Quality control procedures for Genome-Wide SNP data

Originally presented at the BroadE Workshop on Statistical Genetics

June 8th, 2015

**updated July 2020*

Daniel Howrigan

BROAD INSTITUTE

MGH 1811

CHGR

STANLEY CENTER FOR PSYCHIATRIC RESEARCH

Why QC?

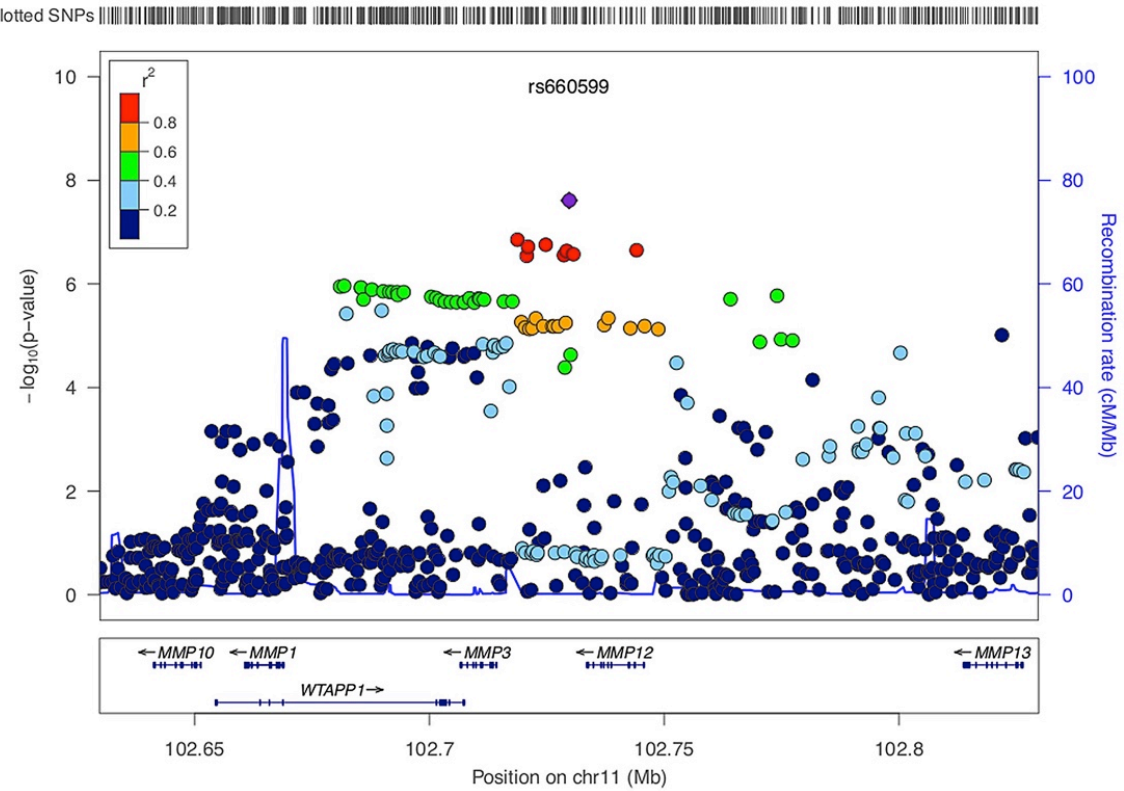
- Identifying real genetic associations relies on high quality data
- Even a low rate of error is detrimental
- Example:
  - 1 million markers are tested for association
  - 0.1% of markers are poorly genotyped
  - Up to 1,000 significant markers will show a false-positive association

SNP genotyping isn't a perfect science…

…but could strict QC throw away true associations?

- Properties of linkage disequilibrium reduce the loss of signal sensitivity when removing SNPs

- Strict multiple testing correction requires very large samples - no single sample will drive a signal



Levels of QC

- Broad strategies (first pass)
  - Poorly genotyped samples / SNP markers
  - Related or duplicated samples (population-based data)
  - Deviations from Hardy-Weinberg

- Complex strategies
  - Batch effects
  - Quality differences between samples
  - Contamination
  - Genetic considerations

  …

Sample QC

- Poorly genotyped individuals
  - Indications of sample mix-up (sex check or ancestry match)
  - Poor quality DNA (high number of failed SNP calls)
  - Contaminated DNA (unusual levels of heterozygosity)

- Related individuals
  - Family-based and population-based samples require different experimental designs
  - Related individuals can bias test statistics across the whole-genome
  - In family-based association: Mendelian errors used as QC

SNP QC

- Poorly genotyped SNPs
  - Poor primer design / nonspecific DNA binding (high number of failed SNP calls)
  - Poor clustering of genotype intensities (deviation from HWE)
  - Uninformative SNPs (too rare or mono-allelic)
- Follow-up on association signals
  - No QC protocol will eliminate all instances of genotyping error
  - Important to re-analyze original intensity of significant associations (whenever possible)

QC Steps overview

- Sample (or individual) QC
  - Confirming sex from genotype
  - Poorly genotyped (high SNP Missingness)
  - Deviation in heterozygosity
  - Related or duplicated samples
  - Population outliers*

- SNP QC
  - Poorly genotyped (high SNP Missingness)
  - Rare or mono-allelic SNPs
  - Deviation from Hardy-Weinberg Equilibrium (HWE)

** detailed in upcoming section*

Papers detailing GWAS QC

- Anderson CA, et al. Data quality control in genetic case-control association studies. *Nature Protocols* 5, 1564–1573 (2010)*

- Turner S, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet.* Jan; Chapter 1: Unit1.19 (2011)

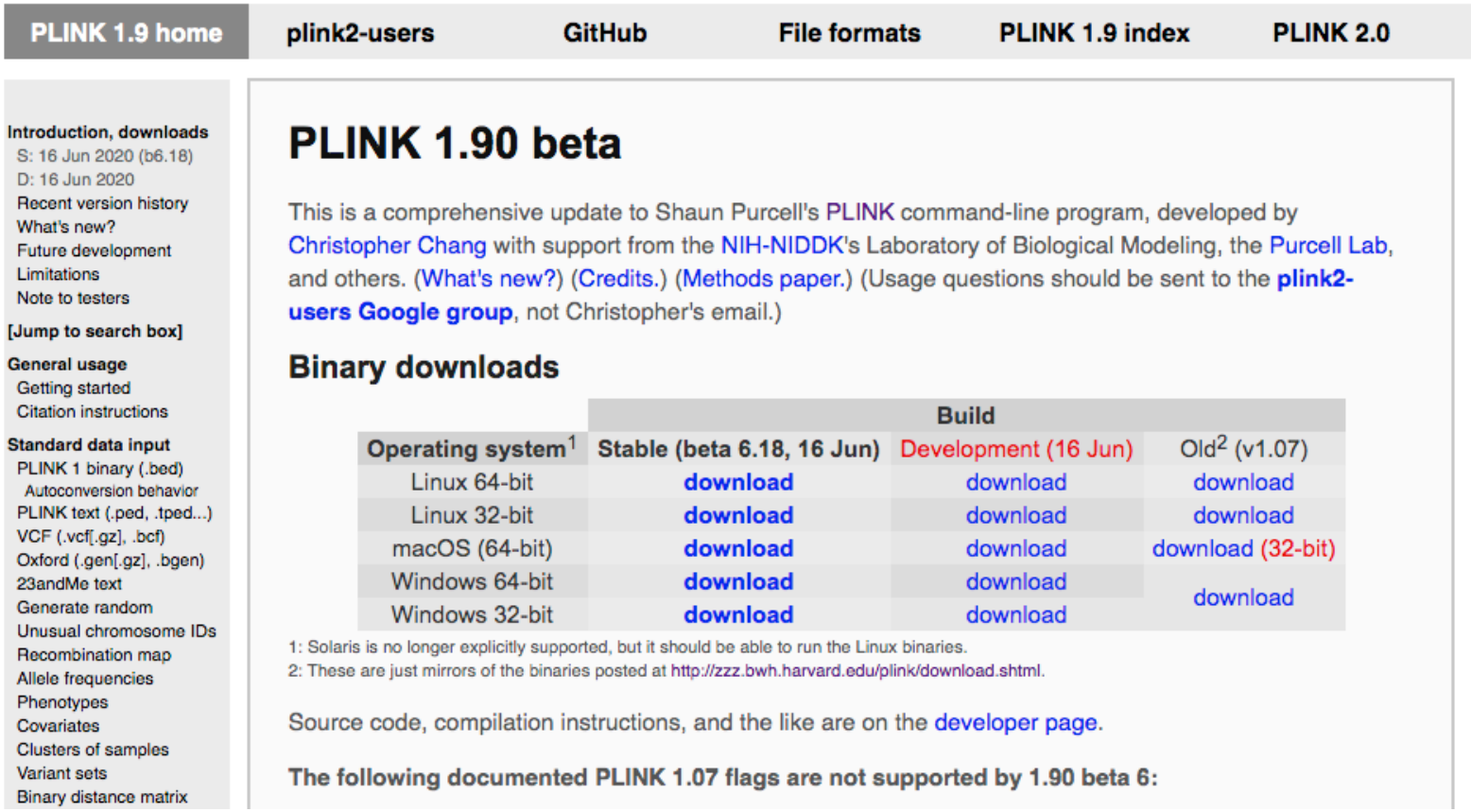
- Winkler TW, et al. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* 9, 1192–1212 (2014)

** paper available in tutorial directory*

GWAS QC Practical

Primary genetic software used - PLINK 1.9

https://www.cog-genomics.org/plink/1.9/



PLINK 1.9 home | plink2-users | GitHub | File formats | PLINK 1.9 index | PLINK 2.0

Introduction, downloads
S: 16 Jun 2020 (b6.18)
D: 16 Jun 2020
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Getting started
Citation instructions
Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf[.gz], .bcf)
Oxford (.gen[.gz], .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix

PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab, and others. (What's new?) (Credits.) (Methods paper.) (Usage questions should be sent to the **plink2-users Google group**, not Christopher's email.)

Binary downloads

	Build		
Operating system[1]	Stable (beta 6.18, 16 Jun)	Development (16 Jun)	Old[2] (v1.07)
Linux 64-bit	**download**	download	download
Linux 32-bit	**download**	download	download
macOS (64-bit)	**download**	download	download (32-bit)
Windows 64-bit	**download**	download	download
Windows 32-bit	**download**	download	

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at http://zzz.bwh.harvard.edu/plink/download.shtml.

Source code, compilation instructions, and the like are on the developer page.

The following documented PLINK 1.07 flags are not supported by 1.90 beta 6:

Getting started..

Log onto Broad server

```
ssh [username]@login.broadinstitute.org
```

Open interactive node

```
use UGER
ish
```

Create graphing directory

```
chmod 755 .
mkdir -p ~/private_html
mkdir -p ~/private_html/atgu_workshop
```

View in web browser (username/password required)

www.internal.broadinstitute.org/~[username]

Getting started..

Create working directory

```
cd ~
mkdir -p atgu_workshop/
cd atgu_workshop/
```

Create link to hapmapEA PLINK data

```
ln -s /web/personal/howrigan/workshop/hapmapEA.bed hapmapEA.bed
ln -s /web/personal/howrigan/workshop/hapmapEA.bim hapmapEA.bim
ln -s /web/personal/howrigan/workshop/hapmapEA.fam hapmapEA.fam
```

Copy over R scripts

```
cp /web/personal/howrigan/workshop/*R .
```

Ensure that PLINK and R are loaded

```
use PLINK
use R-3.3
```

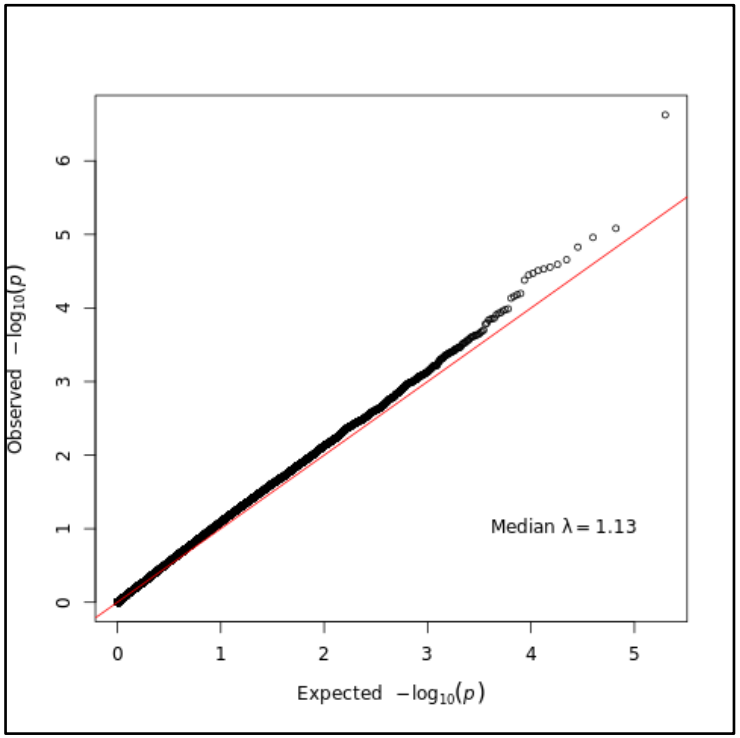
Run case/control association on raw data

```
plink \
--bfile hapmapEA \
--logistic \
--out hapmapEA
```

Create QQ plot and view

```
Rscript QQplot.R hapmapEA.assoc.logistic
```

```
mv hapmapEA.assoc.logistic_QQ.png ~/private_html/atgu_workshop
```



QC Steps overview

- Sample (or individual) QC
  - Confirming sex from genotype
  - Poorly genotyped (high SNP Missingness)
  - Deviation in heterozygosity
  - Related or duplicated samples
  - Population outliers*

- SNP QC
  - Poorly genotyped (high SNP Missingness)
  - Rare or mono-allelic SNPs
  - Deviation from Hardy-Weinberg Equilibrium (HWE)

Checking reported sex with genotypic sex

```
Command line:

plink --bfile hapmapEA \
--check-sex \
--out hapmapEA
```

hapmapEA.log file output (NOTE that no X chromosome SNPs exist here!)

```
PLINK v1.90p 64-bit (29 May 2015)          https://www.cog-genomics.org/plink2
(C) 2005-2015 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hapmapEA.log.
Options in effect:
  --bfile /humgen/gsa-hphome1/sek/gina/br
  --check-sex
  --out hapmapEA

7982 MB RAM detected; reserving 3991 M
Allocated 125 MB successfully, after large
100310 variants loaded from .bim file.
200 people (101 males, 99 females) loade
200 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculati
Before main variant filters, 200 founders a
Calculating allele frequencies... done.
Total genotyping rate is 0.996374.
100310 variants and 200 people pass filter
Among remaining phenotypes, 100 are ca
Error: --check-sex/--impute-sex requires
locus.
```

example of .sexcheck output

FID	IID	PEDSEX	SNPSEX	STATUS	F
T304	T30411	1	1	OK	0.9857
A0641C	06410021C	1	1	OK	0.9841
T06013	T2601310	2	2	OK	-0.06164
T01533	T2153321	1	1	OK	0.9841
T330	T33021	1	1	OK	0.9867
T191	T19120	2	2	OK	0.01155
T329	T32911	1	1	OK	0.9839
T07981	T2798111	1	1	OK	0.9822
A0601C	06010021C	1	1	OK	0.9858
A1008C	10080011C	1	1	OK	0.9817
A0880C	08800331C	1	1	OK	0.9818
T00894	T2089420	2	2	OK	0.01927
A0701C	07010011C	1	1	OK	0.9807
T02911	T2291121	1	1	OK	0.9851
T00588	T2058811	1	2	PROBLEM	-0.3396
A0805C	08050031C	1	1	OK	0.9821
T07755	T2775520	2	2	OK	-0.09906
T03676	T2367611	1	1	OK	0.9845
T082	T08220	2	1	PROBLEM	0.9833

QC Steps overview

- Sample (or individual) QC
  - Confirming sex from genotype
  - Poorly genotyped (high SNP Missingness)
  - Deviation in heterozygosity
  - Related or duplicated samples
  - Population outliers*

- SNP QC
  - Poorly genotyped (high SNP Missingness)
  - Rare or mono-allelic SNPs
  - Deviation from Hardy-Weinberg Equilibrium (HWE)

Checking the genotyping rate across samples and SNPs

```
plink --bfile hapmapEA \
--missing \
--out hapmapEA
```

```
head hapmapEA.imiss
```

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
NA20505	NA20505	N	122	100310	0.001216
NA20504	NA20504	N	1406	100310	0.01402
NA20506	NA20506	N	204	100310	0.002034
NA20502	NA20502	N	847	100310	0.008444
NA20528	NA20528	N	219	100310	0.002183
NA20531	NA20531	N	96	100310	0.000957
NA20534	NA20534	N	338	100310	0.00337
NA20535	NA20535	N	182	100310	0.001814
NA20586	NA20586	N	214	100310	0.002133

Per sample

```
head hapmapEA.lmiss
```

CHR	SNP	N_MISS	N_GENO	F_MISS
1	rs12565286	6	200	0.03
1	rs12124819	8	200	0.04
1	rs4970383	0	200	0
1	rs13303118	0	200	0
1	rs35940137	0	200	0
1	rs2465136	1	200	0.005
1	rs2488991	0	200	0
1	rs3766192	0	200	0
1	rs10907177	0	200	0

Per SNP marker

Checking the heterozygosity rate across samples

```
plink --bfile hapmapEA \
--het \
--out hapmapEA
```

```
head hapmapEA.het
```

FID	IID	O(HOM)	E(HOM)	N(NM)	F
NA20505	NA20505	71394	7.124e+04	100188	0.005246
NA20504	NA20504	69260	7.027e+04	98904	-0.03512
NA20506	NA20506	71344	7.118e+04	100106	0.005509
NA20502	NA20502	71086	7.068e+04	99463	0.01405
NA20528	NA20528	70901	7.117e+04	100091	-0.009297
NA20531	NA20531	71064	7.126e+04	100214	-0.00685
NA20534	NA20534	71197	7.108e+04	99972	0.003942
NA20535	NA20535	71263	7.12e+04	100128	0.002156
NA20586	NA20586	71089	7.119e+04	100096	-0.003495

AWK one-liner to get outlier samples from F statistic

```
awk '$6>0.03 {print $1,$2}' hapmapEA.het | awk 'NR>1' > het_rem.txt
```

Running sample QC using genotyping rate and heterozygosity in R

```
Rscript imiss-vs-het.R
```

```
## === imiss-vs-het.R

imiss <- read.table('hapmapEA.imiss',h=T)
het <- read.table('hapmapEA.het',h=T)

het$P_HET <- (het$N.NM. - het$O.HOM.) / het$N.NM.
upper_3sd <- mean(het$P_HET) + 3*sd(het$P_HET)
lower_3sd <- mean(het$P_HET) - 3*sd(het$P_HET)

xlabels <- c('1e-4','0.001','0.01','0.1','1')

pdf('imiss-vs-het.pdf')
plot(log10(imiss$F_MISS),het$P_HET,xlab='log10(Proportion of missing
genotypes)',ylab='Proportion Heterozygous',xlim=c(-4,0),ylim=c(0,0.5))
axis(side=1,labels=F)
mtext(xlabels,side=1,at=c(-4,-3,-2,-1,0),line=1)

abline(h=upper_3sd,col='red',lty=2)
abline(h=lower_3sd,col='red',lty=2)

abline(v=log10(0.03),col='red',lty=2)

dev.off()

## create list of individuals to remove from data
imiss_rem <- subset(imiss,imiss$F_MISS > 0.03)[,1:2]
het_rem <- subset(het,het$P_HET > upper_3sd | het$P_HET < lower_3sd)[,1:2]
indiv_rem <- rbind(imiss_rem,het_rem)
write.table(indiv_rem,'fail-imisshet-qc.txt',col=F,row=F,quo=F,sep='\t')

## === END of script
```

- Read in data

- Heterozygosity measure

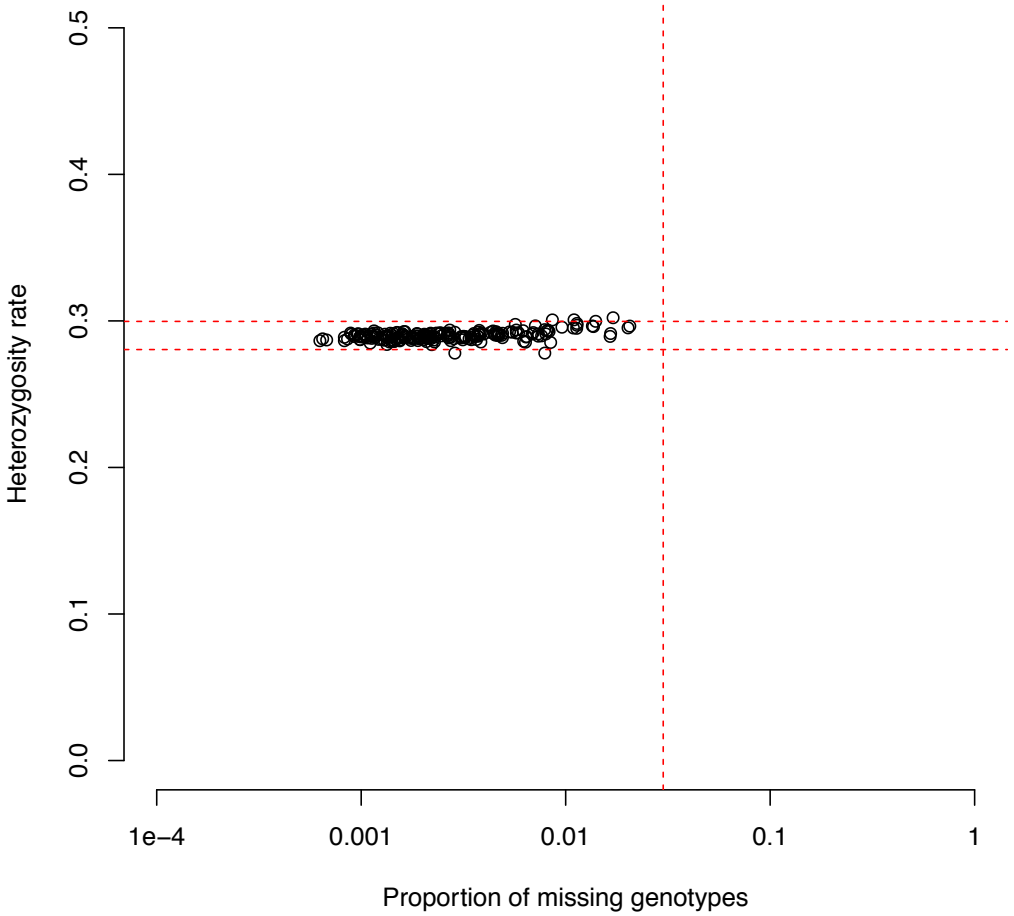
- Construct graph

- Apply QC cutoffs
  het outliers (> 3 SD)
  Missing > 3% of SNPs

- Write out failed samples to file

Viewing per-individual QC results

```
cp imiss-vs-het.pdf ~/private_html/atgu_workshop
```



QC Steps overview

- Sample (or individual) QC
  - Confirming sex from genotype
  - Poorly genotyped (high SNP Missingness)
  - Deviation in heterozygosity
  - Related or duplicated samples
  - Population outliers*
- SNP QC
  - Poorly genotyped (high SNP Missingness)
  - Rare or mono-allelic SNPs
  - Deviation from Hardy-Weinberg Equilibrium (HWE)

Estimating genotypic relatedness from SNPs

First identify a subset of SNPs not in LD

```
plink --bfile hapmapEA \
--remove fail-imisshet-qc.txt \
--indep-pairwise 50 5 0.2 \
--out hapmapEA
```

Use this subset to estimate genotypic relatedness

```
plink --bfile hapmapEA \
--remove fail-imisshet-qc.txt \
--extract hapmapEA.prune.in \
--genome \
--out hapmapEA
```

Checking genotype relatedness across samples

```
head hapmapEA.genome | column -t  | less -S
```

FID1	IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	PI_HAT	PHE	DST	PPC	RATIO
NA20505	NA20505	NA20506	NA20506	UN	NA	0.9872	0.0000	0.0128	0.0128	-1	0.771435	0.3446	1.9712
NA20505	NA20505	NA20502	NA20502	UN	NA	0.9888	0.0096	0.0016	0.0064	-1	0.770233	0.3950	1.9808
NA20505	NA20505	NA20528	NA20528	UN	NA	0.9733	0.0267	0.0000	0.0133	-1	0.770068	0.2922	1.9606
NA20505	NA20505	NA20531	NA20531	UN	NA	0.9789	0.0205	0.0006	0.0109	-1	0.770976	0.7407	2.0479
NA20505	NA20505	NA20534	NA20534	UN	NA	0.9602	0.0398	0.0000	0.0199	-1	0.772123	0.3046	1.9631
NA20505	NA20505	NA20535	NA20535	UN	NA	0.9650	0.0350	0.0000	0.0175	-1	0.771054	0.6510	2.0285
NA20505	NA20505	NA20586	NA20586	UN	NA	0.9728	0.0272	0.0000	0.0136	-1	0.770687	0.4281	1.9869
NA20505	NA20505	NA20756	NA20756	UN	NA	0.9675	0.0325	0.0000	0.0163	-1	0.770762	0.6902	2.0365
NA20505	NA20505	NA20760	NA20760	UN	NA	0.9344	0.0656	0.0000	0.0328	0	0.770978	0.8856	2.0904

	Probability of Sharing IBD Alleles		
Relative Pair	π_0	π_1	π_2
MZ Twins	0	0	1
Full Sibs	0.25	0.50	0.25
Parent-Offspring	0	1	0
First Cousin	0.75	0.25	0
Grandparent-Grandchild	0.50	0.50	0
Half-Sibs	0.50	0.50	0
Avuncular	0.50	0.50	0

cutoff: PI_HAT > 0.10

Removing related individuals

```
Rscript ibd-remove.R
```

- 1st step: Remove individuals related to multiple samples
- 2nd step: Remove individual with lower genotyping rate
- write FID/IID list to file (fail-ibd-qc.txt)

Combine failed sample lists and write new .bed/.bim files

```
cat fail-imisshet-qc.txt fail-ibd-qc.txt > fail-
indiv-qc.txt

plink --bfile hapmapEA \
--remove fail-indiv-qc.txt \
--make-bed \
--out hapmapEA_sampleQC
```

QC Steps overview

- Sample (or individual) QC
  - Confirming sex from genotype
  - Poorly genotyped (high SNP Missingness)
  - Deviation in heterozygosity
  - Related or duplicated samples
  - Population outliers*

- SNP QC
  - Poorly genotyped (high SNP Missingness)
  - Rare or mono-allelic SNPs
  - Deviation from Hardy-Weinberg Equilibrium (HWE)

Checking SNP genotyping rate by phenotype

```
plink --bfile hapmapEA_sampleQC \
--test-missing \
--out hapmapEA_sampleQC
```

```
head hapmapEA_sampleQC.missing
```

CHR	SNP	F_MISS_A	F_MISS_U	P
1	rs12565286	0.03125	0.03093	1
1	rs12124819	0.05208	0.03093	0.4974
1	rs2465136	0	0.01031	1
1	rs4970357	0	0.02062	0.4974
1	rs11466691	0	0.01031	1
1	rs11466681	0.01042	0.01031	1
1	rs34945898	0.03125	0	0.1211
1	rs715643	0.05208	0.02062	0.2787
1	rs13306651	0.01042	0.03093	0.6211

cutoff: *p* < 0.001

AWK one-liner to see if any SNPs reach cutoff *p*-value

```
awk ' {print $5} ' hapmapEA_sampleQC.missing | sort -g | head -n20
```

Checking Hardy-Weinberg Equilibrium

```
plink --bfile hapmapEA_sampleQC \
--hardy \
--out hapmapEA_sampleQC
```

```
head hapmapEA_sampleQC.hwe
```

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	rs12565286	ALL	C	G	0/17/170	0.09091	0.08678	1
1	rs12565286	AFF	C	G	0/6/87	0.06452	0.06243	1
1	rs12565286	UNAFF	C	G	0/11/83	0.117	0.1102	1
1	rs12124819	ALL	G	A	0/77/108	0.4162	0.3296	6.919e-05
1	rs12124819	AFF	G	A	0/41/50	0.4505	0.3491	0.004878
1	rs12124819	UNAFF	G	A	0/36/58	0.383	0.3096	0.02001
1	rs4970383	ALL	A	C	10/68/115	0.3523	0.352	1
1	rs4970383	AFF	A	C	3/36/57	0.375	0.3418	0.5488
1	rs4970383	UNAFF	A	C	7/32/58	0.3299	0.3618	0.401

cutoff: $p < 0.001$

Checking minor allele frequencies

```
plink --bfile hapmapEA_sampleQC \
--freq \
--out hapmapEA_sampleQC
```

```
head hapmapEA_sampleQC.frq
```

CHR	SNP	A1	A2	MAF	NCHROBS
1	rs12565286	C	G	0.04545	374
1	rs12124819	G	A	0.2081	370
1	rs4970383	A	C	0.228	386
1	rs13303118	G	T	0.4093	386
1	rs35940137	A	G	0.03886	386
1	rs2465136	C	T	0.3125	384
1	rs2488991	G	T	0.1528	386
1	rs3766192	C	T	0.4585	386
1	rs10907177	G	A	0.1554	386

cutoff: MAF < 1%

Calculating principal components (using LD-pruned SNP set)

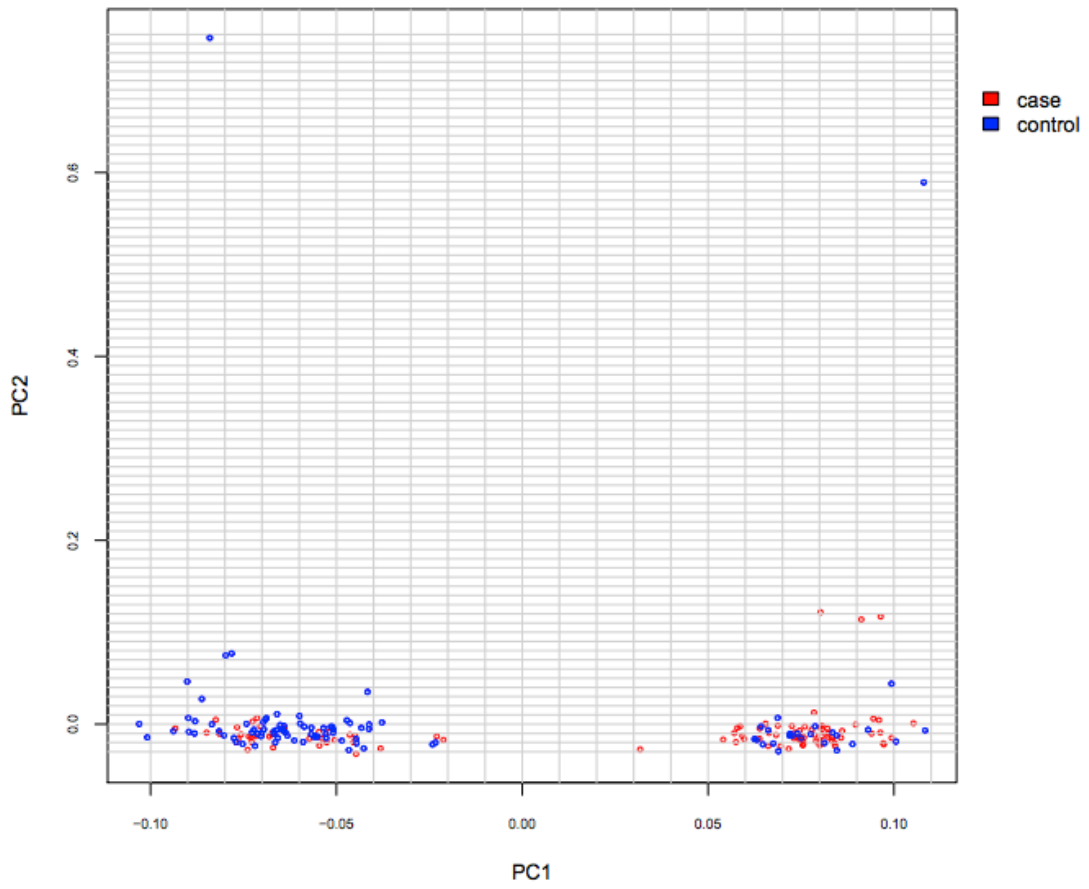
```
plink \
--bfile hapmapEA_sampleQC \
--extract hapmapEA.prune.in \
--pca \
--out hapmapEA_sampleQC
```

```
PLINK v1.90b1g 64-bit (10 Jun 2014)       https://www.cog-genomics.org/plink2
(C) 2005-2014 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hapmapEA_sampleQC.log.
128714 MB RAM detected; reserving 64357 MB for main workspace.
100310 variants loaded from .bim file.
193 people (96 males, 97 females) loaded from .fam.
193 phenotype values loaded from .fam.
--extract: 73226 variants remaining.
Using up to 15 threads (change this with --threads).
Calculating allele frequencies... done.
Total genotyping rate is 0.996554.
73226 variants and 193 people pass filters and QC.
Among remaining phenotypes, 96 are cases and 97 are controls.
Relationship matrix calculation complete.
--pca: Results saved to hapmapEA_sampleQC.eigenval and
hapmapEA_sampleQC.eigenvec .
```

Plotting principal components

```
Rscript PCA_2d_plot.R
```

```
cp hapmapEA_sampleQC_PCA.pdf ~/private_html/atgu_workshop
```



Remove failed SNPs and write out .bed/.bim file

```
plink --bfile hapmapEA_sampleQC \
--maf 0.01 \
--geno 0.05 \
--hwe 0.001 include-nonctrl \
--make-bed \
--out hapmapEA_genotypeQC
```

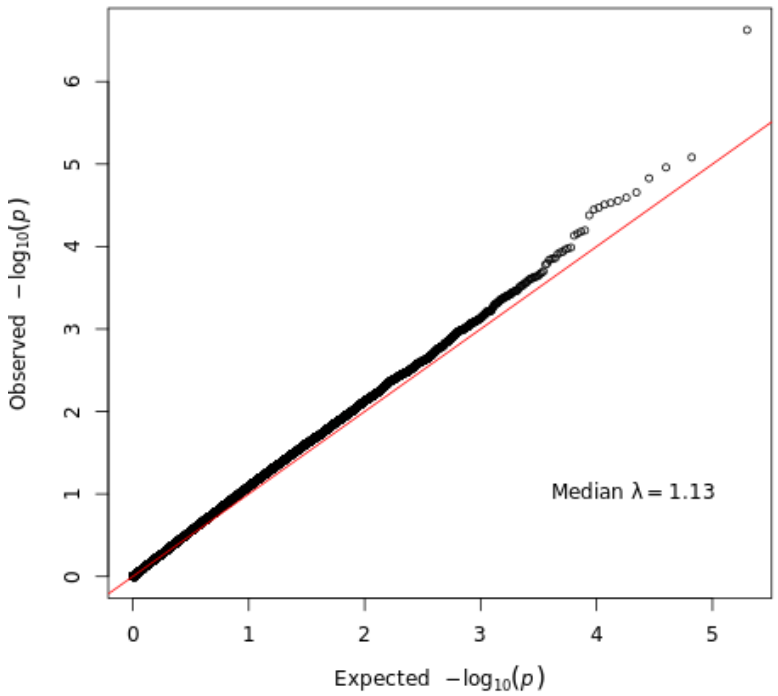
Options in effect:
--bfile hapmapEA_sampleQC
--geno 0.05
--hwe 0.001 include-nonctrl
--maf 0.01
--make-bed
--out hapmapEA_genotypeQC

32232 MB RAM detected; reserving 16116 MB for main workspace.
100310 variants loaded from .bim file.
193 people (96 males, 97 females) loaded from .fam.
193 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 193 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.996591.
53 variants removed due to missing genotype data (--geno).
--hwe: 139 variants removed due to Hardy-Weinberg exact test.
2980 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
97138 variants and 193 people pass filters and QC.
Among remaining phenotypes, 96 are cases and 97 are controls.
--make-bed to hapmapEA_genotypeQC.bed + hapmapEA_genotypeQC.bim +
hapmapEA_genotypeQC.fam ... done.

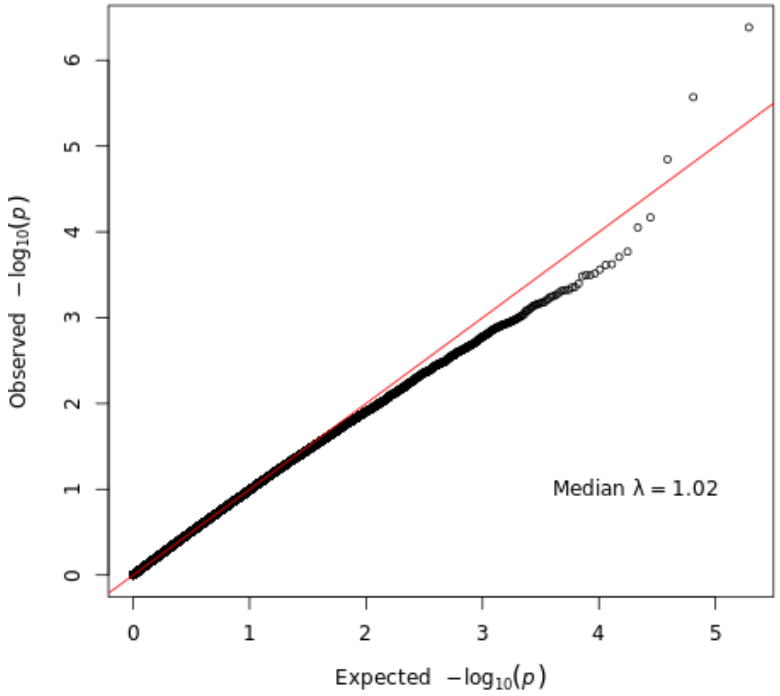
Re-run association

```
plink \
--bfile hapmapEA_genotypeQC \
--logistic \
--covar hapmapEA_sampleQC.eigenvec \
--parameters 1-4 \
--hide-covar \
--out hapmapEA_genotypeQC
```

Pre-QC



Post-QC



QC parameters are not hard and fast rules…

- Stricter QC
  - Low quality DNA / genotyping
  - Multi-ethnic cohorts

- More lenient QC
  - Fine-mapping follow-up
  - Well-designed experiment

Recent QC protocols


ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

PGC Schizophrenia, Nature 2014

Individual QC:

- sex check
- genotyping rate > 98%
- Heterozygosity *F* statistic < +/- 0.2
- Pi-hat relatedness < 0.2

SNP QC:

- genotyping rate > 98%
- Differential genotyping rate < 2% difference
- HWE in controls (p > 1e-6)
- HWE in cases (p > 1e-10)
- MAF > 0 (prior to imputation)