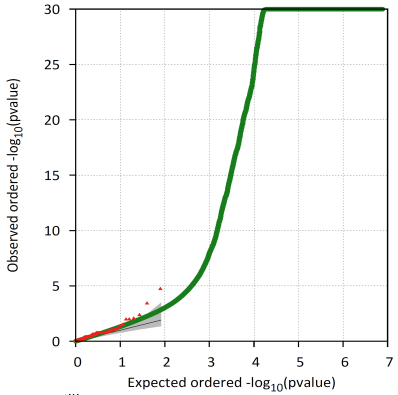
1 PC



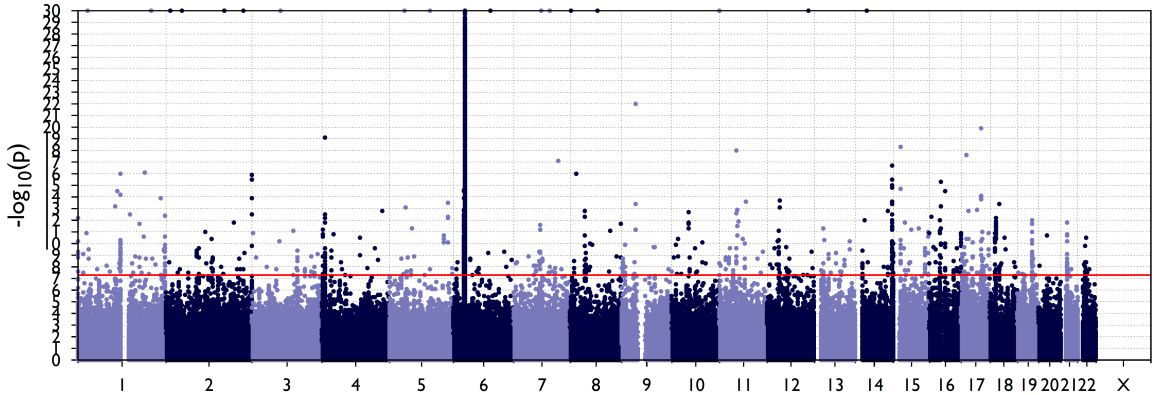


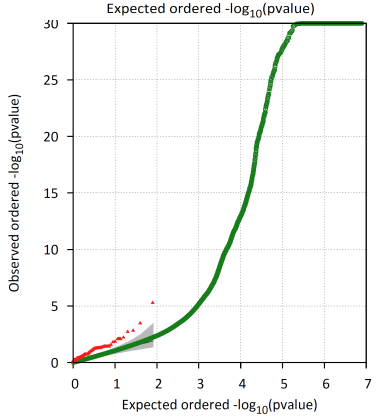
3 PCs





5 PCs

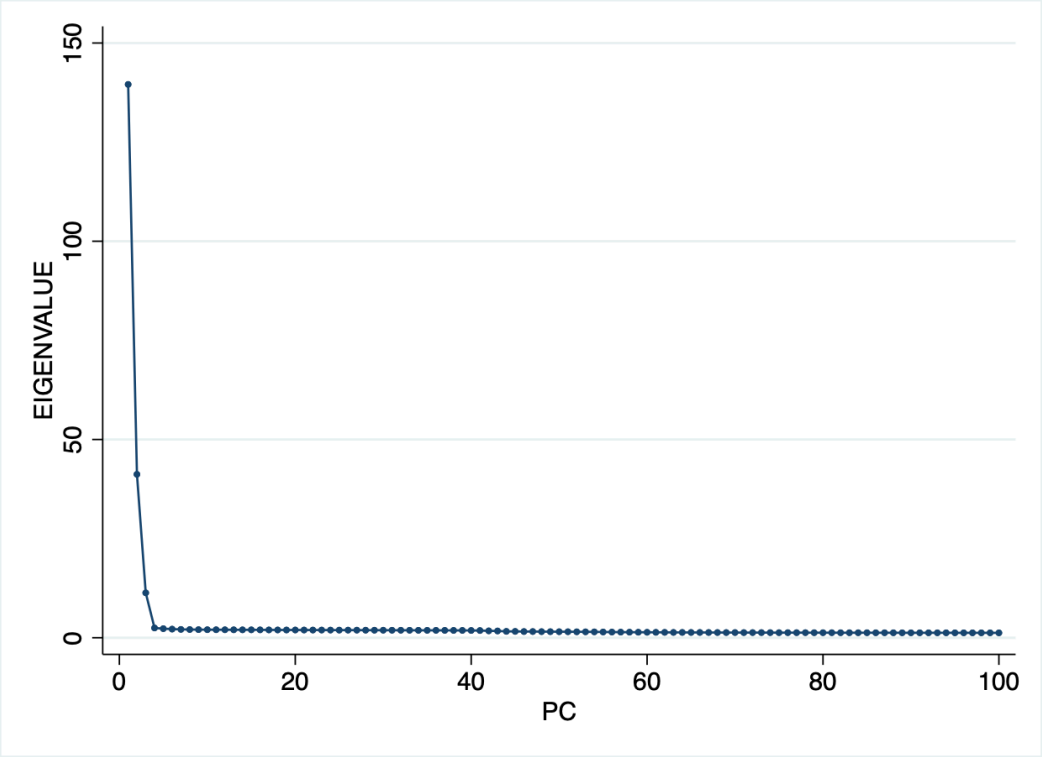




Number of significant PCs

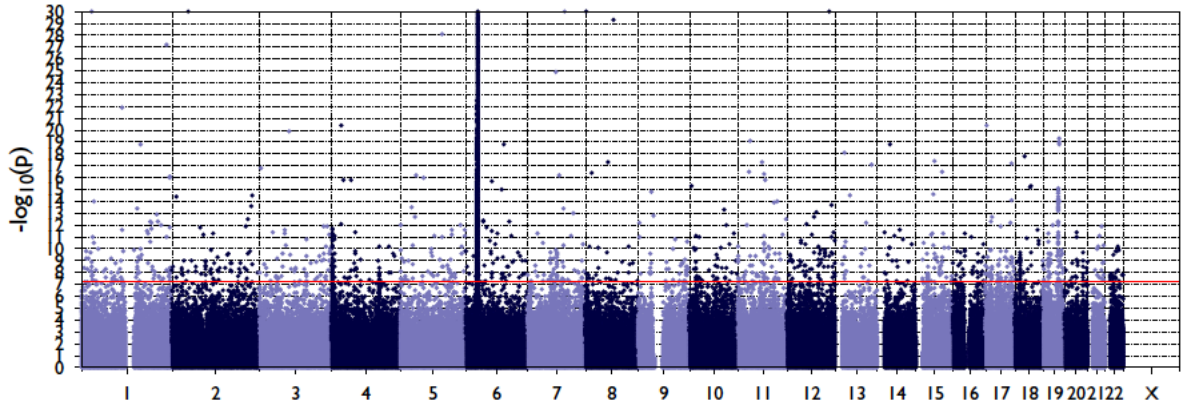
- Formal analysis of number of significant PCs
  - 4 [Minimum Average Partial test - Shriner, 2012]
  - 5 [Tracy-Widom test – Tracy & Widom, 1994; Patterson et al, 2006)
- Distribution of eigenvalues and scree plot

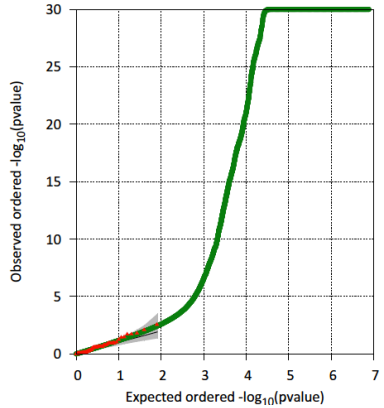
PC	EIGENVALUE
PC1	139.55840
PC2	41.21807
PC3	11.32979
PC4	2.46755
PC5	2.29273
PC6	2.17892
PC7	2.10505
PC8	2.07520
PC9	2.04779
PC10	2.03696



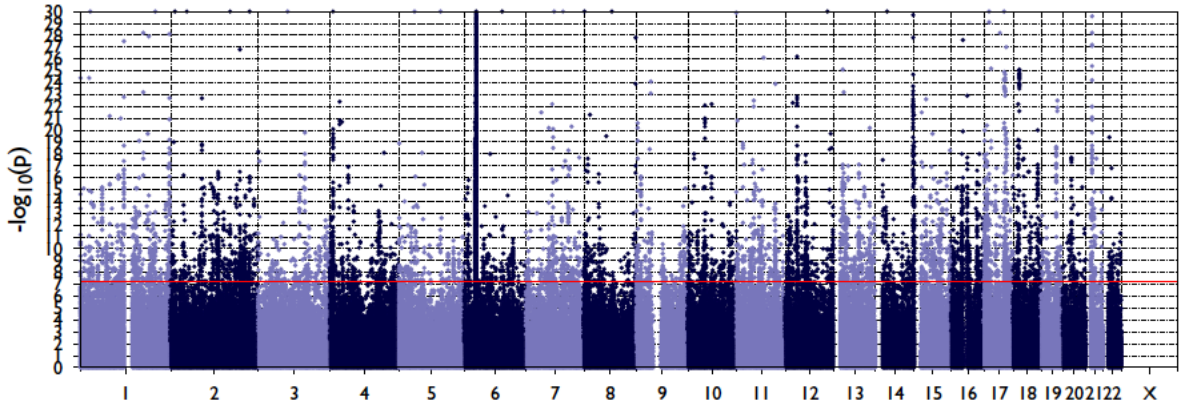
Exploratory analysis controlling for GRM and 4 PCs

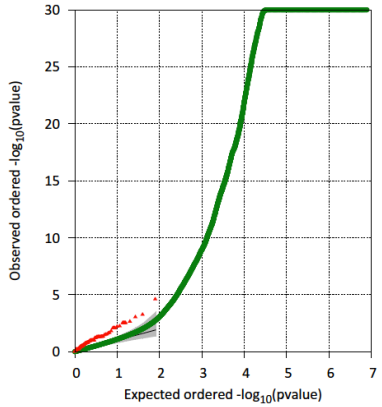
4 PCs only
Wald test



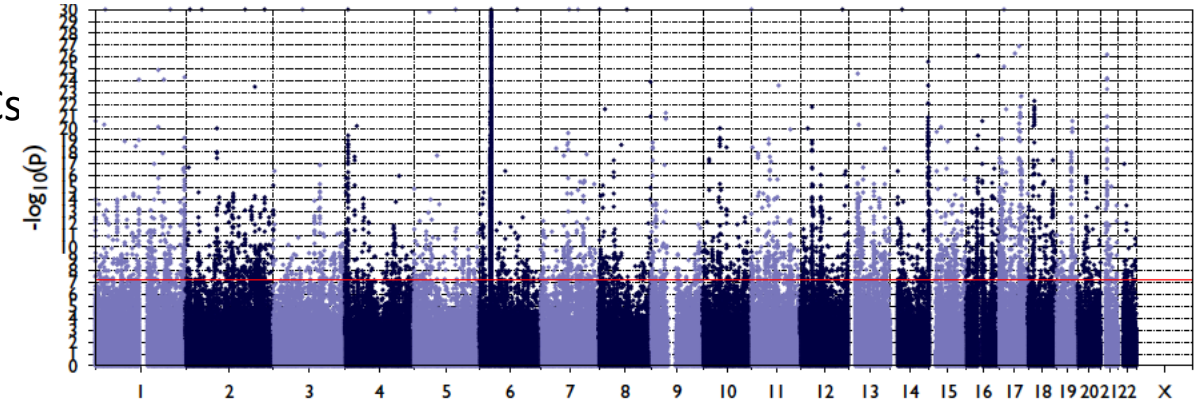


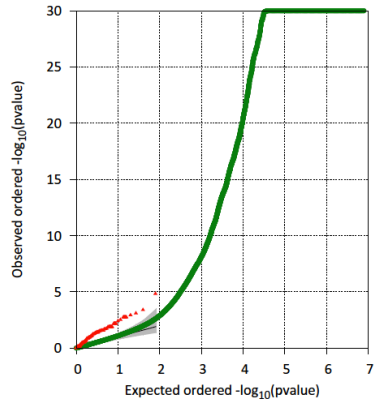
GRM only
emmax



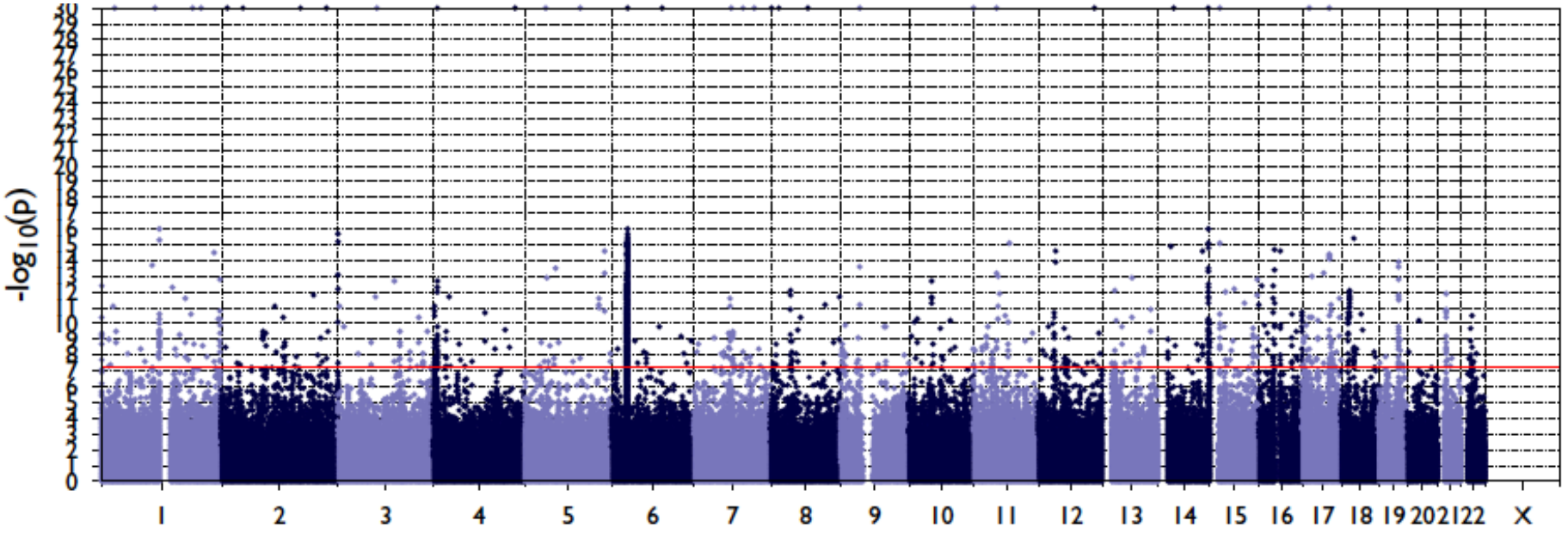


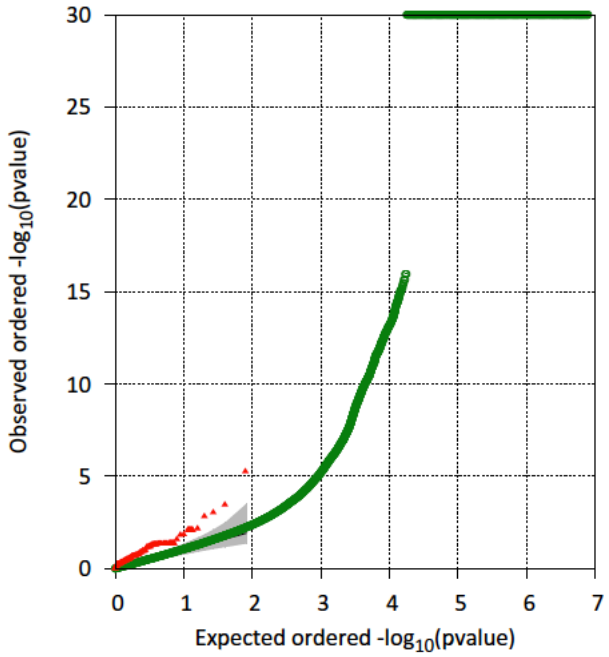
GRM and 4 PCs
emmax





Analysis using Firth bias-corrected logistic model [adjusting for 4 PCs]





Summary points

- Successful selection of 1000G controls matched for ancestry
- As expected, there is significant population stratification
- Per population group analysis shows group specific hits outside chr. 6 locus [SSNS Asian analysis needs checking]
- Number of significant PCs to adjust for in association tests 4-5 but first 5 PCs gives better QQ and Manhattan plots than any model including kinship matrix
- Joint analysis still not satisfactory even after adjustment for computed kinship matrix and PCs

Next steps

- Troubleshooting of SSNS Asian analysis
- Explore joint analysis further
- Meta-analysis of individual population group summary statistics
- Analysis with mixed logistic model [e.g. GMMAT]
- Analysis with method that accounts for case-control imbalance (SAIGE)
- Annotation of top hits
- *In silico* replication of novel hits