Introduction to Genome-Wide Association Studies (GWAS)

2020 ATGU welcome workshop

Presenter: Daniel Howrigan

Data group leader – Neale Lab

Slides adopted from:

Boulder Colorado Stat Gen Workshop (*Lucia Colodro Conde, Katrina Grasby, Shaun Purcell, Abdel Abdellaoui, Sarah Medland*)

Genetics course slides from Abdel Abdellaoui @dr_appie

Lecture Format

- Part 1 (~40 minutes)
  - Goals of GWAS
  - What does the data look like?
  - GWAS Quality Control (QC)
  - 5 min breakout session
- Part 2 (~40 minutes)
  - Relatedness checking
  - Population stratification
  - Principal components analysis (PCA)
  - Imputation
  - 5 min breakout session
- Part 3 (~40 minutes)
  - Association testing
  - Meta-analysis
  - Polygenic Scoring
  - 5 min breakout session
- Preparation for module 3 / additional reading / lingering questions

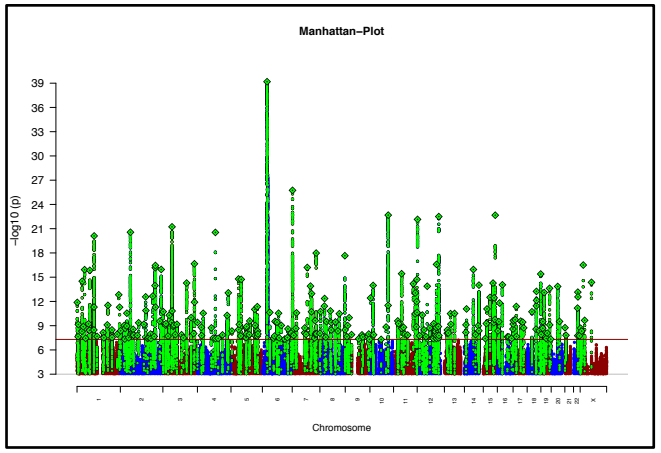
Lecture Format

- Part 1 (~40 minutes)
  - Goals of GWAS
  - What does the data look like?
  - GWAS Quality Control (QC)
  - 5 min breakout session
- Part 2 (~40 minutes)
  - Relatedness checking
  - Population stratification
  - Principal components analysis (PCA)
  - Imputation
  - 5 min breakout session
- Part 3 (~40 minutes)
  - Association testing
  - Meta-analysis
  - Polygenic Scoring
  - 5 min breakout session
- Preparation for module 3 / additional reading / lingering questions

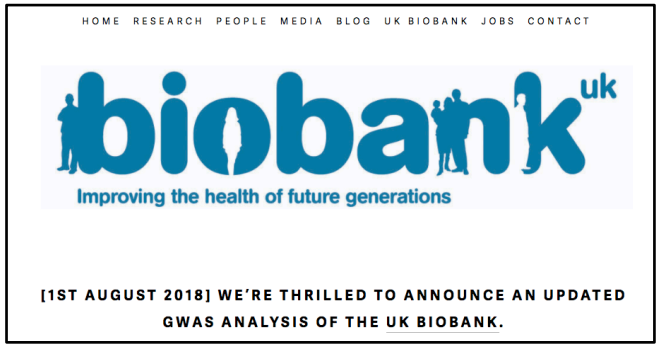
Goals of Genome Wide Association Studies

- 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

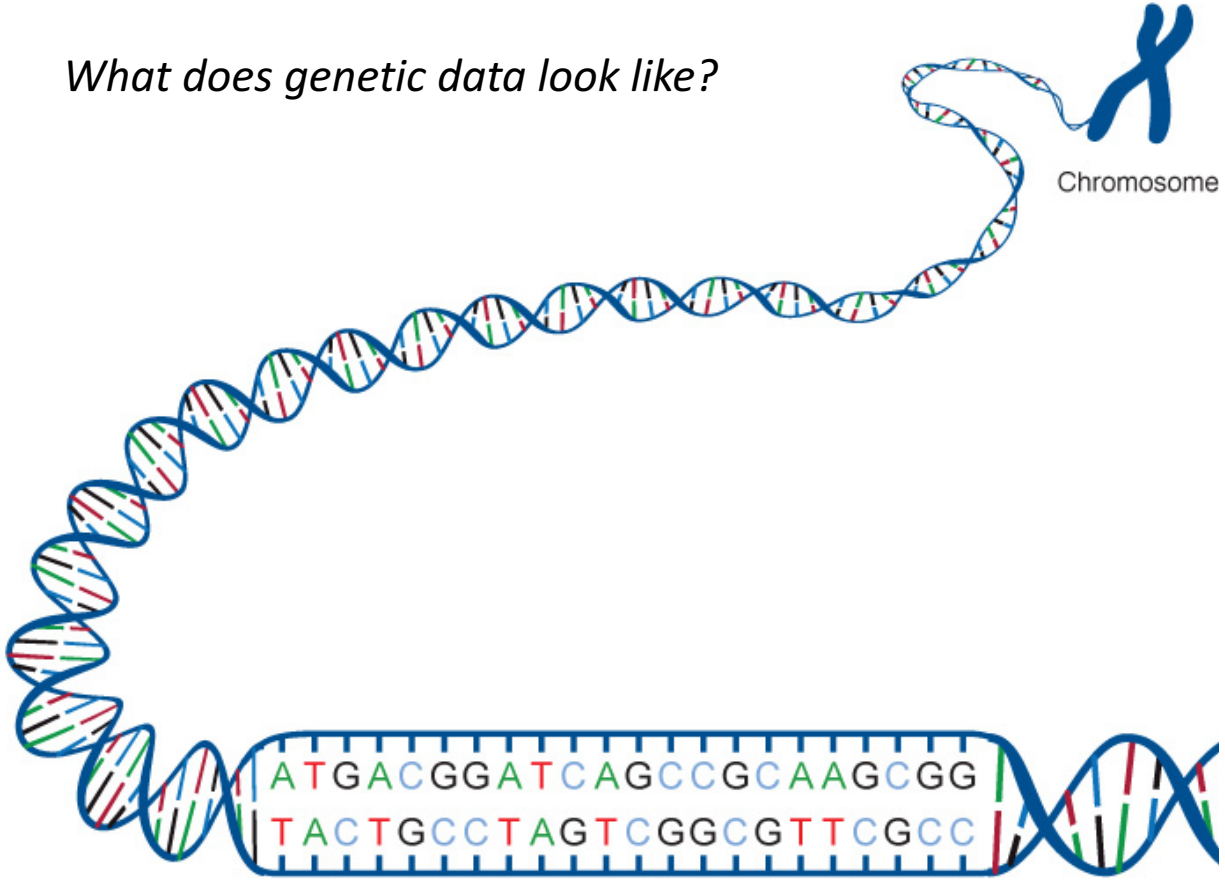
GWAS of Schizophrenia



GWAS of ~4,200 traits



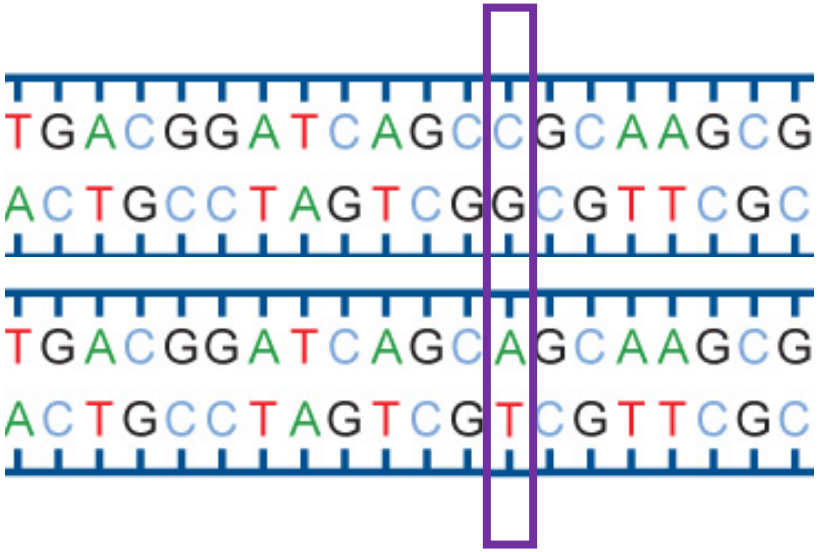
What does genetic data look like?



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

Single Nucleotide Polymorphism

SNP

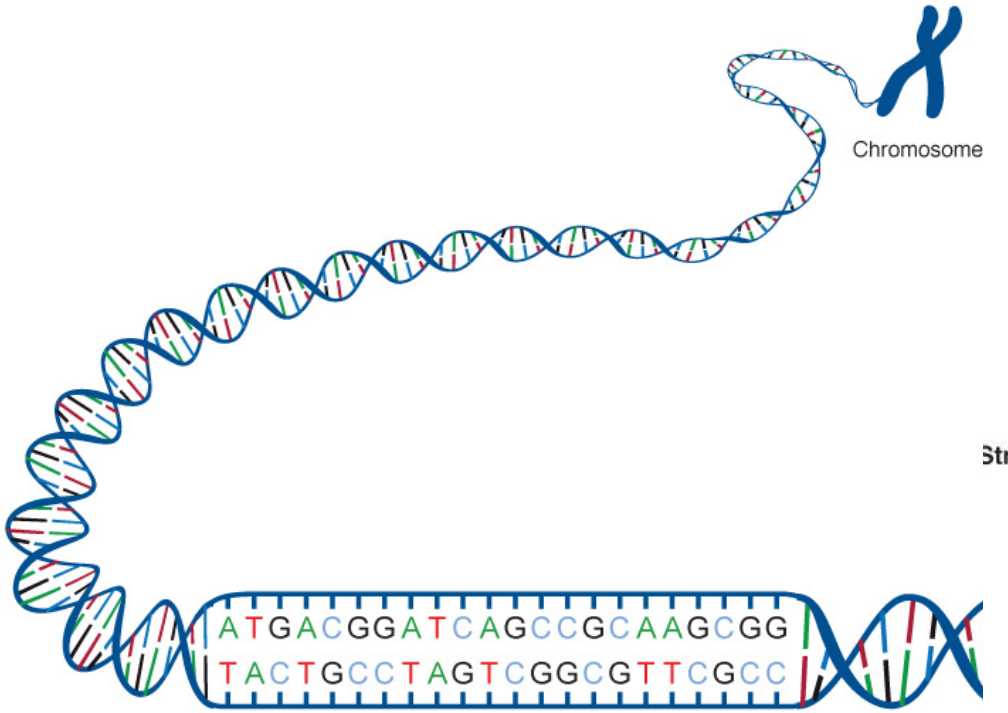


Allele 1 = C
Allele 2 = A
Bi-allelic combinations = C/C, C/A, A/A

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
- deletions

'indels'

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

- There are three chip-manufacturers: Illumina, Affymetrix & Perlegen

<u>Affymetrix:</u>

6.0 chip

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

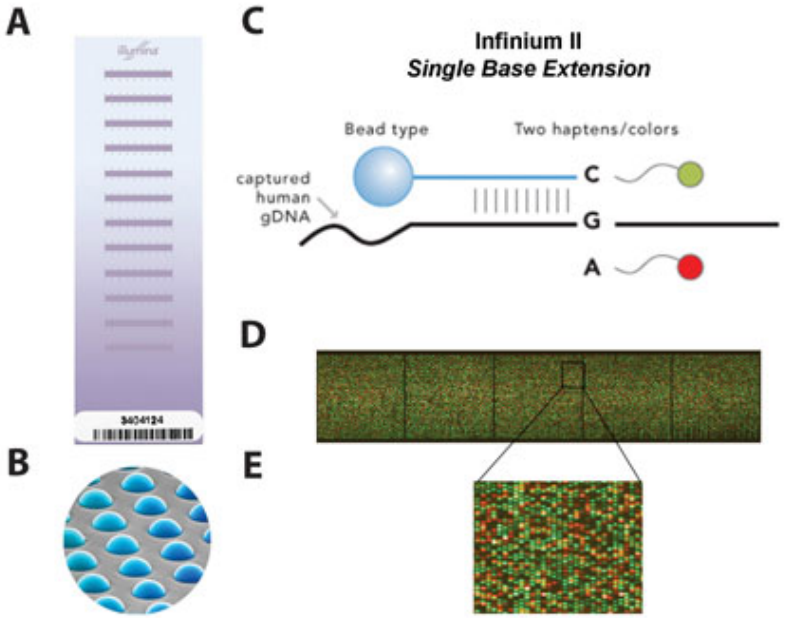
<u>Illumina:</u>

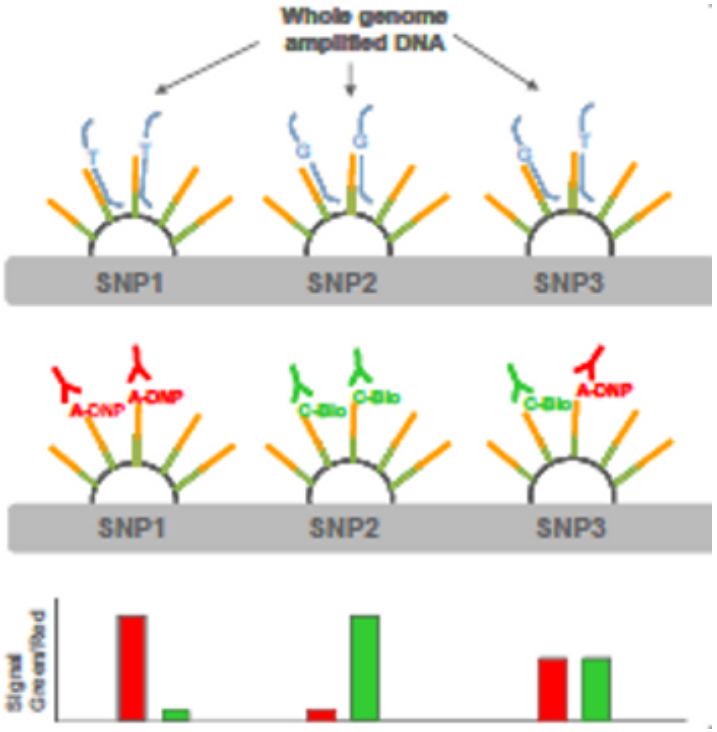
Human1M BeadChip

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

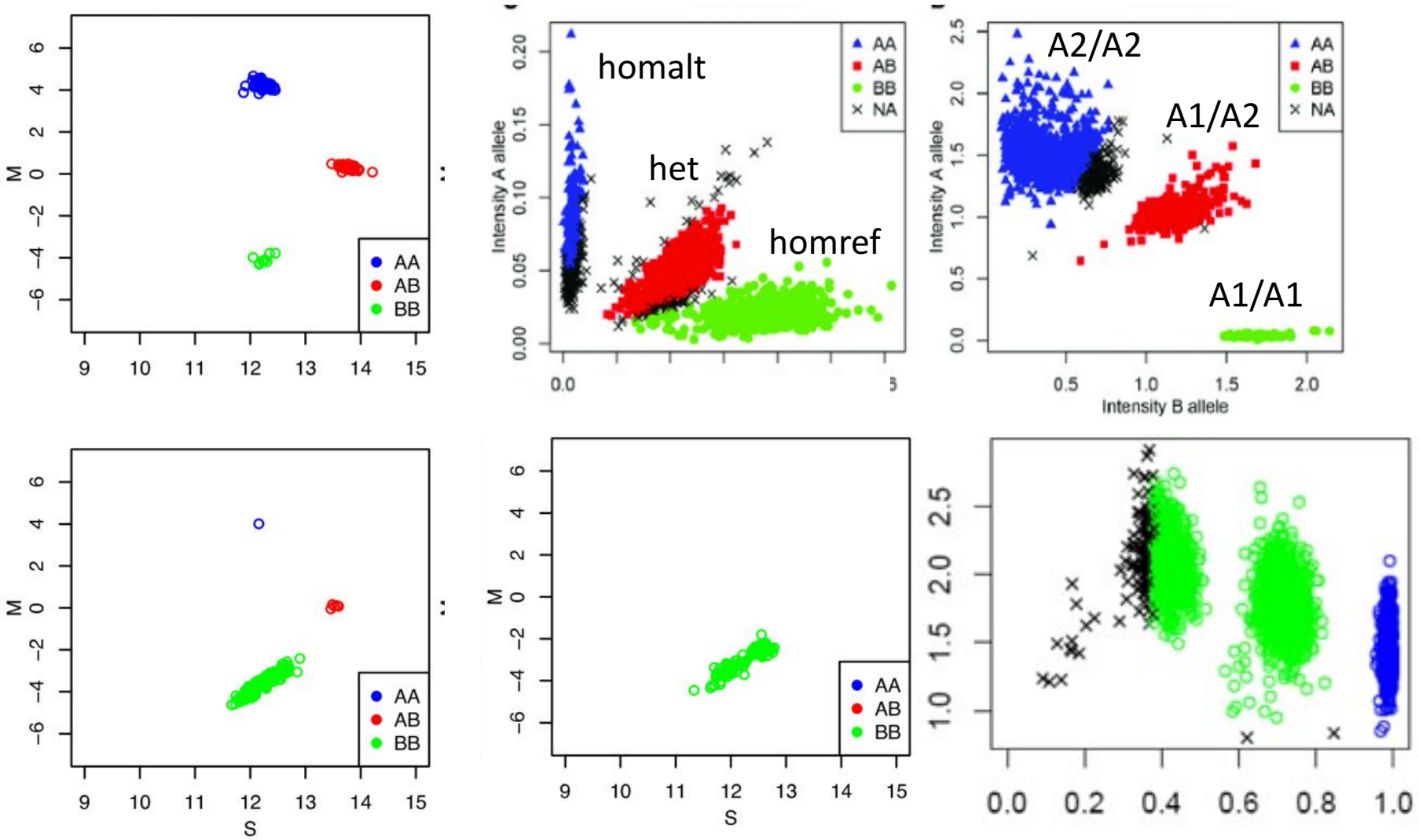
- Intensity measures are produced for both alleles. Genotypes are assigned based on clustering of these two intensities.

From DNA to data

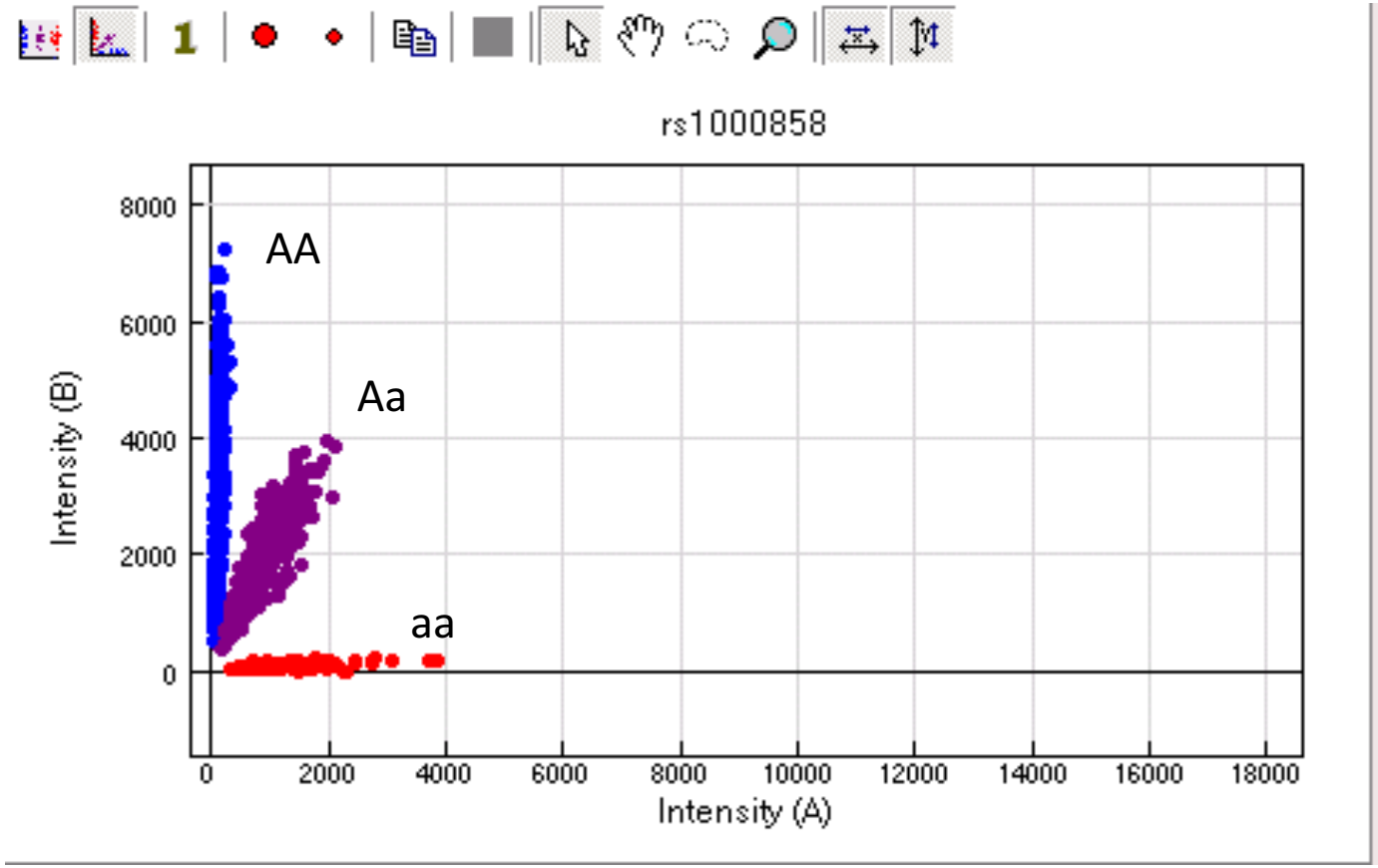




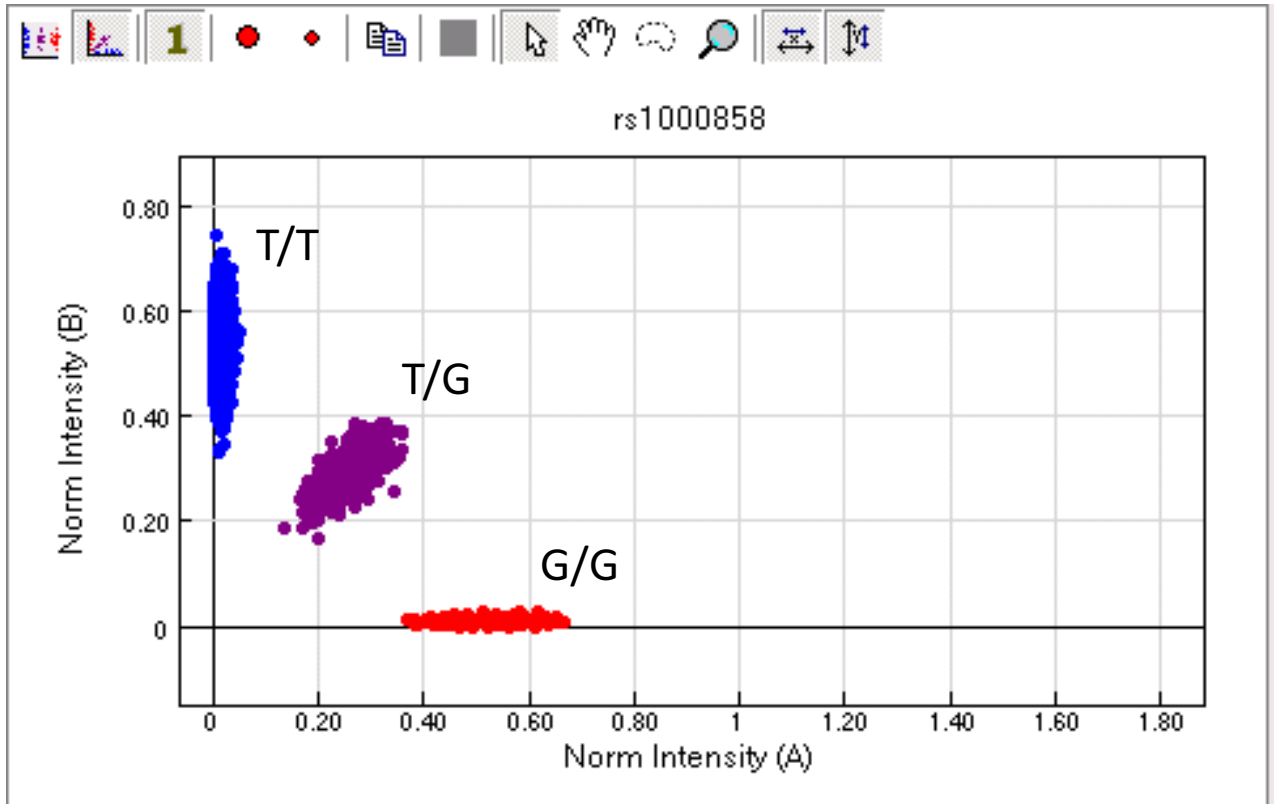
Genotype Intensities



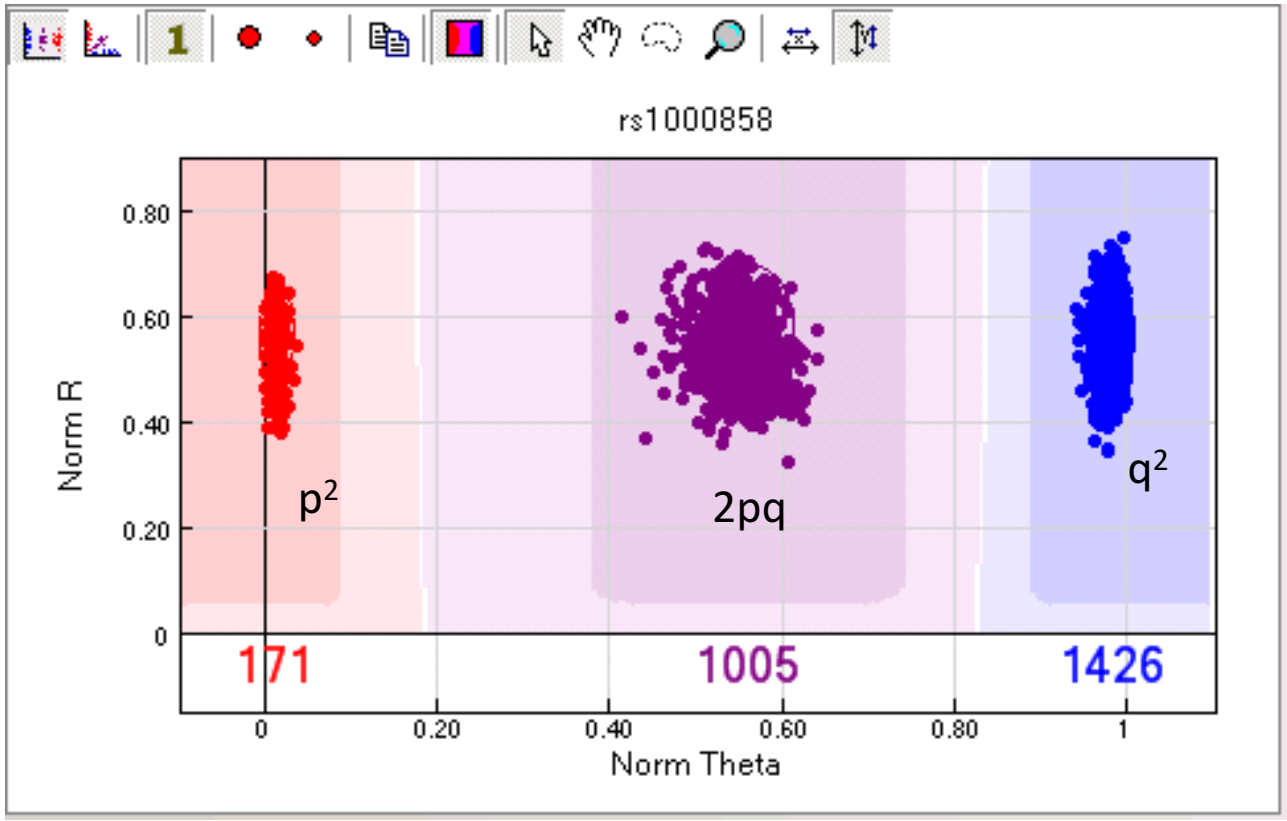
Good SNP (Illumina chip)



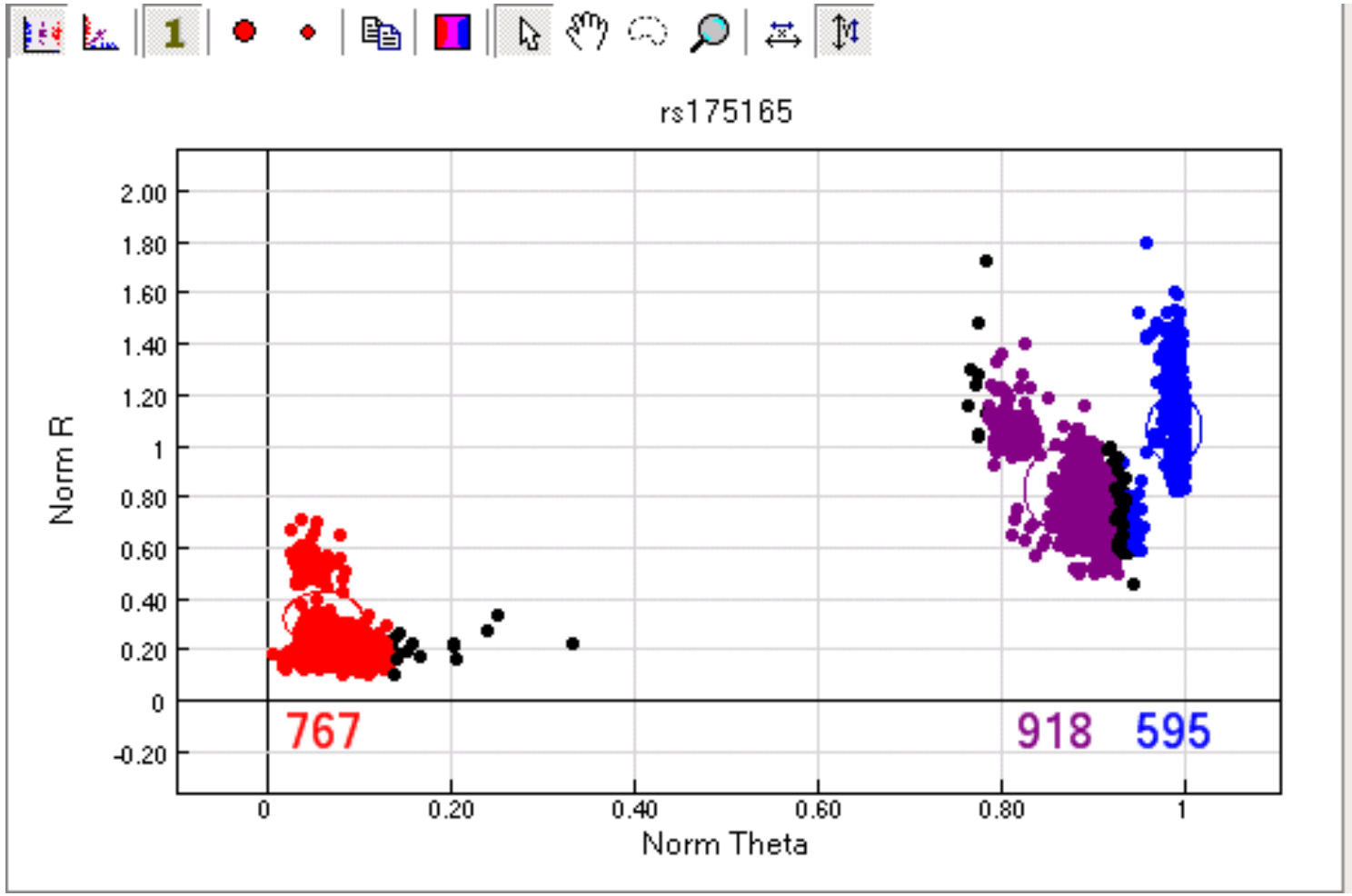
Same SNP, normalized intensities



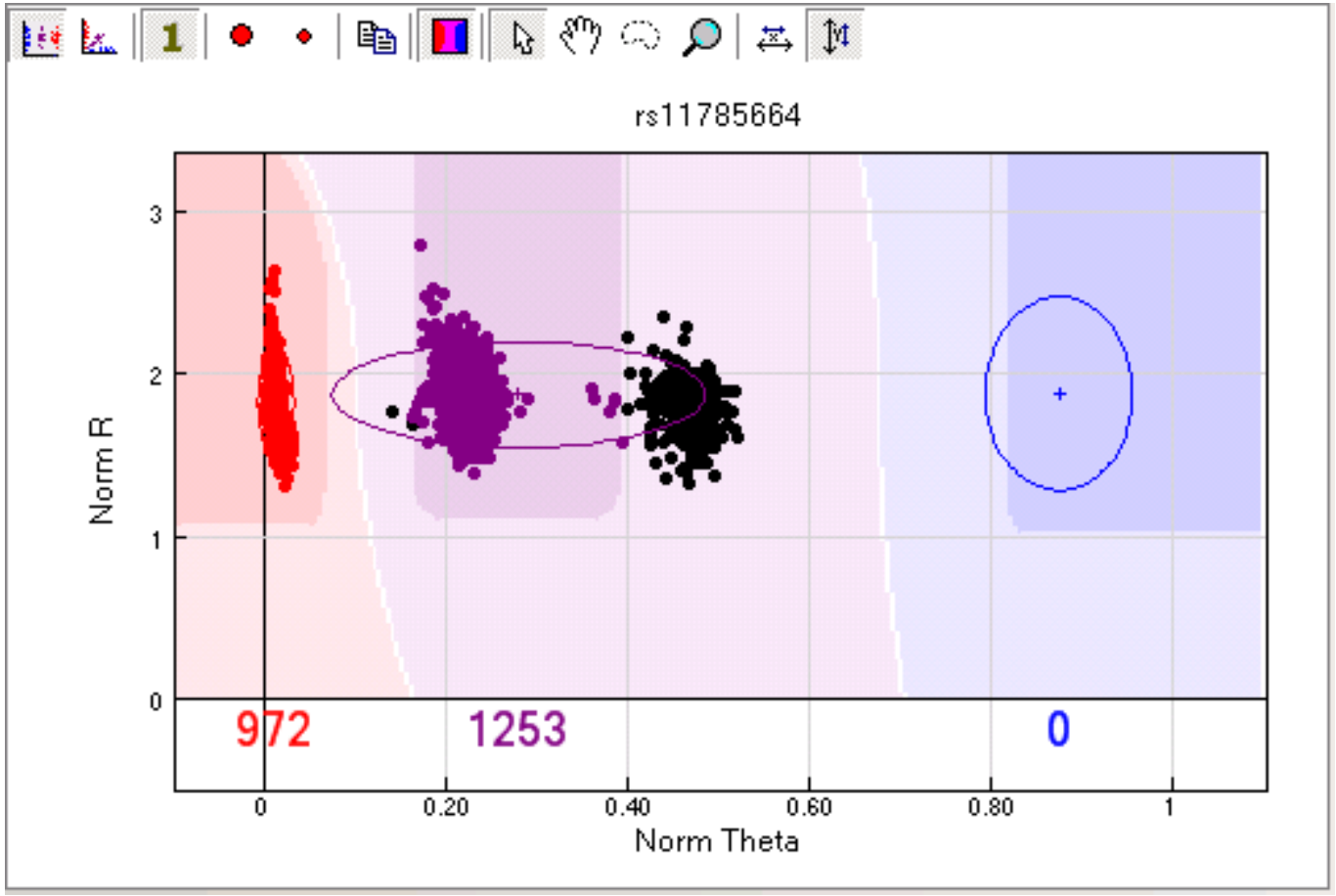
Same SNP, different view



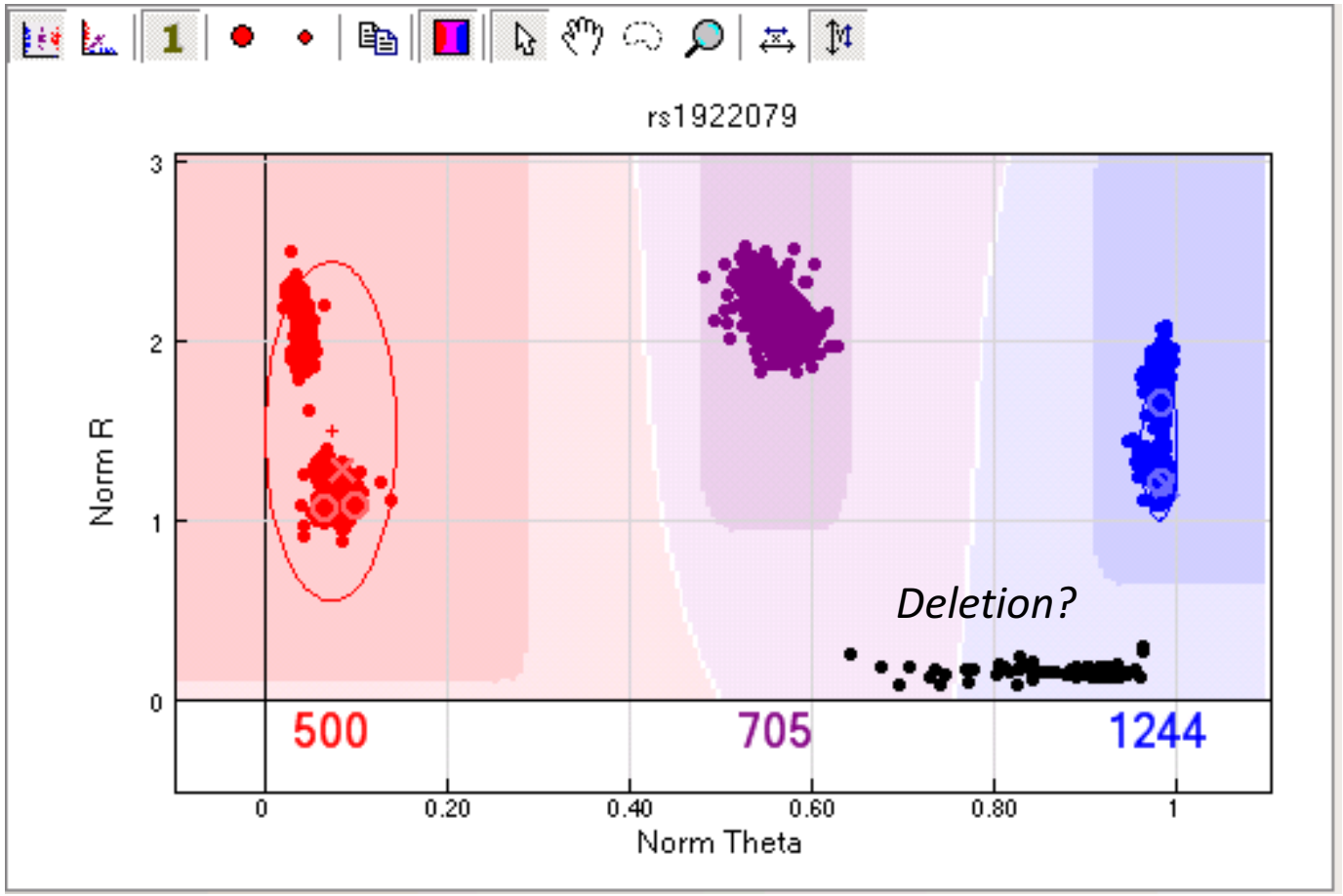
Bad SNP


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

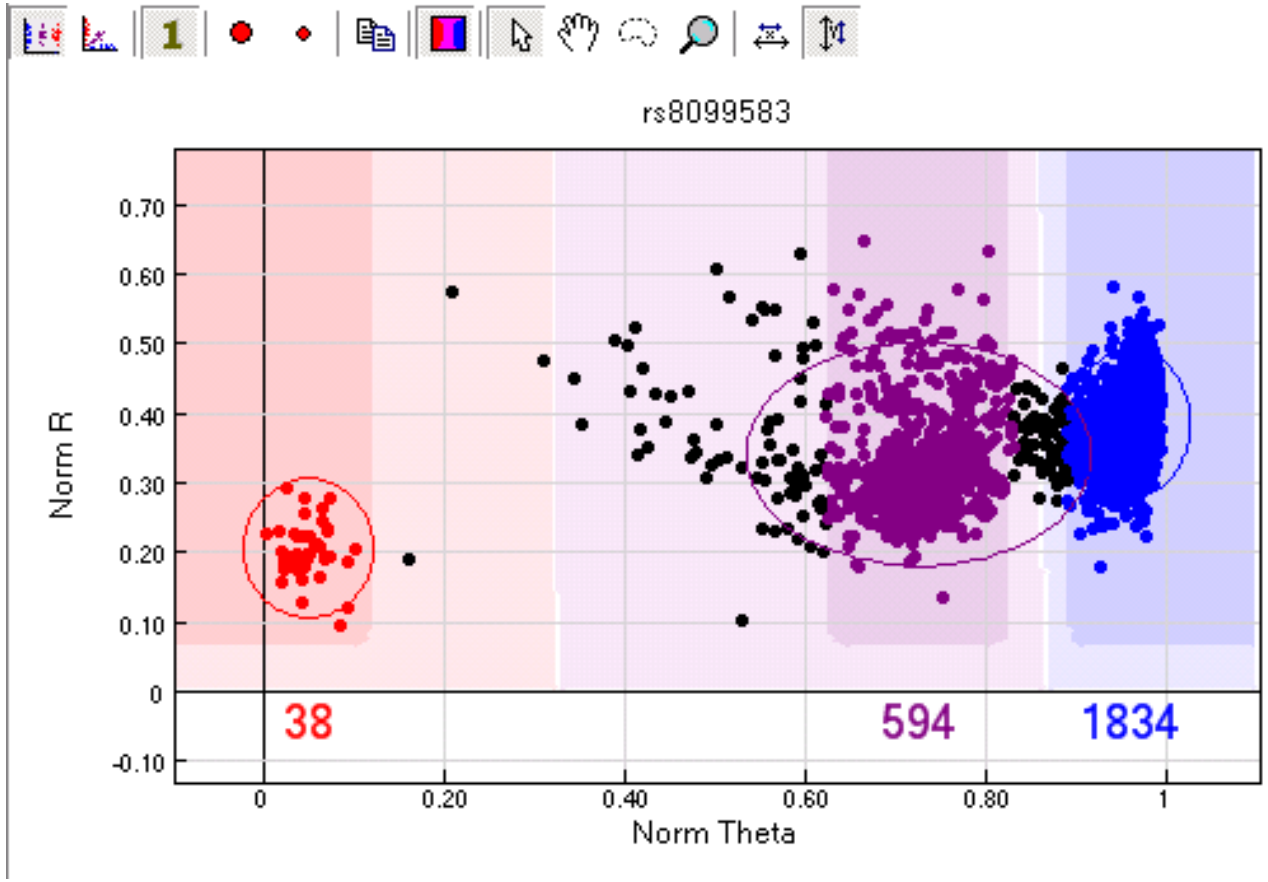
Another bad SNP



Another bad SNP



Another bad SNP



PLINK data format of GWAS data

Subjects

.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
CHR = chromosome
POS = position
A1 = 0 allele
A2 = 1 allele

Genetic variants

.bim file (or .map file)

CHR	POS	SNP ID	A1	A2
1	11852412	rs45496998	A	G
1	11853994	rs116620395	G	C
1	11854457	rs4846051	A	G
1	11854476	rs1801131	G	T
1	11854500	rs200137991	A	C
1	11854823	rs121434296	A	G
1	11855218	1:11855218	G	A
1	11855218	rs121434297	G	A
1	11856328	rs190090719	G	A
1	11856378	rs1801133	A	G
1	11857788	rs17421511	A	G
1	11859046	GSA-rs375817840	A	G
1	11859636	GSA-rs74683406	A	G
1	11861223	rs121434295	T	C
1	11862778	rs17367504	G	A
1	11863022	seq-rs201618781	T	C
1	11863038	rs138189536	A	G
1	11863562	chr1-11863562	A	G
1	11865250	GSA-rs3753583	A	G
1	11870279	GSA-rs34994762	G	A
1	11886226	rs202066883	G	C

Genotype data

.ped file

P1	A	A	A	C	C	G	T	T	A	A	T	T
P2	A	C	A	A	C	G	G	T	A	C	T	T
P3	C	C	A	C	G	G	T	T	A	A	T	T
P4	C	C	A	A	G	G	G	T	A	A	T	T

.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
  - Deviations from Hardy-Weinberg
  - 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
  - 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)
  - 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
  - Important to re-analyze original intensity of significant associations (whenever possible)
  - For meta-analysis, examining heterogeneity of SNP effect

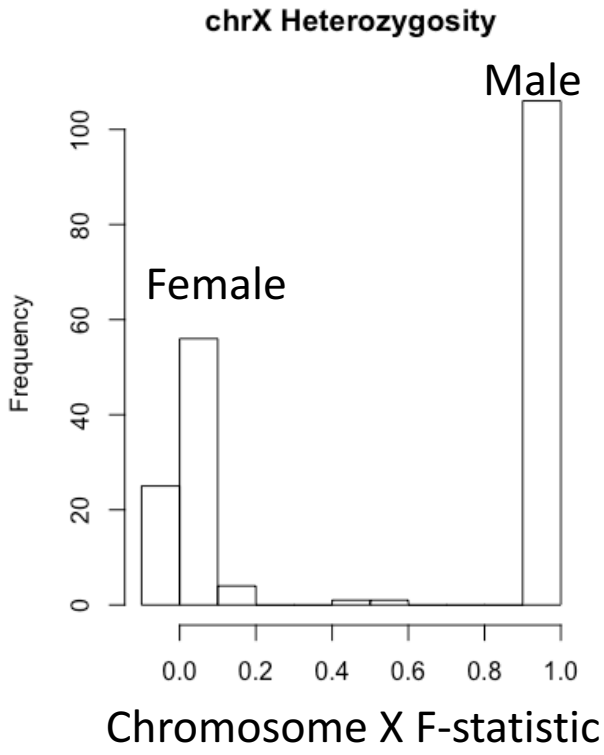
Preliminary QC steps

- 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

Confirming genetic sex

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

Female sex = X/X
Male sex = X/Y



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 (or “missingness”)

- 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

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 frequencies 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
```

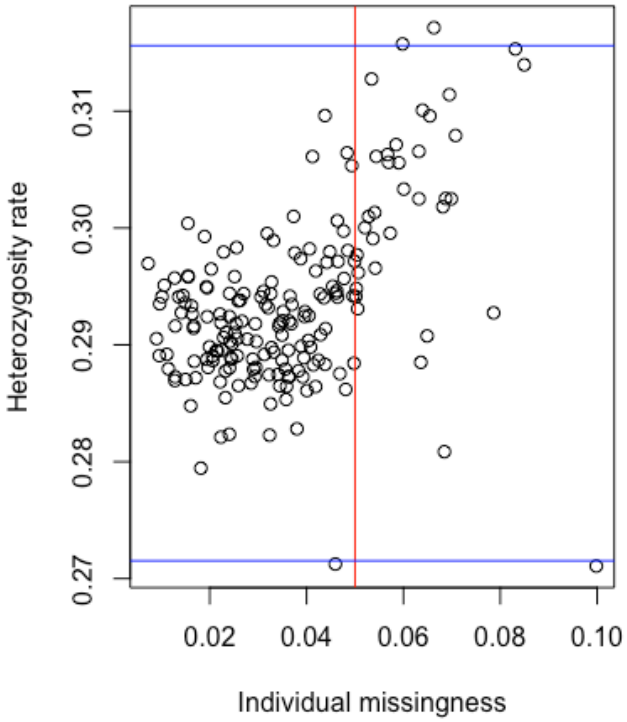
which contains the fields, one row per person in the file:

```
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.

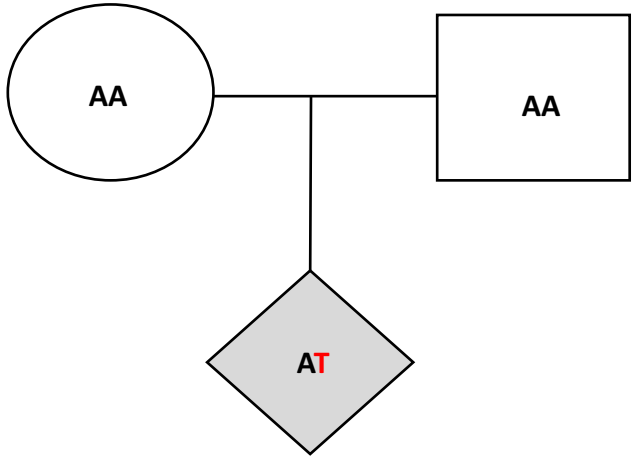
Note The estimate of F can sometimes be negative. Often this will just reflect random sampling error, but a result that is strongly negative (i.e. an individual has *fewer* homozygotes than one would expect by chance at the genome-wide level) can reflect other factors, e.g. sample contamination events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



Mendel errors

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

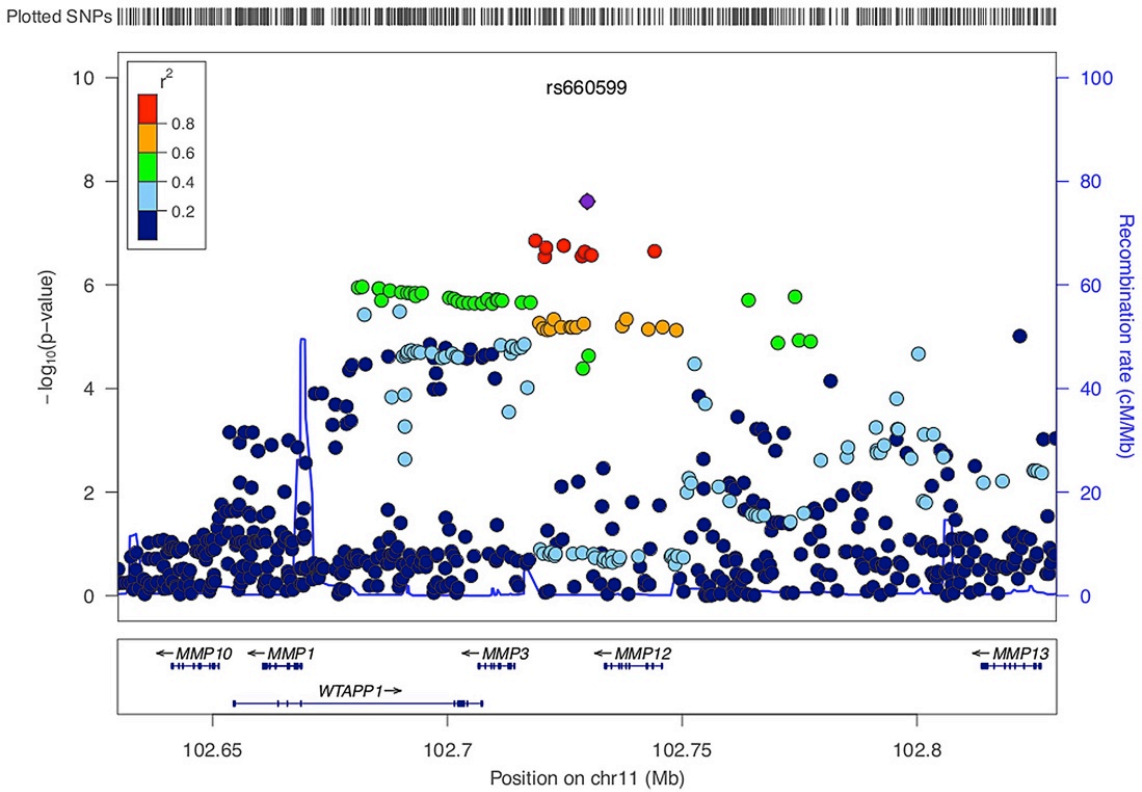
--mendel scans the dataset for Mendel errors, writing a set of reports to plink{.mendel,.imendel,.fmendel,.lmendel}. Haploid and mitochondrial data are ignored. The errors are classified as follows, where '1' refers to the A1 (usually minor) allele and '2' refers to A2:

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

- 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
- LD <u>must</u> be taken into account when examining genetic relatedness, population stratification, and interpreting association



Breakout session (5 min)

- Breakout into small groups
- Introduce yourself to everyone
- Person with earliest letter in their FIRST name will be the note taker
  - E.g. Aaron is the note taker, not Zenia
  - E.g. Aaron is the note taker, not Abel
- Ask any questions you have:
  - “I didn’t understand what …. meant”
  - “I’m confused by the sample/SNP difference in heterzygosity”
- Note takers relay unanswered questions from the breakout session

Lecture Format

- Part 1 (~40 minutes)
  - Goals of GWAS
  - What does the data look like?
  - GWAS Quality Control (QC)
  - 5 min breakout session
- Part 2 (~40 minutes)
  - Relatedness checking
  - Population stratification
  - Principal components analysis (PCA)
  - Imputation
  - 5 min breakout session
- Part 3 (~40 minutes)
  - Association testing
  - Meta-analysis
  - Polygenic Scoring
  - 5 min breakout session
- Preparation for module 3 / additional reading / lingering questions

Genetic Relatedness

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

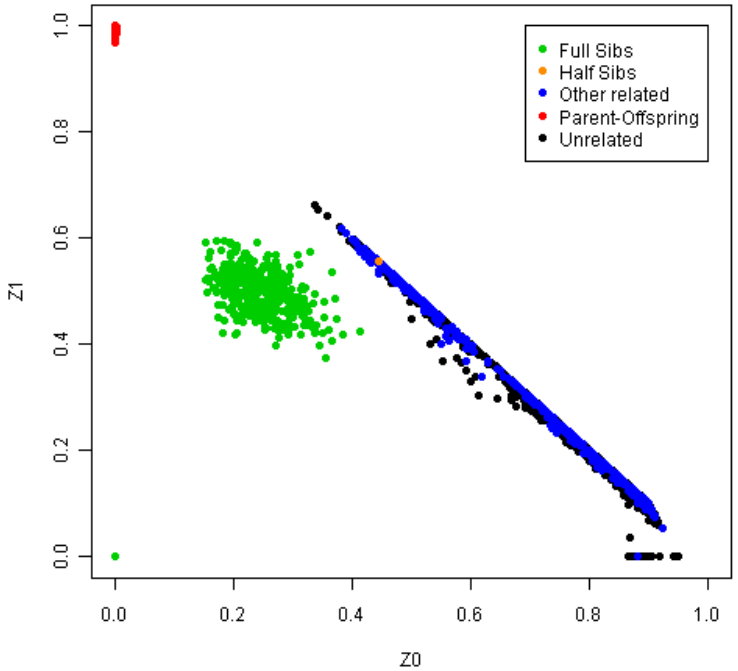
- Question: How much does a pair of samples share 0, 1, or both alleles?
- Identical twins: Shares both alleles across entire genome (barring mutation events)
- Requires using LD-pruned SNPs for accurate estimates
  - Want each SNP to be an “independent” marker
- Used to both “confirm” and “filter” related individuals

Checking genotype relatedness across samples

Example of .genome file in PLINK

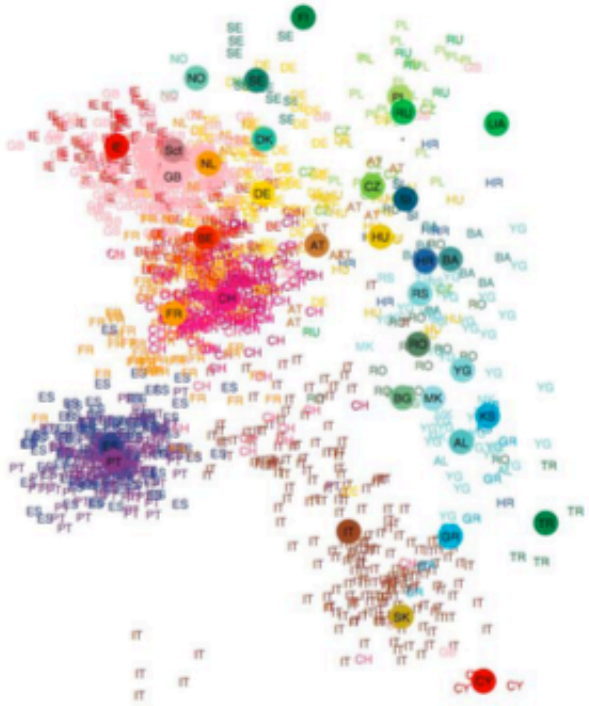
FID1	IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	PI_HAT	PHE	DST	PPC	RATIO
NA20505	NA20505	NA20506	NA20506	UN	NA	0.9872	0.0000	0.0128	0.0128	-1	0.771435	0.3446	1.9712
NA20505	NA20505	NA20502	NA20502	UN	NA	0.9888	0.0096	0.0016	0.0064	-1	0.770233	0.3950	1.9808
NA20505	NA20505	NA20528	NA20528	UN	NA	0.9733	0.0267	0.0000	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

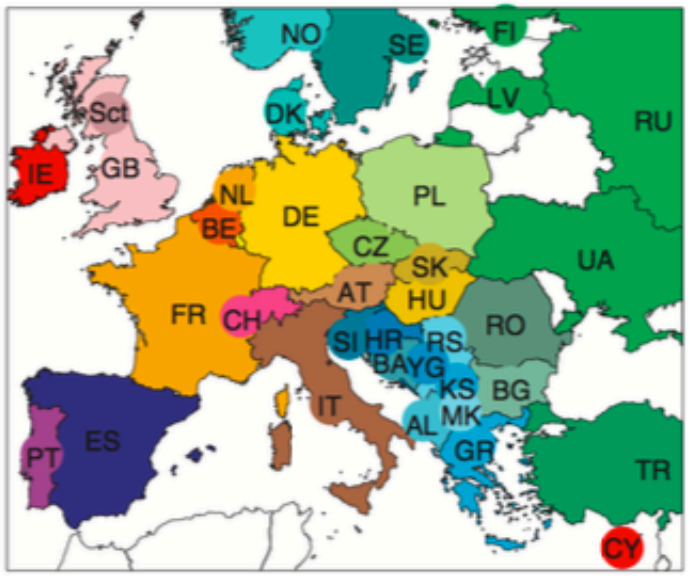
	Probability of Sharing IBD Alleles		
Relative Pair	π_0	π_1	π_2
MZ Twins	0	0	1
Full Sibs	0.25	0.50	0.25
Parent-Offspring	0	1	0
First Cousin	0.75	0.25	0
Grandparent-Grandchild	0.50	0.50	0
Half-Sibs	0.50	0.50	0
Avuncular	0.50	0.50	0



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





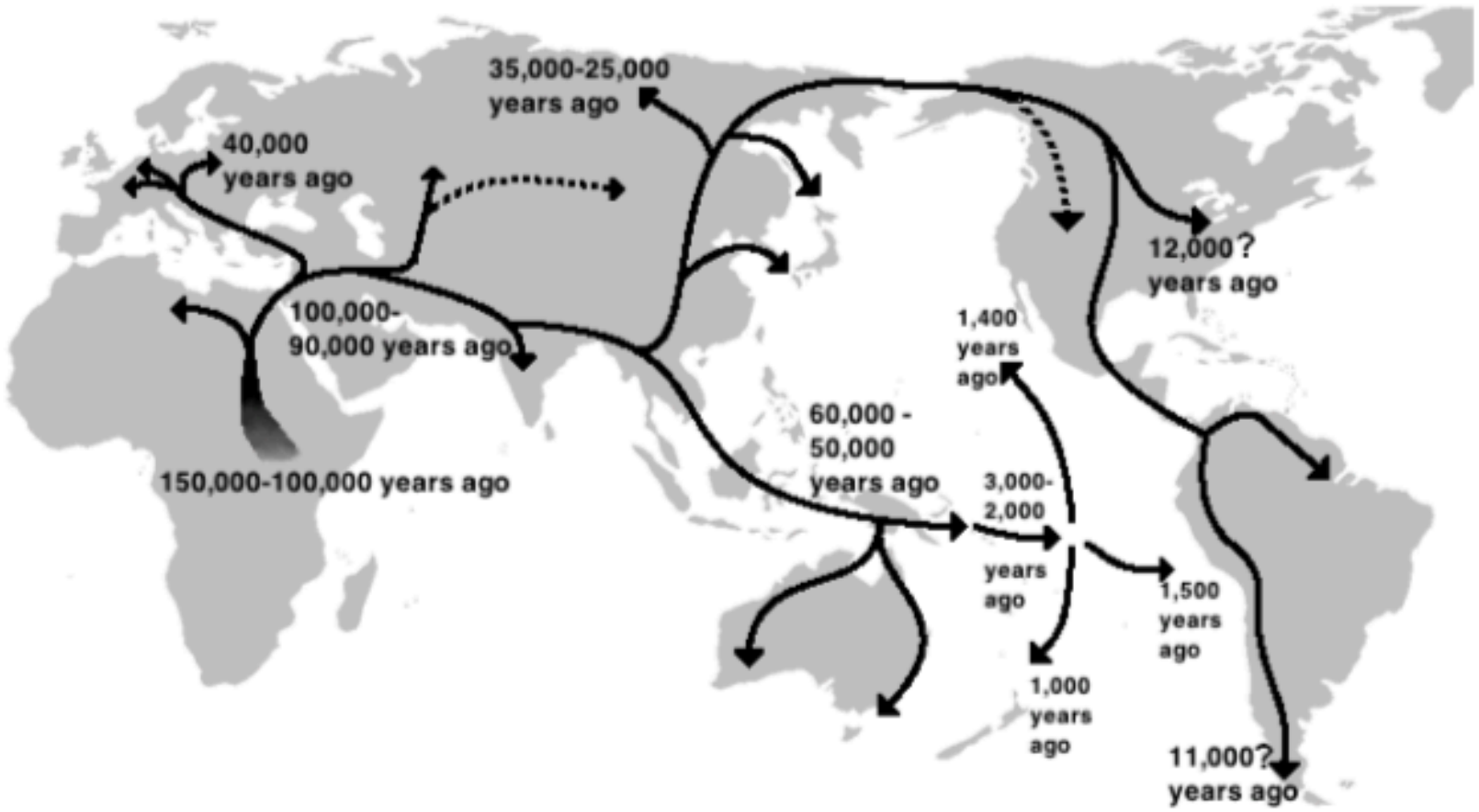
Population Stratification

Abdel Abdellaoui

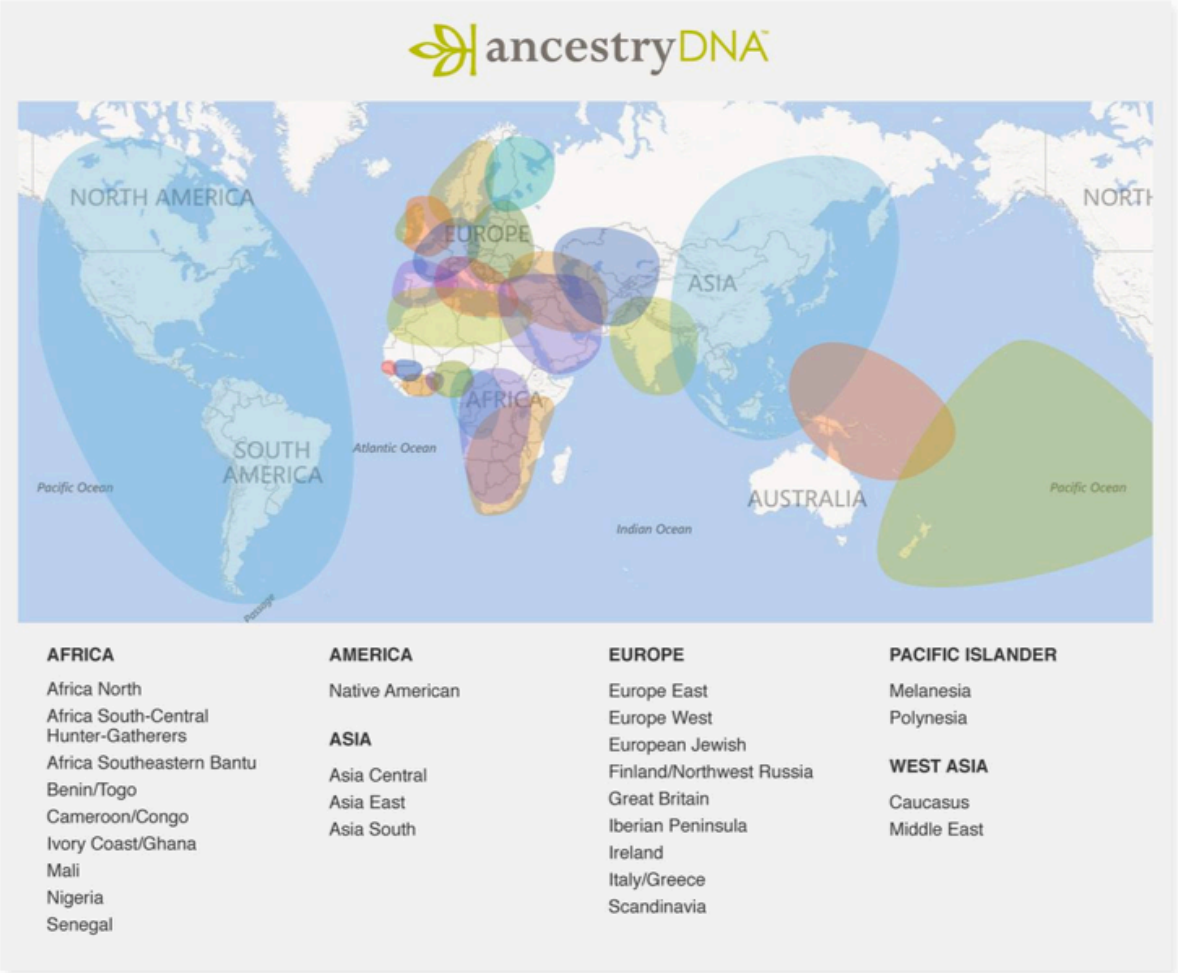
Department of Psychiatry, Amsterdam UMC, University of Amsterdam

Amsterdam UMC
Universitair Medische Centra

E-mail: a.abdellaoui@amc.nl


35,000-25,000
years ago
40,000
years ago
12,000?
years ago
100,000-
90,000 years ago
1,400
years
ago
60,000 -
50,000
years ago
150,000-100,000 years ago
3,000-
2,000
years
ago
1,500
years
ago
1,000
years
ago
11,000?
years ago

Largest patterns of genetic variation = ancestry



88% of GWAS participants is of European descent

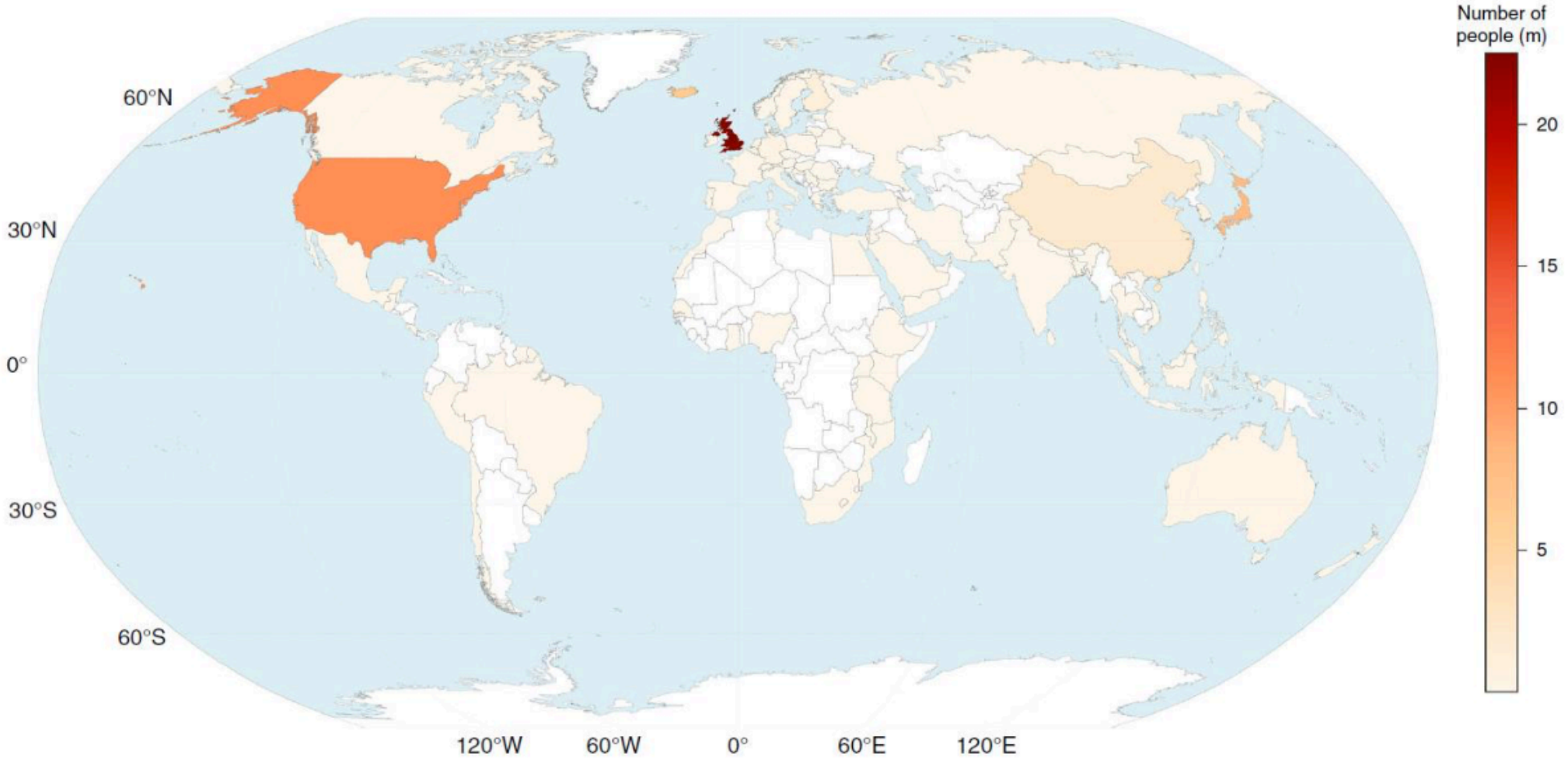


Fig. 3 A Choropleth Map of the Concentration of GWAS Participant Recruitment. A choropleth map (Robinson projection) detailing the geographic recruitment of GWAS participants. Source: NHGRI-EBI GWAS Catalog, Natural Earth (v4.0.0) and the CIA World Factbook. Replication material provides a per-capita population adjusted version

Population stratification

- Population stratification = a systematic difference in allele frequencies between (sub)populations due to different ancestry.
- Can cause false positives if the trait values also differ between the (sub)populations.

Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear allele frequency difference between Americans and Chinese

Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear difference between Americans and Chinese in proportion of "cases" and "controls"

Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Sample 1 + 2 = Americans + Chinese: $\chi^2=34.2$, $p=4.9 \times 10^{-9}$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	640	340	980
Allele 2	400	100	500
Total	1040	440	1480

Dealing with population stratification

Ways to deal with population stratification:

- Genomic Control (GC)
- Principal Component Analysis
- Within Family Association
- Mixed Linear Modeling

nature genetics

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

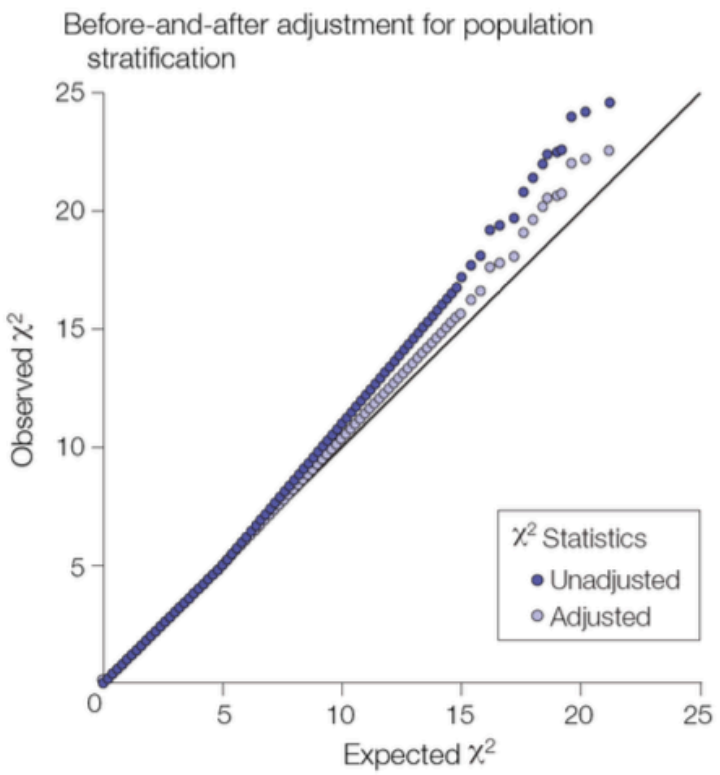
nature genetics

Advantages and pitfalls in the application of mixed-model association methods

Jian Yang[1,2,8], Noah A Zaitlen[3,8], Michael E Goddard[4,9], Peter M Visscher[1,2,9] & Alkes L Price[5–7,9]

Genomic Control (GC)

- Population stratification can result in higher test statistics (= lower *p*-values)
- The genomic control method estimates the factor with which the test statistics are inflated due to population stratification → λ
- Dividing by λ cancels this effect out for all SNPs:
  - Unadjusted: $\lambda\chi^2$
  - Adjusted: χ^2



Genomic Control (GC)

- **λ** is measured by dividing the **median** of the distribution of the chi-square statistics from the **actual tests** by the **median** of the chi-square distribution **under the null**.
- Then, GC applies its correction by dividing the actual association test chi-square statistic results by this λ, thus making these results appropriately more pessimistic.
- GC is too conservative if the trait is **highly polygenic** (i.e. the median test statistic does not represent the null distribution).
- **LD Score regression** can be used to estimate a more powerful and accurate correction factor than GC.

European Journal of Human Genetics (2011) 19, 807–812

www.nature.com/ejhg

ARTICLE

Genomic inflation factors under polygenic inheritance

Jian Yang*,[1], Michael N Weedon[2], Shaun Purcell[3,4], Guillaume Lettre[5], Karol Estrada[6], Cristen J Willer[7], Albert V Smith[8], Erik Ingelsson[9], Jeffrey R O'Connell[10], Massimo Mangino[11], Reedik Mägi[12], Pamela A Madden[13], Andrew C Heath[13], Dale R Nyholt[1], Nicholas G Martin[1], Grant W Montgomery[1], Timothy M Frayling[2], Joel N Hirschhorn[3,14,15], Mark I McCarthy[12,16], Michael E Goddard[17], Peter M Visscher[1] and the GIANT Consortium

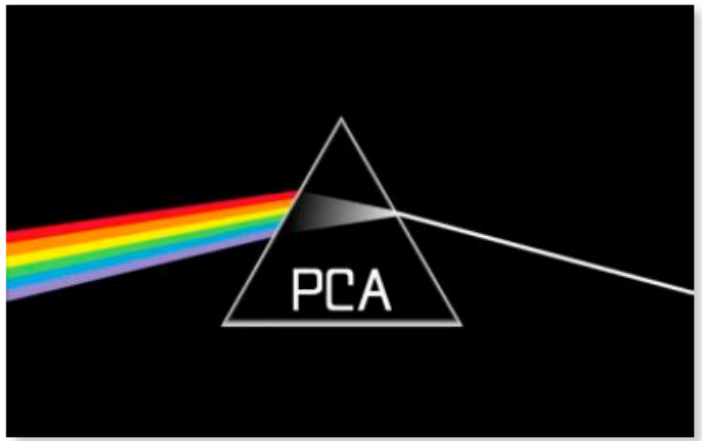
nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6], Schizophrenia Working Group of the Psychiatric Genomics Consortium[7], Nick Patterson[1], Mark J Daly[1–3], Alkes L Price[1,4,8] & Benjamin M Neale[1–3]

Principal Component Analysis (PCA)

- PCA is a statistical method for exploring large number of measurements (e.g., SNPs) by reducing the measurements to fewer principal components (PCs) that explain the main patterns of variation:
  - The first PC is the mathematical combination of measurements that accounts for the largest amount of variability in the data.
  - The second PC (uncorrelated with the first) accounts for the second largest amount of variability.
  - Etc...



nature genetics

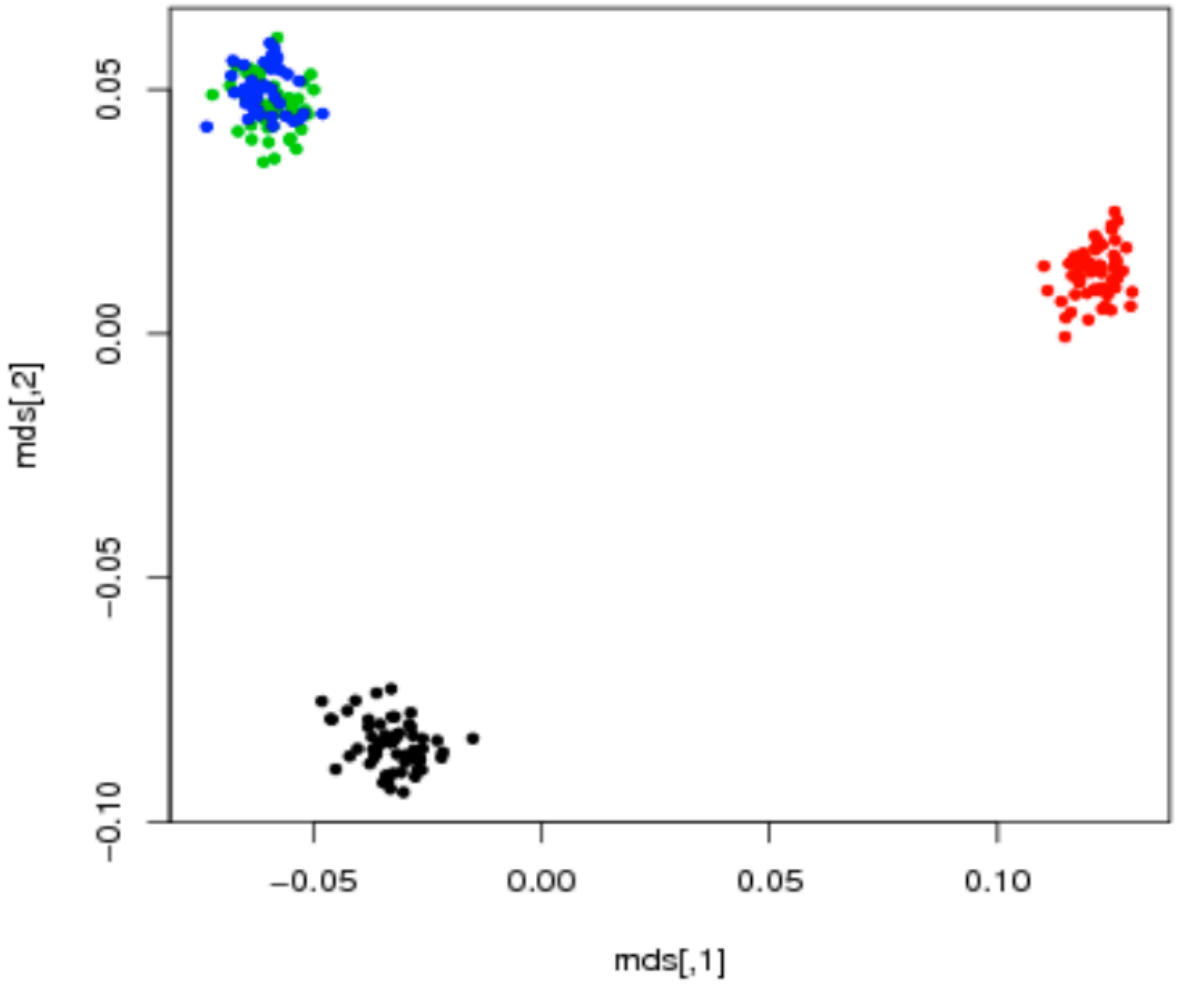
Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

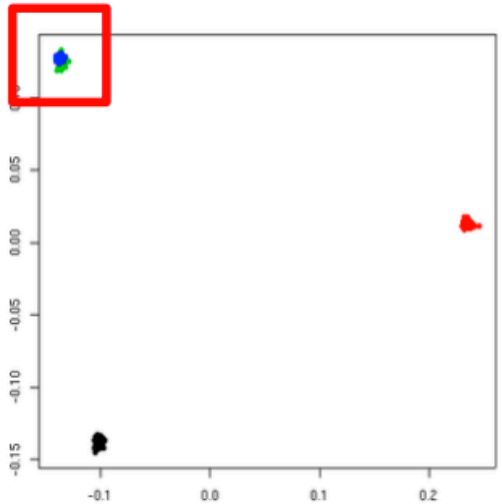
Principal Component Analysis (PCA)

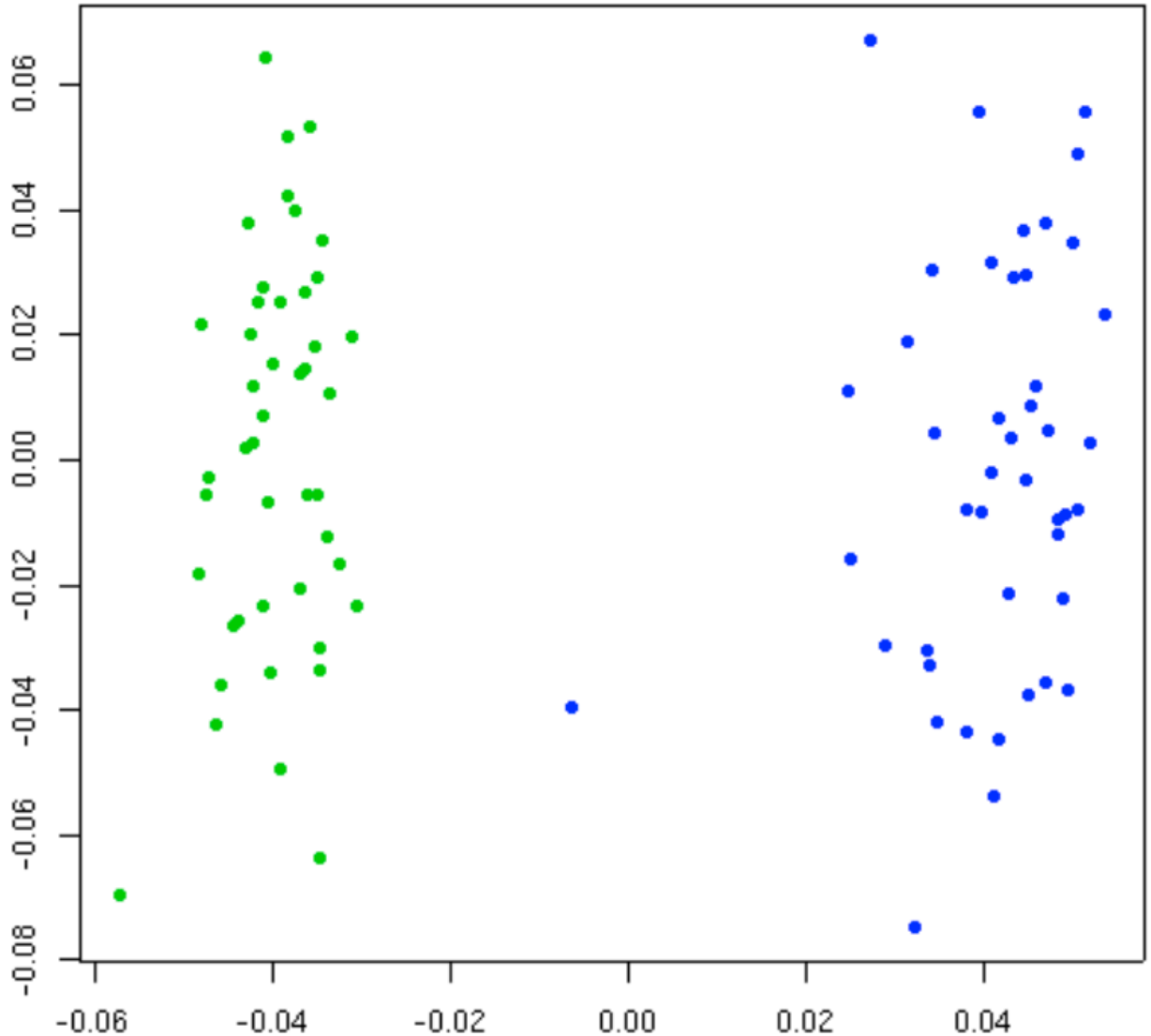
CEPH/European
Yoruba
Han Chinese
Japanese



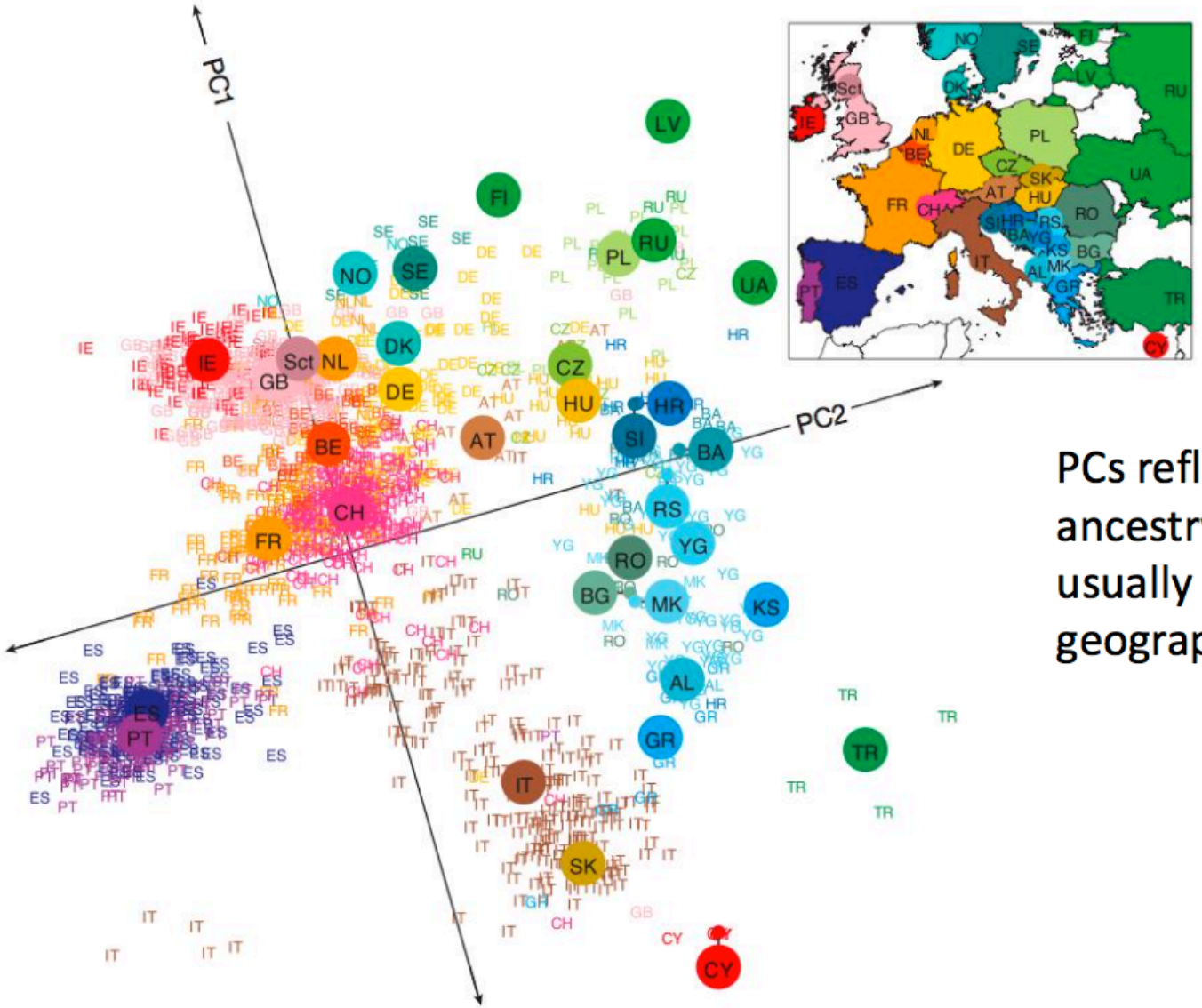
Principal Component Analysis (PCA)

Han Chinese
Japanese





Fine-scale genetic variation reflects geography

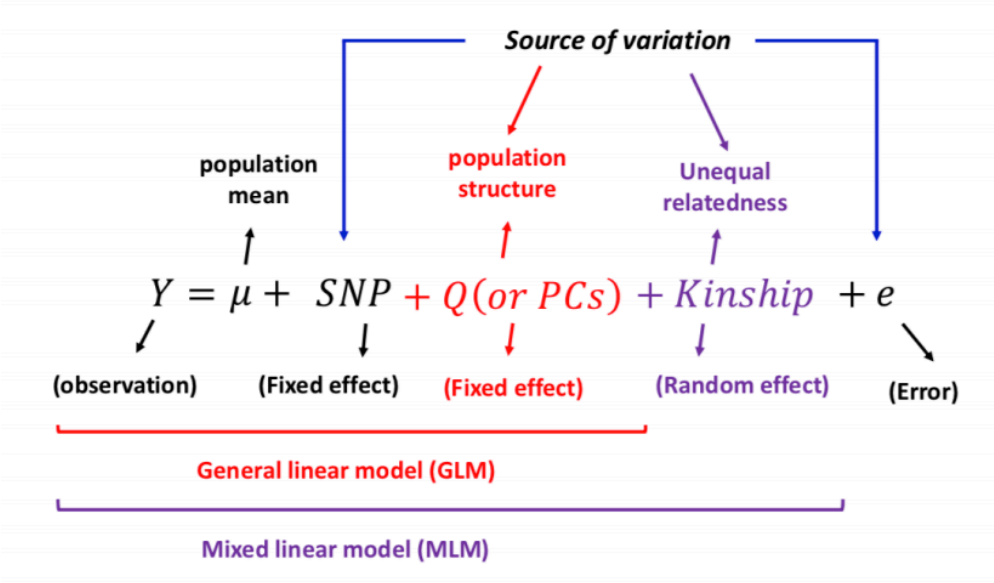


PCs reflecting ancestry differences usually correlate with geography.

Genes mirror geography within Europe (Novembre et al, 2008)

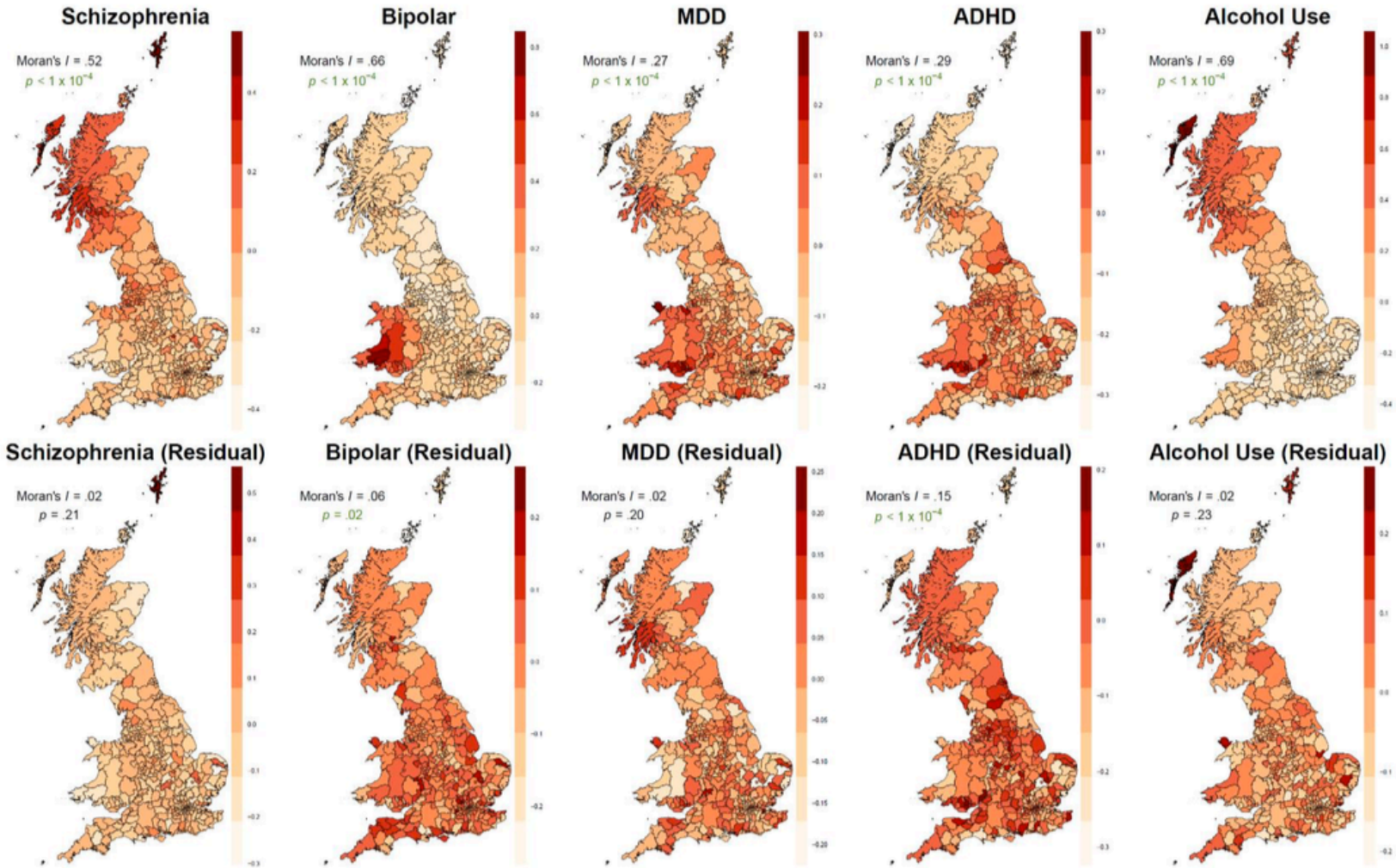
Using PCs in GWAS studies

- Include as covariates in a regression model
- PCs that associate to phenotype very important to include
- Logistic regression sensitive to inclusion of many PCs
- Linear regression more robust
- Mixed linear models can replace PCs with genetic relatedness (GRM) matrix
  - Adding PCs as well still seems to help..



Ancestry differences in Great Britain

- Polygenic scores, before and after regressing out 100 PCs

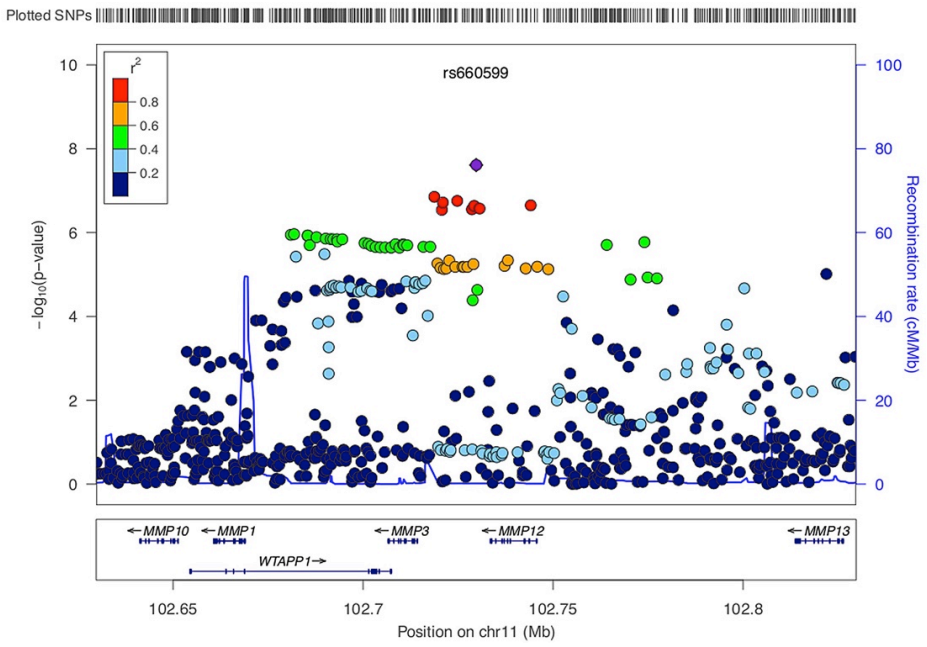


Genetic Consequences of Social Stratification in Great Britain (Abdellaoui et al, 2018)

Phasing and Imputation

Imputation

…but I want to analyze more SNPs!!!!



Impute: *"represent as being done, caused, or **possessed**"*

Main goal: Using local Linkage Disequilibrium (LD) patterns to infer the genotype of a SNP not on your array

Main process: Map your GWAS array SNPs to whole-genome sequence data (i.e. "reference panels") to impute SNPs not on your array

Reference panels / Haplotypes

HapMap (haplotype map) Project

270 whole-genome sequenced samples:

30 parent-offspring trios of the Yoruba from Ibadan, Nigeria (YRI)
30 trios of Utah residents with European ancestry (CEU)
45 individuals from Beijing, China (CHB)
45 individuals from Tokyo, Japan (JPT)

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*.

Reference panels / Haplotypes

1000 Genomes Project

Phase 1: 1,092 individuals from 14 populations..

Phase 3: 2,504 individuals from 26 populations (~500 samples form each 5 continental ancestry groups, with ~5 populations for each group)

Population		Code	Population Color	Continental Group Color	Analysis Panel	Phase 1	Phase 3
African ancestry							
Esan in Nigeria	Esan	ESN			AFR		99
Gambian in Western Division, Mandinka	Gambian	GWD			AFR		113
Luhya in Webuye, Kenya	Luhya	LWK			AFR	97	99
Mende in Sierra Leone	Mende	MSL			AFR		85
Yoruba in Ibadan, Nigeria	Yoruba	YRI			AFR	88	108
African Caribbean in Barbados	Barbadian	ACB			AFR/AMR		96
People with African Ancestry in Southwest USA	African-American SW	ASW			AFR/AMR	61	61
Americas							
Colombians in Medellin, Colombia	Colombian	CLM			AMR	60	94
People with Mexican Ancestry in Los Angeles, CA, USA	Mexican-American	MXL			AMR	66	64
Peruvians in Lima, Peru	Peruvian	PEL			AMR		85
Puerto Ricans in Puerto Rico	Puerto Rican	PUR			AMR	55	104
East Asian ancestry							
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX			EAS		93
Han Chinese in Beijing, China	Han Chinese	CHB			EAS	97	103
Southern Han Chinese	Southern Han Chinese	CHS			EAS	100	105
Japanese in Tokyo, Japan	Japanese	JPT			EAS	89	104
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV			EAS		99
European ancestry							
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU			EUR	85	99
British in England and Scotland	British	GBR			EUR	89	91
Finnish in Finland	Finnish	FIN			EUR	93	99
Iberian Populations in Spain	Spanish	IBS			EUR	14	107
Toscani in Italia	Tuscan	TSI			EUR	98	107
South Asian ancestry							
Bengali in Bangladesh	Bengali	BEB			SAS		86
Gujarati Indians in Houston, TX, USA	Gujarati	GIH			SAS		103
Indian Telugu in the UK	Telugu	ITU			SAS		102
Punjabi in Lahore, Pakistan	Punjabi	PJL			SAS		96
Sri Lankan Tamil in the UK	Tamil	STU			SAS		102
Total						**1092**	**2504**

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*.
The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*.

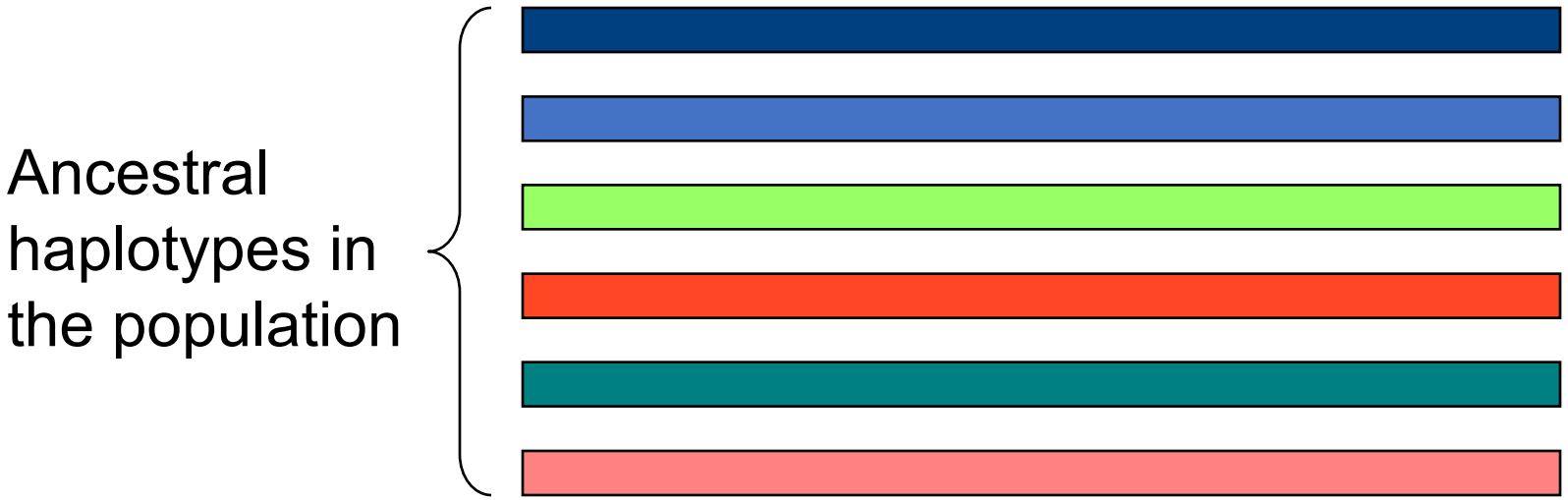
The Haplotype Reference Consortium (HRC)

nature genetics

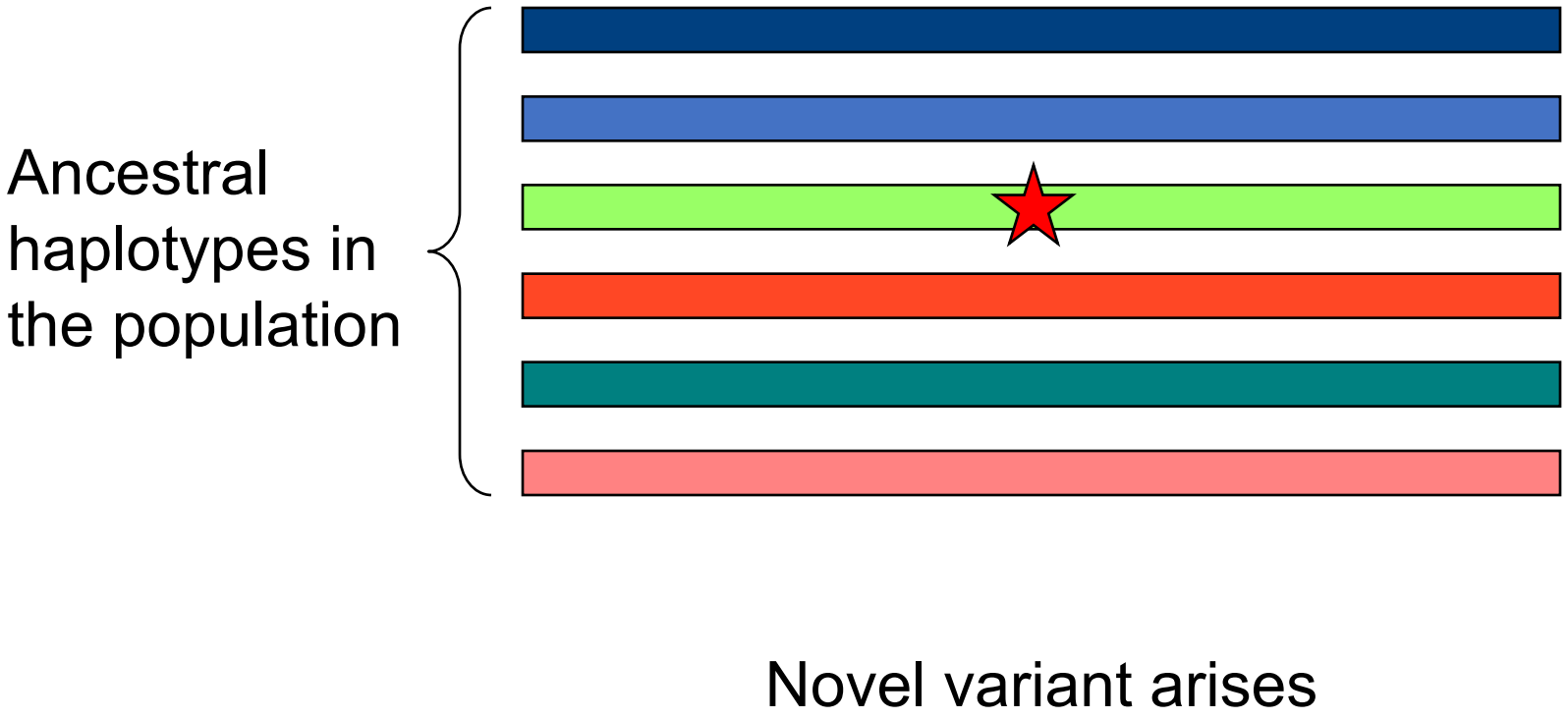
A reference panel of 64,976 haplotypes for genotype imputation

The Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*.

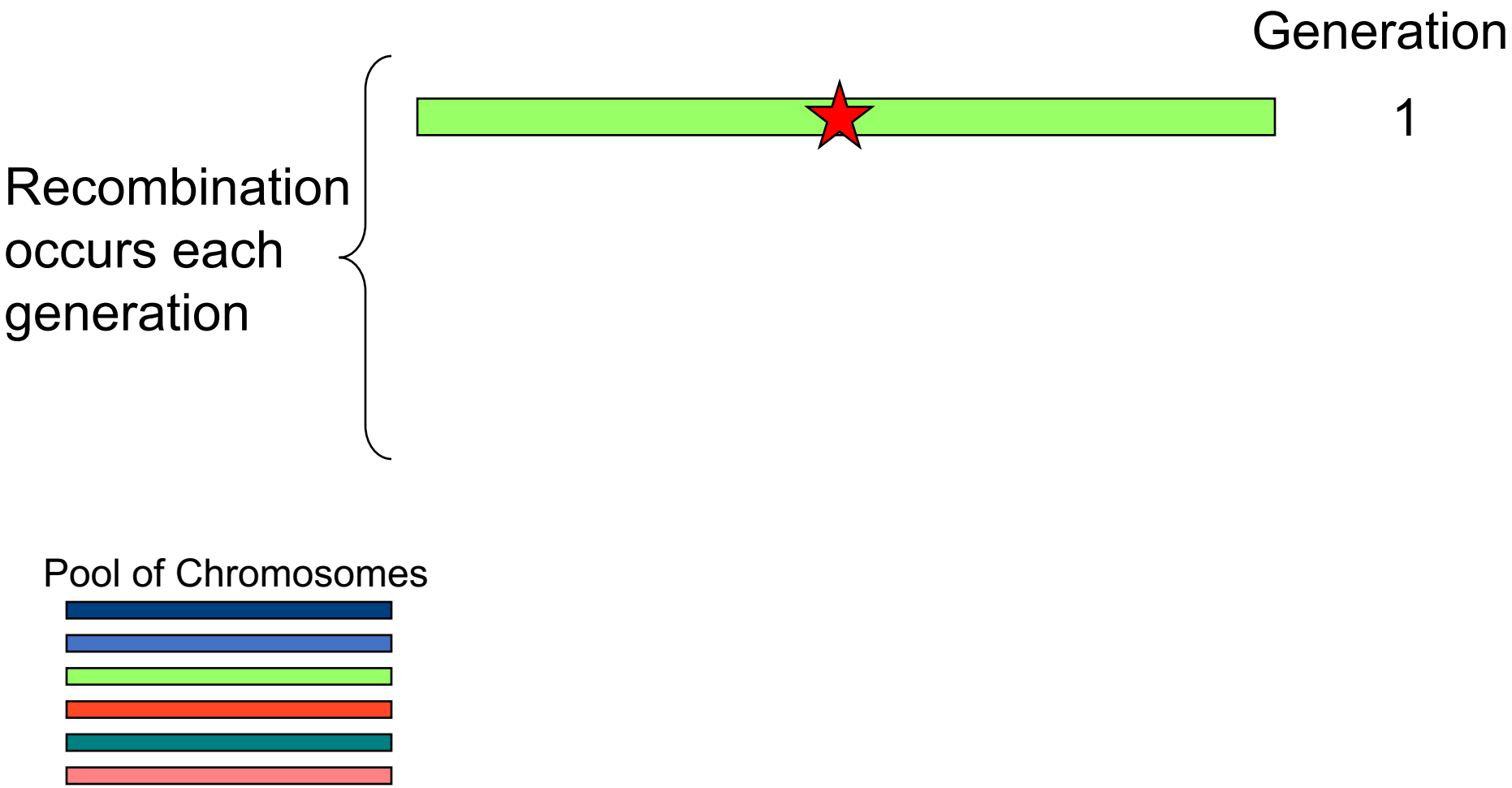
Linkage disequilibrium



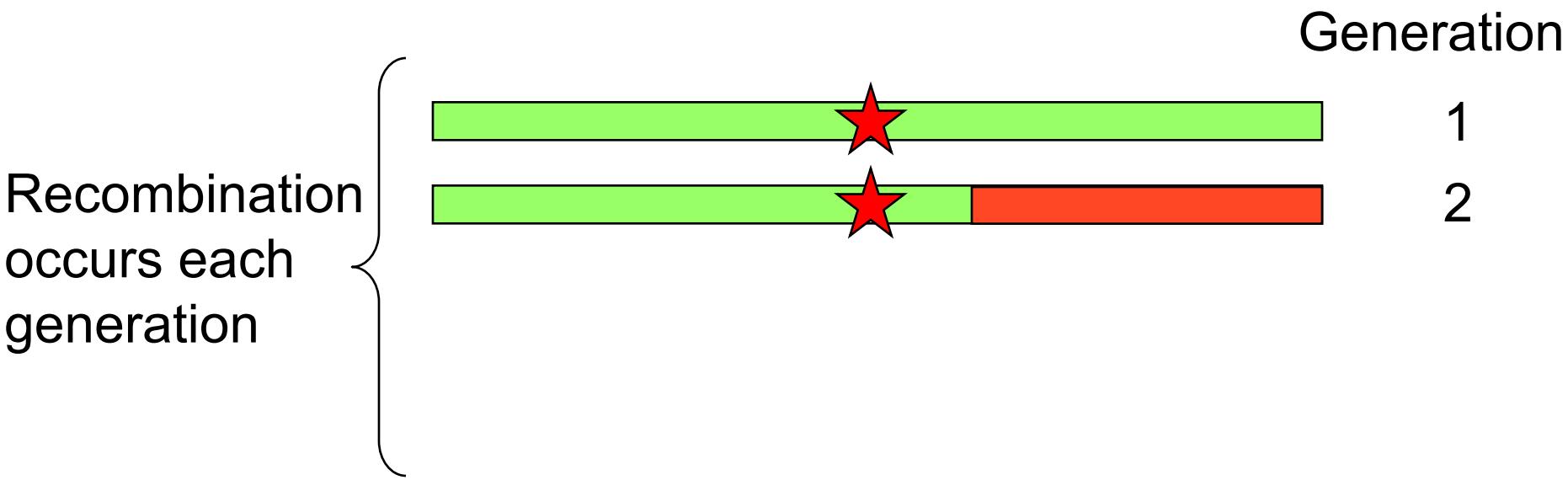
Linkage disequilibrium

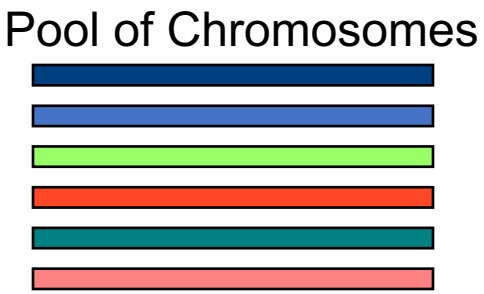


Linkage disequilibrium

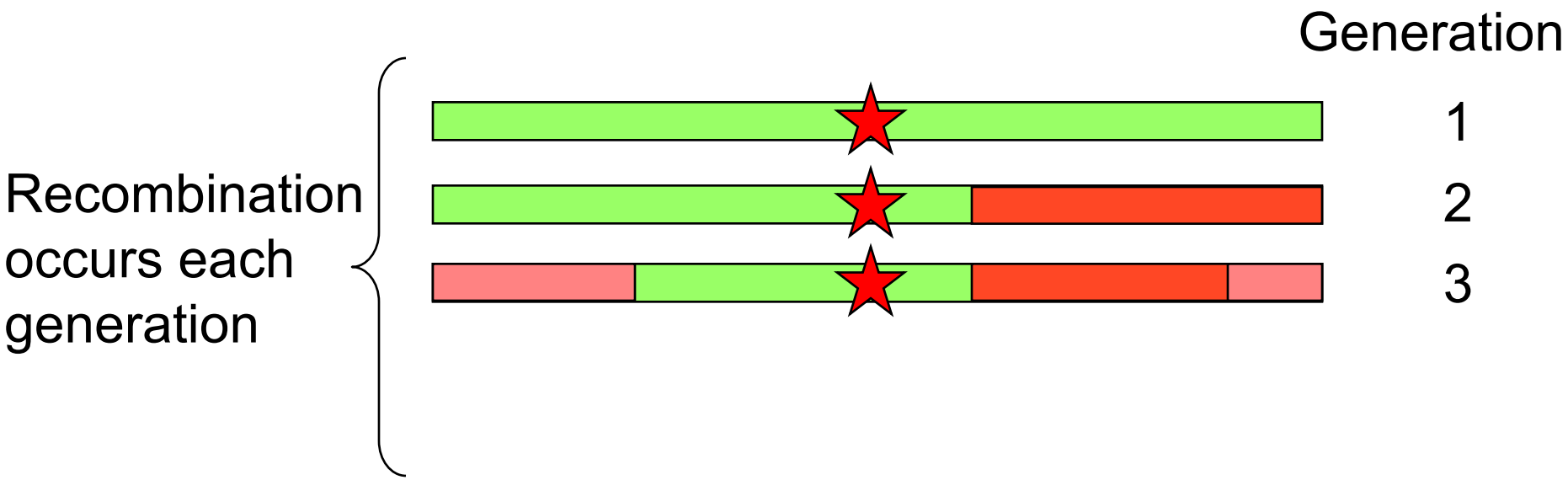


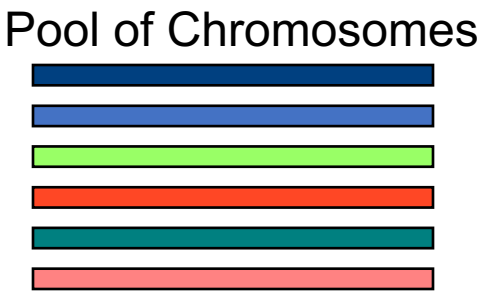
Linkage disequilibrium



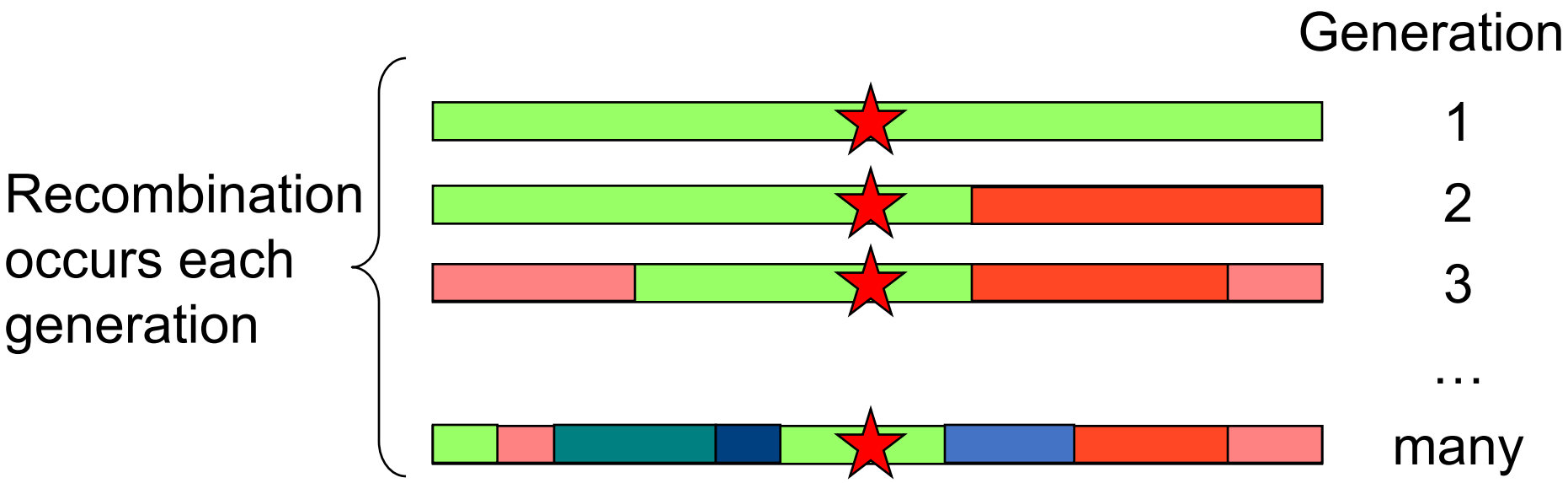


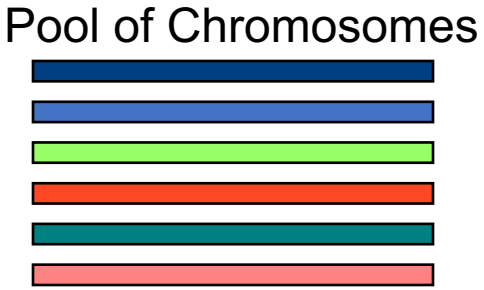
Linkage disequilibrium



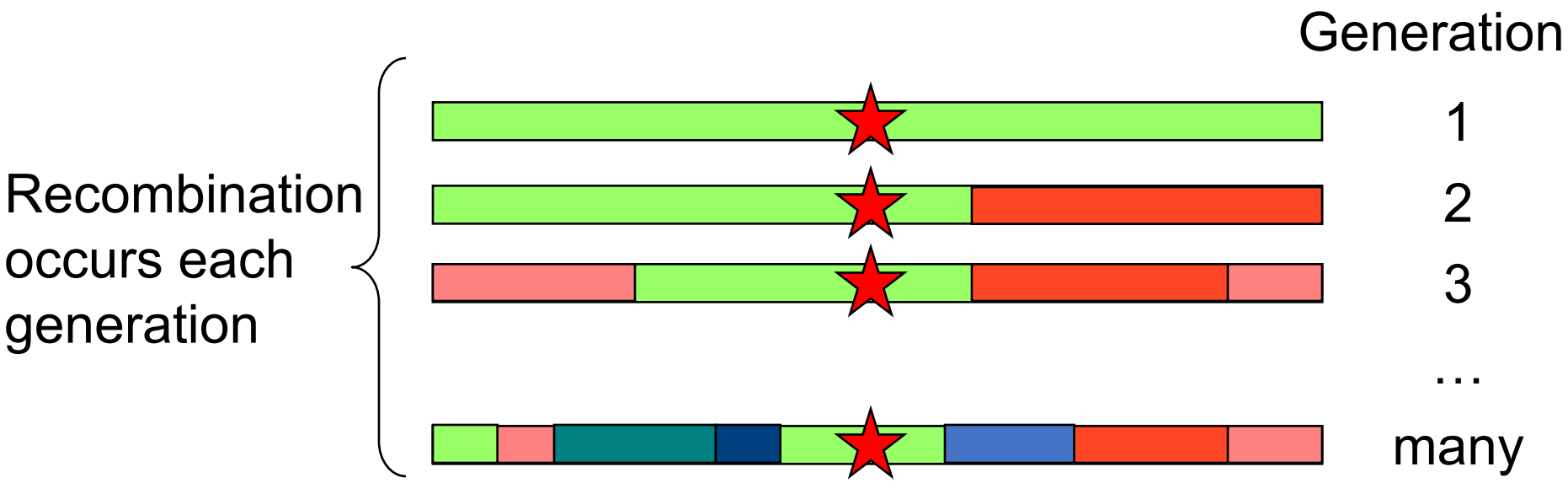


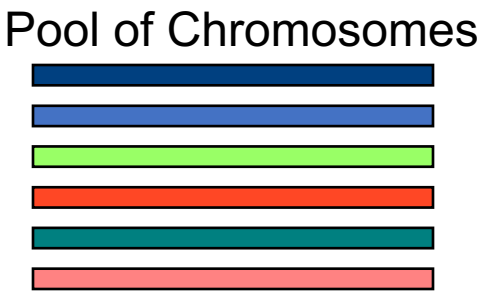
Linkage disequilibrium





Linkage disequilibrium





Chromosomes are a patchwork of the ancestral haplotypes, but local LD still persists

What is phasing

- In this context it is really Haplotype Estimation
- We take genotype data and try to reconstruct the haplotypes
  - Can use reference data to improve this estimation

Heterozygous genotypes at 3 sites

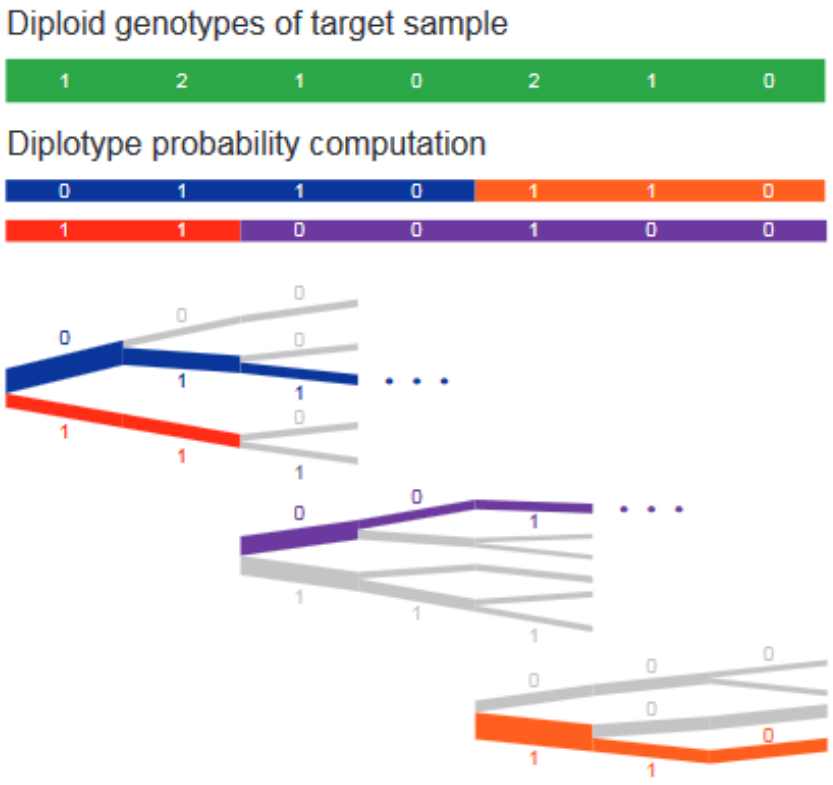
AC TG AT

The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Phasing in Eagle

- Input a target sample and a library of reference haplotypes
- *Selection of conditioning haplotypes.*
- *Generation of HapHedge data structure.*
- *Exploration of the diplotype space.*



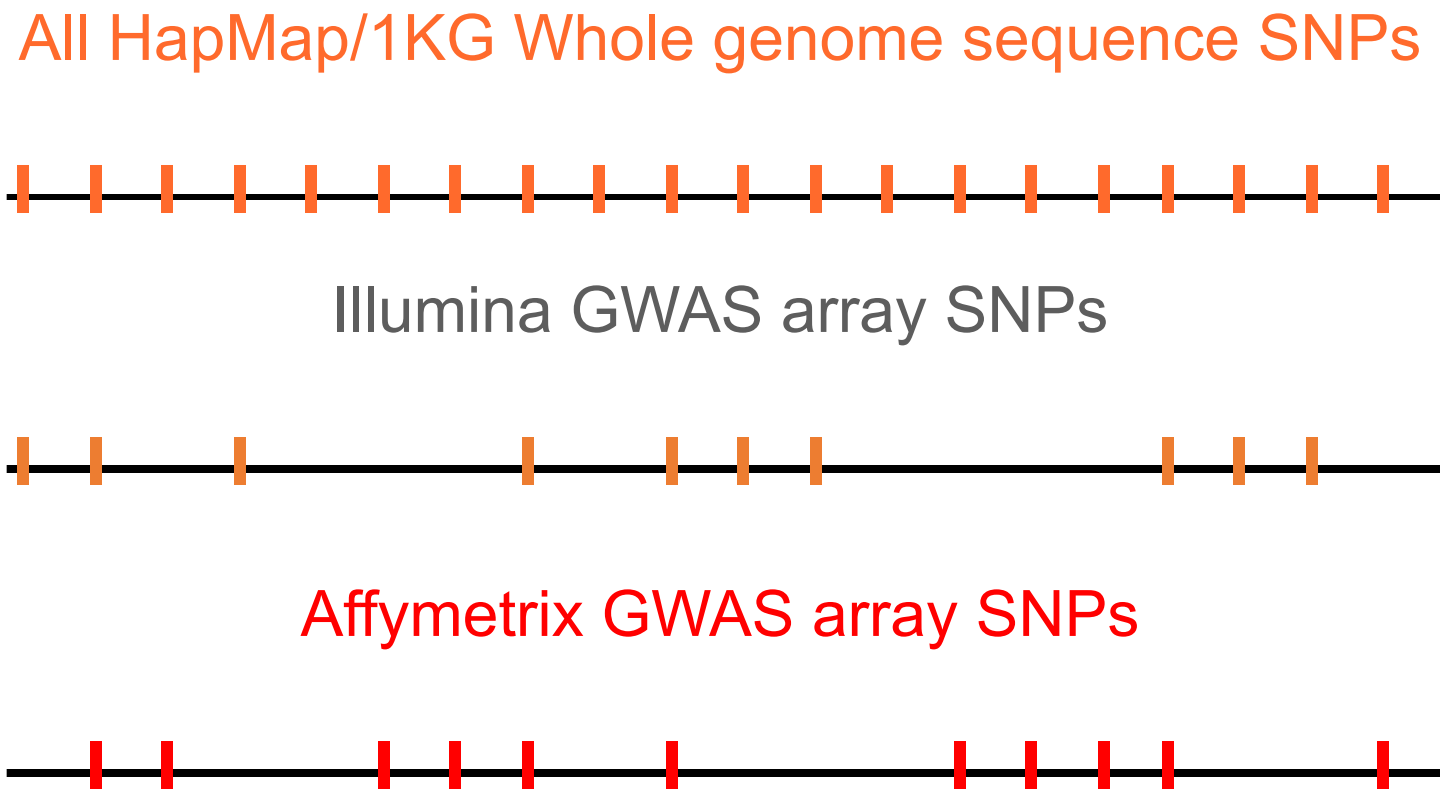
Imputation

All HapMap/1KG Whole genome sequence SNPs

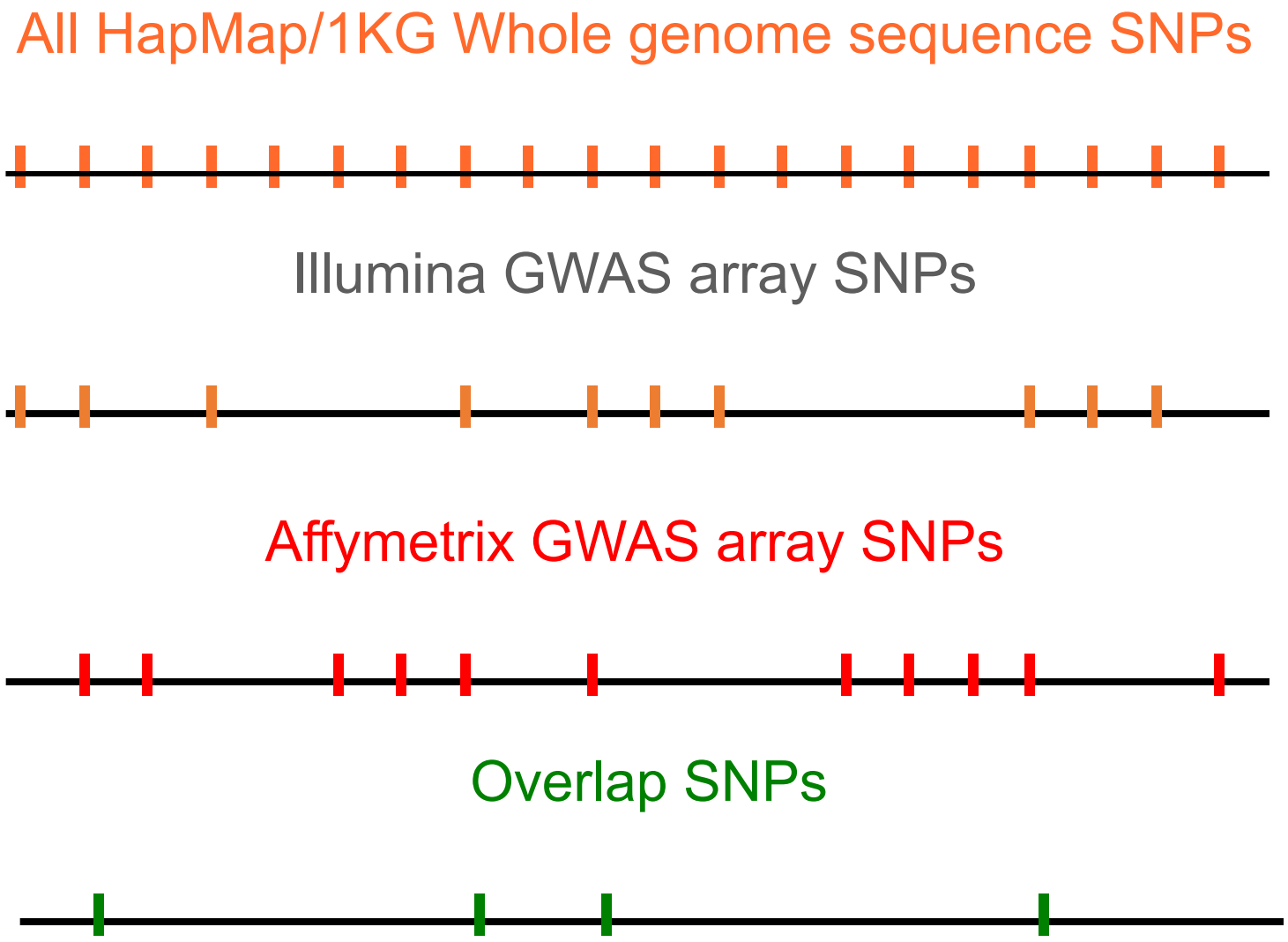
Imputation

All HapMap/1KG Whole genome sequence SNPs

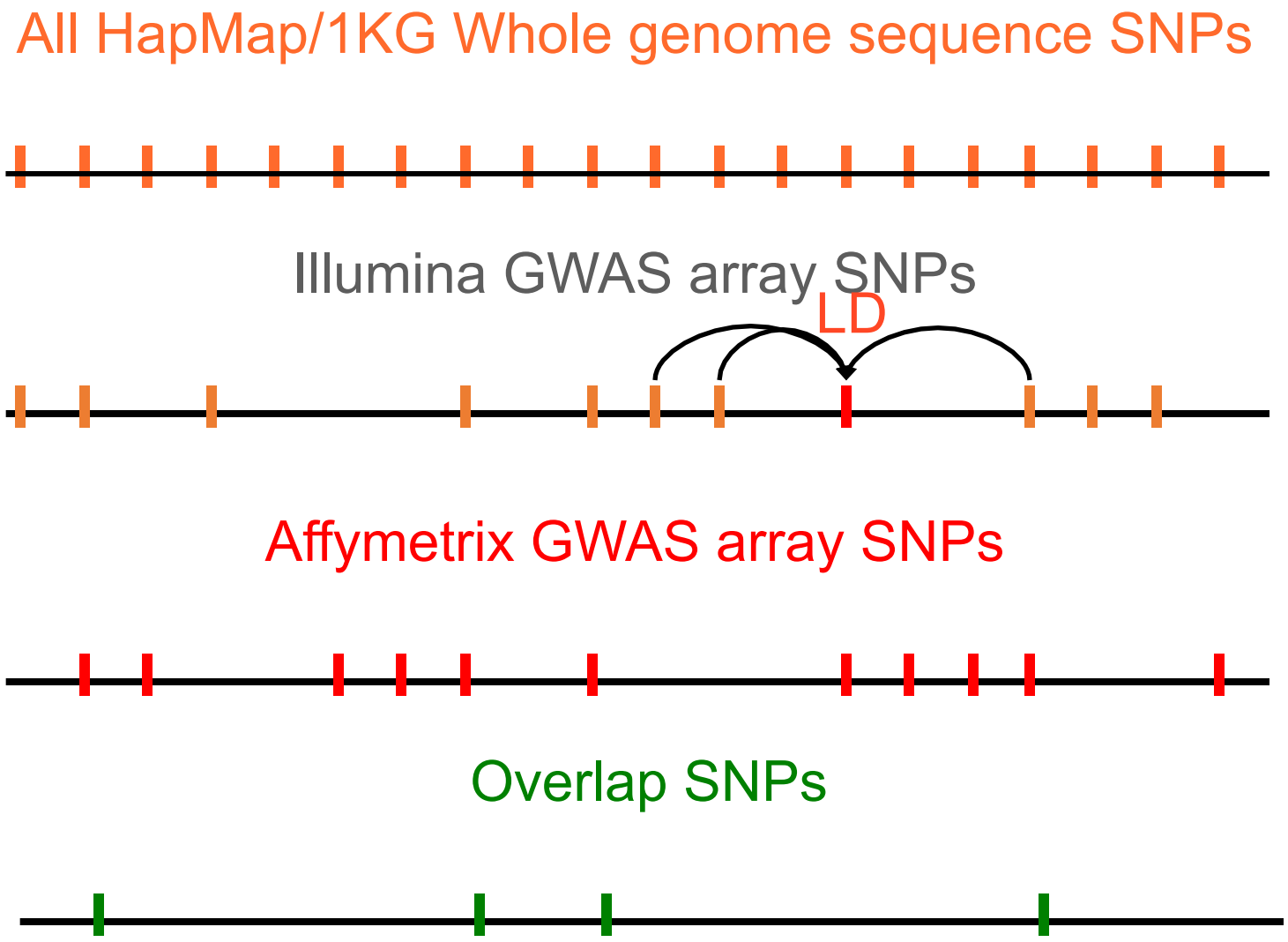
Imputation


All HapMap/1KG Whole genome sequence SNPs
Illumina GWAS array SNPs
Affymetrix GWAS array SNPs

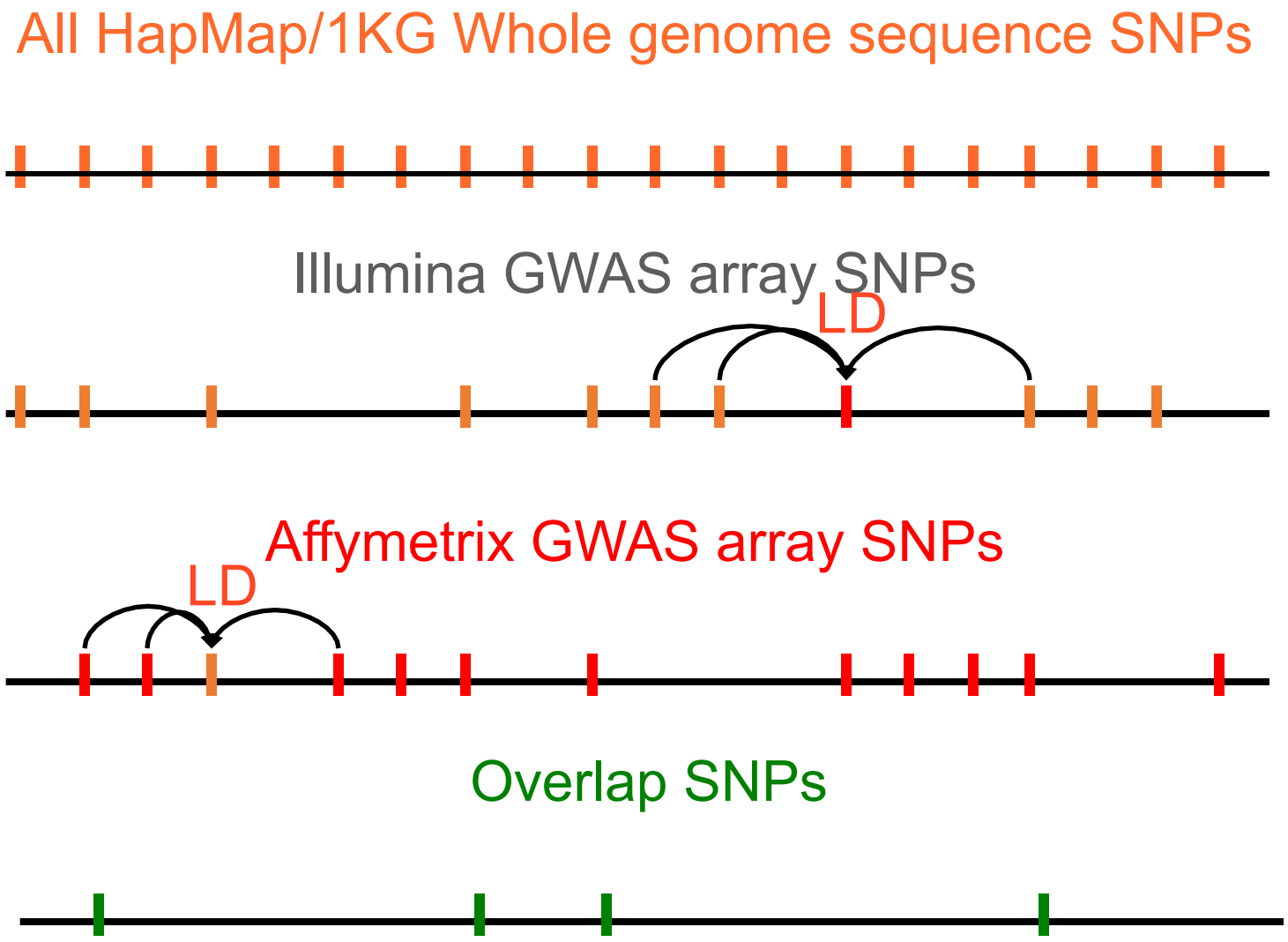
Imputation


All HapMap/1KG Whole genome sequence SNPs
Illumina GWAS array SNPs
Affymetrix GWAS array SNPs
Overlap SNPs

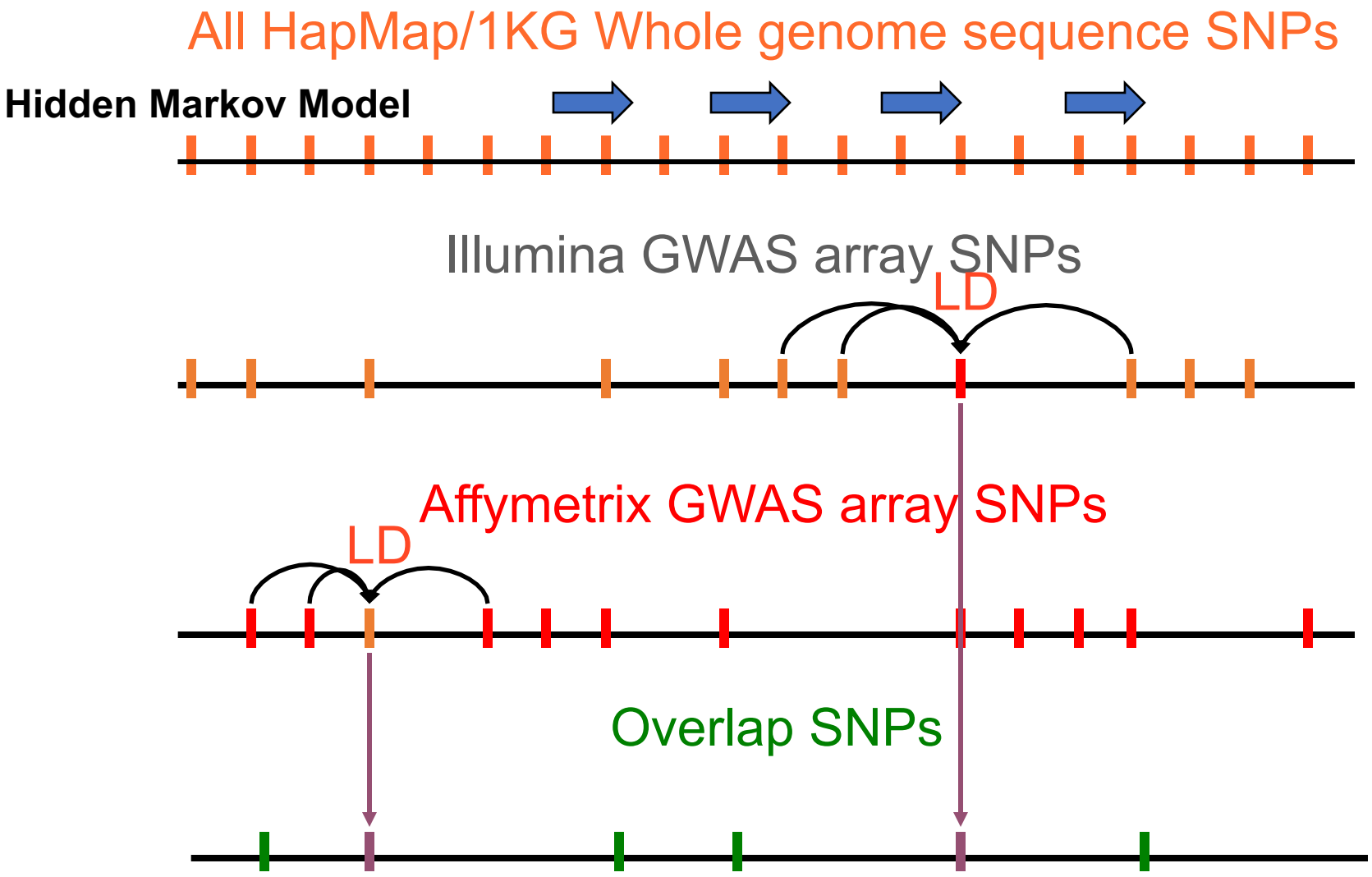
Imputation


All HapMap/1KG Whole genome sequence SNPs
Illumina GWAS array SNPs
LD
Affymetrix GWAS array SNPs
Overlap SNPs

Imputation


All HapMap/1KG Whole genome sequence SNPs
Illumina GWAS array SNPs
LD
Affymetrix GWAS array SNPs
LD
Overlap SNPs

Imputation


All HapMap/1KG Whole genome sequence SNPs
Hidden Markov Model
Illumina GWAS array SNPs
LD
Affymetrix GWAS array SNPs
LD
Overlap SNPs

Imputation output and performance

SNP INFO file:

Main Metric (Rsq)

SNP	Al1	Al2	Freq1	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose1	Dose2
1:10583	G	A	0.79288	0.20712	0.79288	-0.00000	-	-	-	-	-	-
1:10611	C	G	0.97889	0.02111	0.97889	0.00000	-	-	-	-	-	
1:13302	C	T	0.86280	0.13720	0.86280	-0.00000	-	-	-	-	-	-
1:13327	G	C	0.96042	0.03958	0.96042	-0.00000	-	-	-	-	-	-
1:95207182	T	C	0.99547	0.00453	0.99547	0.10108	-	-	-	-	-	-
1:95207382	T	T	1.00000	0.00000	1.00000	0.00000	-	-	-	-	-	-
1:95207442	C	T	0.62754	0.37246	0.99999	1.00507	Genotyped	0.98810	0.99822	0.99645	0.99484	0.00421
1:95207524	G	A	0.78061	0.21939	1.00000	1.00511	Genotyped	1.00059	1.00000	1.00000	0.99924	0.00083
1:95207532:TG_T	R	D	0.78620	0.21380	0.99441	0.97729	-	-	-	-	-	-
1:95207558	C	T	0.99399	0.00601	0.99399	0.05165	-	-	-	-	-	-
1:95207633	A	C	0.93366	0.06634	0.99998	1.00482	Genotyped	0.94847	0.99901	0.99802	0.99621	0.00372
1:95207846	G	T	0.98937	0.01063	0.98942	0.31316	-	-	-	-	-	-

Imputation quality evaluation

Minimac hides each of the genotyped SNPs in turn and then calculates 3 statistics:

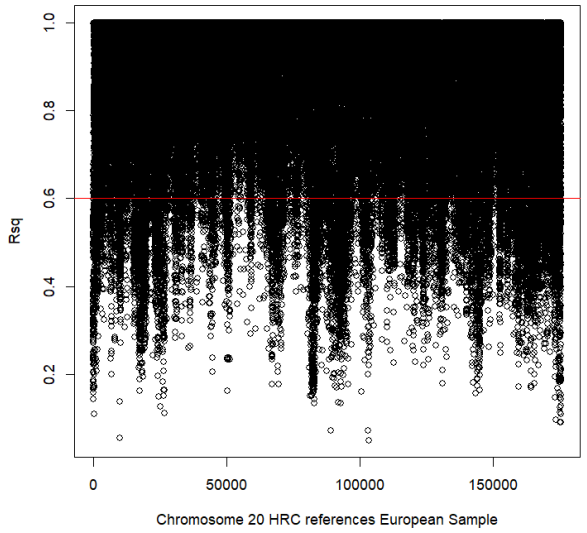
- looRSQ - this is the estimated rsq for that SNP (as if SNP weren't typed).
- empR - this is the empirical correlation between true and imputed genotypes for the SNP. If this is negative, the SNP alleles are probably flipped.
- empRSQ - this is the actual R2 value, comparing imputed and true genotypes.

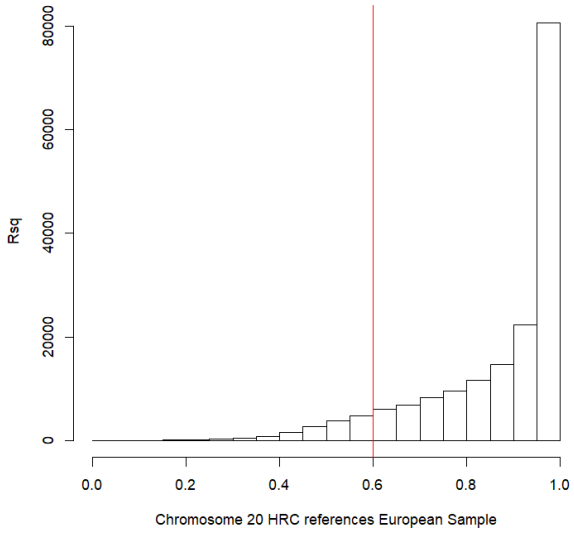
These statistics can be found in the *.info file

Be aware that, unfortunately, imputation quality statistics are not directly comparable between different imputation programs (MaCH/minimac vs. Impute vs. Beagle etc.).

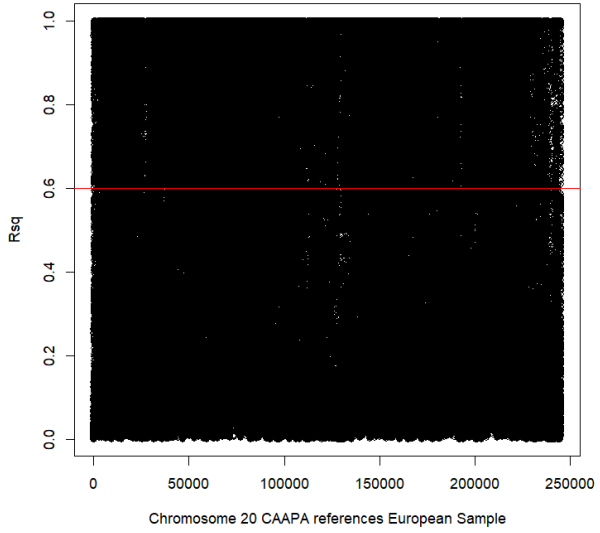
Imputation output and performance

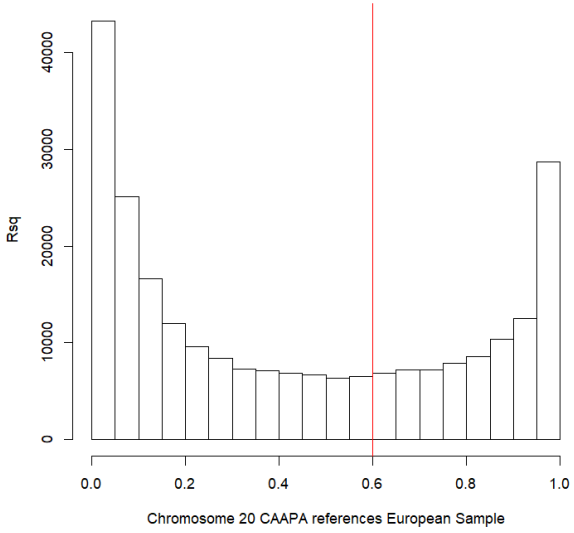
Good imputation





Bad imputation





Phasing/Imputation software

- Imputation programs
  - *IMPUTE2*

    https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
  - *MaCH / minimac*

    http://genome.sph.umich.edu/wiki/Minimac

- Also need to *Phase* data to distinguish haplotypes
  - *Shapeit*

    www.shapeit.fr
  - *Beagle*

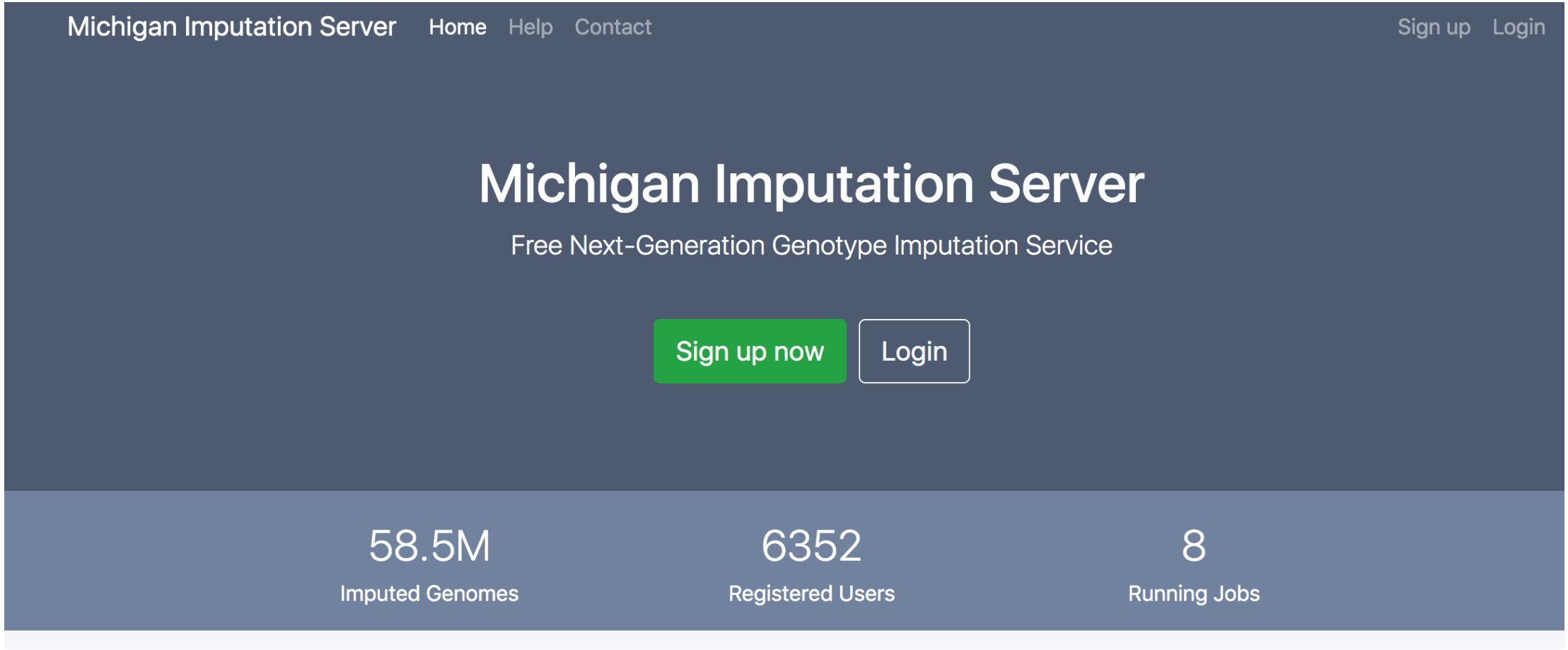
    http://faculty.washington.edu/browning/beagle/beagle.html
  - *Eagle / Eagle2*

    https://data.broadinstitute.org/alkesgroup/Eagle/

- Overall a very computationally expensive process

Imputation Services - Michigan

https://imputationserver.sph.umich.edu/index.html#!



Imputation Services - Sanger

https://imputation.sanger.ac.uk/

Sanger Imputation Service **Beta**

Home | About | Instructions | Resources | Status

Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the Wellcome Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click here to learn more and follow us on Twitter.

Before you start

Be sure to read through the instructions.

You will need to set up a free account with Globus and have Globus Connect running at your institute or on your computer to transfer files to and from the service.

Ready to start?

If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the about page. See also our Privacy and Security statement.

Full name

Organisation

Email address

What is this

Globus user identity

News

@sangerimpute

30/1/2017
Support for chromosome X has been added to all pipelines. PBWT has been updated to increase imputation accuracy of dosages and fix some bugs. See ChangeLog.

31/10/2016
New African Genome Resources panel with 9,912 haplotypes (6,230 African) is now available.

11/04/2016
Thanks to EAGLE2, we can now return **phased data**. The HRC panel has been updated to r1.1 to fix a known issue. See ChangeLog for more details.

Breakout session (5 min)

- Breakout into small groups
- Introduce yourself to everyone
- Person with LATEST letter in their FIRST name will be the note taker
  - E.g. Zenia is the note taker, not Aaron
  - E.g. Abel is the note taker, not Aaron
- Ask any questions you have:
  - “I didn’t understand what …. meant”
  - “I’m confused by the concept of phasing”
- Note takers relay unanswered questions from the breakout session

Lecture Format

- Part 1 (~40 minutes)
  - Goals of GWAS
  - What does the data look like?
  - GWAS Quality Control (QC)
  - 5 min breakout session
- Part 2 (~40 minutes)
  - Relatedness checking
  - Population stratification
  - Principal components analysis (PCA)
  - Imputation
  - 5 min breakout session
- Part 3 (~40 minutes)
  - Association testing
  - Meta-analysis
  - Polygenic Scoring
  - 5 min breakout session
- Preparation for module 3 / additional reading / lingering questions

Association testing

Association testing

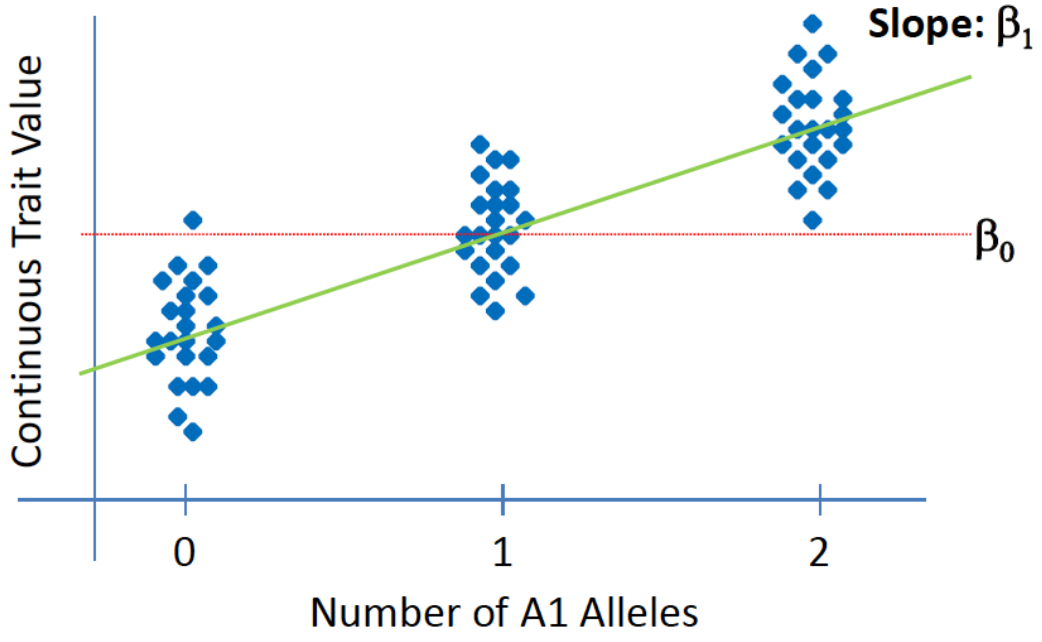
- *Main question*: Does the phenotype examined associate/correlate with the genetic variant?

Linear Regression Equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Logistic Regression Equation

$$\ln\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Tests of SNP association

- Case/control:
  - Chi-square test on contingency table
  - Fisher’s exact test
  - Cochran-Mantel-Haenszel test
  - Cochran-Armitage trend test
  - Logistic regression

- Case/control & quantitative traits:
  - Permutation

Chi-square test

- plink --assoc = chi-square test on alleles
- Null hypothesis: alleles are independent of disease state

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- n = the number of allele – disease combinations (= 4)
- O_i = an observed frequency
- E_i = an expected (theoretical) frequency, asserted by the null hypothesis of no independence between allele and disease
- X^2 = the test statistic that asymptotically approaches a χ^2 distribution

Expected case count of allele 1:

0.6 * 30 = 18

x^2 stat = 0
p=1

	Cases	Controls	Total
Allele 1	18	12	30
Allele 2	42	28	70
Total	60	40	100

Fisher's exact test

- plink --fisher = Fisher's exact test
- Null hypothesis: alleles are independent of disease state
- Should be used instead of the chi-square test if ≥ 1 cells have ≤ 5 observations.
- More computationally expensive than chisq test

	Cases	Controls	Total
Allele 1	a	b	a+b
Allele 2	c	d	c+d
Total	a+c	b+d	n

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Cochran-Mantel-Haenszel (CMH) test

- plink --mh = Cochran-Mantel-Haenszel test
- Comparable to chi-square test, but within different groups (such as different subpopulations to correct for stratification)

Pop 1	Cases	Controls
Allele 1	a	b
Allele 2	c	d

Pop 2	Cases	Controls
Allele 1	a	b
Allele 2	c	d

$$\chi^2_{MH} = \frac{\{|\sum[a-(a+b)(a+c)/n|-0.5\}^2}{\sum(a+b)(a+c)(b+d)(c+d)/(n^3-n^2)}$$

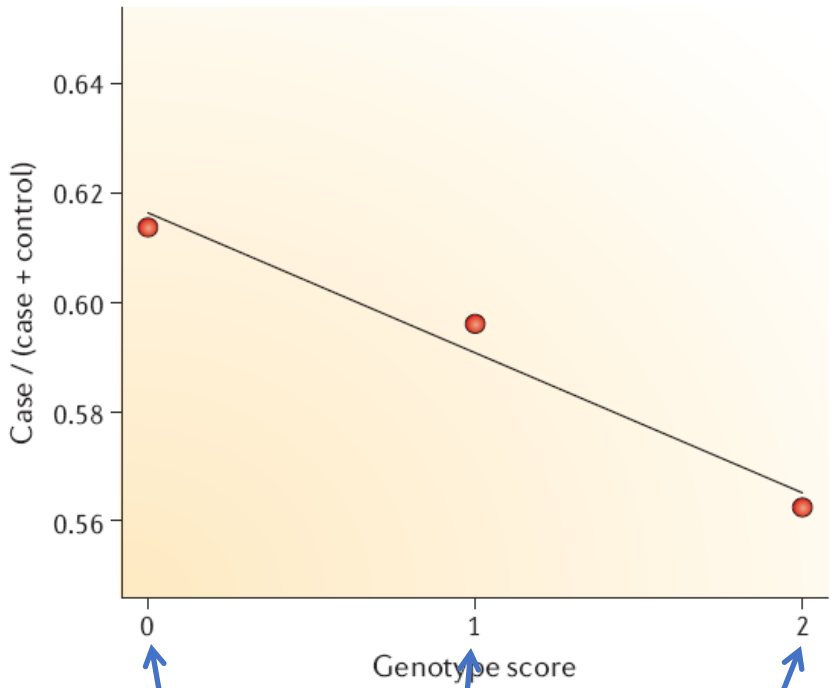
Cochran-Armitage trend test

- plink --model =
  - Cochran-Armitage trend test
  - Allelic test: D vs d
  - Genotypic test: DD vs Dd vs dd
  - Test for dominant effect of D: (DD & Dd) vs dd
  - Test for recessive effect of D: DD vs (Dd & dd)

Chi-square tests

Cochran-Armitage trend test

- Cochran-Armitage trend test
- Null hypothesis: the line has zero slope
  - Does not assume Hardy-Weinberg equilibrium (HWE)
  - Assumes that there are additive effects
  - More conservative than the chi-square test



	Cases	Controls	Total
AA	11	7	18
Aa	37	25	62
aa	50	39	89
Total	98	71	164

Nr. of A alleles	0	1	2
Prop. of Cases	.615 (11/18)	.596 (37/62)	.561 (50/89)

Logistic regression

- plink --logistic = logistic regression (= regression analysis for categorical data)
  - A useful way to describe the relationship between one or more risk factors (alleles + covariates) and a binary trait (case/control).
  - Allows testing of allelic, genotypic, dominant & recessive effects.

$$ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

- Plink gives the *p*-value and the odds ratio (OR) of the risk factor
- OR = e^{β}

Case/control odds ratio

- Odds Ratio (OR) = a measure of effect size, describing the strength of association between two binary data values (alleles 1 & 2 – case & control status).

	Cases	Controls
Allele 1	a	b
Allele 2	c	d

$$OR = \frac{a \times d}{b \times c}$$

- An OR of 1.2 for example, means that the odds (<u>not the probability!</u>) of getting the disease increases with a factor of 1.2 if you carry the risk allele (odds = $\frac{P}{1-P}$).

	Cases	Controls
Allele 1	120	100
Allele 2	100	100

$$OR = 1.2$$

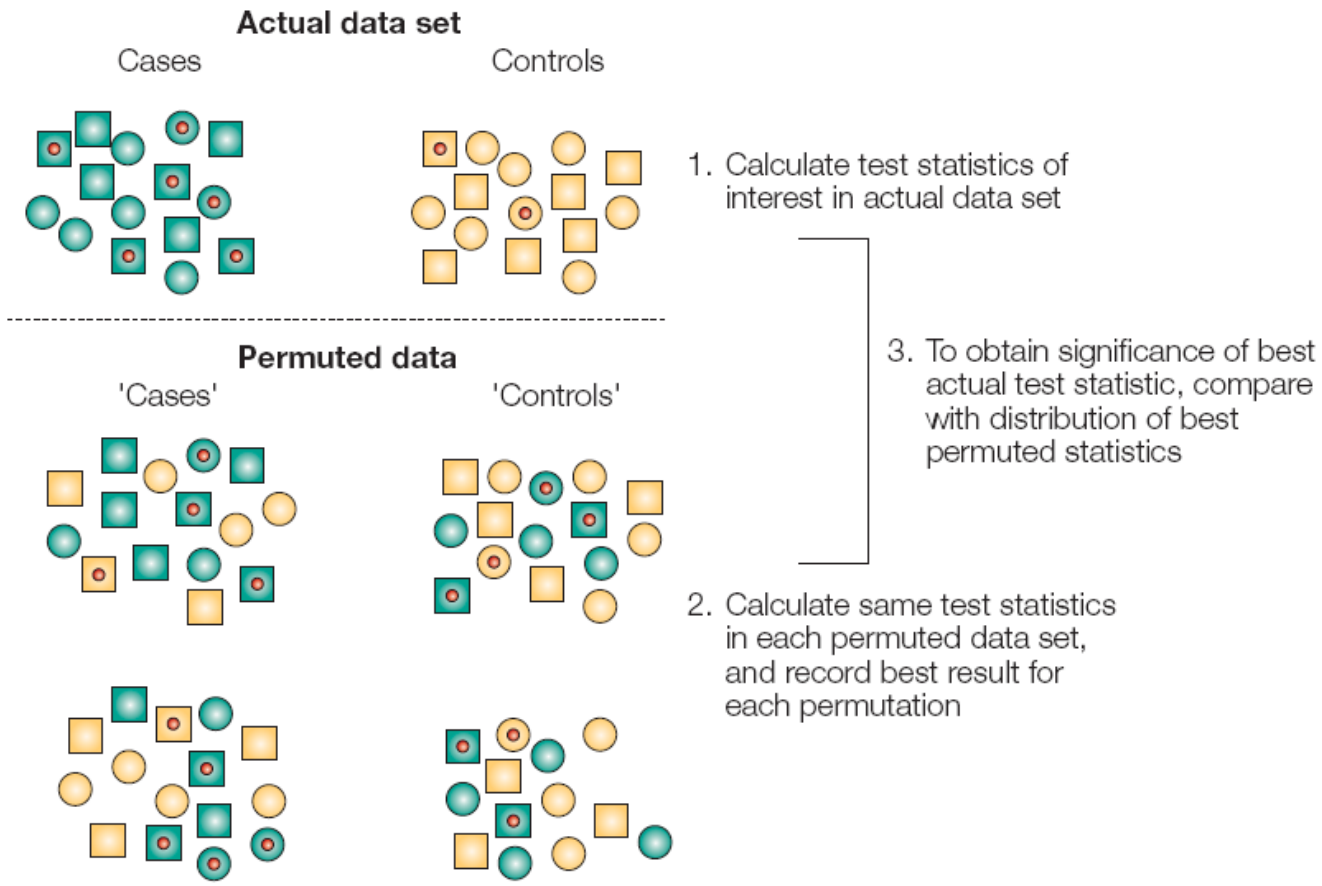
Case/control phenotype

- No *a priori* hypothesis:
  - Chi square test genotypic (2×3): --model
  - Logistic regression – genotypic (allows covariates): --logistic
- Additive effects:
  - Cochran-Armitage test (doesn't assume HWE) (2×2): --model
  - Chi square test allelic (large sample size) (2×2): --assoc
  - Fisher's exact test allelic (small sample size) (2×2): --fisher
  - Logistic regression - allele test (allows covariates): --logistic
- Dominant effects:
  - Chi square test genotypic (2×2): --model
  - Logistic regression – dominance test (allows covariates): --logistic
- Recessive effects:
  - Chi square test genotypic (2×2): --model
  - Logistic regression – recessive test (allows covariates): --logistic

Permutation

- This empirical method evaluates how often a given *p*-value would arise by chance if the study were repeated without any true associations.

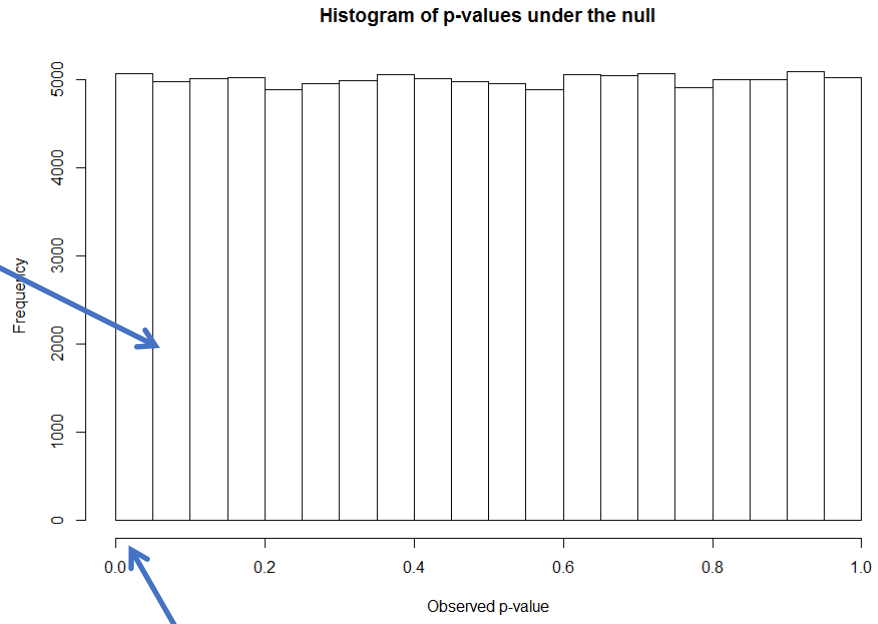
- Square = male
- Circle = female
- Green = affected
- Yellow = healthy
- Red circle = disease susceptibility allele



Permutation

- How is the empirical *p*-value calculated?
  - (rank of the *p*-value of the real dataset) / (nr of permutations)

Distribution of *p*-values or test statistics for 1000 samples with randomly swapped phenotypes



Suppose the *p*-value or test statistic of the real association test is ranked at nr 3 (i.e. there are 2 *p*-values that are more significant)

Empirical *p*-value = 3/1000 = .003

Permutation

- Advantages
  - Does not assume that the phenotype is normally distributed
  - Does not assume HWE
  - Better for rare alleles and small sample sizes
  - Empirical *p*-values can be corrected for multiple testing, while preserving the correlational structure between all SNPs (= less conservative than Bonferroni correction = less false negatives)
  - Allows for association analyses within clusters (which allows you to correct for population stratification and other confounding variables)
- Disadvantage:
  - It can take a very long time to compute…

Permutation

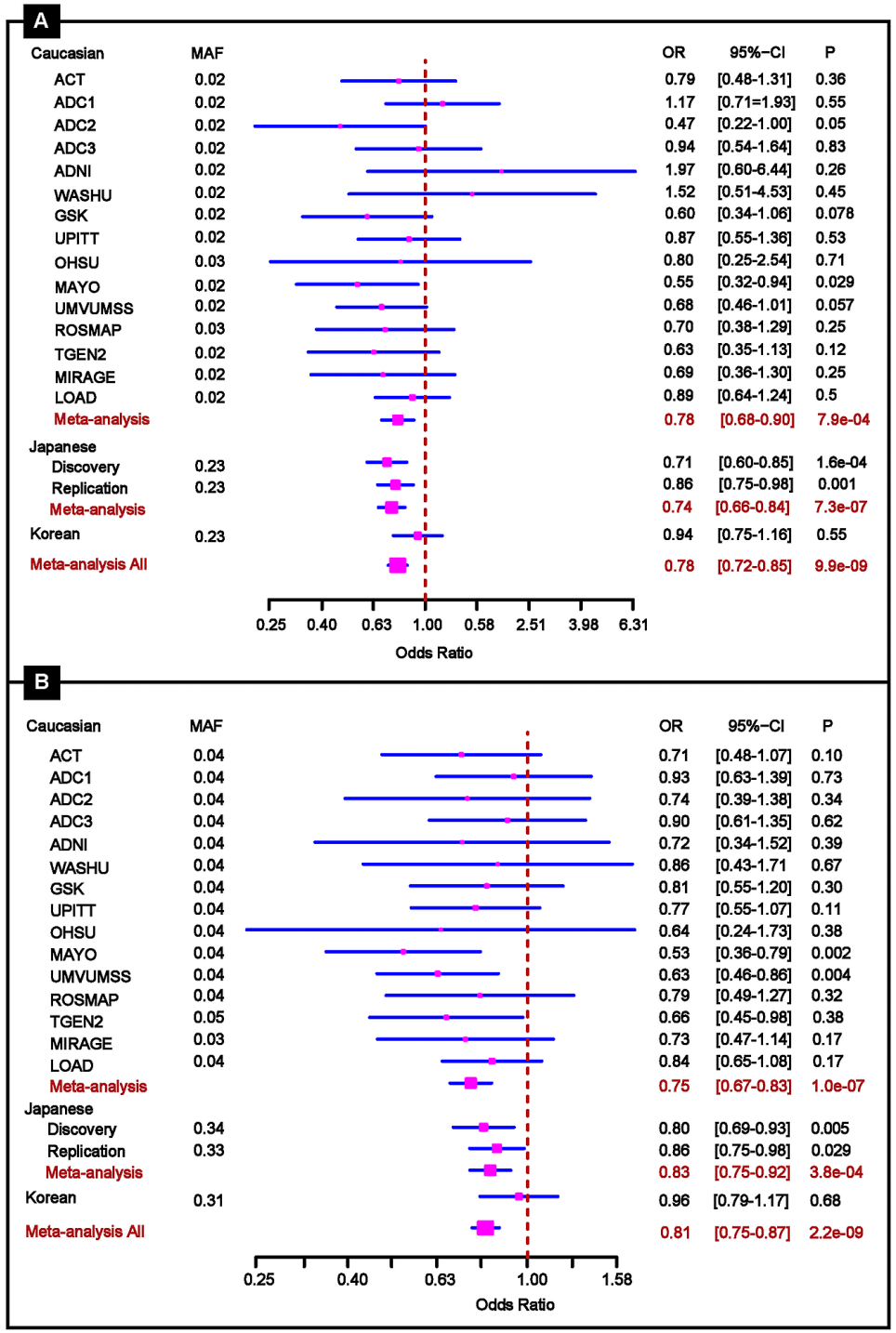
- Plink can do two kinds of permutation:
  - Adaptive: permutations of SNPs that are not likely to be significant are stopped prematurely. The advantage is that the permutation procedure does not have to take as long.
  - max(T): all permutations are performed for all SNPs. The advantage is that this allows for the calculation of a *p*-value that is corrected for multiple testing.

GWAS
Meta-Analysis

Meta-analysis

Goal: Combine separate studies to increase power to discover SNP associations

- Evaluate summary statistics (quicker/lighter)
- Examine potential study bias



Significance - Weighted Z

$$Z = \frac{\sum_{i=1}^{m} w_i Z_i}{\sqrt{\sum_{i=1}^{m} w_i^2}}$$

larger sample = larger weight

where $w_i = \sqrt{n_i}$

and $Z_i = \sqrt{\chi_i^2}$

The test statistic Z_i can be obtained from two-tailed *p*-values and the direction of effect, or one-tailed *p*-values, using the inverse normal distribution function

Effect size - Weighted β

larger sample ->
smaller standard error ->
larger weight

$$\hat{\beta} = \frac{\sum_{i=1}^{m} w_i \hat{\beta}_i}{\sum_{i=1}^{m} w_i}$$

$$w_i = \frac{1}{\sigma_i^2}$$

$$SE^* = \sqrt{\frac{1}{\sum_{i=1}^{m} \frac{1}{\sigma_i^2}}}$$

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
2	exm-rs10199914	239896861	2884	-0.1214	0.02747	0.006729	-4.419	1.029e-05
1	exm83171	111490837	2884	-0.8702	0.19780	0.006668	-4.398	1.130e-05
10	exm836623	79601934	2884	-2.5020	0.58020	0.006413	-4.313	1.666e-05
9	exm790074	134321955	2881	0.6023	0.14470	0.005981	4.162	3.246e-05
3	exm305671	44636284	2884	-2.8950	0.71070	0.005726	-4.074	4.744e-05
2	exm266787	219854997	2884	-2.3490	0.58040	0.005653	-4.048	5.307e-05
8	exm733014	145736215	2883	-1.8150	0.44970	0.005625	-4.037	5.552e-05
10	exm853248	105194086	2884	-0.1116	0.02788	0.005535	-4.005	6.355e-05
14	exm1091009	23939305	2884	-0.8758	0.22010	0.005466	-3.980	7.069e-05
9	exm736710	4618014	2883	1.7840	0.44980	0.005428	3.965	7.512e-05
4	exm419283	113352955	2884	-2.8020	0.71080	0.005362	-3.941	8.291e-05
22	exm1595949	26286807	2884	-2.7700	0.71090	0.005240	-3.896	9.988e-05
2	exm201502	71887715	2884	-1.4610	0.38030	0.005098	-3.843	1.243e-04

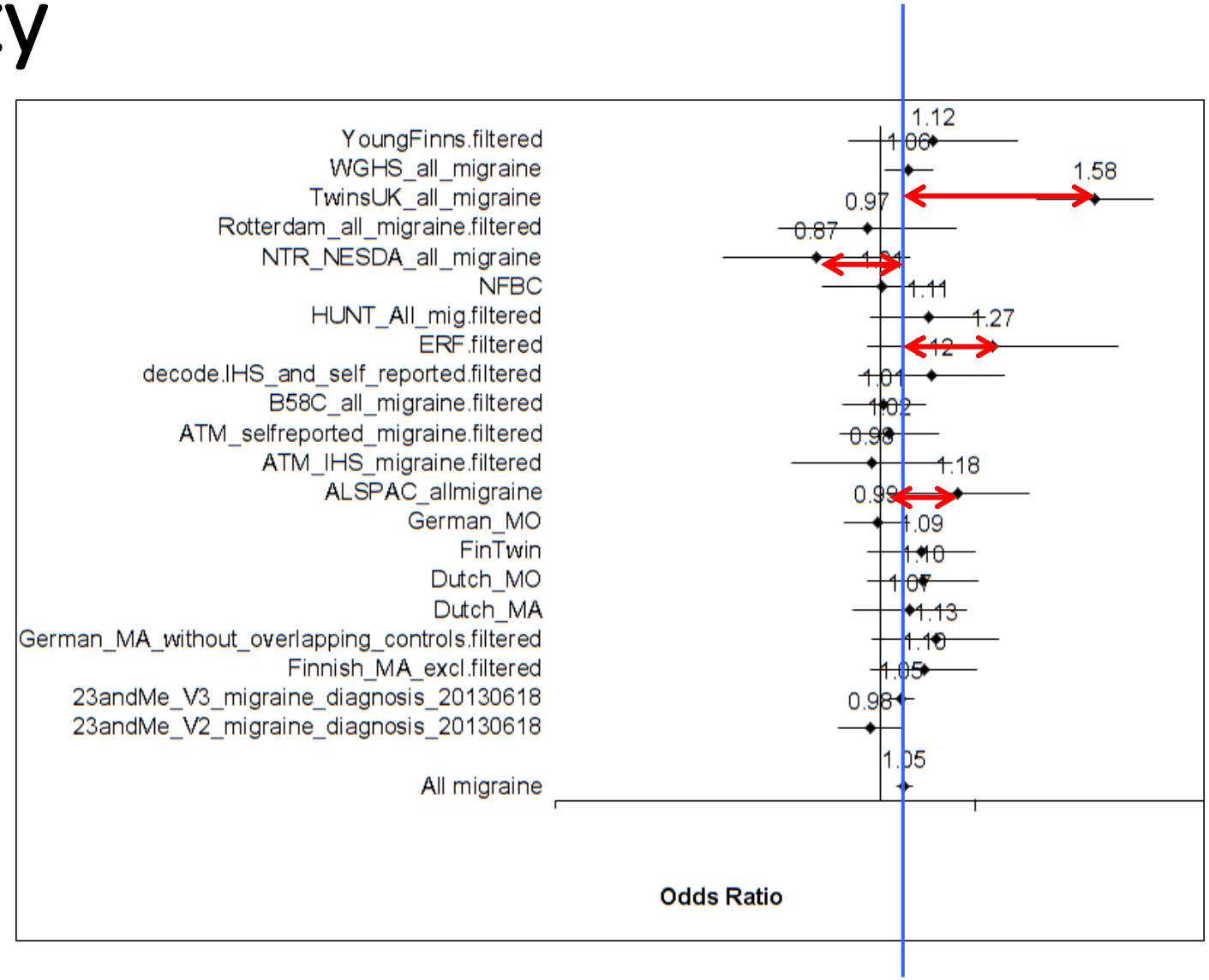
Test for Heterogeneity
Cochran's Q

$$\hat{\beta} = \frac{\sum_{i=1}^{m} w_i \hat{\beta}_i}{\sum_{i-1}^{m} w_i}$$

test of distance from the weighted mean

$$Q = \sum_{i=1}^{m} w_i \left(\hat{\beta}_i - \hat{\beta}\right)^2 \sim \chi^2_{m-1}$$

$$I^2 = 100 \times \frac{Q - (m-1)}{Q}$$



Meta-analysis software: METAL

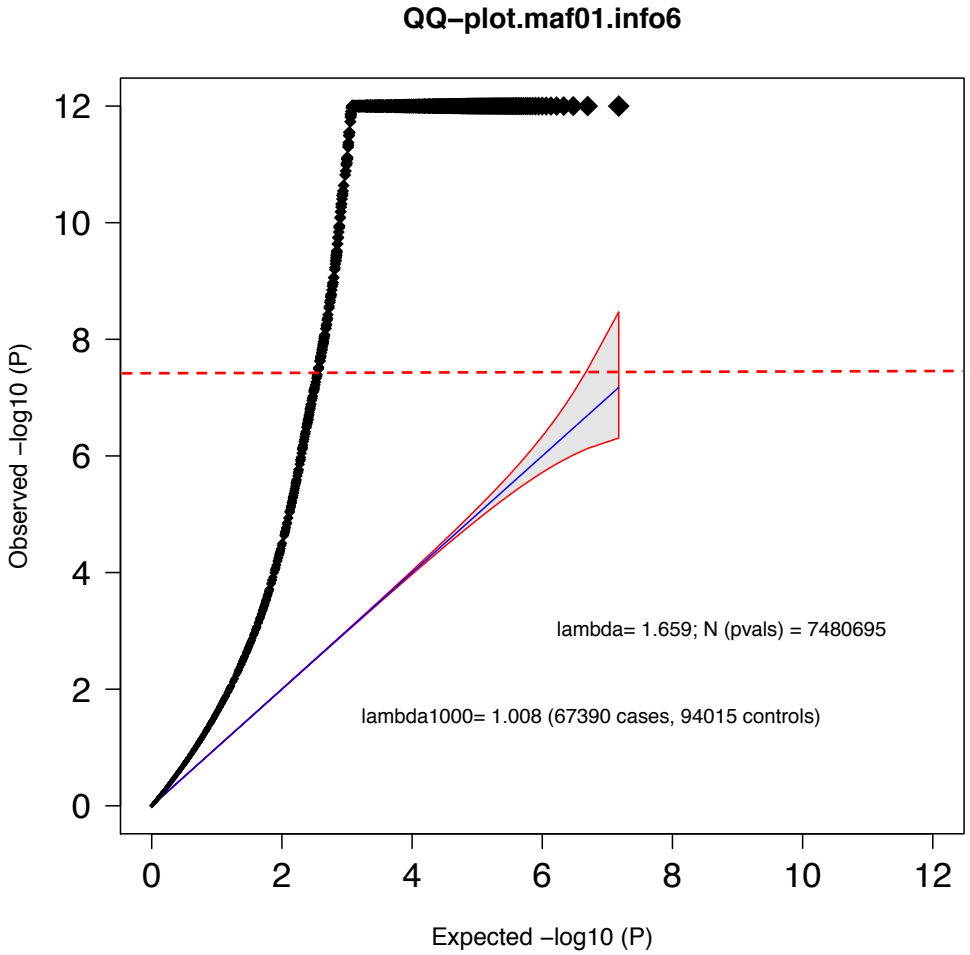
http://www.sph.umich.edu/csg/abecasis/metal/

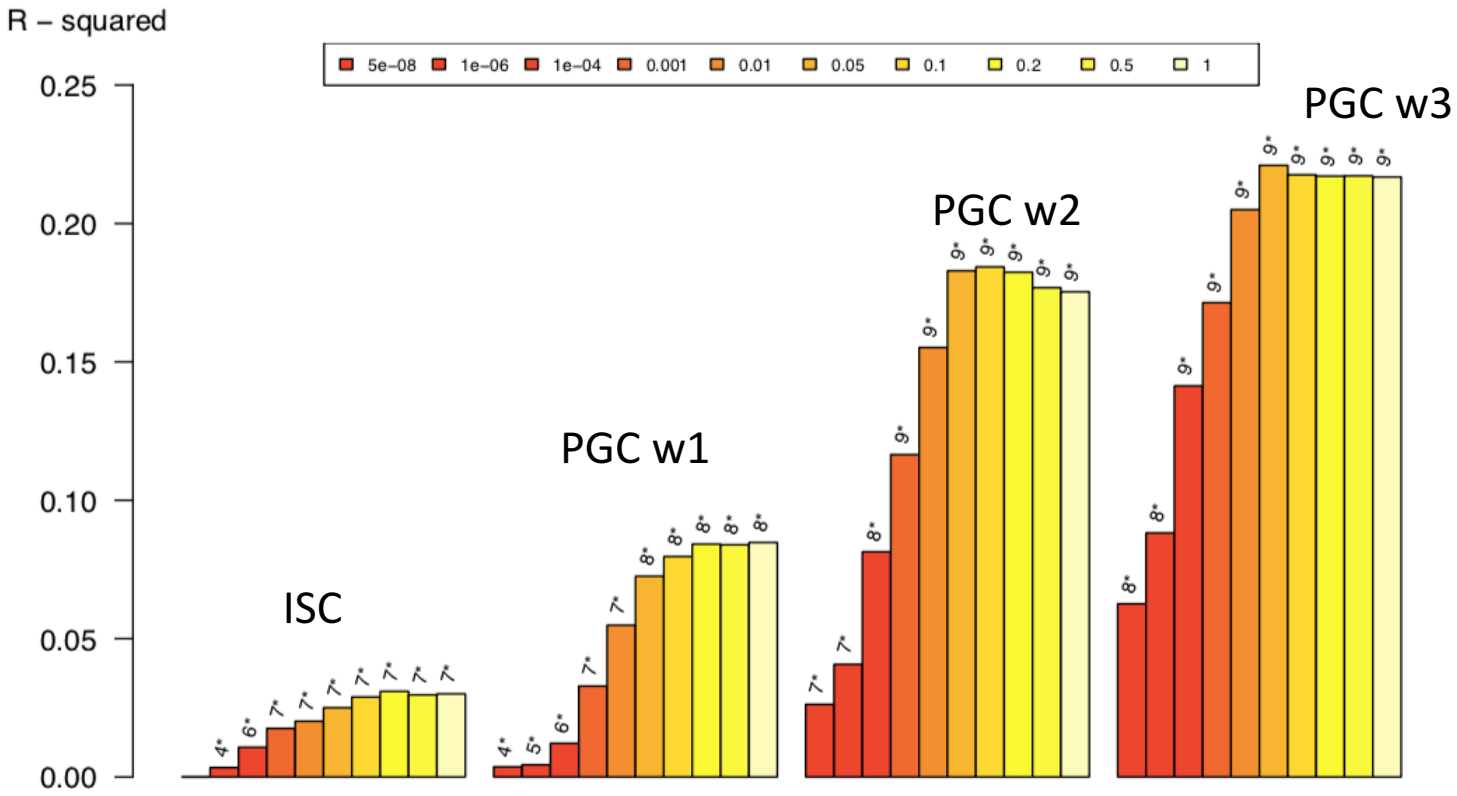
Documentation can be found at the metal wiki:
https://genome.sph.umich.edu/wiki/METAL

Polygenic Scores

Polygenic scores – adding up the effects

From PGC SCZ wave 3



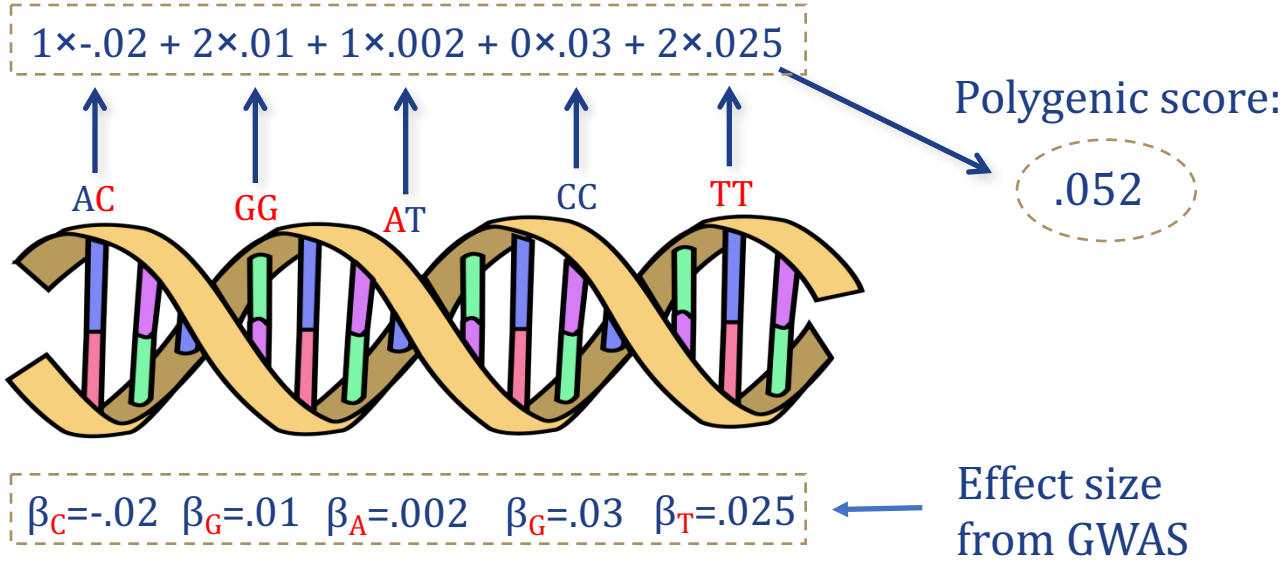


R-square = proportion of variance in phenotype explained by score

Polygenic Scores (PGS or PS)

Polygenic Scores capture (part of) someone's genetic "risk" by summing all risk alleles weighted by the effect sizes estimated in a Genome-Wide Association Study (GWAS)

Also known as **polygenic risk scores (PRS), genetic risk score (GRS)**, or **genome-wide score (GS)**



Polygenic Scores

http://zzz.bwh.harvard.edu/plink/profile.shtml

- By summing the collective effect sizes of many SNPs you can quantify part of the genetic "risk" in an **independent** dataset

- Polygenic Scores generally improve when adding SNPs that individually didn't reach genome-wide significance

Basic usage

The basic command to generate a score is the `--score` option, e.g.

```
./plink --bfile mydata --score myprofile.raw
```

which takes as a parameter the name of a file (here `myprofile.raw`) that describes the scoring system. This file has the format of one or more lines, each with exactly three fields

```
SNP ID
Reference allele
Score (numeric)
```

for example

```
SNPA   A    1.95
SNPB   C    2.04
SNPC   C   -0.98
SNPD   C   -0.24
```

These scores can be based on whatever you want. One choice might be the log of the odds ratio for significantly associated SNPs, for example. Then, running the command above would generate a file

```
plink.profile
```

with one individual per row and the fields:

```
FID     Family ID
IID     Individual ID
PHENO   Phenotype for that
CNT     Number of non-missing SNPs used for scoring
CNT2    The number of named alleles
SCORE   Total score for that individual
```

nature genetics
LETTERS
https://doi.org/10.1038/s41588-018-0183-z

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4]*

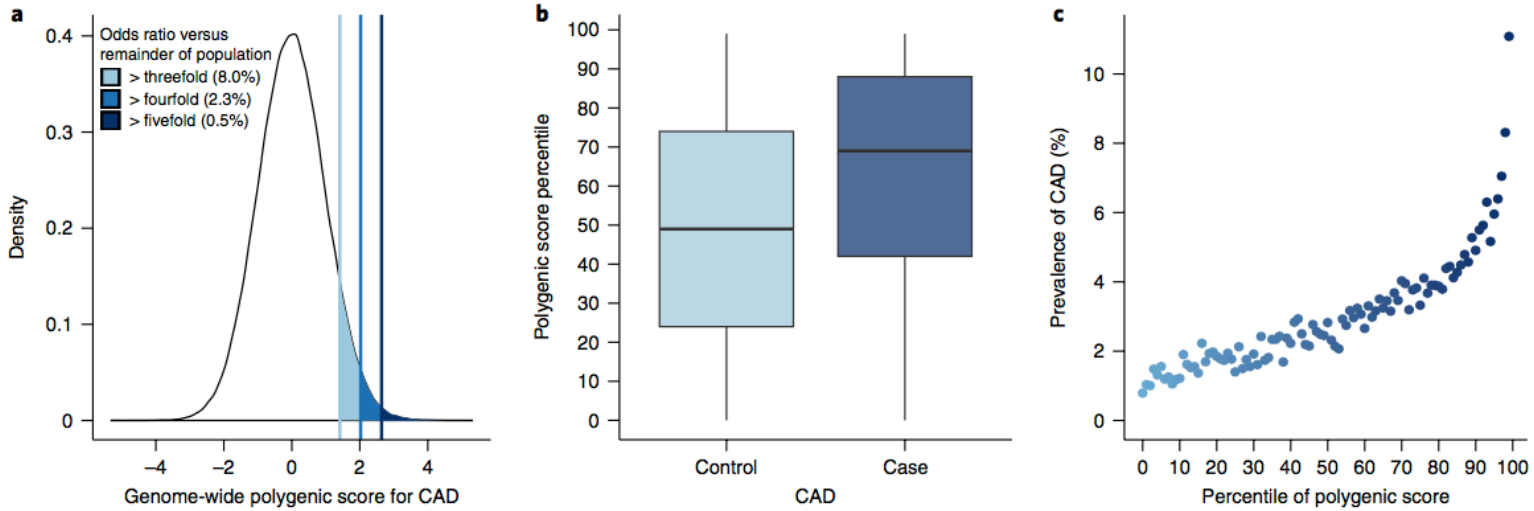
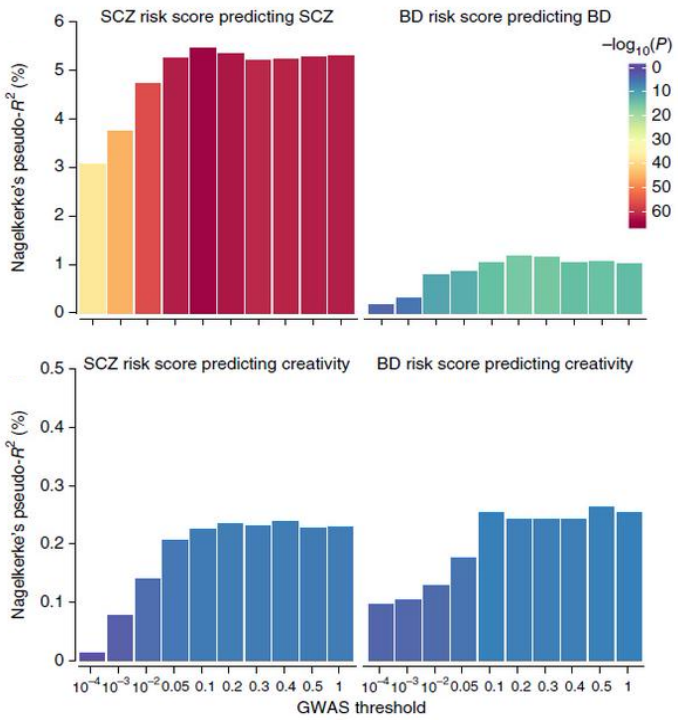


Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset ($n = 288{,}978$). The *x* axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

nature neuroscience

Polygenic risk scores for schizophrenia and bipolar disorder predict creativity

Robert A Power[1,2], Stacy Steinberg[1], Gyda Bjornsdottir[1], Cornelius A Rietveld[3], Abdel Abdellaoui[4], Michel M Nivard[4], Magnus Johannesson[5], Tessel E Galesloot[6], Jouke J Hottenga[4], Gonneke Willemsen[4], David Cesarini[7], Daniel J Benjamin[8], Patrik K E Magnusson[9], Fredrik Ullén[10], Henning Tiemeier[11], Albert Hofman[11], Frank J A van Rooij[11], G Bragi Walters[1], Engilbert Sigurdsson[12,13], Thorgeir E Thorgeirsson[1], Andres Ingason[1], Agnar Helgason[1,13], Augustine Kong[1], Lambertus A Kiemeney[6], Philipp Koellinger[14], Dorret I Boomsma[4], Daniel Gudbjartsson[1], Hreinn Stefansson[1] & Kari Stefansson[1,13]



Breakout session (5 min)

- Breakout into small groups
- Introduce yourself to everyone
- Person with LATEST letter in their LAST name will be the note taker
  - E.g. Zerick is the note taker, not Adelson
  - E.g. Abraham is the note taker, not Adelson
- Ask any questions you have:
  - “I didn’t understand what …. meant”
  - “I’m confused by what is being added in the polygenic score”
- Note takers relay unanswered questions from the breakout session

Lecture Format

- Part 1 (~40 minutes)
  - Goals of GWAS
  - What does the data look like?
  - GWAS Quality Control (QC)
  - 5 min breakout session
- Part 2 (~40 minutes)
  - Relatedness checking
  - Population stratification
  - Principal components analysis (PCA)
  - Imputation
  - 5 min breakout session
- Part 3 (~40 minutes)
  - Association testing
  - Meta-analysis
  - Polygenic Scoring
  - 5 min breakout session
- Preparation for module 3 / additional reading / lingering questions

For the next module…

- Preparation for module 3
  - Access to ATGU wiki:
  - https://sites.google.com/a/broadinstitute.org/atgu
  - Useful UNIX commands
  - https://sites.google.com/a/broadinstitute.org/atgu/getting-started/useful-unix-commands
  - Logging onto Broad servers:
  - https://sites.google.com/a/broadinstitute.org/atgu/getting-started

- Additional reading
  - Papers behind most of the methods used in statistical genetics:
  - https://sites.google.com/a/broadinstitute.org/atgu/core-publication-list
  - 10 years of GWAS discovery: Visscher_GWAS10yrs_AJHG_2017.pdf
  - Genetic architecture of complex traits: Timpson_GeneticArch_NRG_2017.pdf

- Final questions??