Data QC / cleaning in Genome-Wide Association Studies (GWAS)

2023 Statistical Genetics workshop
Presenter: Daniel Howrigan
Data group leader – Neale Lab (MGH, Broad Institute)

Slides adapted from previous workshop presenters:
Lucia Colodro Conde (QIMR), Katrina Grasby (QIMR), Shaun Purcell (HMS)

With help from:
John Kemp (University of Queensland) and Daniel Gustavson (IBG)

Session Outline – genetic data QC

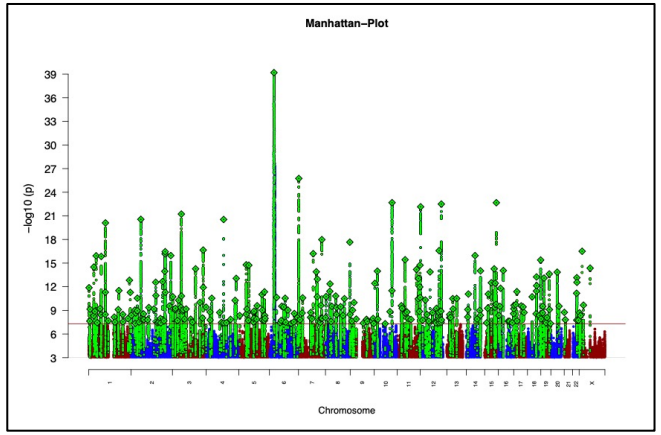
- Lecture portion (~40 minutes)
  - Goals of GWAS
  - What does genetic data look like?
  - GWAS Quality Control (QC)

- Practical portion (~40 minutes)
  - Viewing genotype data
  - Sample and SNP QC
  - Relatedness checking
  - Principal components analysis (PCA)

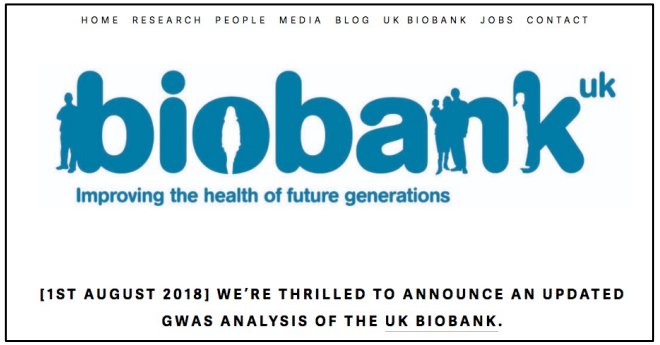
Goals of Genome Wide Association Studies

- Go from trait heritability towards biological mechanism
  - What genes/genetic variants drive heritable differences?
- Genome-wide interrogation
  - Moving away from candidate gene studies
  - Technological advancement and dropping cost
- Flexible application of study design
  - All heritable traits can be studied
  - Biological/mathematical properties of DNA quite robust

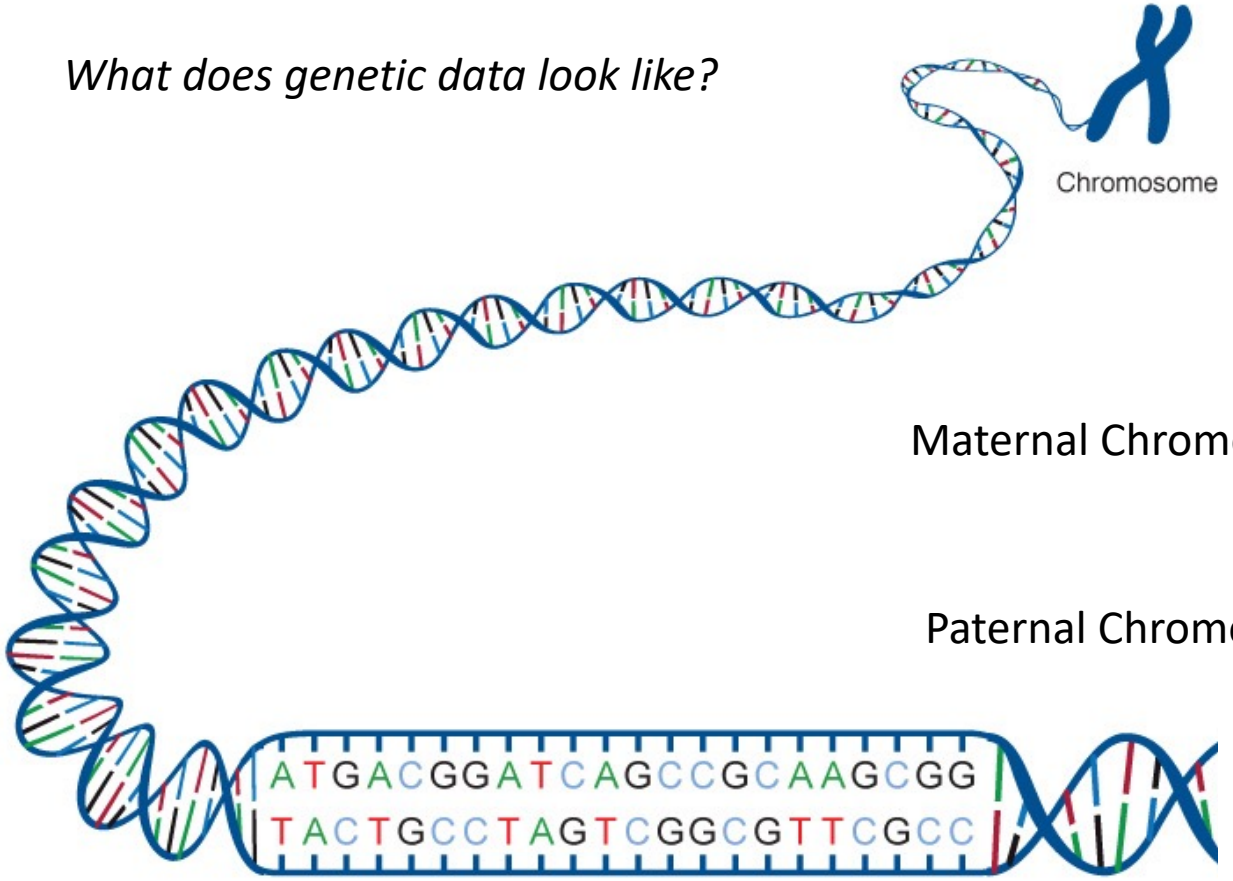
GWAS of Schizophrenia



GWAS of ~4,200 traits



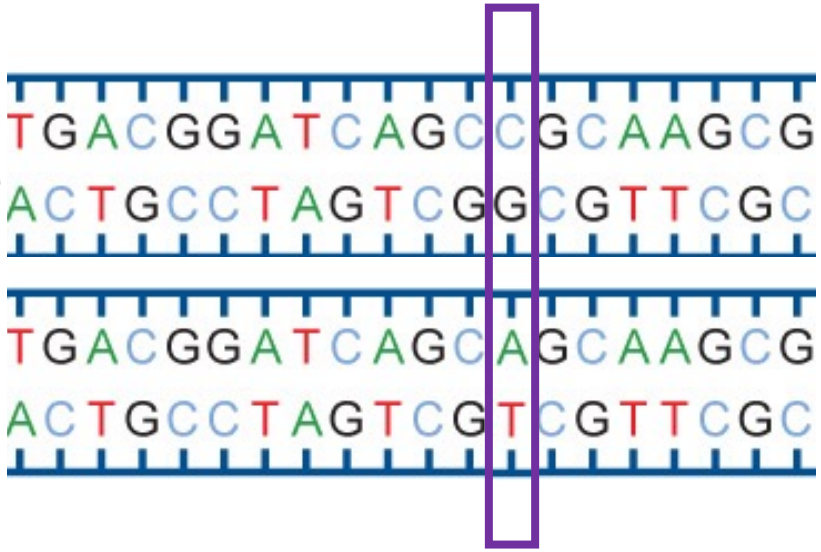
What does genetic data look like?



adenine (A), thymine (T), cytosine (C), guanine (G)

Single Nucleotide Polymorphism

SNP



Allele 1 = C
Allele 2 = A
Bi-allelic combinations = C/C, C/A, A/A

Genetic variation: differences in the sequence of DNA among individuals.

Mutation: a newly arisen variant

Examples of genetic variation

Sequence variation

Single nucleotide
- substitutions
- insertions | 'indels'
- deletions

Structural variation

2bp to 1,000bp
- VNTRs: microsatellites, minisatellites
- indels
- inversions
- di-, tri-, tetranucleotide repeats

1kb to submicroscopic
- copy number variants
- segmental duplications
- inversions, translocations
- copy number variant regions
- microdeletions, microduplications

Microscopic to subchromosomal
- segmental aneusomy
- chromosomal deletions (losses)
- chromosomal insertions (gains)
- chromosomal inversions
- intrachromosomal translocations
- chromosomal abnormality
- heteromorphisms
- fragile sites

Whole chromosomal to whole genome
- interchromosomal translocations
- ring chromosomes, isochromosomes
- marker chromosomes
- aneuploidy
- aneusomy

GWAS

Genotyping on a chip

Affymetrix:

6.0 chip

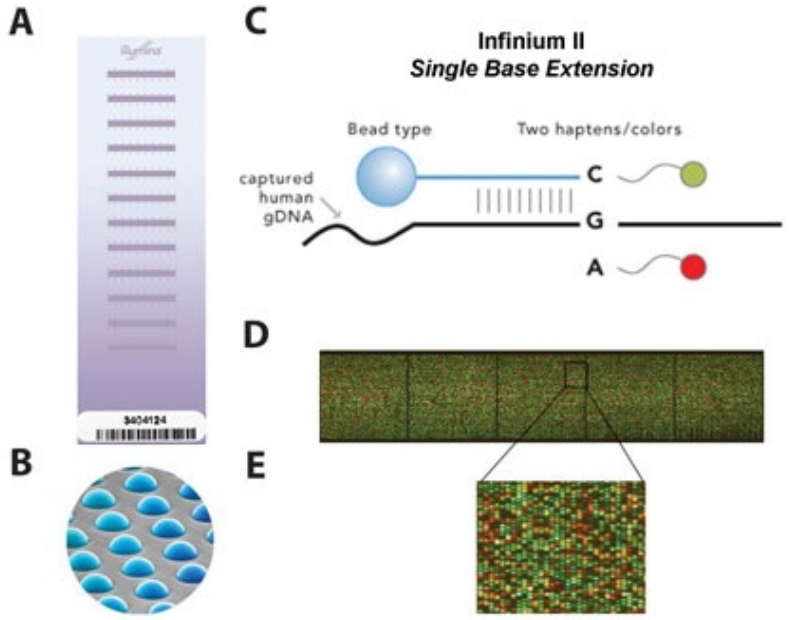
>900,000 SNPs
CNV probes
82% coverage CEU HapMap
Accuracy 99.90%

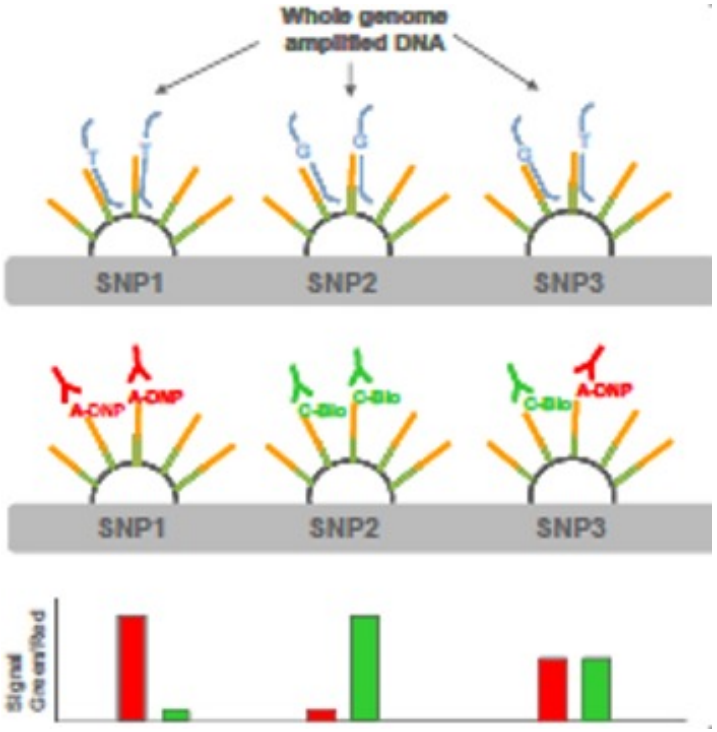
Illumina:

Human1M BeadChip

>1 million SNPs
CNV probes
95% coverage CEU HapMap
Accuracy 99.94%

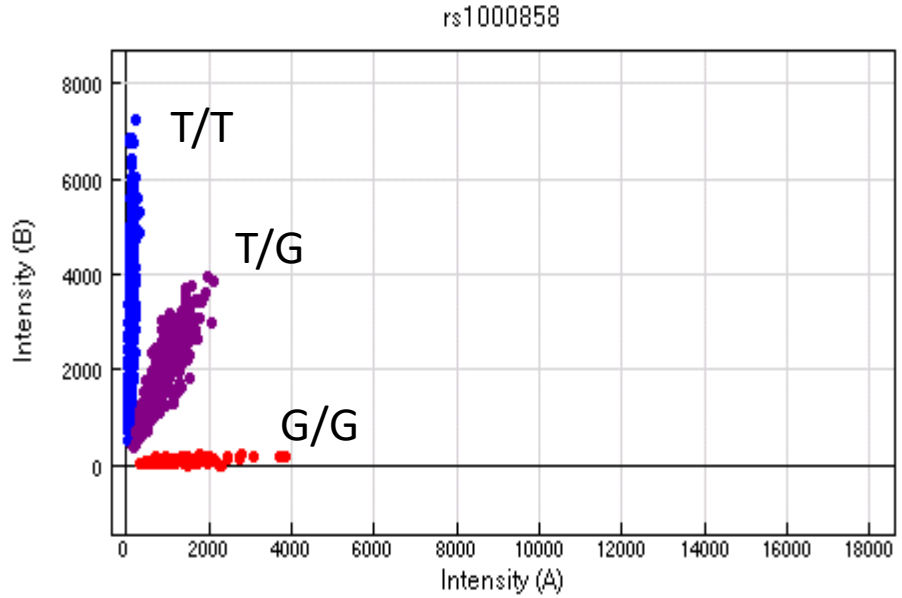
From DNA to data



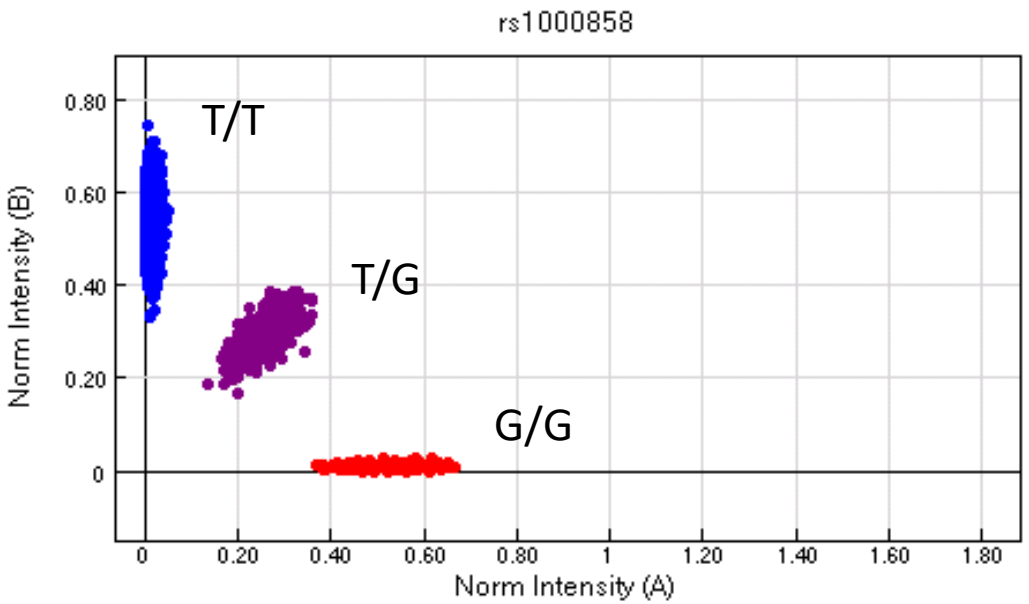


Good SNP (Illumina chip example)

Raw Intensity



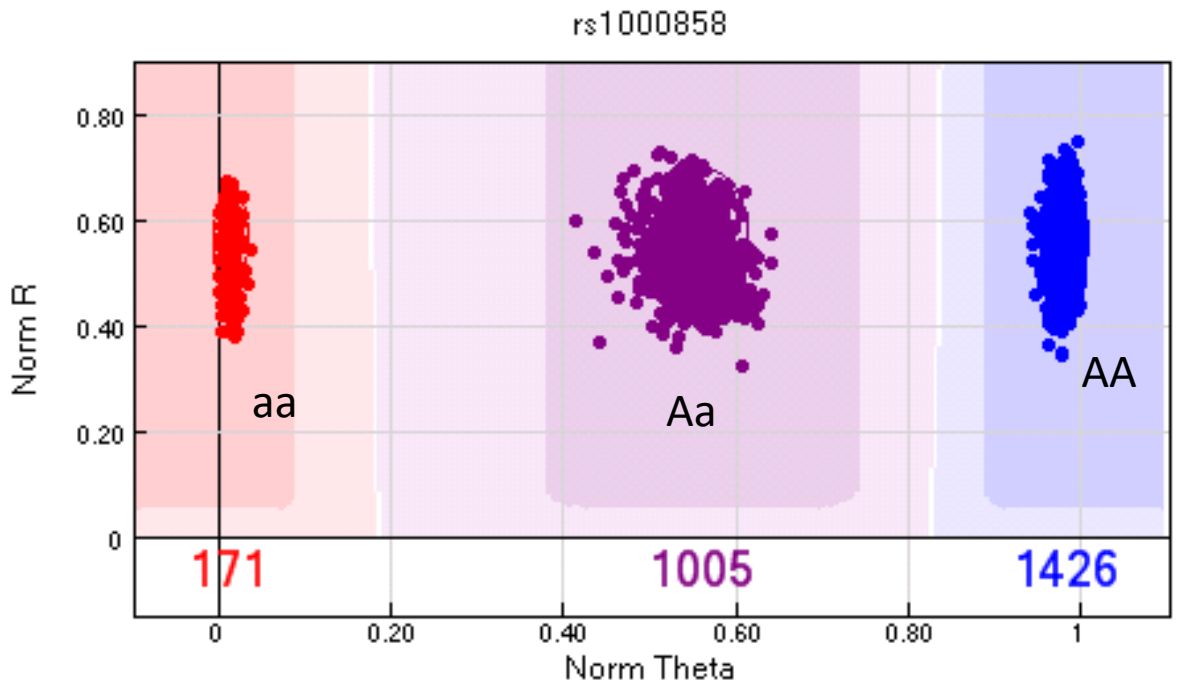
Normalized Intensity



Each dot is an individual genotype

Same SNP, different view

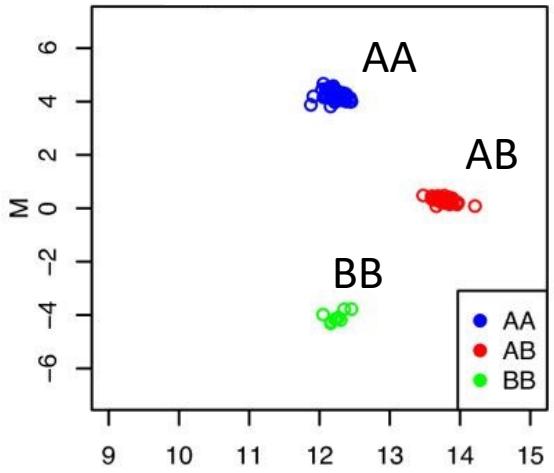
Overall Intensity



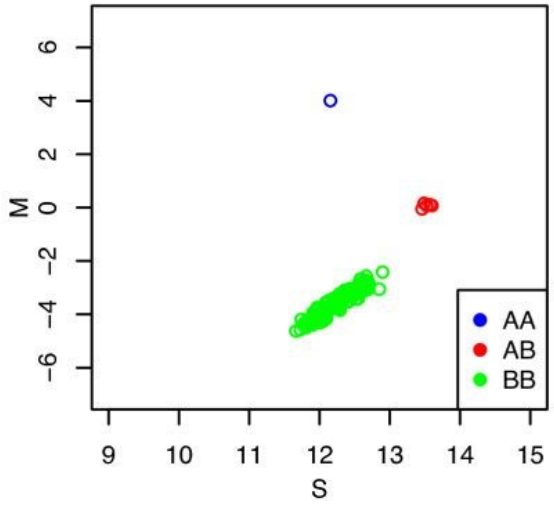
Angular position

SNPs with different allele frequencies

High MAF



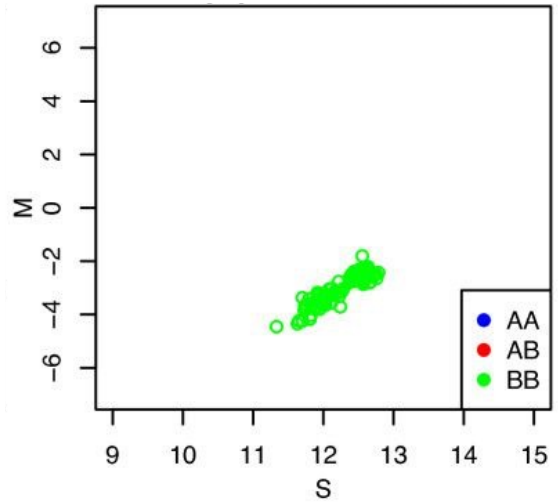
Less common MAF



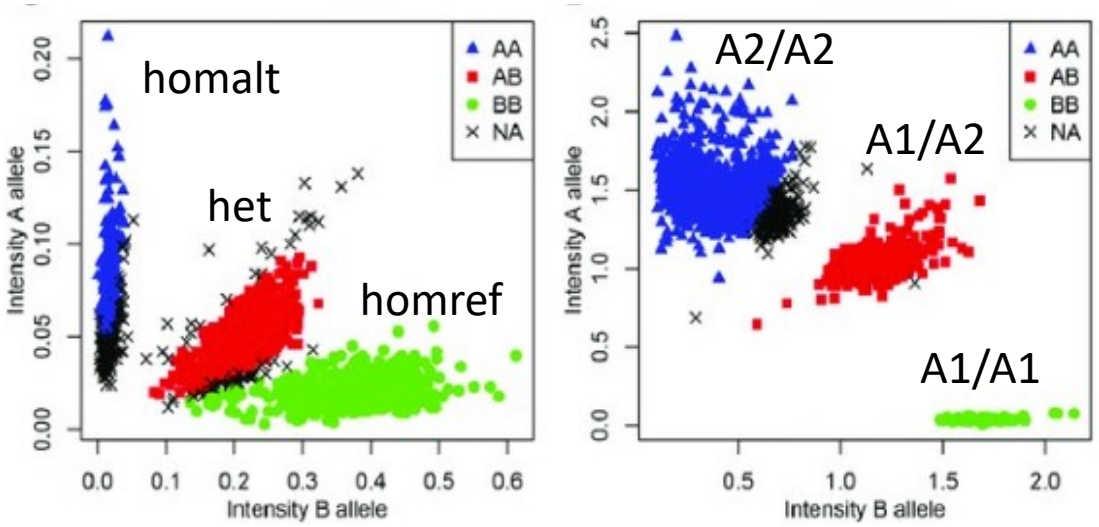
MAF = Minor Allele Frequency

- "Common SNPs" = MAF > 5%? 1%? 0.1?
- "Low Frequency SNPs" = MAF < 1%
- "Ultra-rare variants" = MAF < 1e5 (1 in 100k)

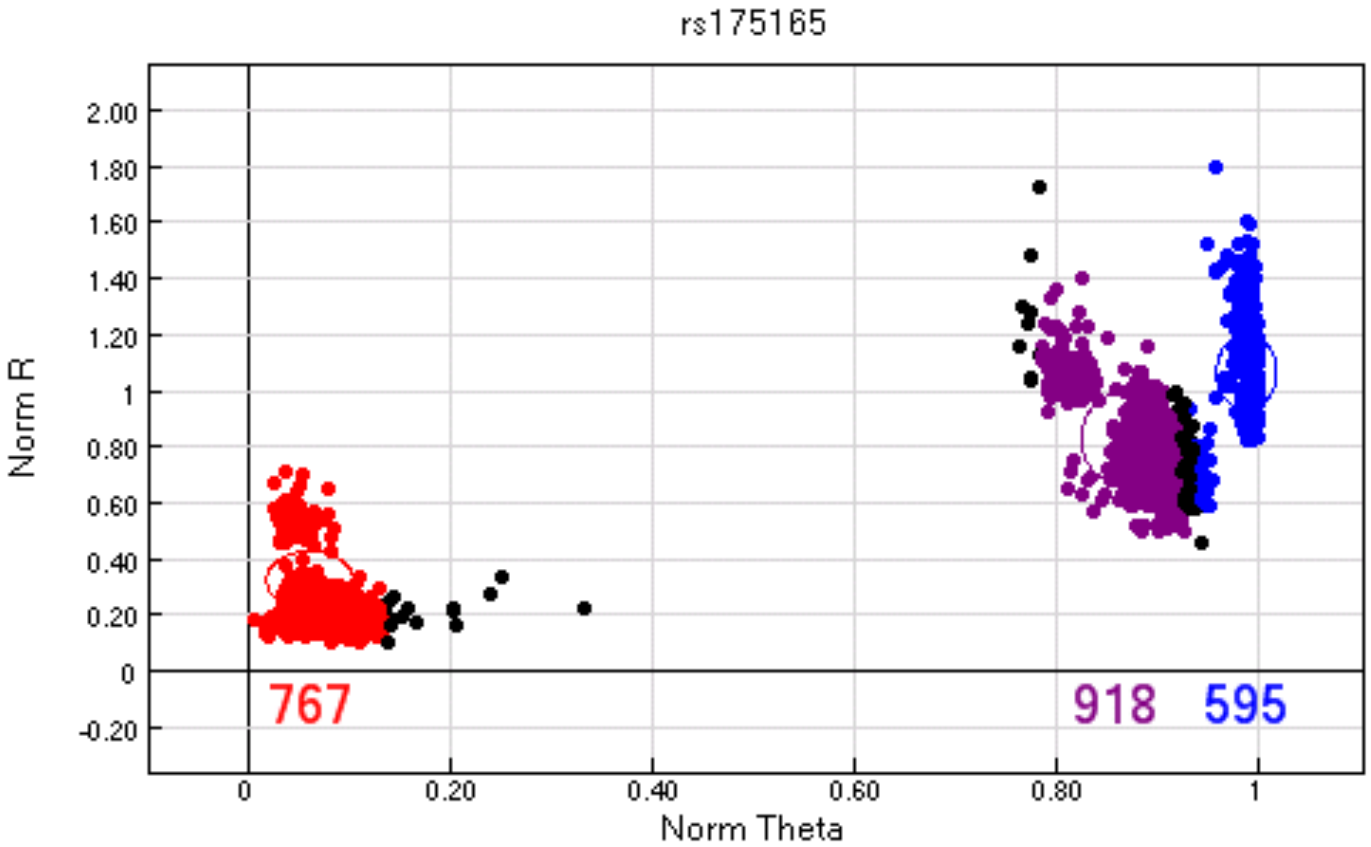
Monoallelic in the sample



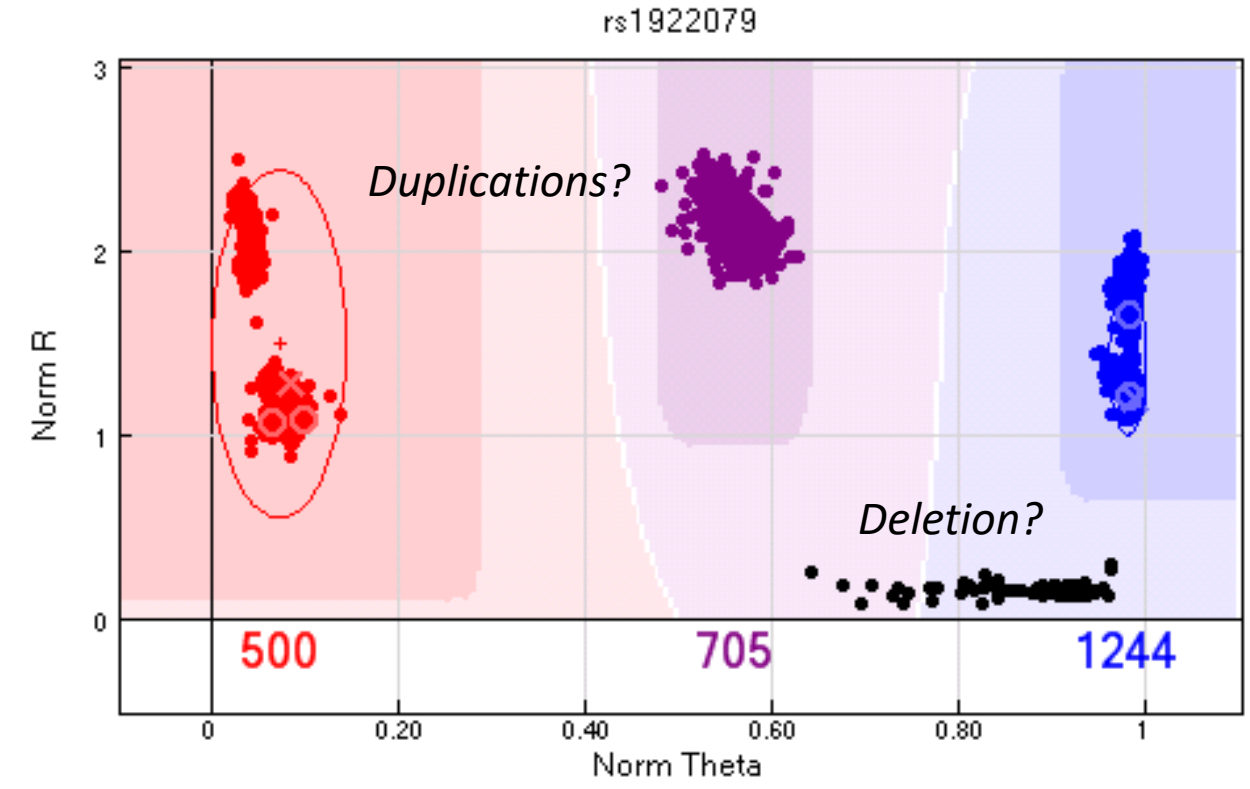
Bad SNP call examples



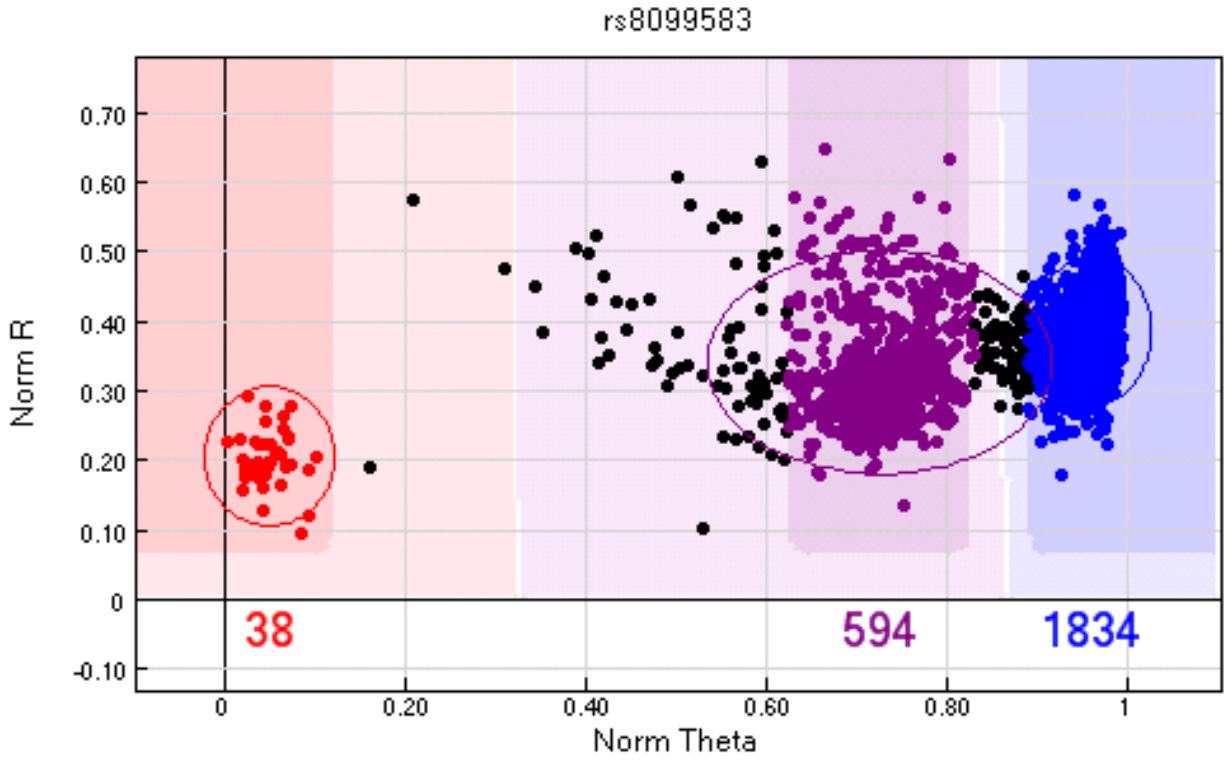
Bad SNP


rs175165
Norm R
Norm Theta
2.00
1.80
1.60
1.40
1.20
1
0.80
0.60
0.40
0.20
0
-0.20
0
0.20
0.40
0.60
0.80
1
767
918
595

Another bad SNP



Another bad SNP



PLINK data format of GWAS data

Samples

.fam file

FID	IID	PID	MID	SEX	AFF
Taiw_1	PT-VXBB	PT-VXES	PT-VXEG	1	2
Taiw_1	PT-VXEG	0	0	2	1
Taiw_1	PT-VXES	0	0	1	1
Taiw_2	PT-VX4A	0	0	1	1
Taiw_2	PT-VX7E	PT-VX4A	PT-VX72	1	2
Taiw_2	PT-VX72	0	0	2	1
Taiw_4	PT-VX6B	0	0	2	1
Taiw_4	PT-VX6N	PT-VX73	PT-VX6B	2	2
Taiw_4	PT-VX73	0	0	1	1
Taiw_5	PT-VX5N	PT-VX5Z	PT-VX6M	2	2

FID = family ID
IID = Individual ID
PID = paternal ID
MID = maternal ID
AFF = affection status

- 1 = control
- 2 = case
- -9 or 0 = unknown

Genetic variants

.bim file (or .map file)

CHR	SNP ID	CM	POS	A1	A2
1	GSA-rs114420996	0	58814	A	G
1	GSA-rs9283150	0	565508	A	G
1	GSA-rs9326622	0	567092	C	T
1	GSA-1:726912	0	726912	G	A
1	GSA-rs116587930	0	727841	A	G
1	rs3131972	0	752721	G	A
1	rs12567639	0	756268	A	G
1	GSA-rs114525117	0	759036	A	G
1	rs12127425	0	794332	A	G
1	GSA-rs79373928	0	801536	G	T
1	GSA-rs72888853	0	815421	C	T
1	rs28444699	0	830181	G	A
1	GSA-1:830731	0	830731	C	T
1	GSA-rs116452738	0	834830	A	G
1	GSA-rs72631887	0	835092	G	T
1	rs4970383	0	838555	A	C

CHR = chromosome
POS = position
CM = Centimorgan (often unused)
A1 = 0 allele
A2 = 1 allele

Genotype data

.ped file

```
WGACON 11 0 0 2 2 T T A A G G
WGACON 12 0 0 1 2 T T A A G G
WGACON 15 0 0 1 2 C C 0 0 0 0
WGACON 17 0 0 1 2 T T A A G G
WGACON 18 0 0 1 2 T T A A G G
WGACON 20 0 0 1 1 T T A A G G
WGACON 22 0 0 2 1 C T G A 0 0
```

compression

.bed file

```
0101010010101010101
1010011101010101010
1101110101001010101
1101001011101101010
1101010101010111010
```

GWAS QC

GWAS Quality Control (QC)

- **GOAL**: Remove bad samples/SNPs, keep good samples/SNPs

- Preliminary strategies (first pass)
  - Poorly genotyped samples / SNP markers
  - Potential genotype/phenotype mismatches
  - Deviation away from expected heterozygosity
  - Related or duplicated samples (population-based data)

- Follow-up strategies
  - Batch effects
  - Quality differences between datasets
  - Comparison with reference data

  …and more

Sample QC

- Poorly genotyped individuals
  - Poor quality DNA (high number of failed SNP calls)
  - Contaminated DNA (unusual levels of heterozygosity)
- Reporting error
  - Indications of sample mix-up (sex check or ancestry match)
- 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)
  - Mendelian errors (if family-based data available)
  - Uninformative SNPs (too rare or mono-allelic)

- Follow-up on association signals
  - No QC protocol will eliminate all instances of genotyping error
  - Re-analyze original intensity of significant associations (whenever possible)
  - For meta-analysis, examining heterogeneity of SNP effect

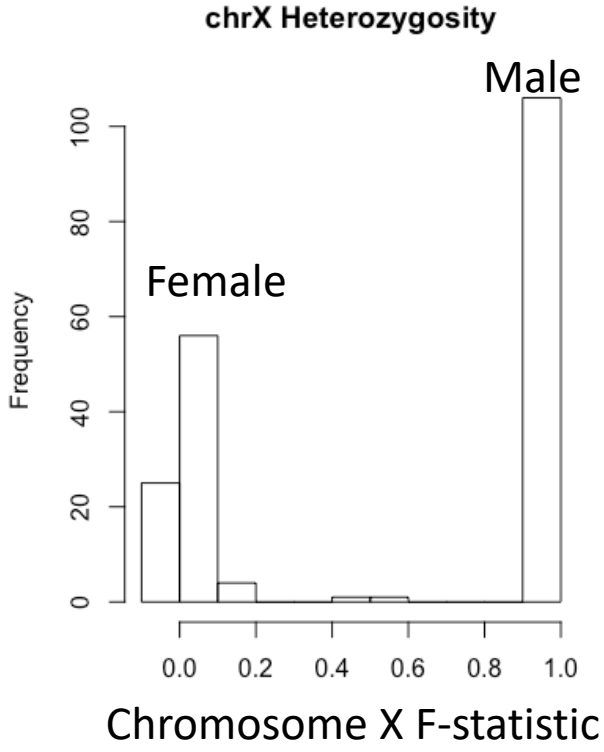
Preliminary QC steps

- **SAMPLE**: Sex-check (chr X heterozygosity)
- **SNP**: Genotyping Call Rate (genotypes missed in individuals)
- **SAMPLE**: Sample Call Rate (individuals missing genotypes)
- **SNP**: Hardy-Weinberg Equilibrium
- **SAMPLE**: Proportion of Heterozygosity
- **SAMPLE/SNP**: Mendelian errors
- **SAMPLE**: Genetic Relatedness

Confirming genetic sex

- Primary question: Is the sample-level data correctly matching the SNP data?

Female sex = XX
Male sex = XY



Example .sexcheck file from PLINK (male=1, female=2)

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

SNP genotyping call rate ("missingness")

Bad SNP design, poor clustering…

- Usually done iteratively
  - Remove SNPs with < 95% call rate
  - Run sample QC
  - Remove SNPs with < 98% call rate
- For case/control data
  - Look at difference in genotyping rate
  - Threshold usually at > 2% call rate difference

Example .lmiss file from PLINK

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

Example .missing file from PLINK

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

Sample genotyping call rate

Low quality DNA, degradation, lab error, contamination

Example .imiss file from PLINK

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

Missing genotypes

To generate a list genotyping/missingness rate statistics:

```
plink --file data --missing
```

This option creates two files:

```
plink.imiss
plink.lmiss
```

which detail missingness by individual and by SNP (locus), respectively. For individuals, the format is:

```
FID             Family ID
IID             Individual ID
MISS_PHENO      Missing phenotype? (Y/N)
N_MISS          Number of missing SNPs
N_GENO          Number of non-obligatory missing genotypes
F_MISS          Proportion of missing SNPs
```

http://zzz.bwh.harvard.edu/plink/summary.shtml#missing

Hardy-Weinberg Equilibrium (HWE)

- A genetic variant is said to be in HWE if the genotype proportions can be predicted by the allele frequencies in the following way:
  - If:
    - f(A1) = p
    - f(A2) = q

    $p + q = 1$
  - Then:
    - $f(A1/A1) = p^2$
    - $f(A1/A2) = 2pq$
    - $f(A2/A2) = q^2$

    $p^2 + 2pq + q^2 = 1$

Example:	In C/T SNP terms:
p = 0.2	C allele freq. = 20%
q = 0.8	T allele freq.= 80%
p2 = 0.04	C/C freq. = 4%
2pq = 0.32	C/T freq. = 32%
q2 = 0.64	T/T freq. = 64%

Testing for deviation from HWE

Deviations from HWE can be caused by:

- Non-random mating (inbreeding, assortative mating, …)
- **Population stratification**
- Mutation
- Limited population size
- Random genetic drift
- Gene flow
- **Genotyping errors**
- Selection (→ may be due to true association!)

Example .hardy output in PLINK

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

So only extreme deviation from HWE ($p < 10^{-6}$) is worrisome.

Proportion of heterozygosity (Fhet)

Inbreeding coefficients

Given a large number of SNPs, in a homogeneous sample, it is possible to calculate inbreeding coefficients (i.e. based on the observed versus expected number of homozygous genotypes).

```
plink --file mydata --het
```

which will create the output file:

```
plink.het
```

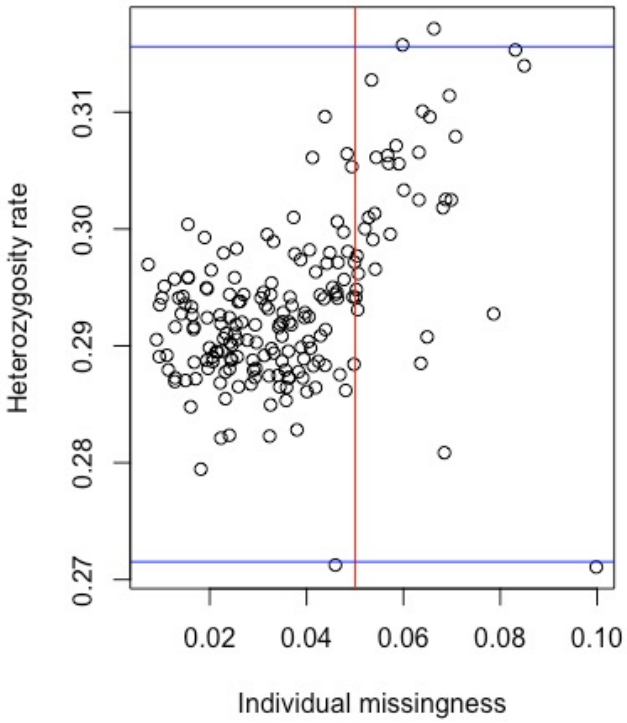
which contains the fields, one row per person in the file:

```
FID       Family ID
IID       Individual ID
O(HOM)    Observed number of homozygotes
E(HOM)    Expected number of homozygotes
N(NM)     Number of non-missing genotypes
F         F inbreeding coefficient estimate
```

This analysis will automatically skip haploid markers (male X and Y chromosome markers).

Note With whole genome data, it is probably best to apply this analysis to a subset that are pruned to be in approximate linkage equilibrium, say on the order of 50,000 autosomal SNPs. Use the `--indep-pairwise` and `--indep` commands to achieve this, described here.

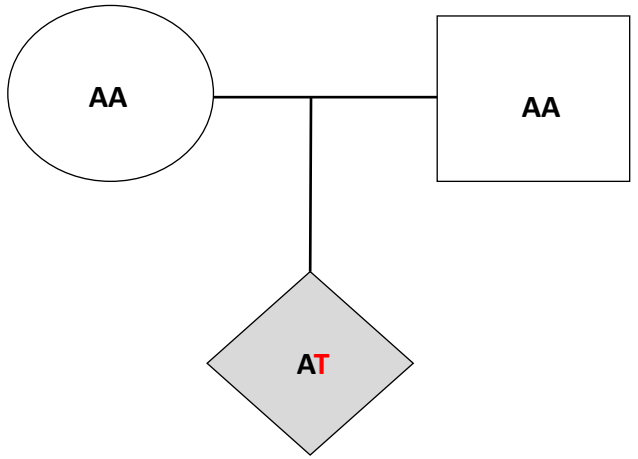
Note The estimate of F can sometimes be negative. Often this will just reflect random sampling error, but a result that is strongly negative (i.e. an individual has *fewer* homozygotes than one would expect by chance at the genome-wide level) can reflect other factors, e.g. sample contamination events perhaps.



http://zzz.bwh.harvard.edu/plink/ibdibs.shtml#inbreeding

Mendelian errors

- Requires parent-offspring data
- Similar to genotyping rate, can be examined at sample and SNP level
- High sample-level mendel error rate
  - Parental uncertainty
- High SNP-level mendel error rate
  - Poor genotype quality



de novo mutation is a type of mendelian error

Mendel errors

```
--mendel ['summaries-only']
--mendel-duos
--mendel-multigen
```

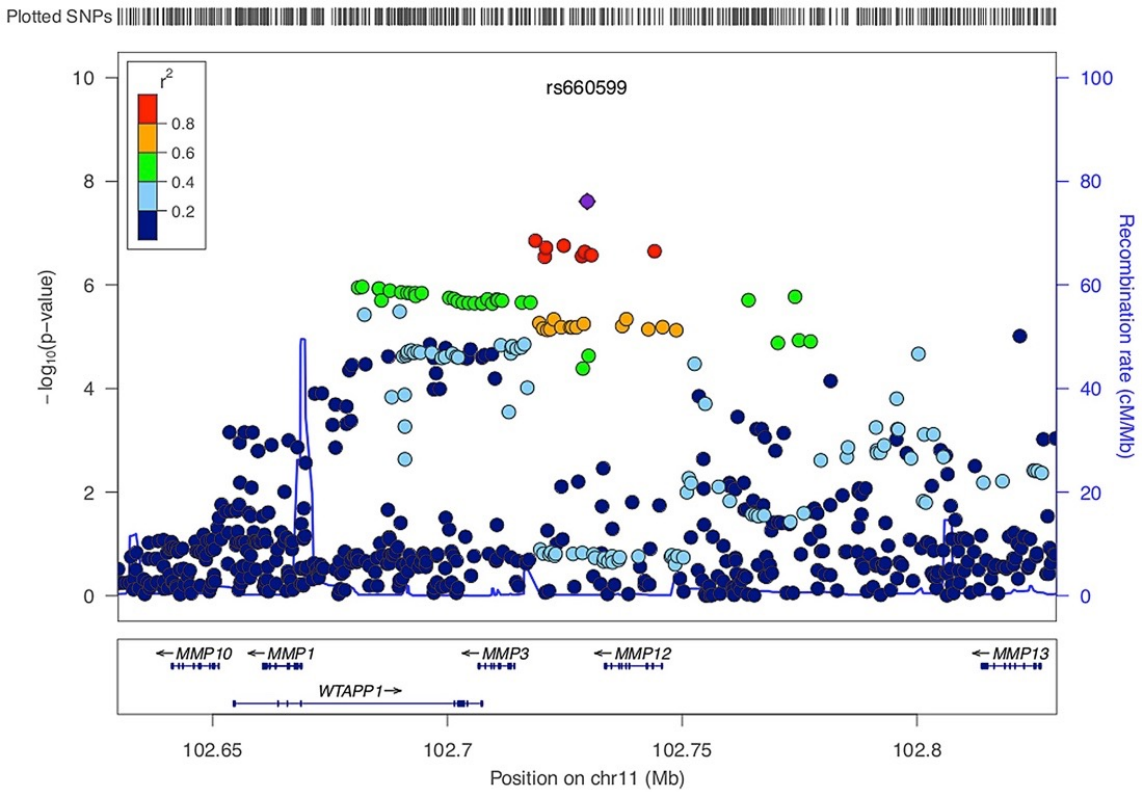
--mendel scans the dataset for Mendel errors, writing a set of reports to plink{.mendel,.imendel,.fmendel,.lmendel}. Haploid and mitochondrial data are ignored. The errors are classified as follows, where '1' refers to the A1 (usually minor) allele and '2' refers to A2:

Code	Pat. genotype	Mat. genotype	Child genotype	Samples implicated
1	11	11	12	all
2	22	22	12	all
3	22	11/12/missing	11	father, child
4	11/12/missing	22	11	mother, child
5	22	22	11	child
6	11	12/22/missing	22	father, child
7	12/22/missing	11	22	mother, child
8	11	11	22	child
9	(Xchr male)	11	22	mother, child
10	(Xchr male)	22	11	mother, child

https://www.cog-genomics.org/plink/1.9/basic_stats#mendel

Linkage disequilibrium (LD) allows us to be more robust with our QC protocols

- TL/DR: "Nearby SNPs are correlated"
- Properties of linkage disequilibrium reduce the loss of signal sensitivity when removing SNPs
- Strict multiple testing correction often requires very large samples - no single sample will drive a signal
- LD must be taken into account when examining genetic relatedness, population stratification, and interpreting association



Genetic relatedness using Identity-By-Descent (IBD) calculation

- Question: How much does a pair of samples share 0, 1, or both alleles?

- Identical twins: Shares both alleles across entire genome (barring mutation events)

- Requires using LD-pruned SNPs for accurate estimates
  - Want each SNP to be an "independent" marker

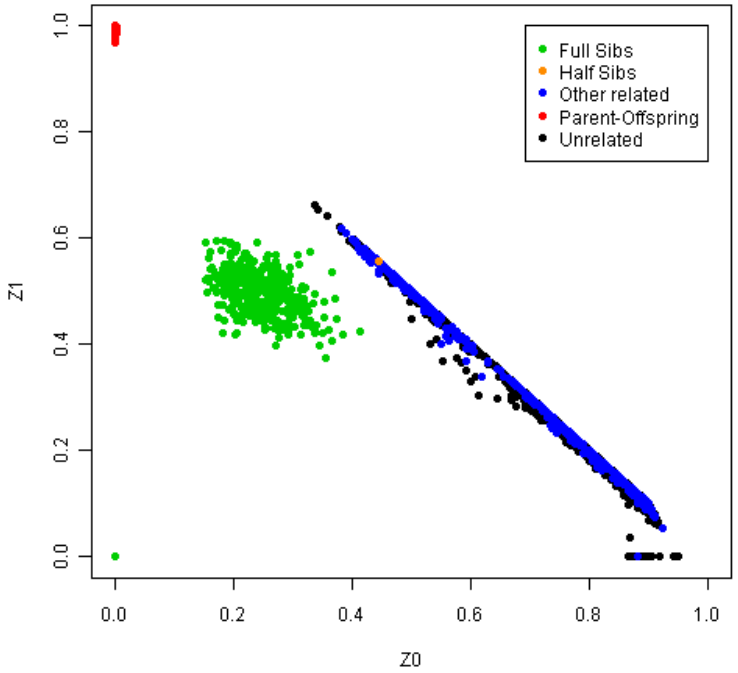
- Used to both "confirm" and "filter" related individuals

Checking genotype relatedness across samples

Example of .genome file in PLINK

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



Using genetic relatedness estimates

- Confirm unrelated or “population-based” sample ascertainment
  - Filter out related samples (pi-hat > 0.2 often used)
  - “Cryptic relatedness” – related individuals identified in ”unrelated” sample
- Confirm family structure (pedigree)
  - Ensure parent-child and sibling relationship
- Watch out for distinct ancestries
  - Can skew IBD estimates and incorrectly identify recent relatedness
  - PCrelate more robust to these patterns https://rdrr.io/bioc/GENESIS/man/pcrelate.html

Session Outline – genetic data QC

- Practical portion (~40 minutes)
  - Data checking
  - Sample and SNP QC
  - Relatedness checking
  - Principal components analysis (PCA)

- Go to: workshop.colorado.edu
  - **Slides + practical:** /faculty/daniel/2023/QC
  - **Terminal:** workshop.colorado.edu/ssh
  - **Rstudio:** workshop.colorado.edu/rstudio

Script that you will be working through:

QC_practical_statgenWorkshop2023.txt

Full path: **/faculty/daniel/2023/QC/QC_practical_statgenWorkshop2023.txt**

Walk through this script and copy/paste commands to the ssh command line

Qualtrics version: **https://ucsas.qualtrics.com/jfe/form/SV_eWpdYL7srw7Cy6W**

Answers to be filled out by a single table member

See the ISGW forum for these and other useful links to start your practical session:

https://isgw-forum.colorado.edu/

1.1 Creating workspace

Create day1 subdirectory (-p creates full path into new directories)
```
mkdir -p ~/day1/QC
```

traverse into new subdirectory
```
cd ~/day1/QC
```

1.2 Copying over genetic dataset

Copy the files to your working subdirectory
```
cp /faculty/daniel/2023/QC/* .
```

Check you have the required files:

```
ls -l
```

HM3.bed
HM3.bim
HM3.fam
QC_practical_BoulderWorkshop2023.R
QC_practical_BoulderWorkshop2023.sh
QC_practical_BoulderWorkshop2023.txt
cc.ped
cc.map

=== Main QC ===

STEP 1. Data and Formats

STEP 2. Check for reported/genotype sex discrepancies

STEP 3. Obtain information on individuals missing SNP data

STEP 4. Variant QC: SNPs missing data; MAF; Hardy-Weinberg

STEP 5. Sample QC: genotype call rate and heterozygosity

STEP 6. LD-pruned SNP set

STEP 7. Sample QC: sex check filtering using LD-pruned SNP set

STEP 8. Sample QC: Checking for cryptic relatedness