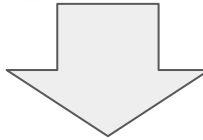


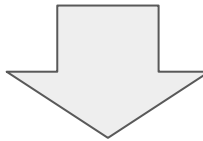
# Quality control in GWAS

Øyvind Helgeland

Genotyping array

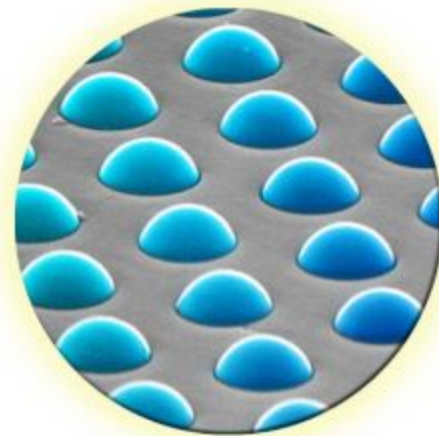
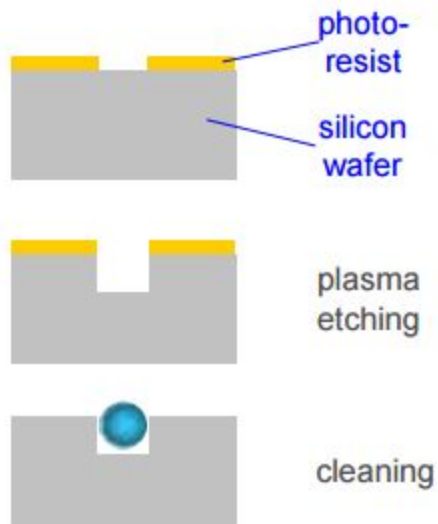


?

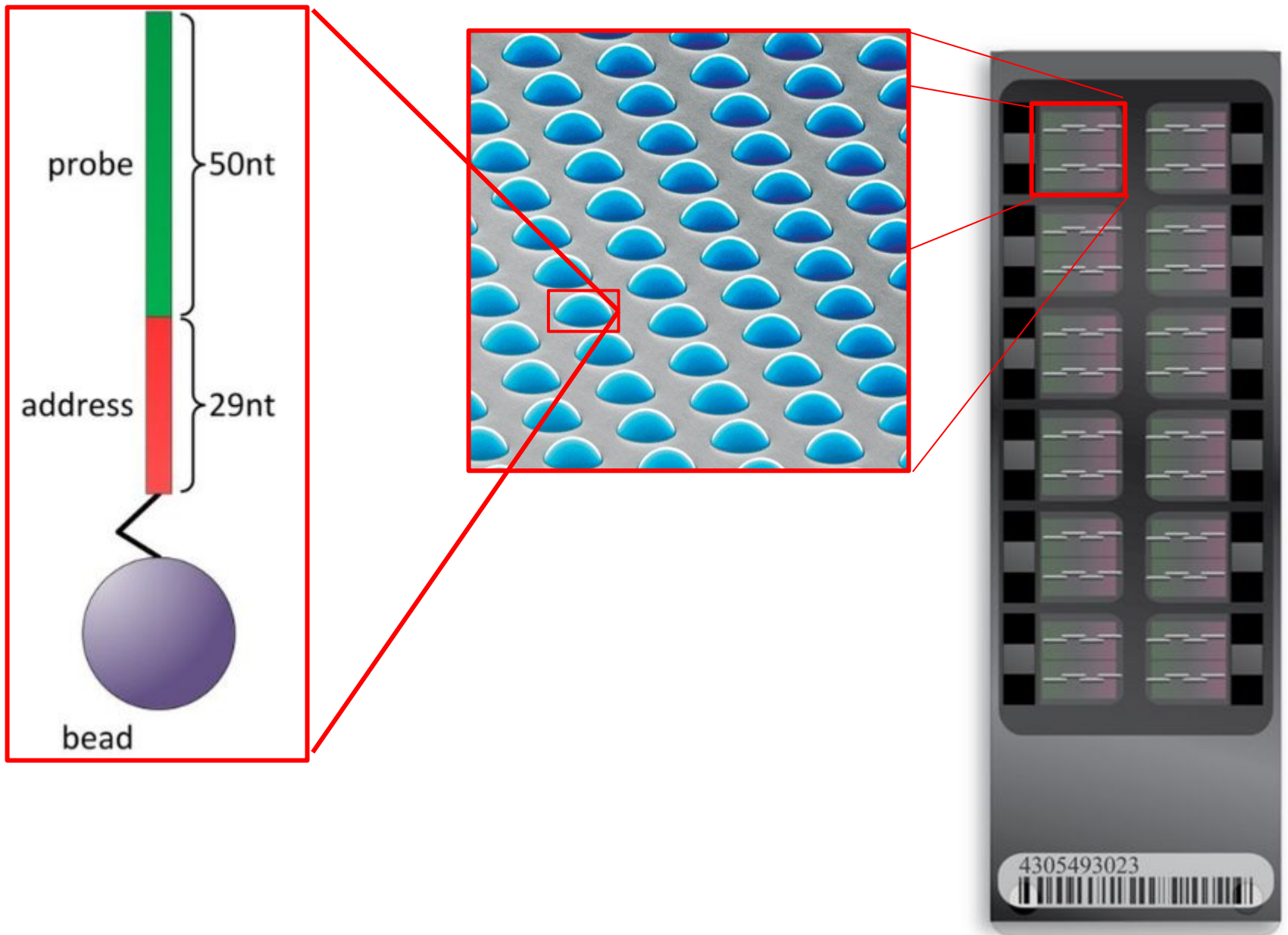


1328	NA06989	0	0	2	-9	A	G	A	C
1377	NA11891	0	0	1	-9	A	G	C	C
1349	NA11843	0	0	1	-9	G	G	C	C
1330	NA12341	0	0	2	-9	A	G	C	C
1444	NA12739	NA12748	NA12749	1	-9	G	G	C	C
1344	NA10850	0	NA12058	2	-9	A	G	A	C
1328	NA06984	0	0	1	-9	G	G	A	C
1463	NA12877	NA12889	NA12890	1	-9	A	G	C	C
1418	NA12275	0	0	2	-9	A	G	C	C
13291	NA06986	0	0	1	-9	G	G	C	C

# Microfabrication of BeadChip Wells

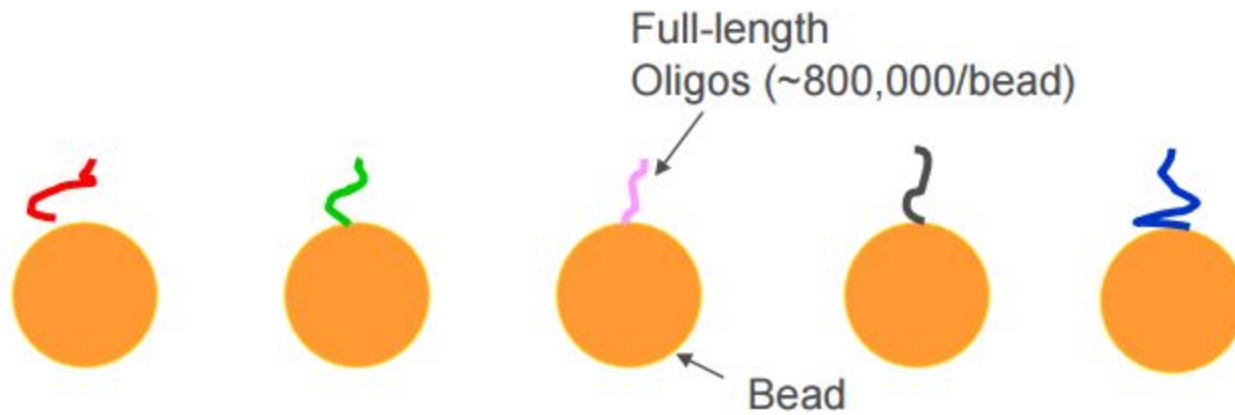


2  $\mu\text{m}$  beads in wells



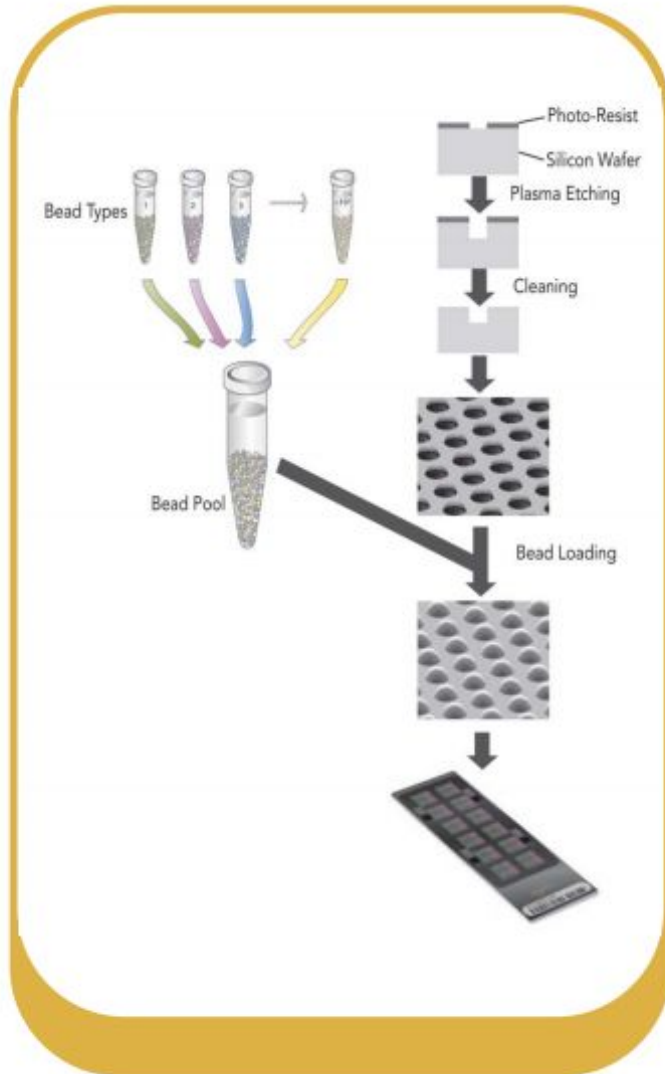
The array / "chip"

# Bead pool

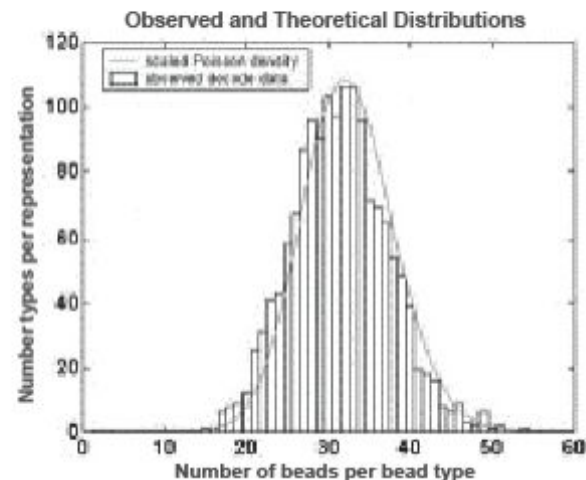


*Redundancy: average of ~15 beads per beadtype*

# Bead Preparation and Array Production

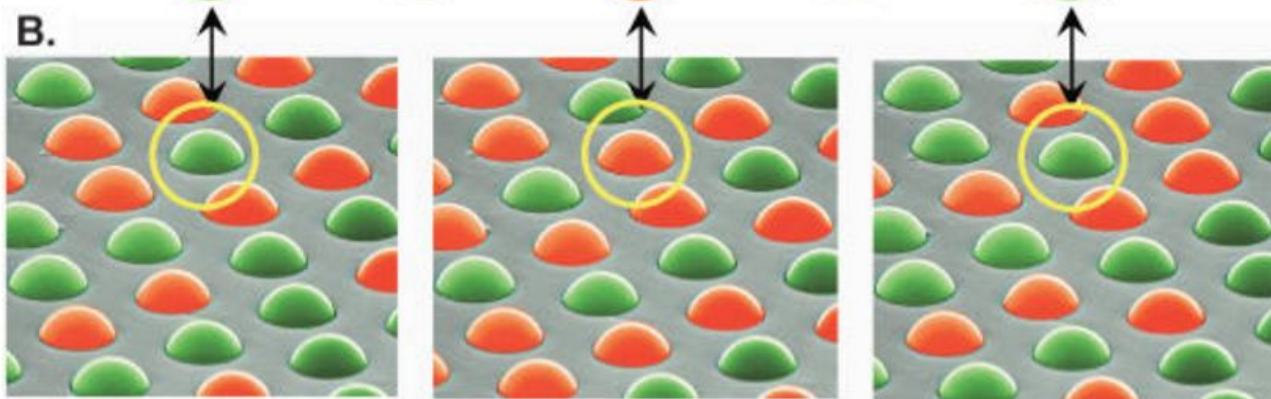


- Unique oligo for each bead type
- Bead Pool can be  $> 1,000,000$  bead types
- Random self-assembly of beads
- Average  $\sim 15$  beads per beadtype
- Functional validation of array





A. Stage 1 → Dehyb → Stage 2 → Dehyb → Stage 3



C.

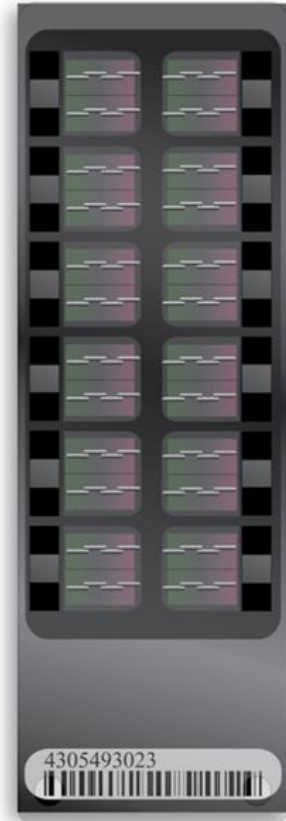
		Stage 1	Stage 2	Stage 3	Signature	Code	Parity Code
Sequences	0				GGG	000	0000
	1				GGR	001	0011
	2				GRG	010	0101
	3				GRR	011	0110
	4				RGG	100	1001
	5				RGR	101	1010
	6				RRG	110	1100
	7				RRR	111	1111



# Genotyping array

## Illumina Human Core Exome Bead Chip

- **rare and common variants** (aka. “combo-chip”)
  - ~ 250.000 common variants
  - ~ 250.000 rare variants (MAF < 1%)
- Two slightly different chips in the ERC/HARVEST-project
  - ~ 21.000 samples genotyped with Human Core Exome **12 ver. 1.1 (aka. MOBA12)**
  - ~12.000 samples genotyped using Human Core Exome **24 ver. 1.0 (aka. MOBA24)**





GENOMIC DNA (200 ng)

## DAY 1



1 Make Amplified DNA

2 Incubate Amplified DNA

## DAY 2



3 Fragment Amplified DNA



4 Precipitate & Resuspend

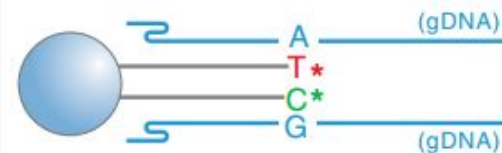


5 Prepare BeadChip



6 Hybridize Samples on BeadChip

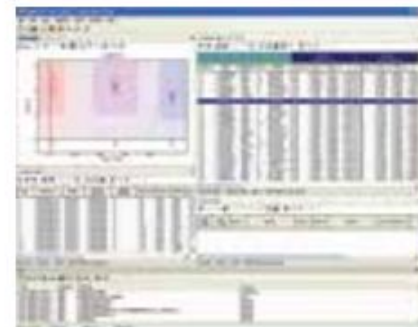
## DAY 3



7 Extend/Stain Samples on BeadChip

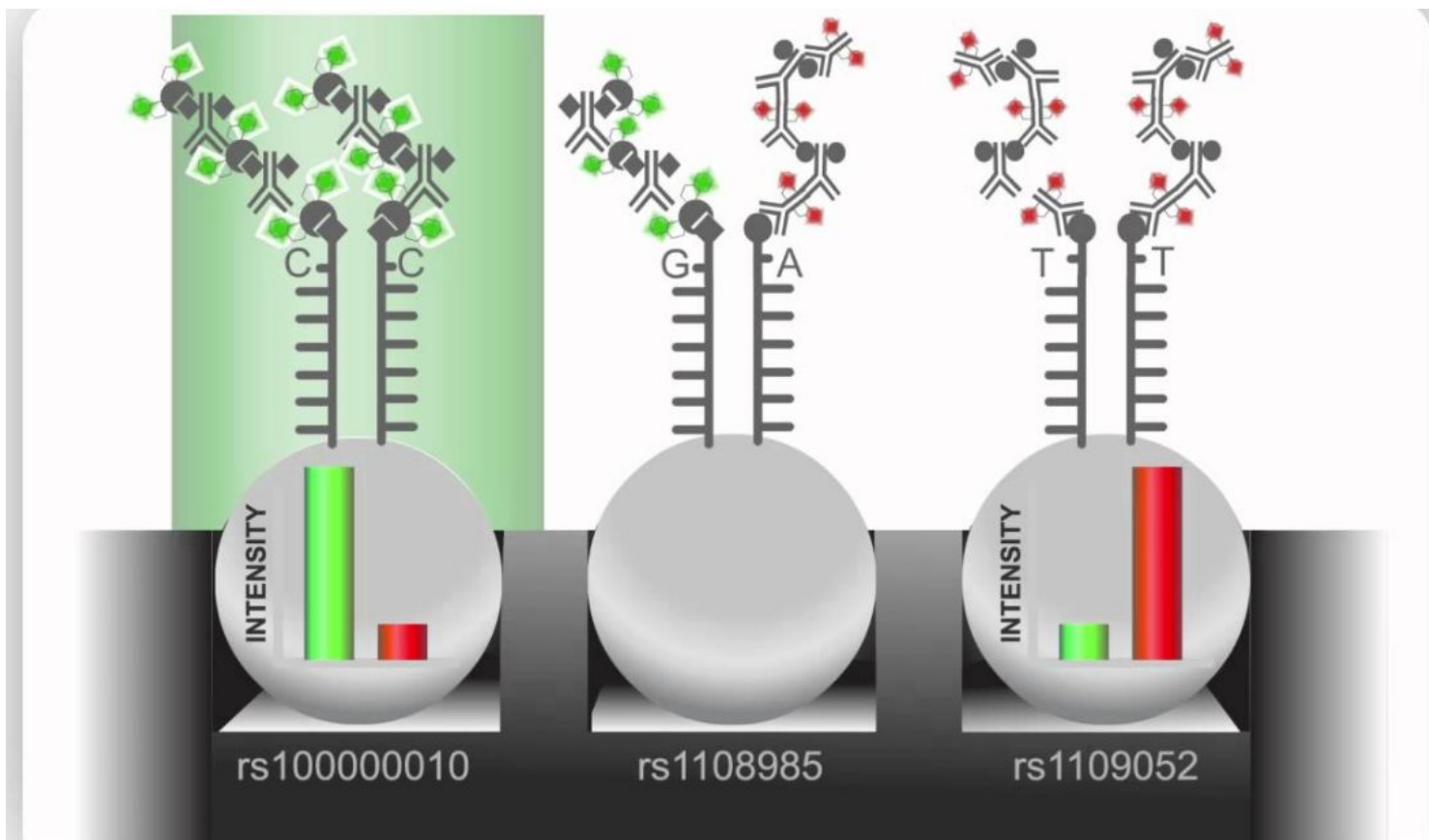


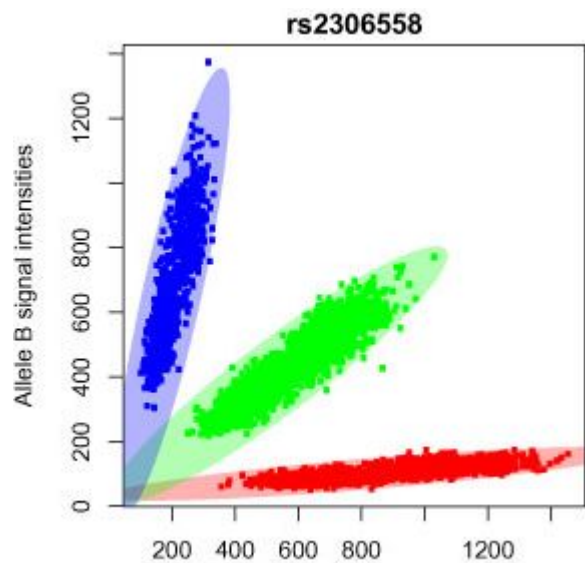
8 Image BeadChip



9 Auto-Call Genotypes and Generate Reports

# Signal intensity





.idat file

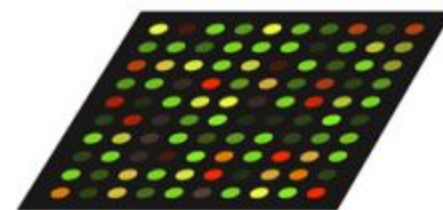
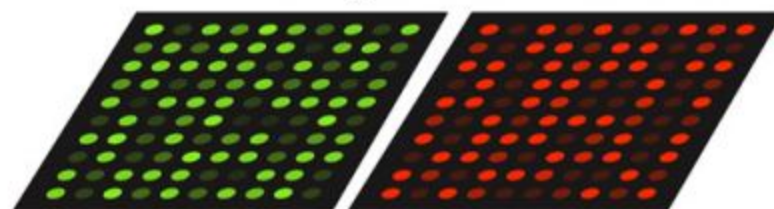


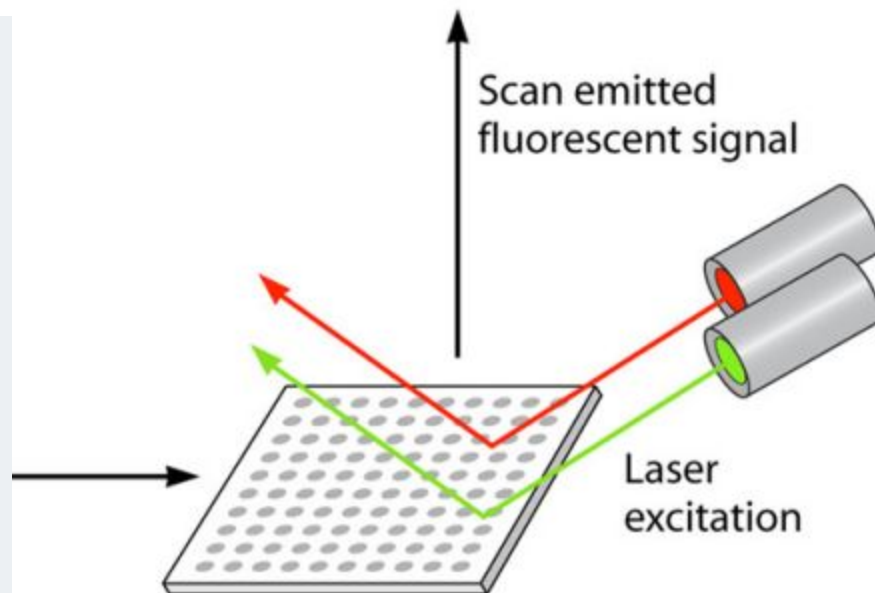
Image analysis

Raw images of each channel

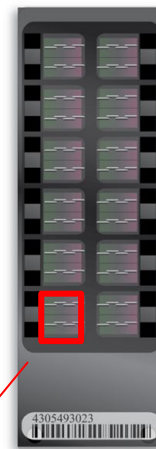
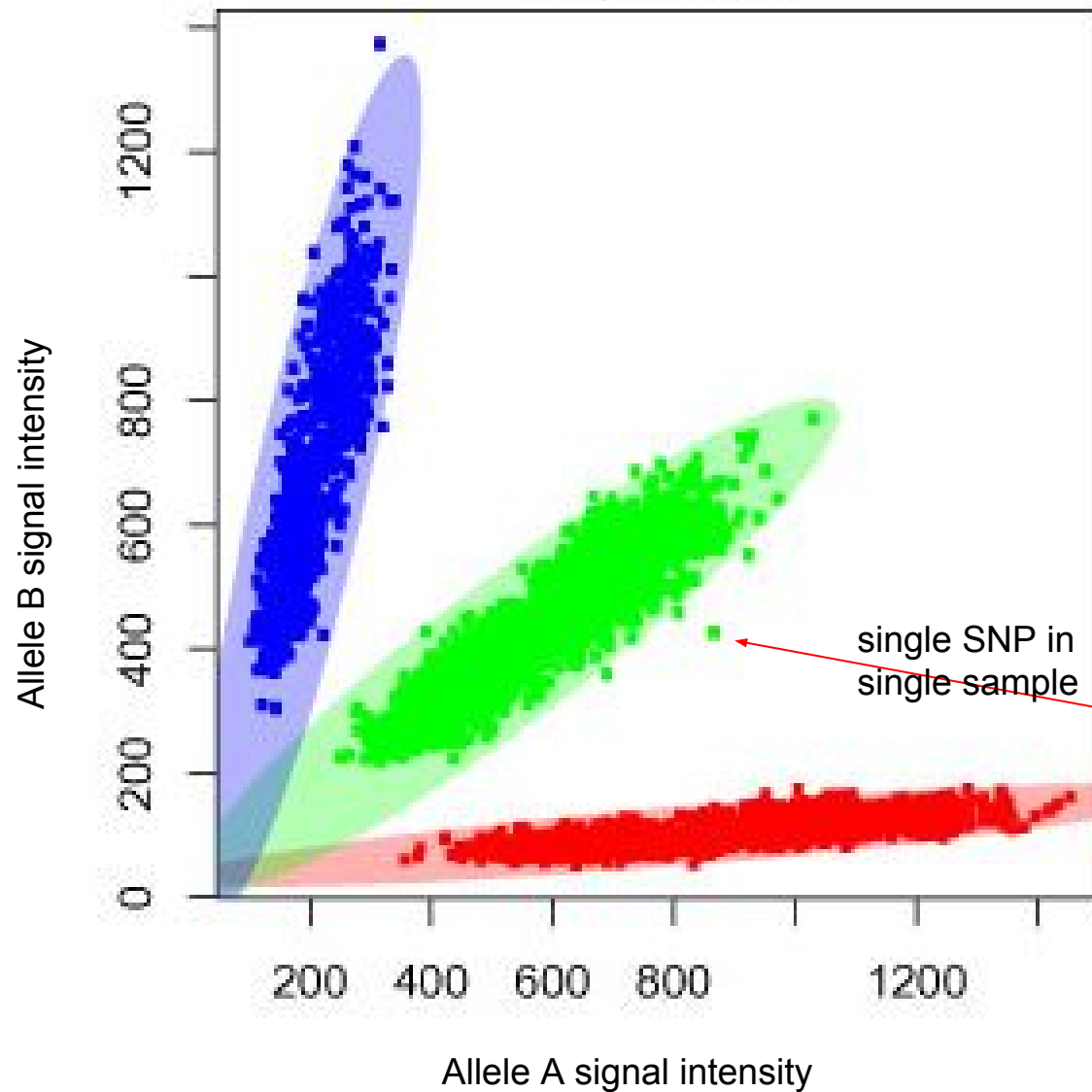


Scan emitted  
fluorescent signal

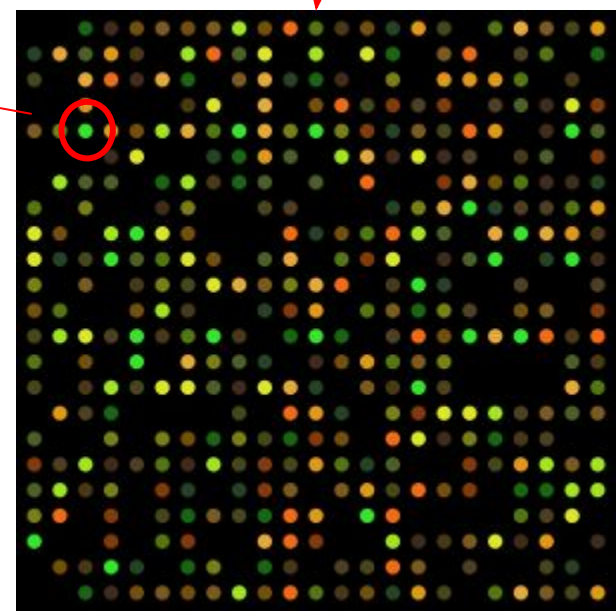
Laser  
excitation



rs2306558



single sample...



# Genotype calling

from signal intensity to genotype call

## What we have:

- The .idat file - with all the signal intensities for each SNP for each sample
- The .dmap - file (got this from manufacturer) containing information on what SNP is located where on the chip.

## What we want:

- To know the genotype for each SNP for each sample

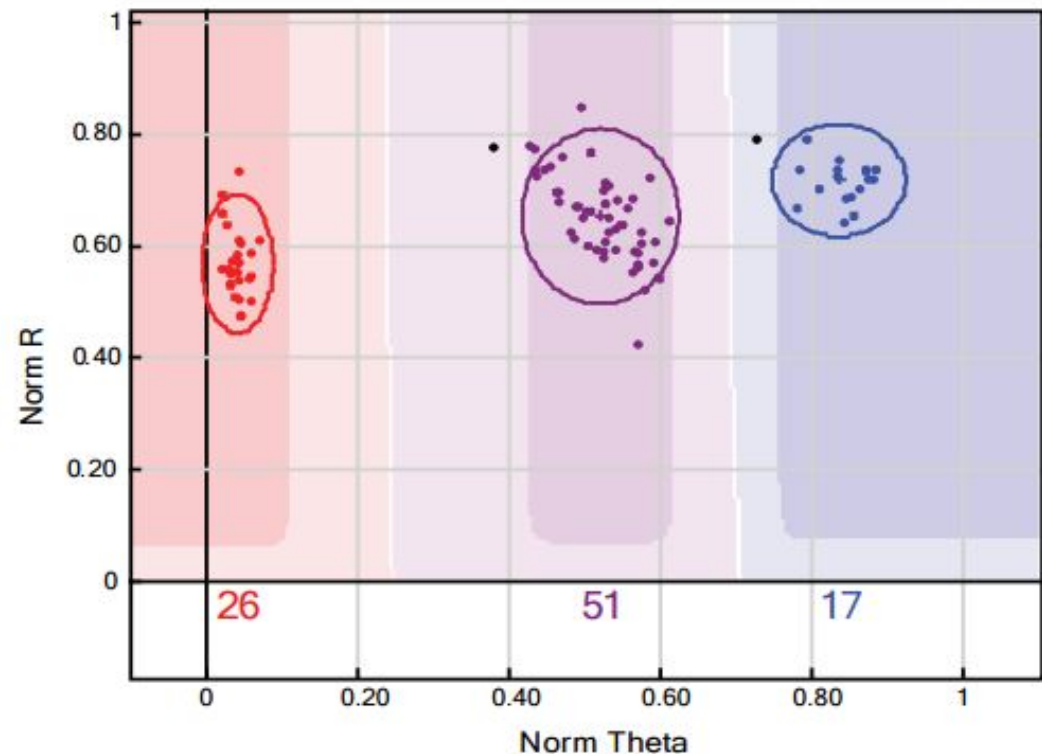
## What we need to do:

- Convert signal intensities to genotypes (aka. calling genotypes)

# Genotype calling

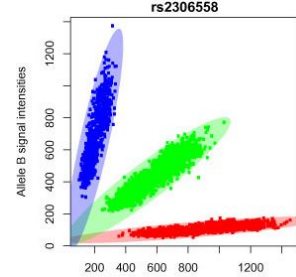
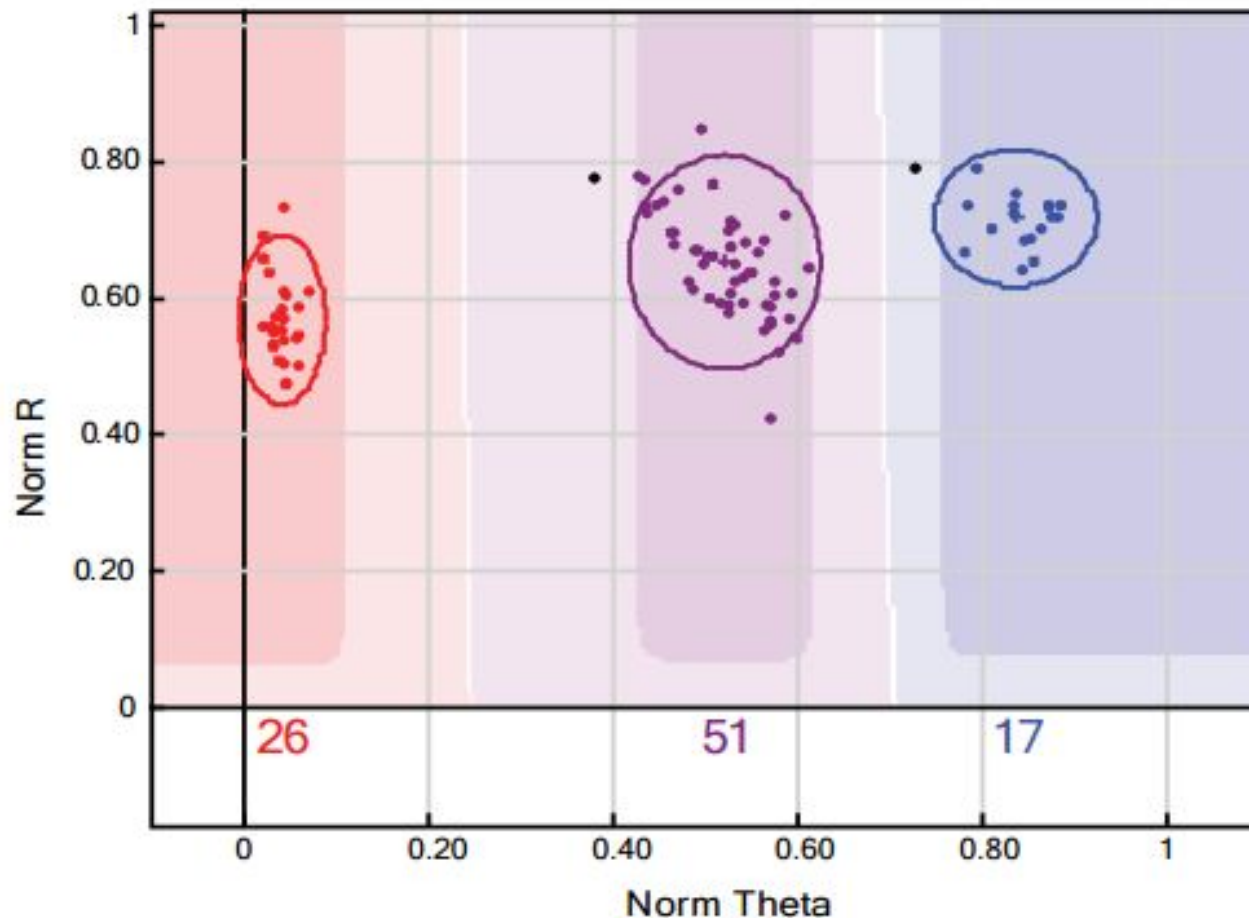
calling genotypes from clustered data

- Illumina's Genome Studio (with modules for array calling) uses GenCall algorithm to automatically cluster, call genotypes, and assign confidence scores.
- GenCall incorporates a clustering algorithm (GenTrain) and a calling algorithm
- Genotyping calls for a specific SNP is made by the calling algorithm, relying on information provided by the GenTrain clustering algorithm
- Clustering based on Illumina provided cluster file or clustered based on the data itself
- Illumina software only available for Windows...





# The more common way of plotting



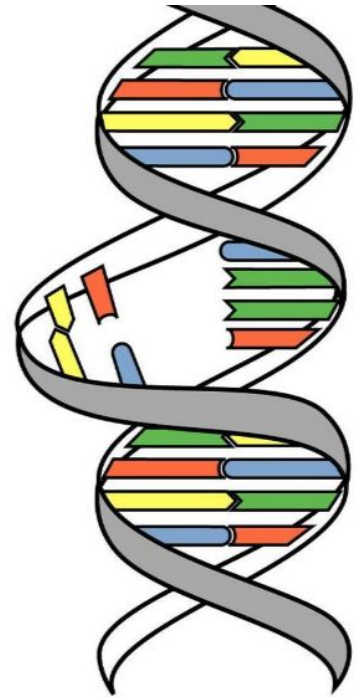
**Theta value** is the same as B allele frequency (ratio between the two different signals). It ranges from 0 to 1 and represents the fraction of bases that are genotyped as the B allele (variant allele). 0 means homozygous reference (AA), 0.5 means heterozygous (AB) and 1 means homozygous variant (BB).

**The R value** represents the fluorescence intensity of that probe of that sample.

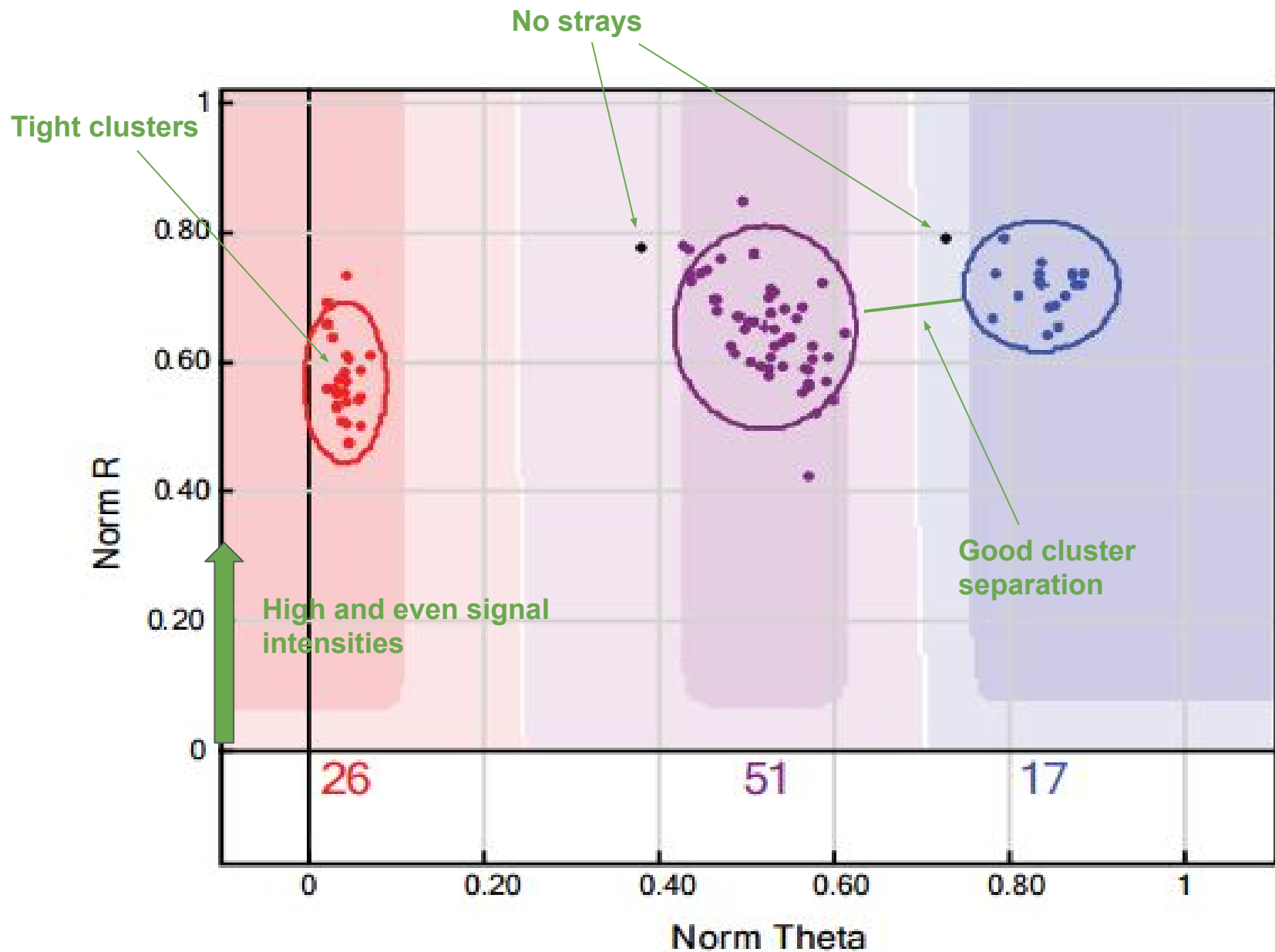
# Clustering - from signal intensity to genotype

There are several answers to why the chips underperform - many of them can be avoided!

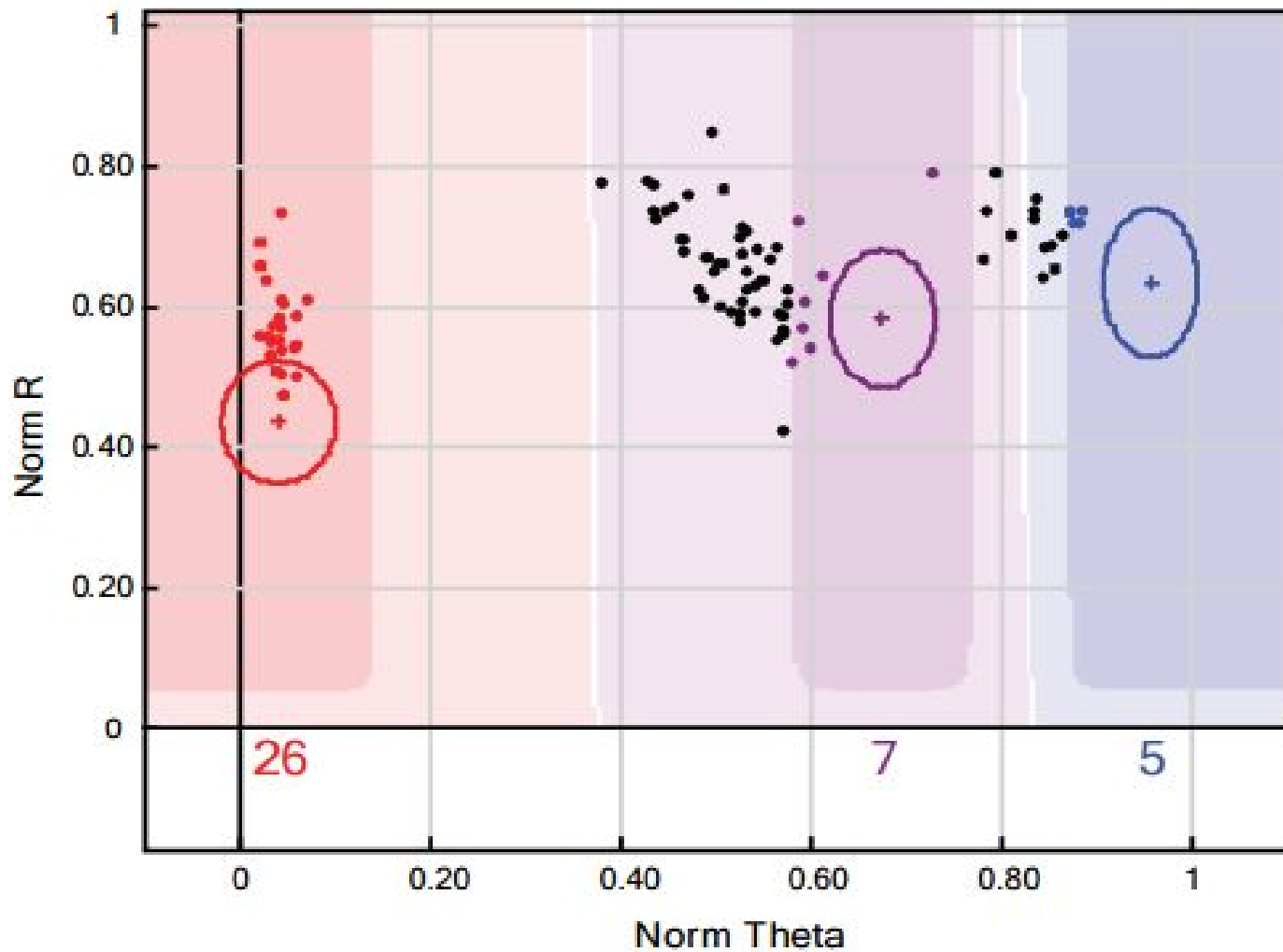
- Poor DNA quality and/or poor handling of the samples
- Contaminated sample
- Bad quality/bad handling assay chemistry
- Interaction between probes in the assay (custom array)
  - Poor hybridization
  - Poor probe specificity
- Poor handling of the chip
  - Dust
  - Hair(!)
  - Fingerprints
- Malfunctioning scanner
- Markers inherently hard to genotype
- ++



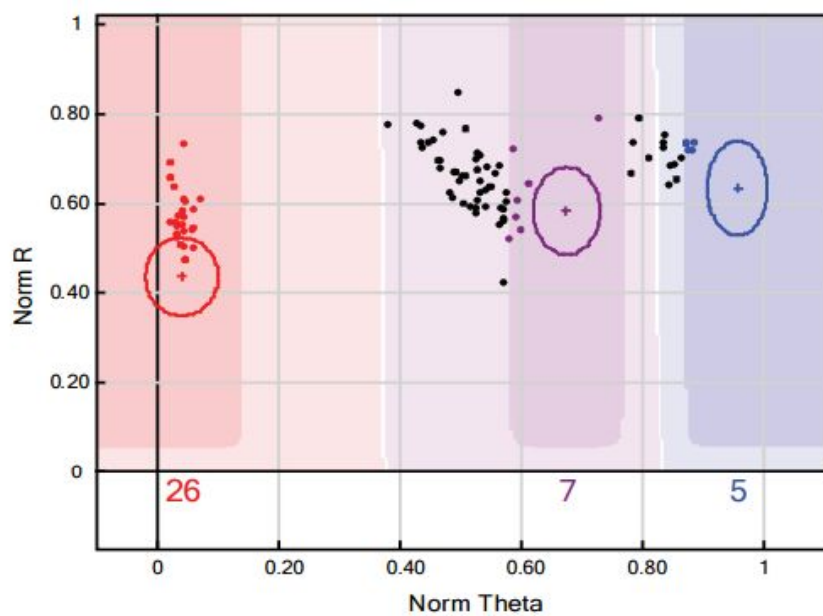
# The ideal plot



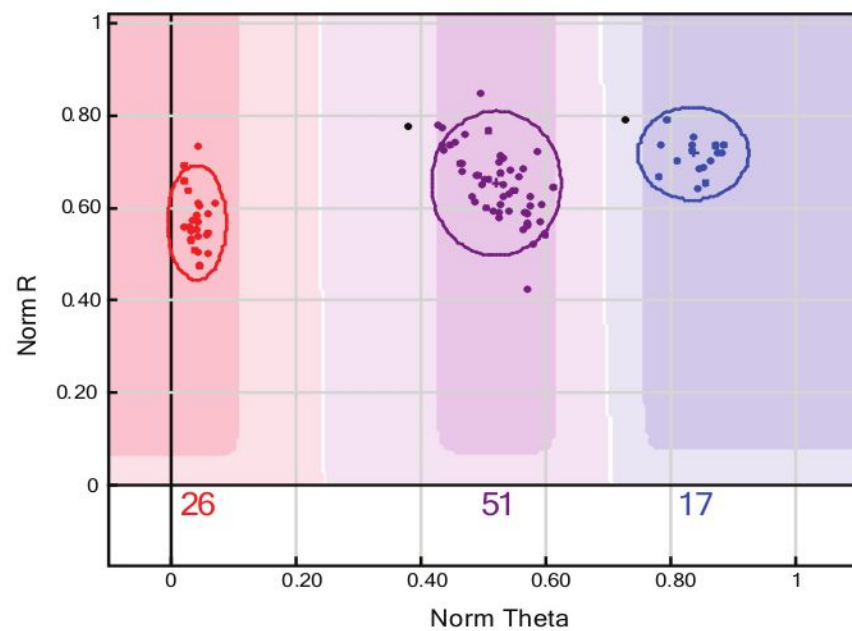
# Cluster example 1



# Cluster example 1 - continued

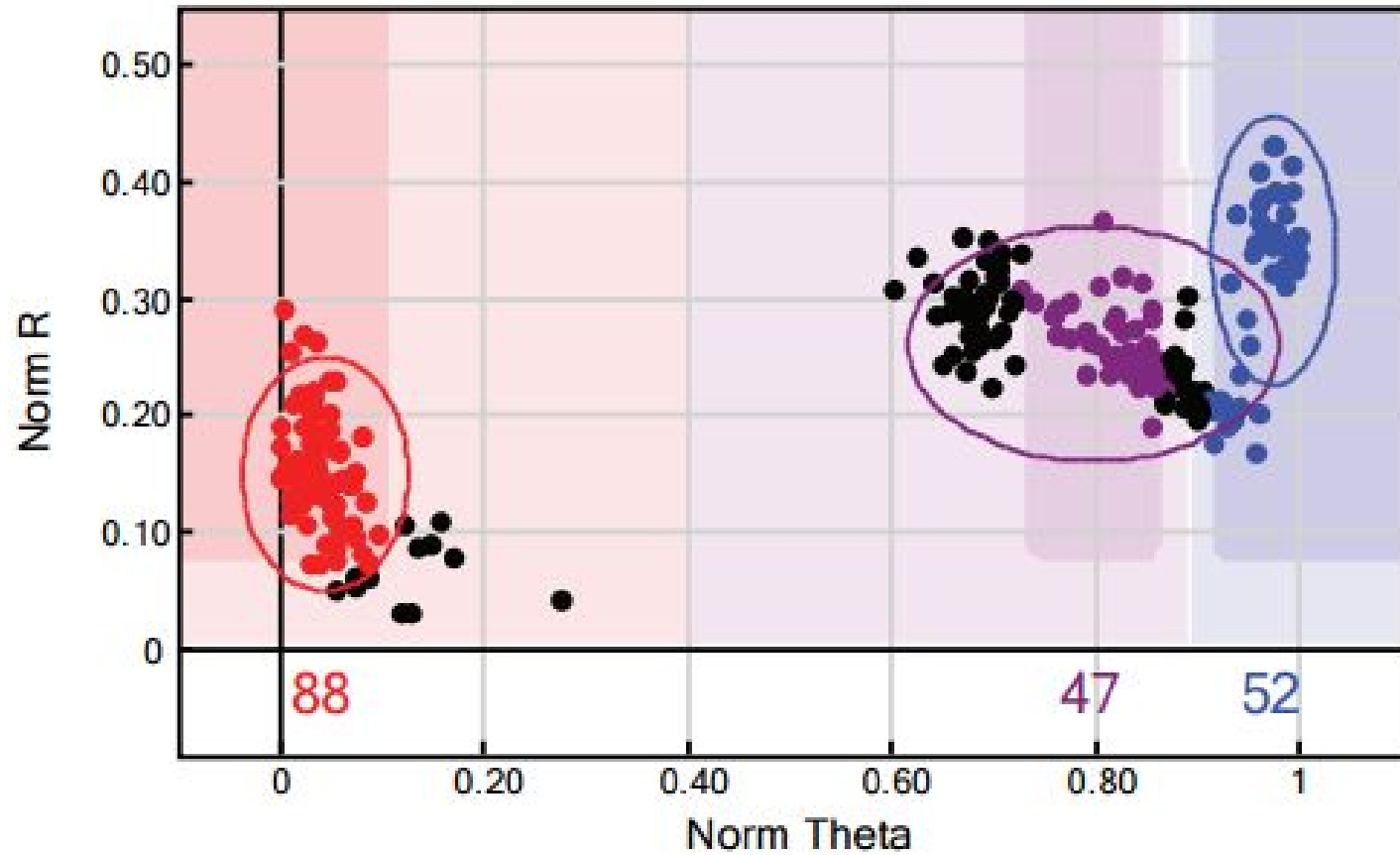


Illumina cluster file

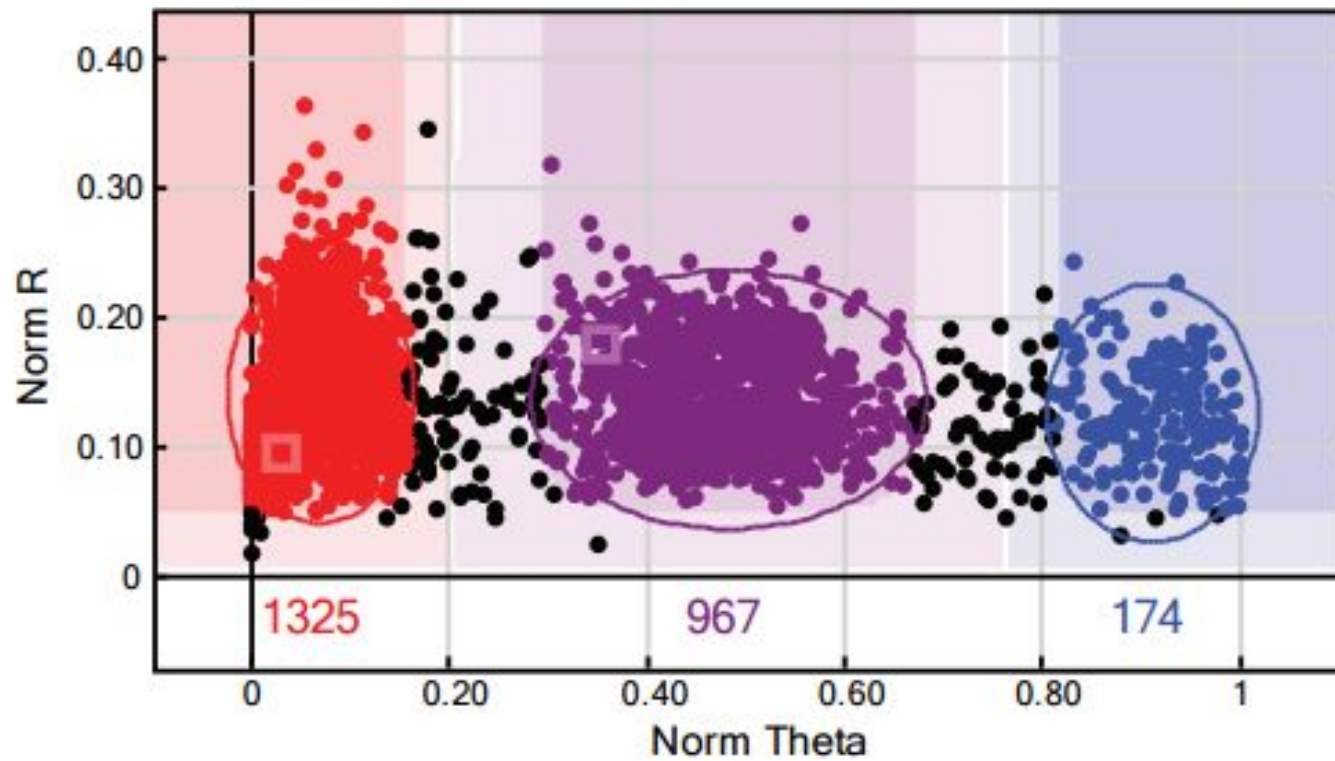


Reclustered data

## Cluster example 2

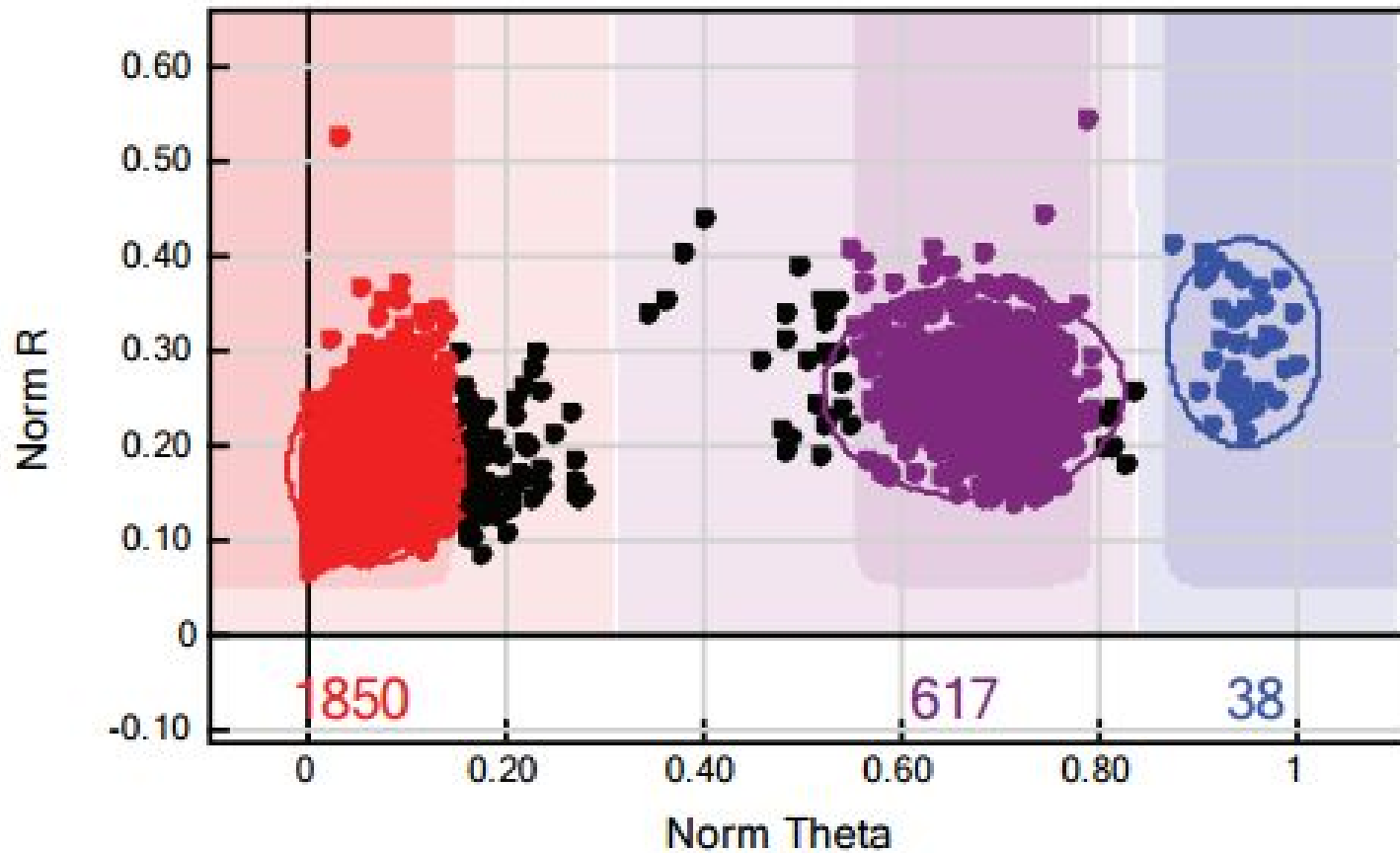


## Cluster example 3

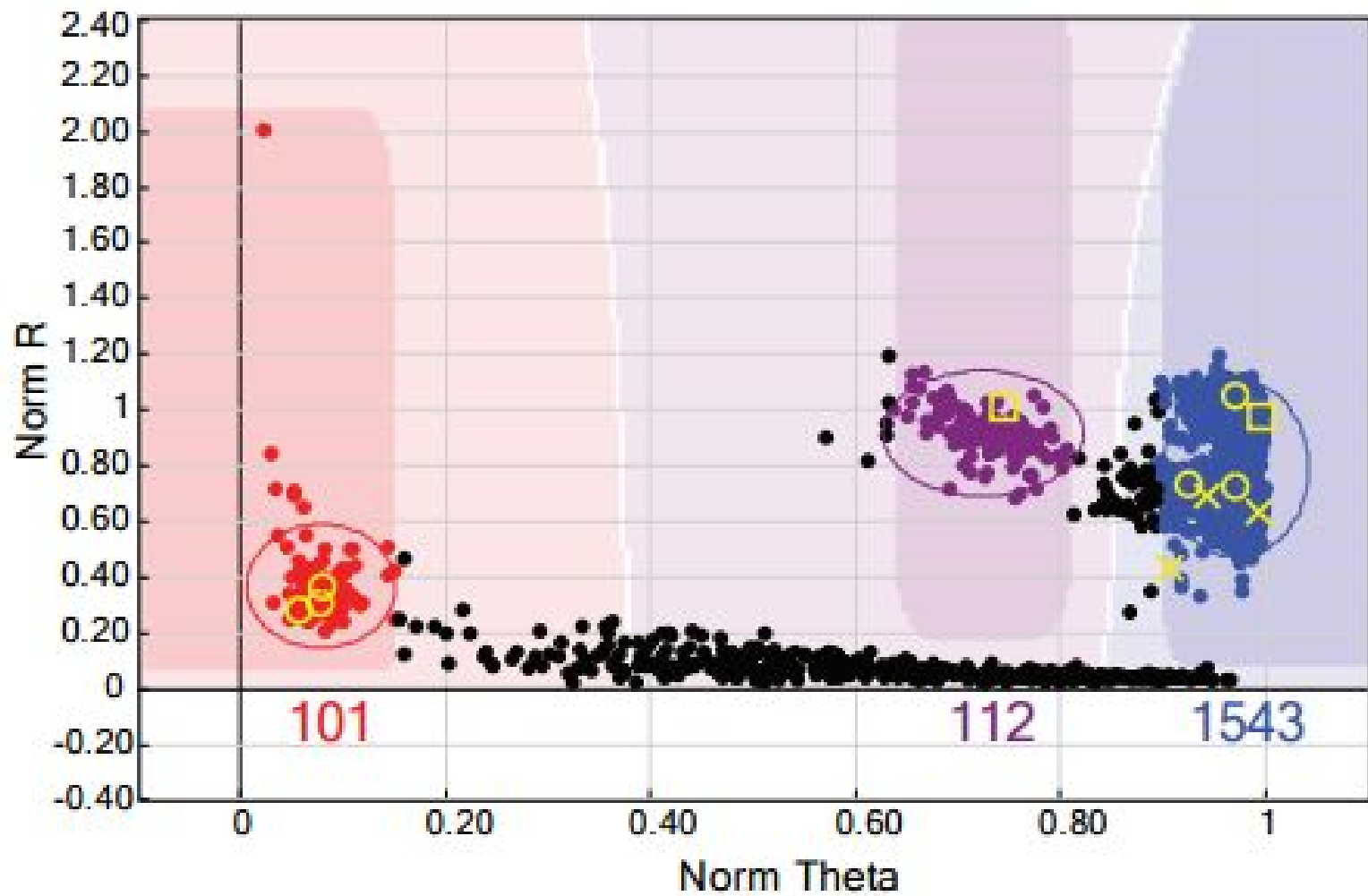




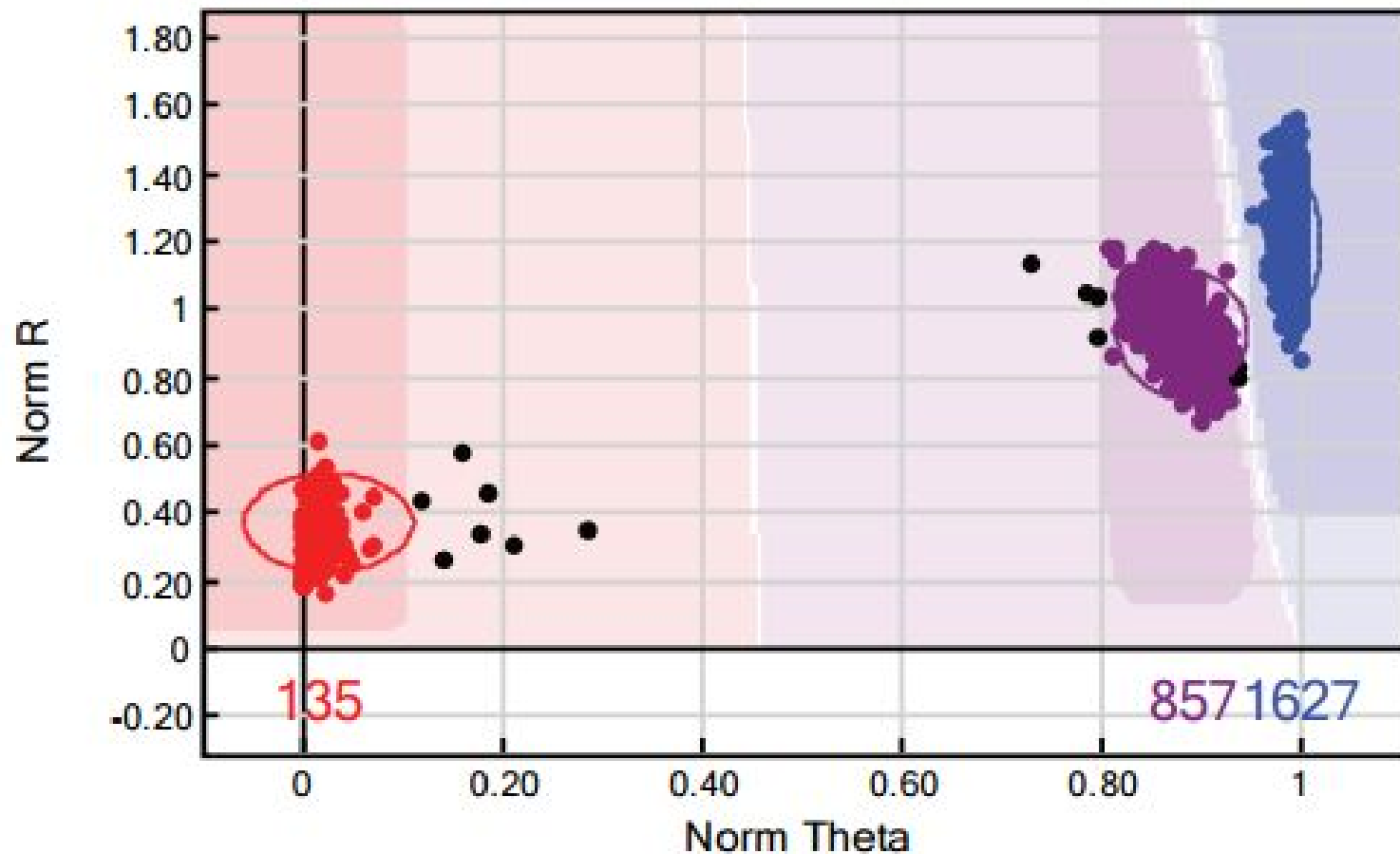
## Cluster example 4



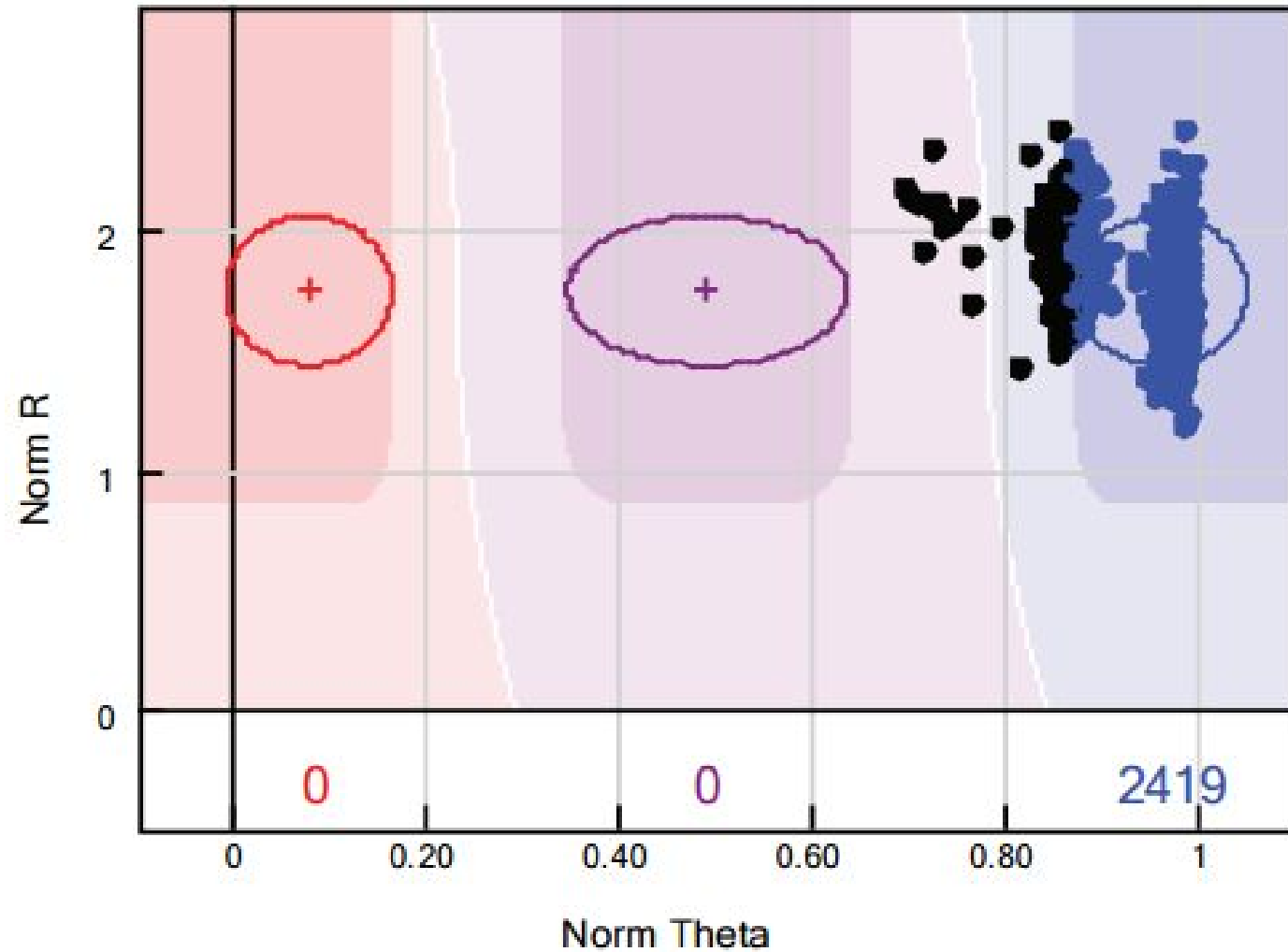
## Cluster example 5



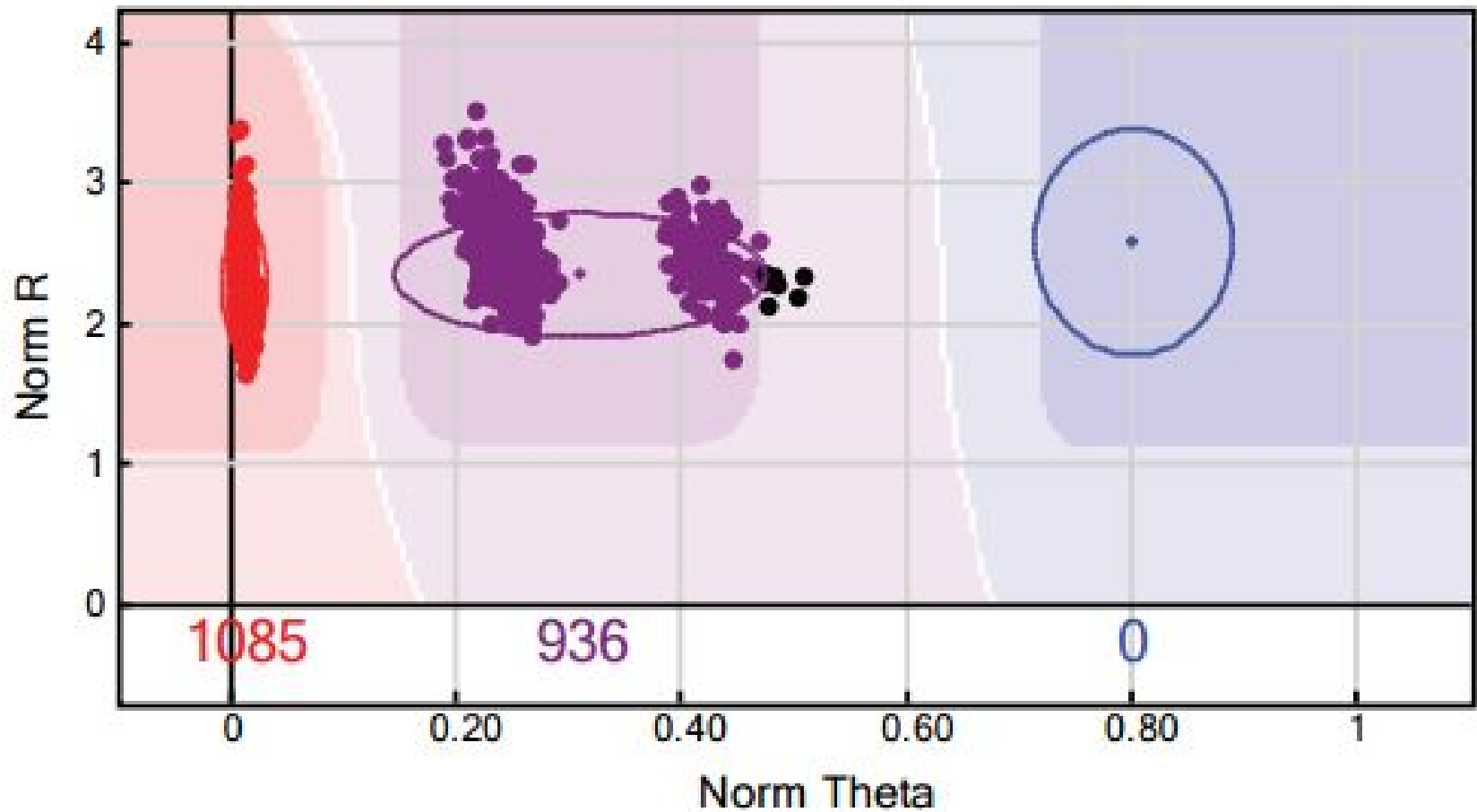
## Cluster example 6



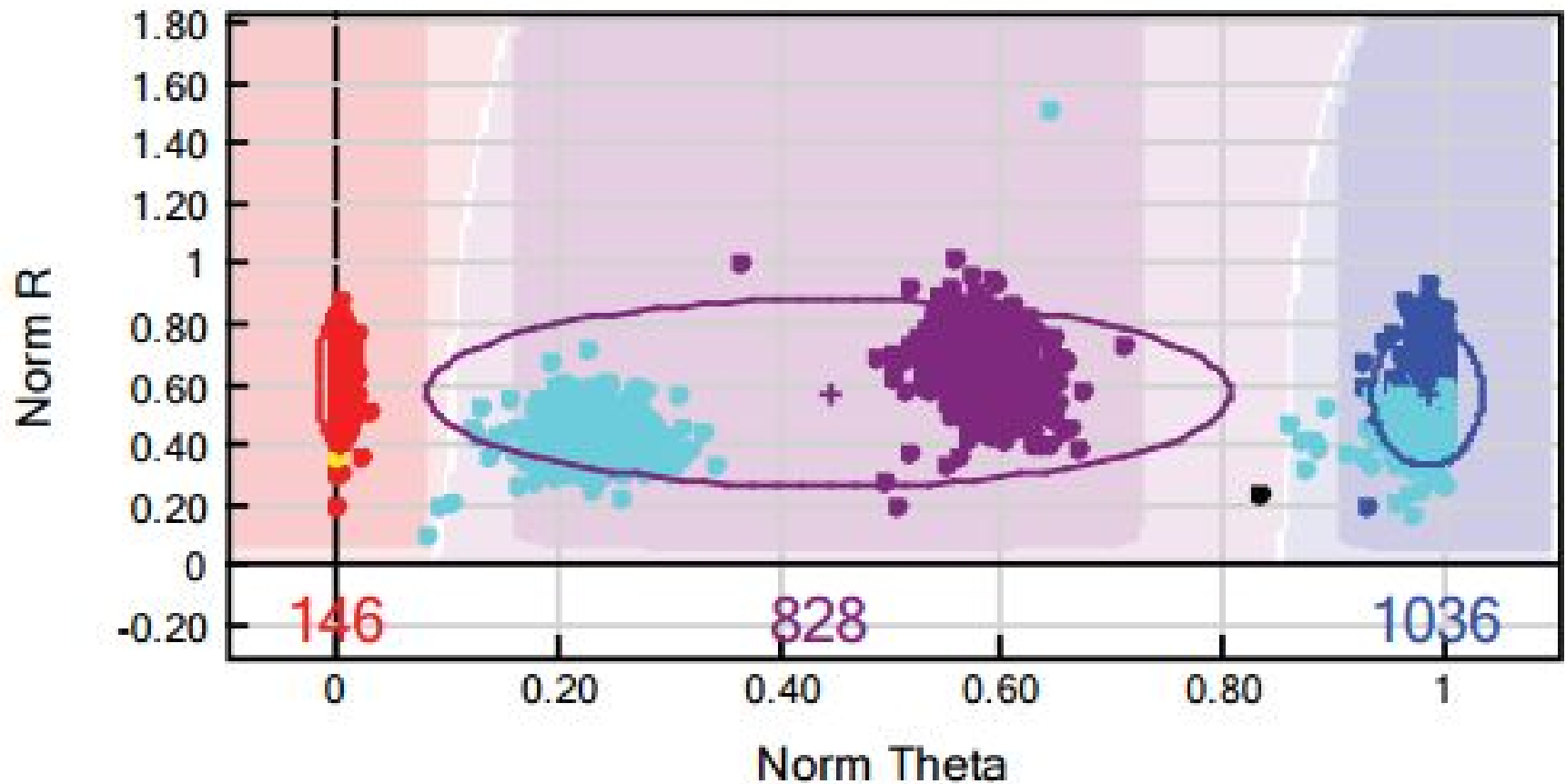
## Cluster example 7



## Cluster example 8

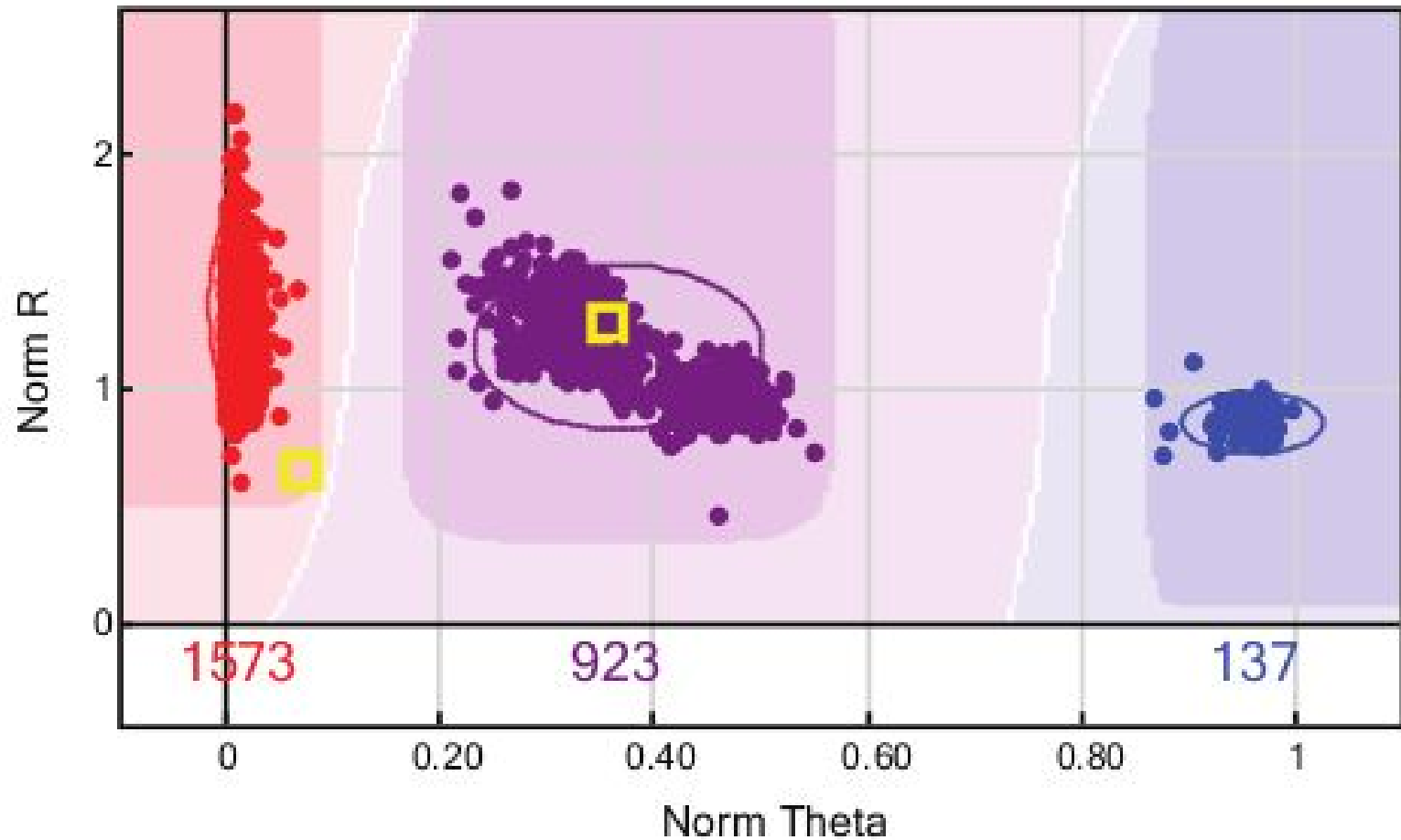


## Cluster example 9



- SNP on X-chromosome
- Turquoise points are males

## Cluster example 10



Yellow square is a replicate error (same sample genotyped twice with differing calls)



# Quality control in GenomeStudio

- Many rely on GenCall to do the genotype calling and run QC from there
- Several “best practice” papers on cluster QC exists
- Proposes thresholds for exclusion based on cluster metrics
- Can be very time/labour intensive
  - Large datasets in Windows can be slow
  - Eyeballing/correcting a lot of clusters
- Usually some, but not all, thresholds are imposed
  - e.g. cluster separation ( $> 0.3$ )
- In MOBA - full vs. light CHARGE protocol (pilot) → little difference

 **PLOS** | ONE Publish About Browse

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium

Megan L. Grove , Bing Yu, Barbara J. Cochran, Talin Haritunians, Joshua C. Bis, Kent D. Taylor, Mark Hansen, Ingrid B. Borecki, L. Adrienne Cupples, Myriam Fornage, Vilmundur Gudnason, Tamara B. Harris, Sekar Kathiresan, [ ... ], Eric Boerwinkle [\[ view all \]](#)

Published: July 12, 2013 • <http://dx.doi.org/10.1371/journal.pone.0068095>

Article	Authors	Metrics	Comments	Related Content
				

Abstract

Introduction

Results

Discussion

Materials and Methods

Acknowledgments

Author Contributions

References

---

Reader Comments (0)

Media Coverage (0)

Figures

### Abstract

Genotyping arrays are a cost effective approach when typing previously-identified genetic polymorphisms in large numbers of samples. One limitation of genotyping arrays with rare variants (e.g., minor allele frequency [MAF]  $< 0.01$ ) is the difficulty that automated clustering algorithms have to accurately detect and assign genotype calls. Combining intensity data from large numbers of samples may increase the ability to accurately call the genotypes of rare variants. Approximately 62,000 ethnically diverse samples from eleven Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium cohorts were genotyped with the Illumina HumanExome BeadChip across seven genotyping centers. The raw data files for the samples were assembled into a single project for joint calling. To assess the quality of the joint calling, concordance of genotypes in a subset of individuals having both exome chip and exome sequence data was analyzed. After exclusion of low performing SNPs on the exome chip and non-overlap of SNPs derived from sequence data, genotypes of 185,119 variants (11,356 were monomorphic) were compared in 530 individuals that had whole exome sequence data. A total of 98,113,070 pairs of genotypes were tested and 99.77% were concordant, 0.14% had missing data, and 0.09% were discordant. We report that joint calling allows the ability to accurately genotype rare variation using array technology when large sample sizes are available and best practices are followed. The cluster file from this experiment is available at [www.chargeconsortium.com/main/exomechip](http://www.chargeconsortium.com/main/exomechip).

# CHARGE consortium - “best practice”

## Best Practices Criteria

All X, Y, XY and MT variants

Call frequency between 0.95 and 0.99

Cluster separation  $< 0.4$

AB frequency  $> 0.6$

AB R mean  $< 0.2$

Het excess  $> 0.1$

Het excess  $< -0.9$

AA theta mean between 0.2 and 0.3

BB theta mean between 0.7 and 0.8

AB theta mean between 0.2 and 0.3

AB theta mean between 0.7 and 0.8

AA theta deviation  $> 0.025$

AB theta deviation  $\geq 0.07$

BB theta deviation  $> 0.025$

AB frequency = 0 and minor allele frequency  $> 0$

AA frequency = 1 and call rate  $< 1$

BB frequency = 1 and call rate  $< 1$

MAF  $< 0.0001$  and call rate  $\neq 1$

Rep error  $> 2$

PPC error  $> 1$

PC error  $> 1$

Variants removed from v1.1 exome chip

Cautious sites

PLINK

# PLINK

- Open-source whole genome association analysis software
- Developed at the Broad Institute
  - v.1 Shaun Purcell - slow and less features, good documentation (latest stable version)
  - v.2 (aka. 1.9 beta) - Christopher Chang - faster, more features, good documentation
- Several other (better) tools available for complex analyses, imputation etc.
- Now mostly used for data handling and basic analyses
- Wide support for various file formats (input and output)
- Multi-platform (Windows, Linux, OSX) - no compilation needed
- Available at - <https://www.cog-genomics.org/plink2/>
- Tip: Spend a day reading the documentation!

## Introduction, downloads

S: 16 May 2016 (b3.37)

D: 16 May 2016

[Recent version history](#)

[What's new?](#)

[Future development](#)

[Limitations](#)

[Note to testers](#)

## [Jump to search box]

## General usage

[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](#)

[Phenotypes](#)

[Covariates](#)

[Clusters of samples](#)

[Variant sets](#)

[Binary distance matrix](#)

[IBD report \(.genome\)](#)

## Input filtering

[Sample ID file](#)

[Variant ID file](#)

[Cluster membership](#)

[Set membership](#)

[Attribute-based](#)

[Chromosomes](#)

[SNPs only](#)

[Simple variant window](#)

[Multiple variant ranges](#)

[Sample/variant thinning](#)

[Covariates \(--filter\)](#)

[Missing genotypes](#)

[Missing phenotypes](#)

[Minor allele frequencies](#)

[Hardy-Weinberg](#)

[Mendel errors](#)

[Quality scores](#)

## 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](#) at Mount Sinai School of Medicine, and others. ([What's new?](#)) ([Credits.](#)) ([Methods paper.](#))

## Binary downloads

Operating system <sup>1</sup>	Build		
	Stable (beta 3.37, 16 May)	Development (16 May)	Old <sup>2</sup> (v1.07)
Linux 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Linux 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
OS X (64-bit)	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Windows 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Windows 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>

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://pngu.mgh.harvard.edu/~purcell/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 3:

- [--qual-geno-scores](#)<sup>3</sup>
- [--segment](#)<sup>4</sup>
- [--dfam](#)
- [--p2](#), [--genedrop](#)
- [--hap](#), [--hap-window](#), [--hap-snp](#)<sup>5</sup>
- [--proxy-assoc](#), [--proxy-impute](#)<sup>5</sup>
- [--cnv-list](#), [--cfile](#), [--gfile](#)
- [--R](#)
- [--id-dict](#), [--id-match](#)<sup>6</sup>
- [--compress](#), [--decompress](#)<sup>7</sup>

Continue using PLINK 1.07 for most of these operations. However, be aware that

# PLINK - Standard input file formats

**PLINK primarily has two “native” formats for storing genotype data**

- Text files (aka. “ped-files”) - all information is stored in plain text
  - Easier to handle when datasets were small (early GWASs)
  - Easy to view/parse data in the console
  - Files can get very big in large datasets with many markers
  - Many third party tools require raw text files (not compliant with PLINK binary datasets)
  - fileset consists of:
    - .ped - containing genotype and phenotype data
    - .map - containing marker information data
- Binary files - genotypes stored as binary data
  - Smaller filesets
  - Faster file operations
  - Third party tools are gradually more supportive of this format
  - fileset consists of:
    - .bed - binary representation of genotypes
    - .bim - text file containing marker information
    - .fam - text file containing pedigree information

# Text files/"ped-files"/ **ped**+map

.ped file: genotype + phenotype

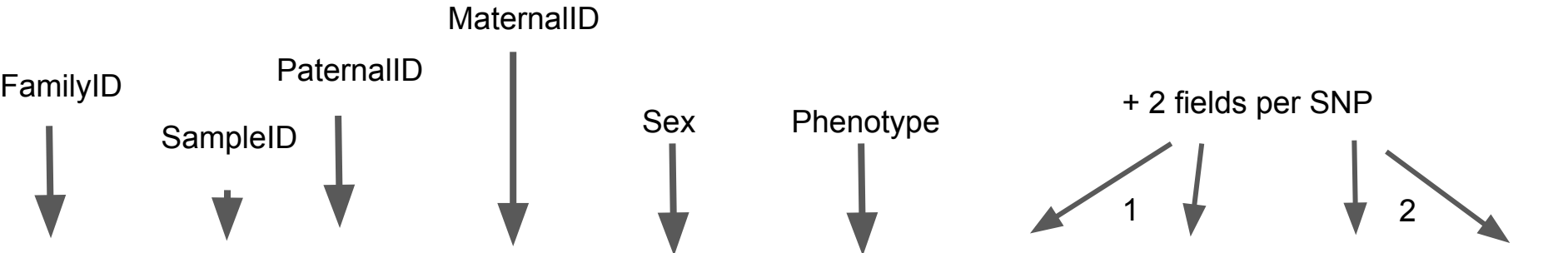


Diagram illustrating the structure of a .ped file, showing the mapping of fields to columns in the data table below:

- FamilyID (points to column 1)
- SampleID (points to column 2)
- PaternalID (points to column 3)
- MaternalID (points to column 4)
- Sex (points to column 5)
- Phenotype (points to column 6)
- + 2 fields per SNP (points to columns 7 and 8, labeled 1 and 2)

1328	NA06989	0	0	2	-9	A	G	A	C
1377	NA11891	0	0	1	-9	A	G	C	C
1349	NA11843	0	0	1	-9	G	G	C	C
1330	NA12341	0	0	2	-9	A	G	C	C
1444	NA12739	NA12748	NA12749	1	-9	G	G	C	C
1344	NA10850	0	NA12058	2	-9	A	G	A	C
1328	NA06984	0	0	1	-9	G	G	A	C
1463	NA12877	NA12889	NA12890	1	-9	A	G	C	C
1418	NA12275	0	0	2	-9	A	G	C	C
13291	NA06986	0	0	1	-9	G	G	C	C



# Text files/"ped-files"/ ped+map

.map file: marker information

Chromosome                  SNP-ID                  start-position (bp)                  end position (bp)



1	rs3131972	0	742584
1	rs4970383	0	828418
1	rs4475691	0	836671
1	rs13302982	0	851671
1	rs28391282	0	894028
1	rs2341354	0	908436
1	rs9777703	0	918699
1	rs1891910	0	922320
1	rs3128117	0	934427
1	rs2465136	0	980280

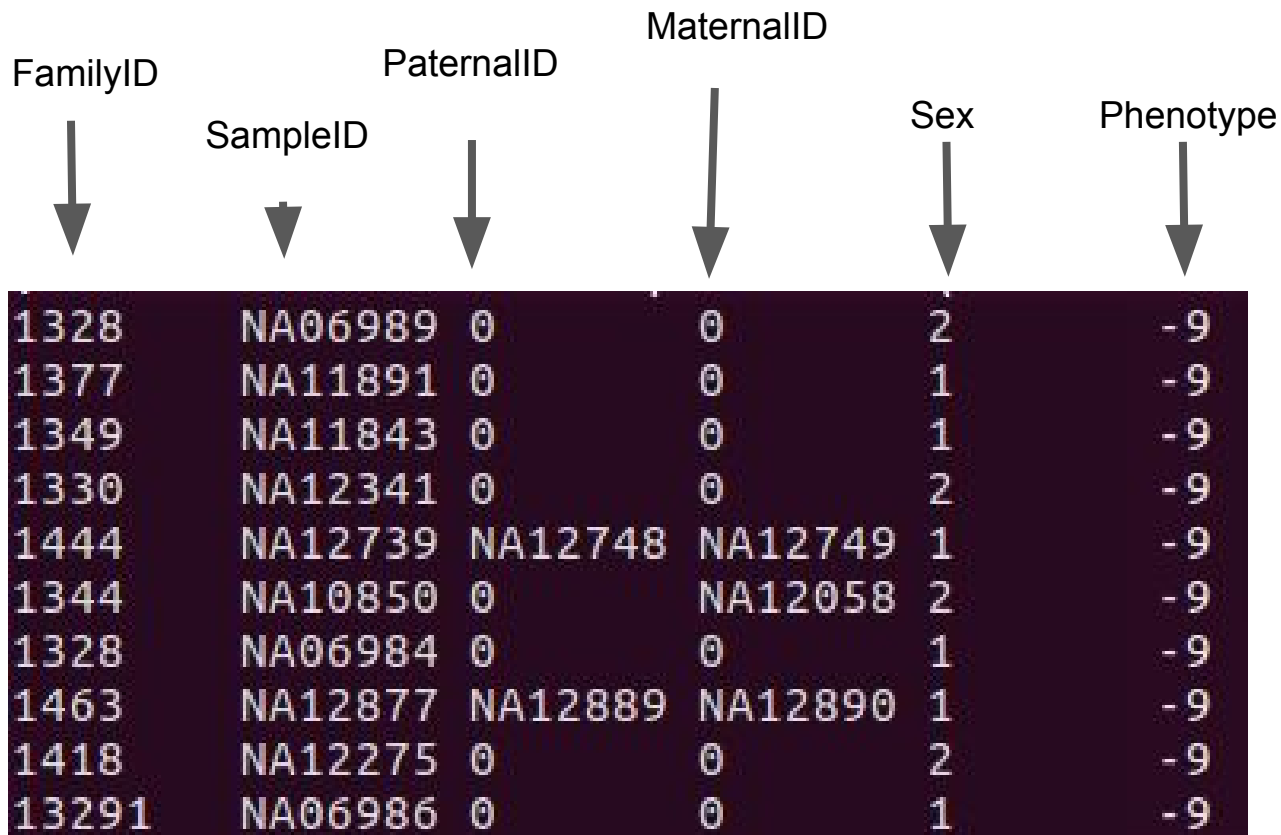
# Binary files/"bed-files"/(bed+bim+fam)

.bim file: marker information

Chromosome	SNP-ID	start-position (bp)	end position (bp)	Allele1	Allele2
↓	↓	↓	↓	↓	↓
1	rs3131972	0	742584	A	G
1	rs4970383	0	828418	A	C
1	rs4475691	0	836671	T	C
1	rs13302982	0	851671	A	G
1	rs28391282	0	894028	A	G
1	rs2341354	0	908436	A	G
1	rs9777703	0	918699	C	T
1	rs1891910	0	922320	A	G
1	rs3128117	0	934427	C	T
1	rs2465136	0	980280	C	T

# Binary files/"bed-files"/(bed+bim+fam)

.fam file: pedigree + phenotype (i.e. ped file without the genotypes)



FamilyID	SampleID	PaternalID	MaternalID	Sex	Phenotype
1328	NA06989	0	0	2	-9
1377	NA11891	0	0	1	-9
1349	NA11843	0	0	1	-9
1330	NA12341	0	0	2	-9
1444	NA12739	NA12748	NA12749	1	-9
1344	NA10850	0	NA12058	2	-9
1328	NA06984	0	0	1	-9
1463	NA12877	NA12889	NA12890	1	-9
1418	NA12275	0	0	2	-9
13291	NA06986	0	0	1	-9

# PLINK general usage

`./plink` **<input data>** **<filtering>** **<analysis>** **<output coding>** **<outputname>**

## **input data:**

- file** for text input (ped+map)
- bfile** for binary input (bed+bim+fam)

## **filtering:**

- geno** filtering genotypes by missingness
- mind** filtering samples by missingness

## **analysis**

- linear** linear regression analysis

## **output coding**

- recode AD** - convert to textfile
- make-bed** - convert to binary file

## **output name**

- out** name of (stem) of output

# PLINK Documentation

the flag

## Generate text fileset

optional modifier in <>

```
--recode <01 | 12> <23 | A | A-transpose | AD | beagle | beagle-nomap |  
bimbam | bimbam-lchr | compound-genotypes | fastphase | fastphase-lchr | HV  
| HV-lchr | lgen | lgen-ref | list | oxford | rlist | structure | transpose  
| vcf | vcf-fid | vcf-iid> <tab | tabx | spacex | bgz | gen-gz> <include-  
alt> <omit-nonmale-y>  
--recode-allele [filename]
```

required params in [ ]

**--recode** creates a new text fileset, after applying sample/variant filters and other operations. By default, the fileset includes a **.ped** and a **.map** file, readable with **--file**.

- The **'12'** modifier causes A1 (usually minor) alleles to be coded as '1' and A2 alleles to be coded as '2', while **'01'** maps A1→0 and A2→1. (PLINK forces you to combine '01' with **{output-missing-genotype}** when this is necessary to prevent missing genotypes indistinguishable from A1 calls.)
- The **'23'** modifier causes a 23andMe-formatted file to be generated. This can only be used on a single sample's data (a one-line **--keep** file may come in handy here). There is currently no special handling of the XY pseudo-autosomal region.
- The **'AD'** modifier causes an **additive (0/1/2) + dominant (het = 1, otherwise 0) component file**, suitable for loading from R, to be generated. **'A'** is the same, except without the dominance component.
  - By default, A1 alleles are counted; this can be customized with **--recode-allele**. **--recode-allele**'s input file should have variant IDs in the first column and allele IDs in the second.
  - By default, the header line for raw files only names the counted alleles. To include the

Additional info on each param below

Introduction to quality control of called genotypes



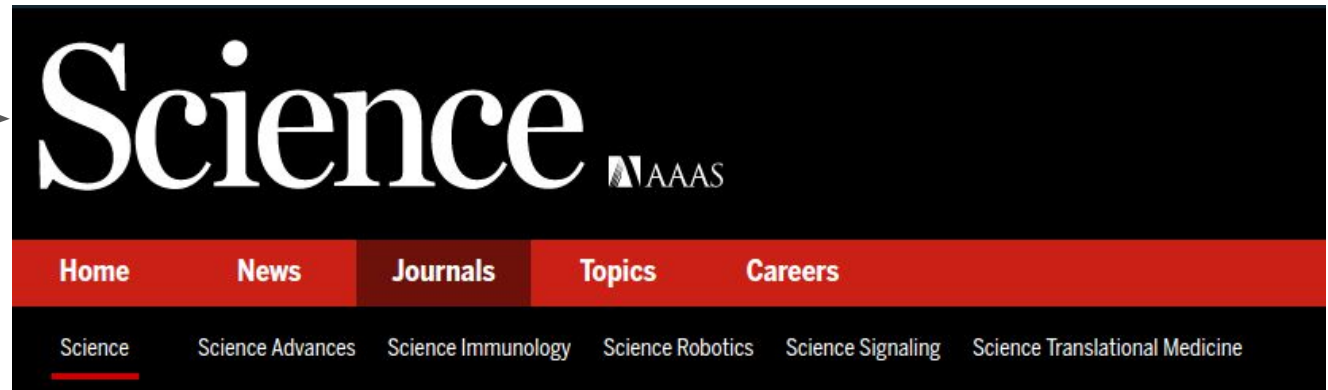
# Why do we need rigorous quality control?

GWAS goal: to test for an allelic frequency difference between cases and controls to find SNPs that affect disease/trait susceptibility

- Most variants detected to date have small effects
- Collectively account for a small fraction of the total genetic variance
- Very large sample sizes are required to identify and validate findings - a lot of room for errors
- Small sources of systematic or random error can cause spurious results or obscure real effects
- Differences in genotypes between cases and control that are not actually associated to the disease/trait in question yields false positive results
- False negative results possible eg. if noise in the dataset “obscures” a true association
  - low quality DNA samples
  - poorly-performing SNP assays
  - too strict cleaning of raw genotype data
- There's a fine line between removing true genetic variance and removing noise in the data

What we want to avoid...

Good



SHARE

REPORT



0

## Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani<sup>1,\*</sup>, Nadia Solovieff<sup>1</sup>, Annibale Puca<sup>2</sup>, Stephen W. Hartley<sup>1</sup>, Efthymia Melista<sup>3</sup>, Stacy Andersen<sup>4</sup>, Daniel A. Dworkis<sup>3</sup>, Jemma B. Wilk<sup>5</sup>, Richard H. Myers<sup>5</sup>, Martin H. Steinberg<sup>6</sup>, Monty Montano<sup>3</sup>, Clinton T. Baldwin<sup>6,7</sup>, Thomas T. Perls<sup>4,\*</sup>

+ Author Affiliations

\*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)

*Science* 01 Jul 2010;

DOI: 10.1126/science.1190532

Article

Figures & Data

Info & Metrics

eLetters

PDF

Not so good



**This article has been retracted. Please see:**  
[Is retracted by - July 22, 2011](#)



# Outline of quality control on called genotypes

Raw intensities → genotype calls



QC on called genotypes

1. Update alleles and positions
2. Exclude samples and markers with low genotyping rates
3. Gender checking
4. Hardy-Weinberg checks
5. Check heterozygosity
6. Check for cryptic relatedness
7. Check ethnicities - PCA/MDS plotting
8. Check duplicate concordance
9. Check for batch effects



**Imputation**

# 1. Update alleles and positions

PLINK: --update-alleles

- Update alleles
  - Output genotypes from scanner differs e.g. AB (Illumina std.), 12, 1234, ATCG
  - Depends on chip manufacturer, software, collaborators (who may have converted)
  - Updating to ATCG as soon as possible is usually a good idea
  - Will Rayner at Oxford University maintains conversion files for many chips
    - <http://www.well.ox.ac.uk/~wrayner/strand/>
- Update positions - ensure build and strand orientation
  - Illumina uses AB TOP/BOT by default
  - Update to FWD/+ strand (Genome Reference Consortium)
  - GRCh37/hg19 still widely used (aka. “build 37”, “b37”)
  - “Lift” data to the desired build version (eg. use resource from Will Rayner page)

# SNP strand orientation and position

- By standard convention, the locations of SNPs are based on their chromosomal position
- This position changes every time a new reference human genome assembly is released
- the latest assembly/"build" is GRCh38, and the previous one, still widely used, is GRCh37 (hg19 - "identical" assembly)
- Genes are read (transcribed) in either the forward or reverse direction
- A SNP can be represented on either of the two strands - deciding on which of the strands to use is not always easy (most use GRC defined "forward")
- If a new build comes along that flips a large segment of a chromosome, the gene direction will change.
- The way a SNP is defined in dbSNP is based on its flanking sequences (Illuminas TOP/BOT algorithm), so between builds:
  - major and minor alleles of a SNP should not change
  - position on the chromosome will change
  - whether the SNP is on the plus or minus strand may change

# Why is unknown strand orientation problematic?

Say that you get your dataset with a SNP rs1004491 (A;G). You don't know if this is on the forward or the reverse strand

- You go to dbSNP and find this SNP to be defined as C;T. [http://www.ncbi.nlm.nih.gov/SNP/snp\\_ref.cgi?searchType=adhoc\\_search&type=rs&rs=rs1004491](http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?searchType=adhoc_search&type=rs&rs=rs1004491)
- Since A→T and G→C you decide to flip this SNP to C;T.
- **This is reasonable and is done frequently**

The next dataset you get you find a SNP rs9999 (A;T). You still don't know the strand orientation

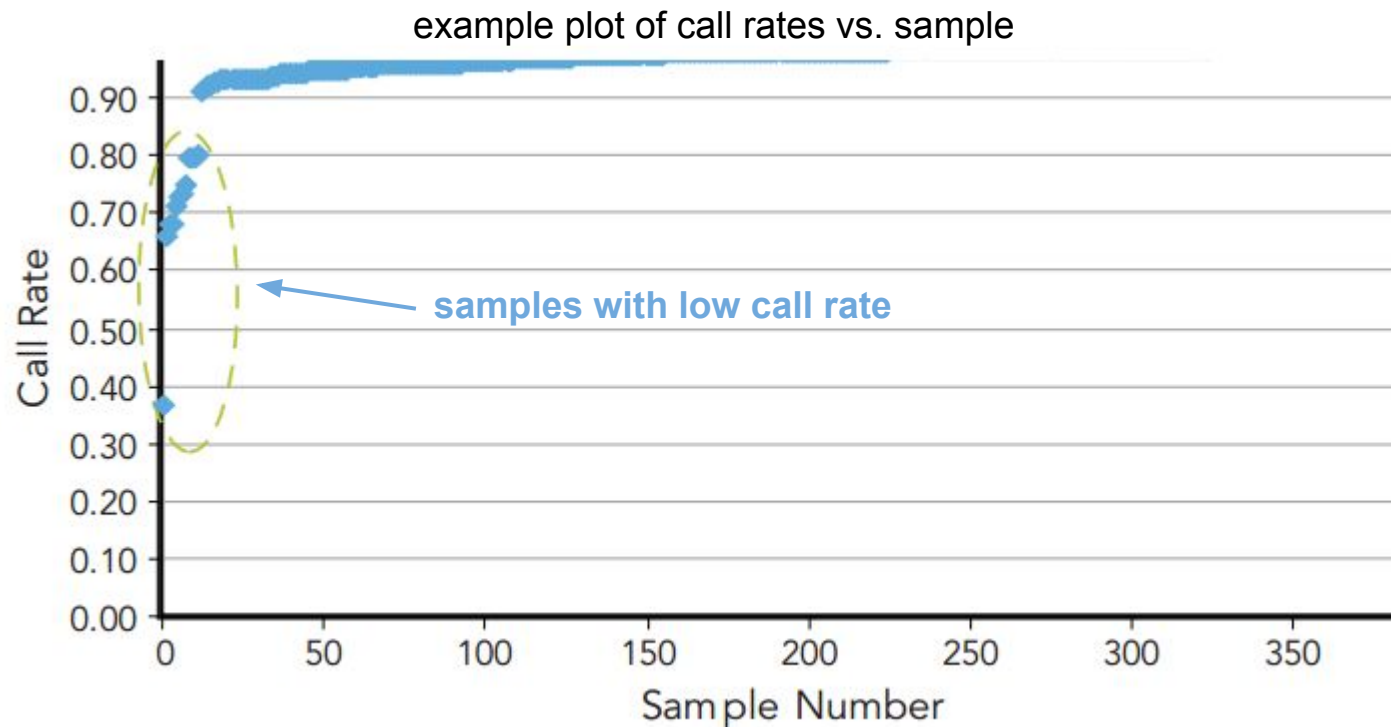
- You go to dbSNP to find that this is an A;T SNP.
- The distribution in your dataset is currently (24% AA, 50 % AT, 26% TT)
- **Flipping this SNP would reverse the distribution**

**NOTE: A/T and G/C are ambiguous and a common source of strand problems**

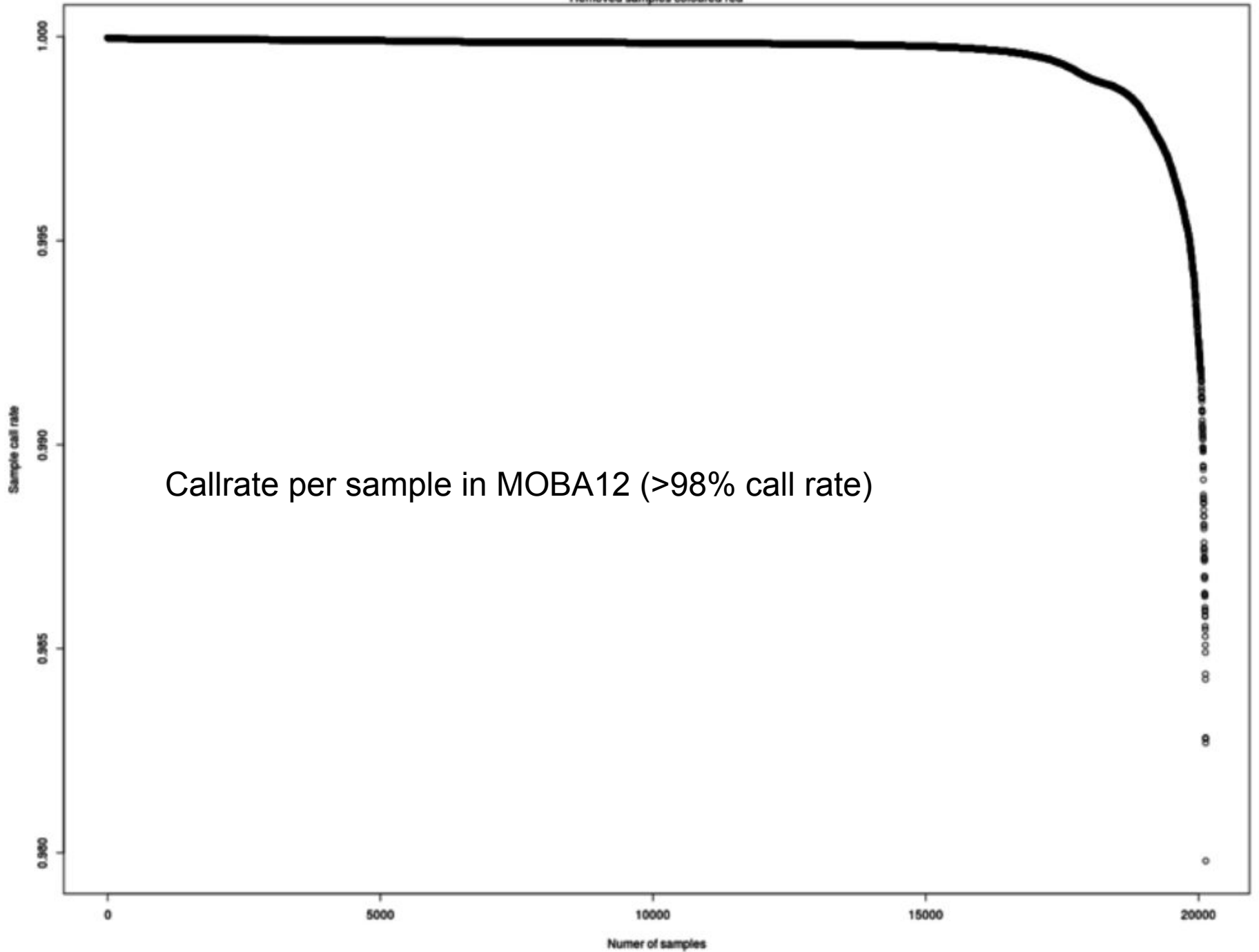
## 2. Remove bad markers and bad samples

PLINK: --mind, --geno

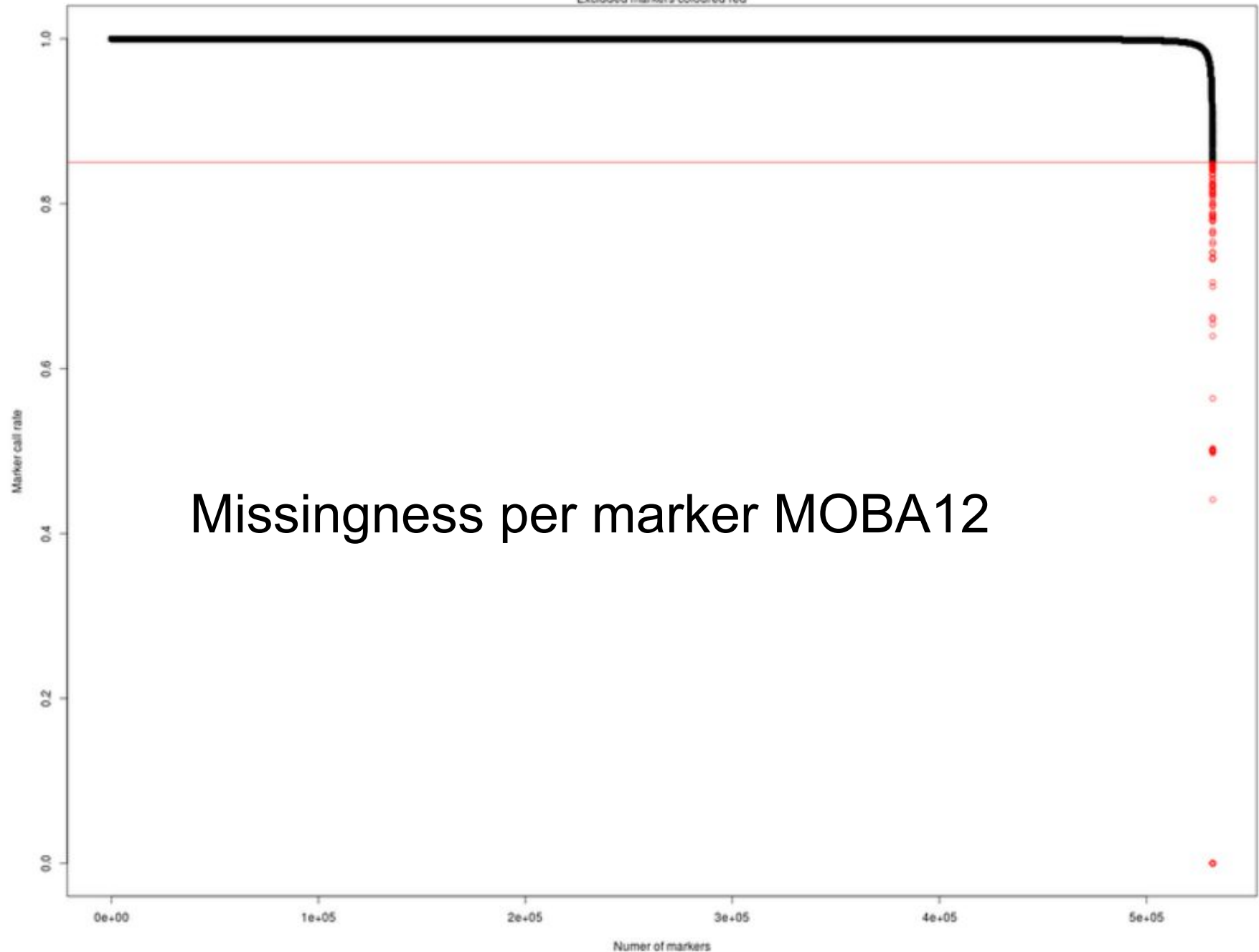
- Remove markers with high missingness rates (low call rate)
  - High missingness → poor marker → difficult to trust the remaining calls for this SNP
- Remove samples with too high missingness (low call rate)
  - High missingness → poor sample → difficult to trust the remaining calls for this sample



Missigness per sample  
Removed samples coloured red



Missigness per marker  
Excluded markers coloured red



Missingness per marker MOBA12

# Remove bad markers and samples

Filters for exclusion on missingness are usually applied in passes to recalculate statistics between removals/passes. This avoids the worst samples to contribute to marker missingness and vice versa.

## Proposed thresholds are:

- **1. pass: --mind 0.15 --geno 0.15**
  - removes all individuals and markers with more than 15% missingness
- **2. pass: --mind 0.05 --geno 0.05**
  - removes all individuals and markers with more than 5% missingness
- **3. pass: --mind 0.02 --geno 0.02**
  - removes all individuals and markers with more than 2% missingness



# 3. Gender checking

PLINK: --check-sex



- Gender misidentification is not uncommon
  - Wrong info reported in questionnaires
  - Mistyping at the biobank
  - Different classification in different data sources:
    - Male=0, Female=1
    - Male=1, Female=0
    - Male=1, Female=2 → PLINK standard (with '0' = unknown)
    - M/F=male/female vs. M/F=mother/father
- Chromosome sex differing from phenotypic sex
  - Males born 46XX (translocation of a tiny section of the sex determining region on Y)
  - Females born 46XY (mutations in the Y chromosome)
- Chromosomal abnormalities
  - Turner syndrome (phenotypic females, missing/partially missing one X-chr. → 45X)
  - Klinefelter syndrome (phenotypic males, extra X-chromosome(s) → 47XXY)
  - Triple-X (phenotypic females, extra X-chromosome(s))
  - XYY (phenotypic males, extra Y-chromosome)

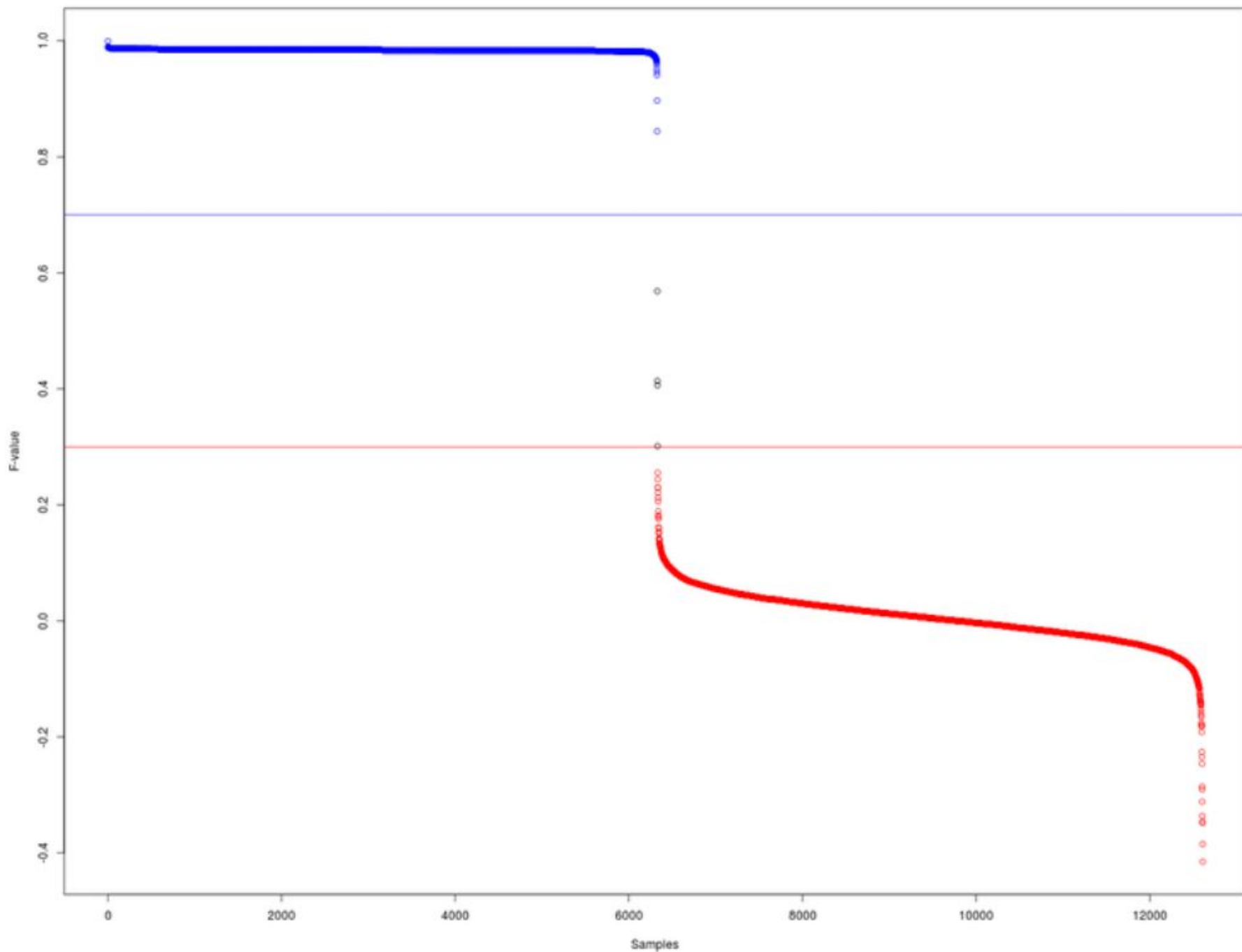
# Gender checking

PLINK : `--check-sex, split-x`

- By default PLINK compares sex assignments in the input dataset with those imputed from X chromosome inbreeding coefficients
- By default only checks homozygosity of the X-markers (also possible to use Y)
- Make sure to split of the Pseudoautosomal region (PAR) (`--split-x`)
  - homologous sequences of nucleotides on the X and Y chromosomes
- The inbreeding coefficient (F) can be interpreted as the fraction of all the X markers that are homozygous
  - < 0.2 yields female calls by default
  - > 0.8 yields male calls by default

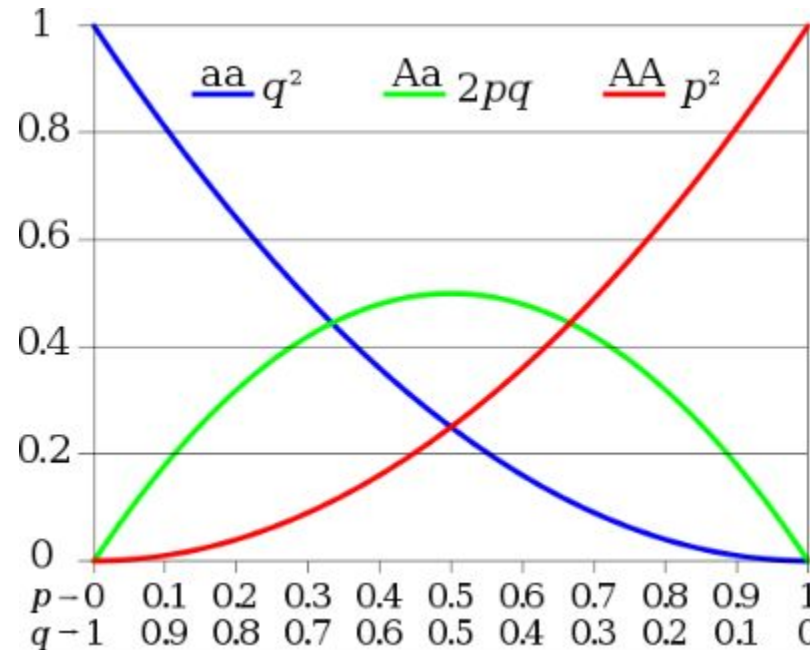
FID	IID	PEDSEX	SNPSEX	STATUS	F
2427	NA19919	1	1	OK	1
2431	NA19916	1	1	OK	1
2424	NA19835	2	2	OK	-0.008176
2469	NA20282	2	2	OK	0.06641
2368	NA19703	1	1	OK	1
2425	NA19902	2	2	OK	0.06505
2425	NA19901	2	2	OK	0.01425
2427	NA19908	1	1	OK	1
2430	NA19914	2	2	OK	0.003262
1349	NA10854	2	1	PROBLEM	0.9968

X chromosome homozygosity estimate (F value)



# Hardy-Weinberg law

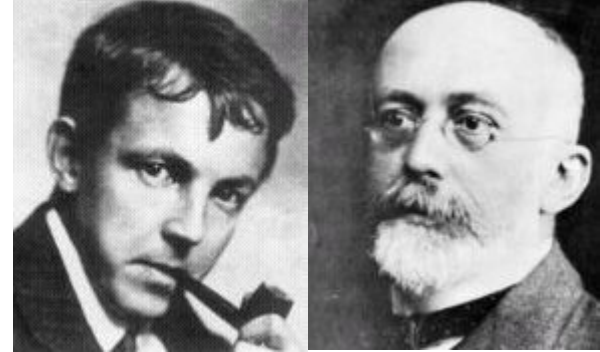
- Hardy-Weinberg law: states that allele and genotype frequencies in a population will remain constant from generation to generation in the *absence of other evolutionary influences*



Hardy-Weinberg proportions for two alleles: the horizontal axis shows the two allele frequencies  $p$  and  $q$  and the vertical axis shows the expected genotype frequencies. Each line shows one of the three possible genotypes.

## 4. Hardy-Weinberg exclusion

PLINK: --hardy, --hwe

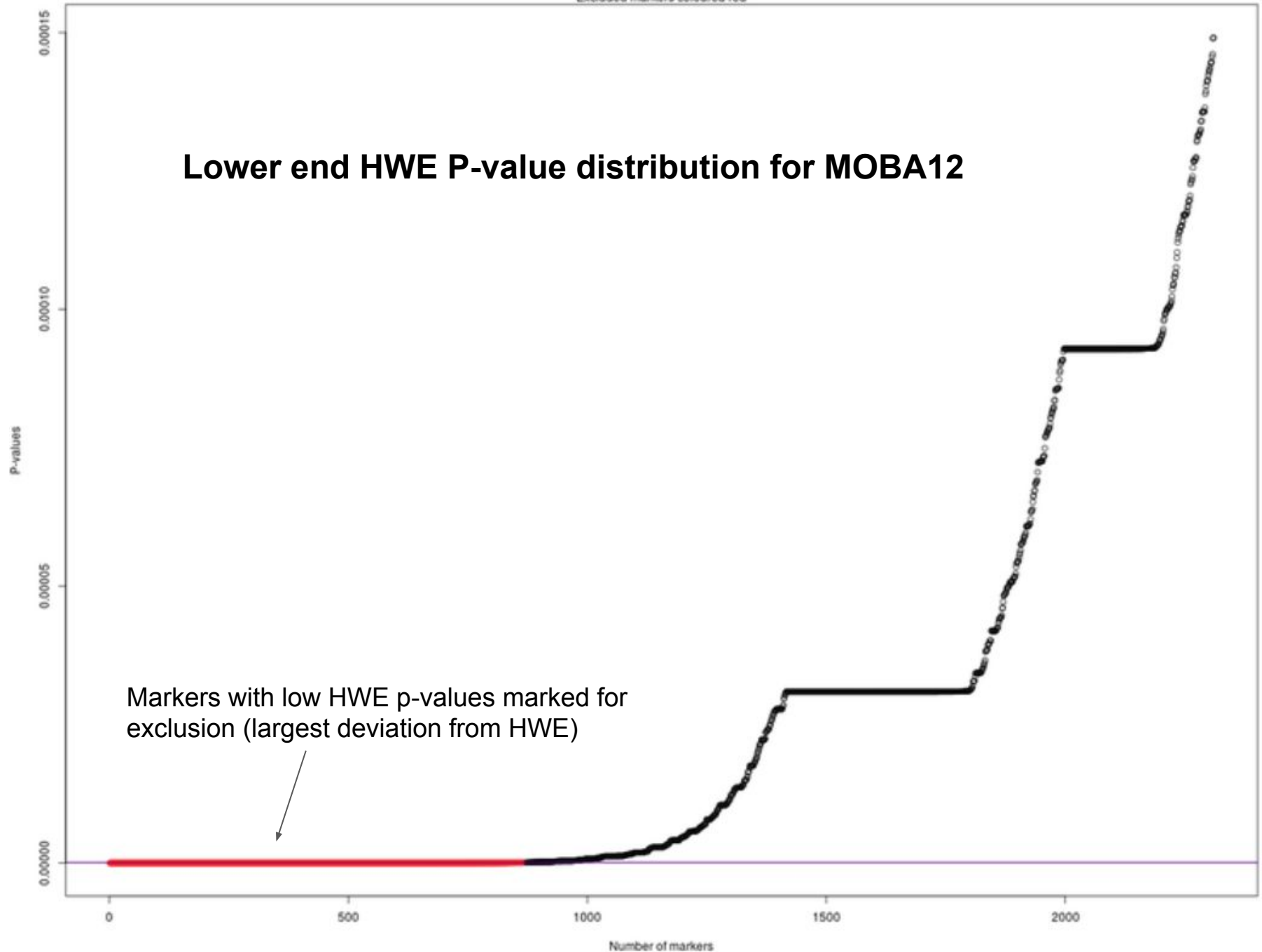


Godfrey Hardy and Wilhelm Weinberg

- PLINK uses an exact test to compare the observed vs. expected number of heterozygous markers for a SNP
- The PLINK documentation recommends a very low P-value threshold for catching genotyping errors primarily
- Recommended values in various pipelines  $10E-4$  to  $10E-6$
- Genuine SNP-trait associations are expected to deviate slightly from Hardy-Weinberg equilibrium
- Beware: HWE in datasets with mixed ethnicities must be used with caution
  - Stratify your data eg. using Structure/fastStructure
- PLINK: --hardy is for generating statistics and --hwe is for excluding markers below HWE P-val. threshold

# HWE test statistics per marker

Excluded markers coloured red

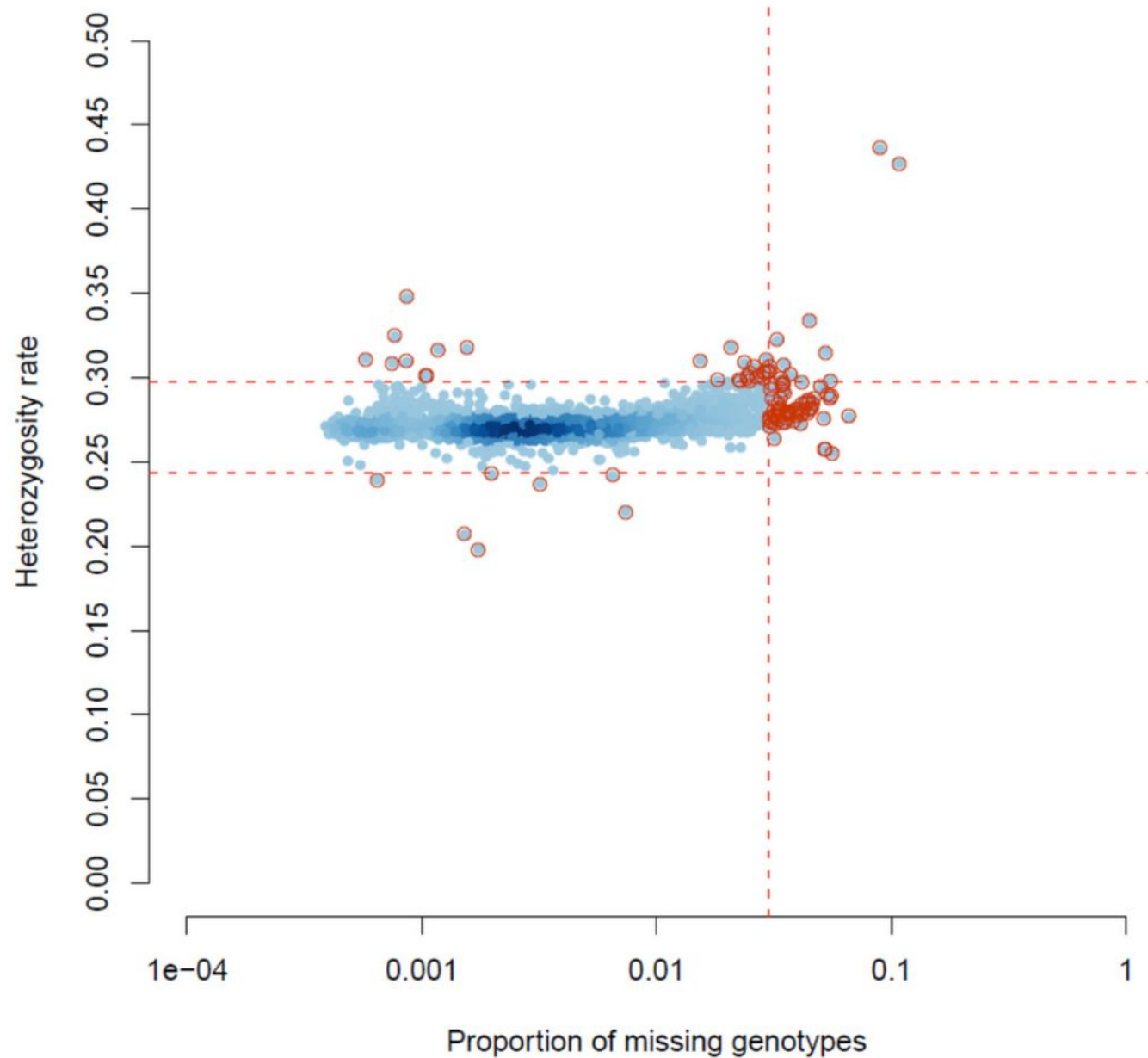


## 5. Heterozygosity checking

PLINK: --het

- Variation in DNA quality can have large effects on genotype call rate and accuracy
- Low quality or low concentration DNA samples usually have below average call rate and genotype accuracy
- Excess heterozygosity (excluding X-chromosomes) may be indicative of a bad or contaminated sample
- Contamination and/or concentration problems tend to deviate the signal towards the heterozygous cluster
- Low heterozygosity might be suggestive of inbreeding
- Note: mean heterozygosity will differ between populations and SNP genotyping panels
- Heterozygosity should be checked in rare and common markers separately (eg. split by MAF 1%)
- Check in each ethnicity separately if large groups (eg. use Structure - later)

# Heterozygosity vs. missingness





## 6. Cryptic relatedness

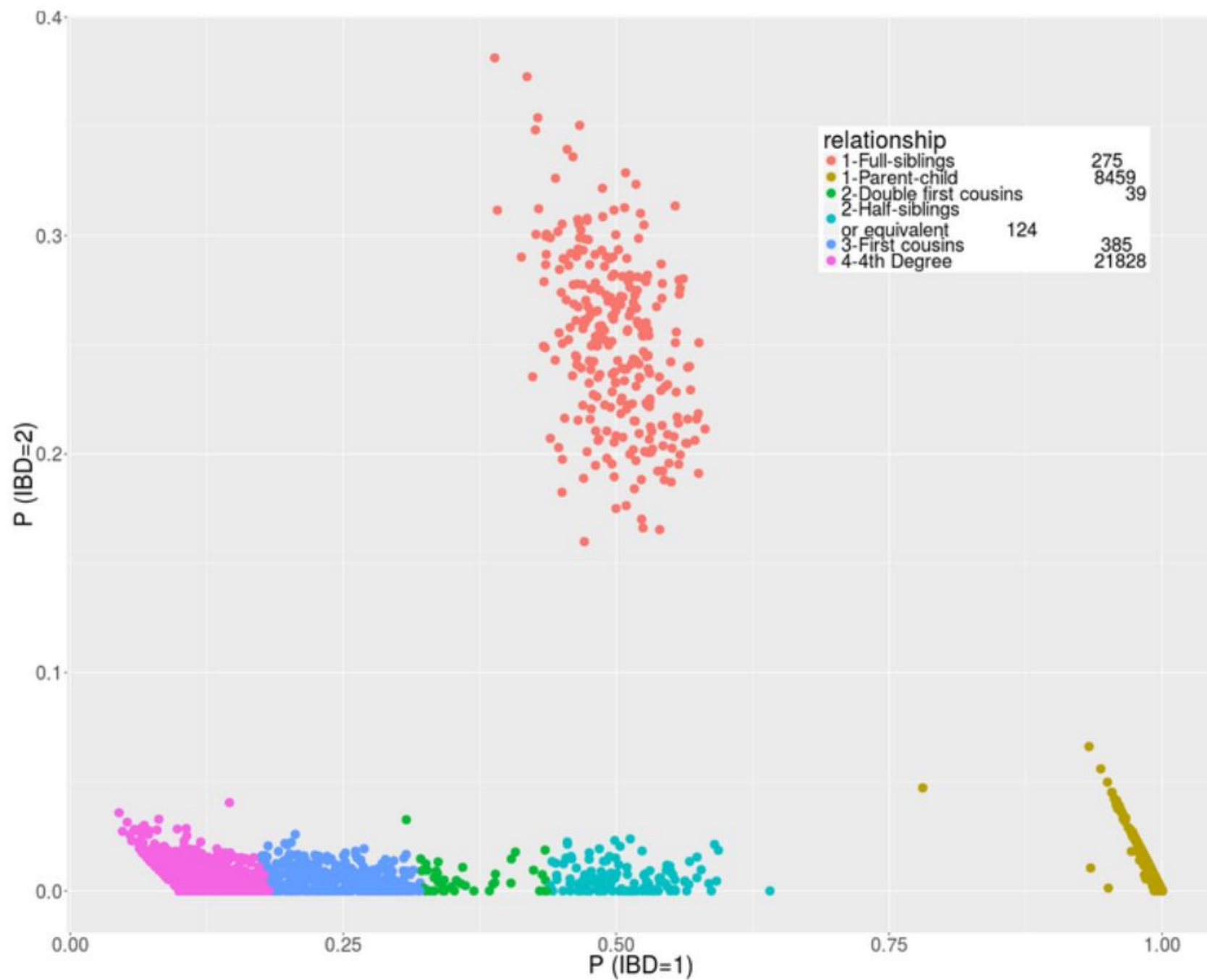
PLINK: --genome

- A basic feature of standard population-based case-control association studies is that all the samples are unrelated
- If duplicates, first- or second- degree relatives are present, a bias may be introduced to the study because the genotypes within families will be over-represented, and thus the sample may no longer be a fair reflection of the allele frequencies in the entire population.
- Important to uncover any hidden relatedness in the dataset
- To measure relatedness PLINK calculates an Identical By State (IBS) metric for all pairs of samples in the dataset.
- Estimates how many alleles are shared between two samples (for all pairs)
- Z0, Z1 and Z2 specifies the sharing of none, one or both alleles.
- The PIHAT denotes the proportion of IBD, i.e.  $P(\text{IBD}=2) + 0.5 * P(\text{IBD}=1)$
- These calculations are not LD-sensitive → Important to LD prune dataset before running --genome

# Cryptic relatedness

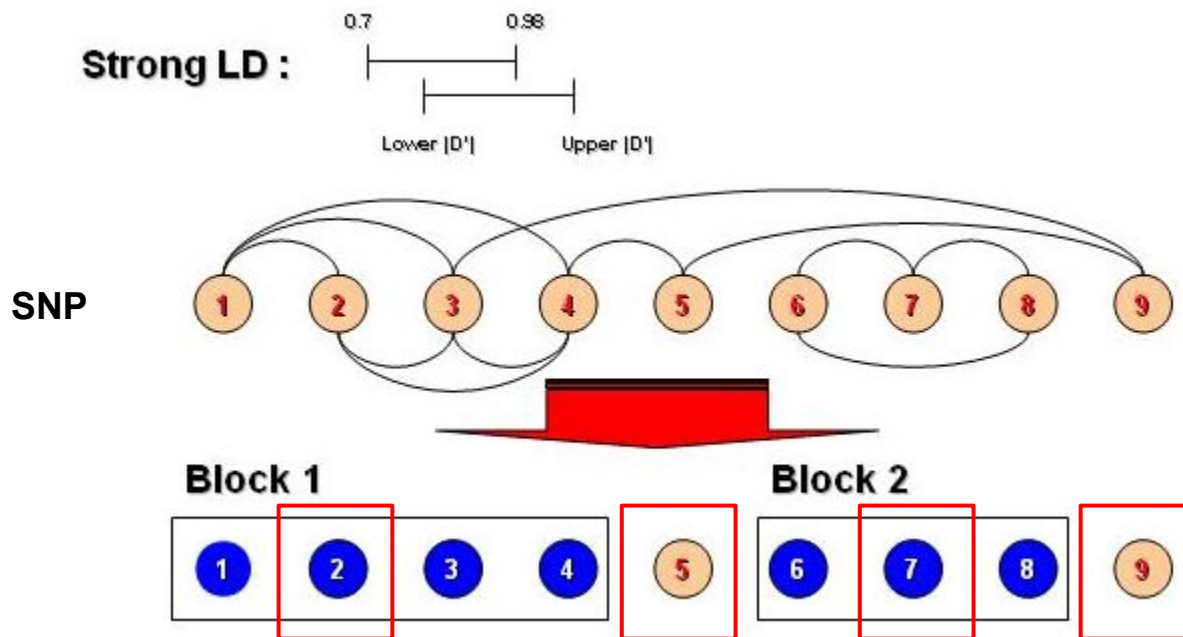
Proposed thresholds for inferring relation between samples

<b>Z0</b>	<b>Z1</b>	<b>Z2</b>	<b>Pihat</b>	<b>Relationship</b>
		<b>&gt;0.9</b>	<b>&gt;0.9</b>	<b>Monozygotic twin, duplicate</b>
	<b>&gt;0.95</b>		<b>0.48-0.53</b>	<b>Parent-child</b>
<b>0.25</b>	<b>0.40-0.60</b>	<b>~0.25</b>	<b>0.40-0.60</b>	<b>Sibling</b>
<b>0.7-0.8</b>		<b>0.2-0.3</b>		<b>1<sup>st</sup> degree cousin</b>
			<b>0.23-0.27</b>	<b>2<sup>nd</sup> degree relation (uncle-nephew, half-siblings, grandparent-grandchild)</b>



# Linkage Disequilibrium pruning (LD-pruning)

- Linkage disequilibrium (LD) is the non-random association of alleles at different loci
- Presence of statistical associations between alleles at different loci that are different from what would be expected if alleles were independently, randomly sampled based on their individual allele frequencies.
- If there is no linkage disequilibrium between alleles at different loci they are said to be in linkage equilibrium.
- LD pruning prunes/reduces the dataset by removing markers with high LD, thus you are left with a subset of markers that are in approximate linkage equilibrium with each other
- In PLINK based on correlations between genotype allele counts (phase is not considered)



Johnson

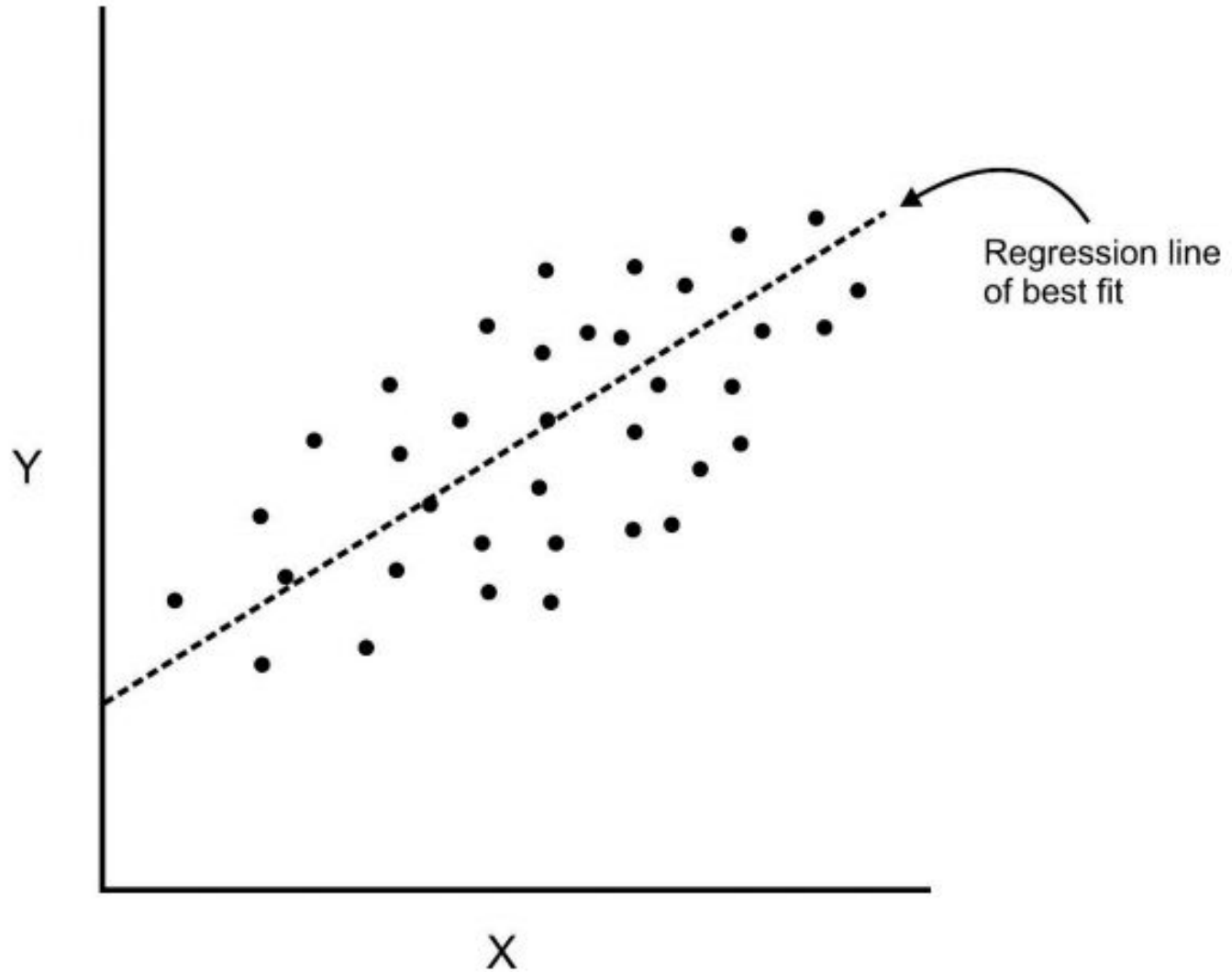
Workload

Smith

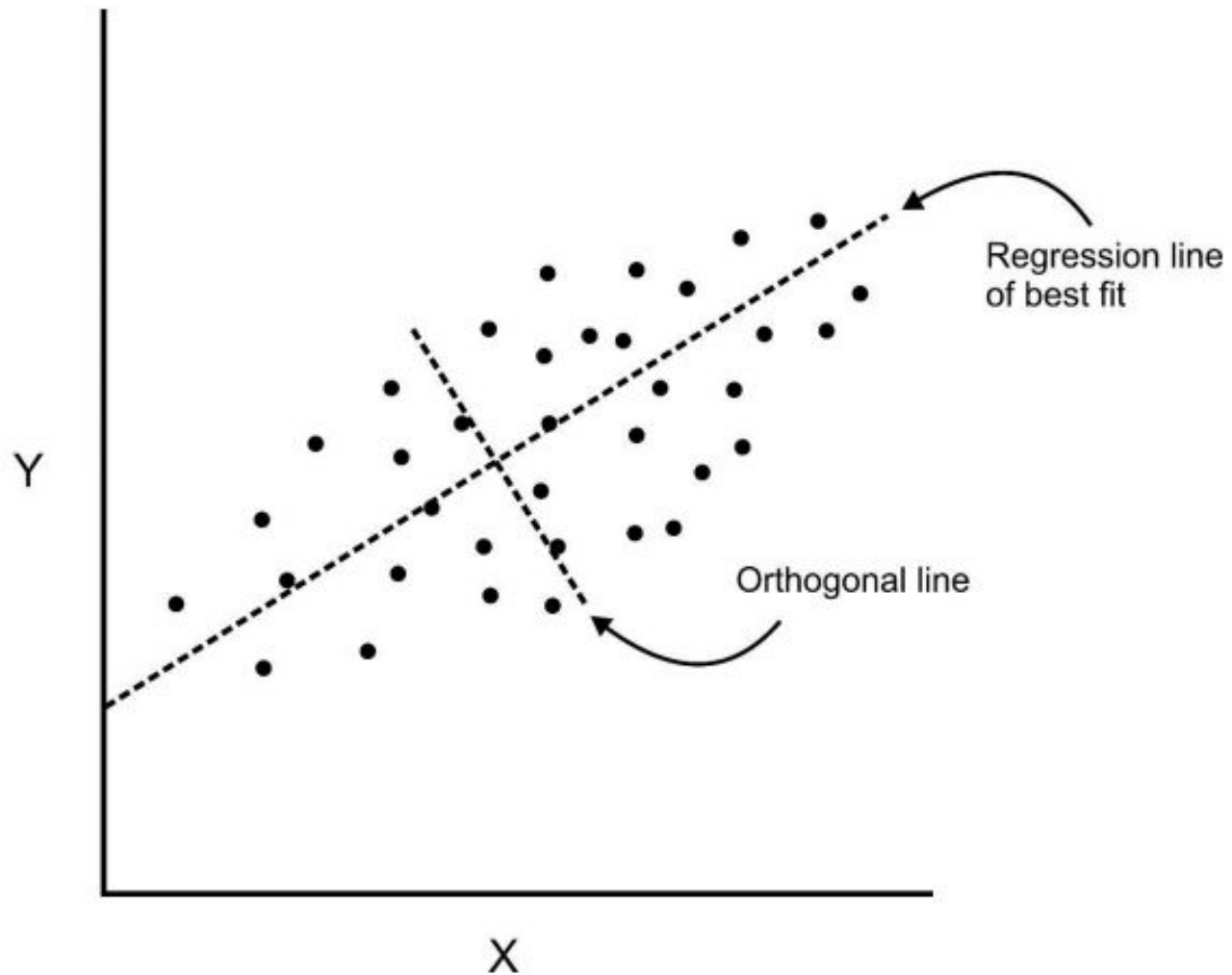
Salary

	Workload	Distance to Work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	1.0
Williams	3.0	0.0	0.8

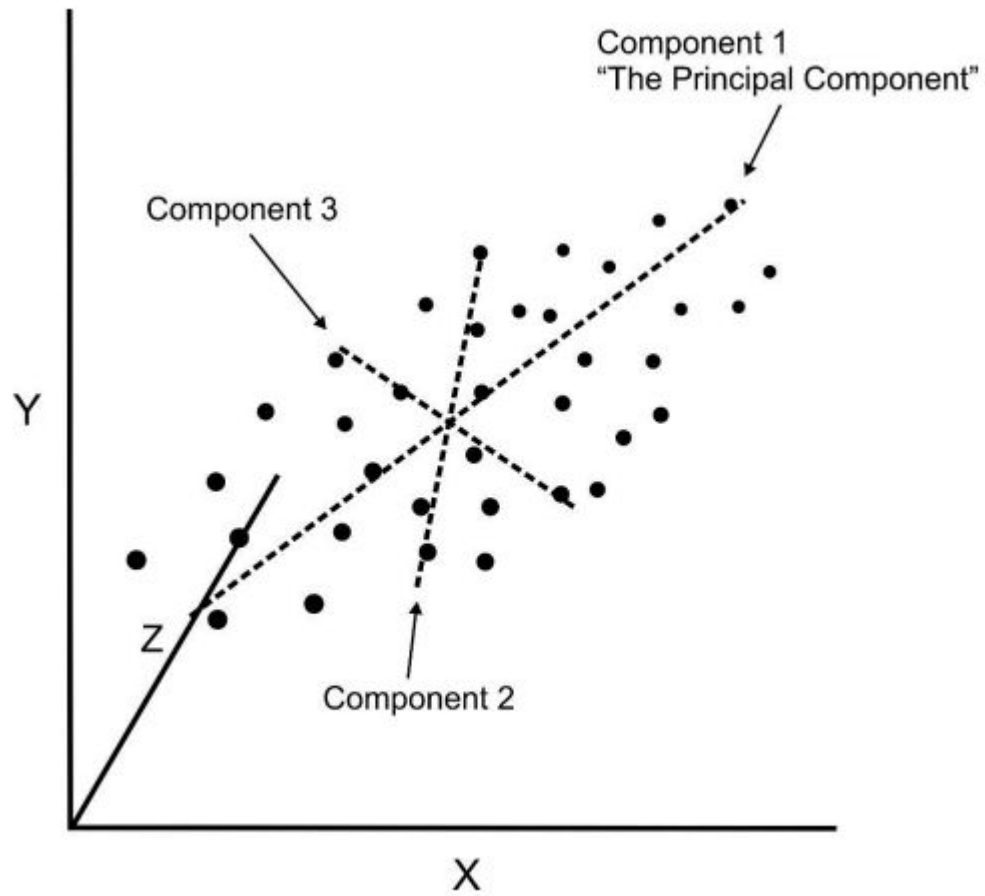
# PCA



# PCA



# PCA



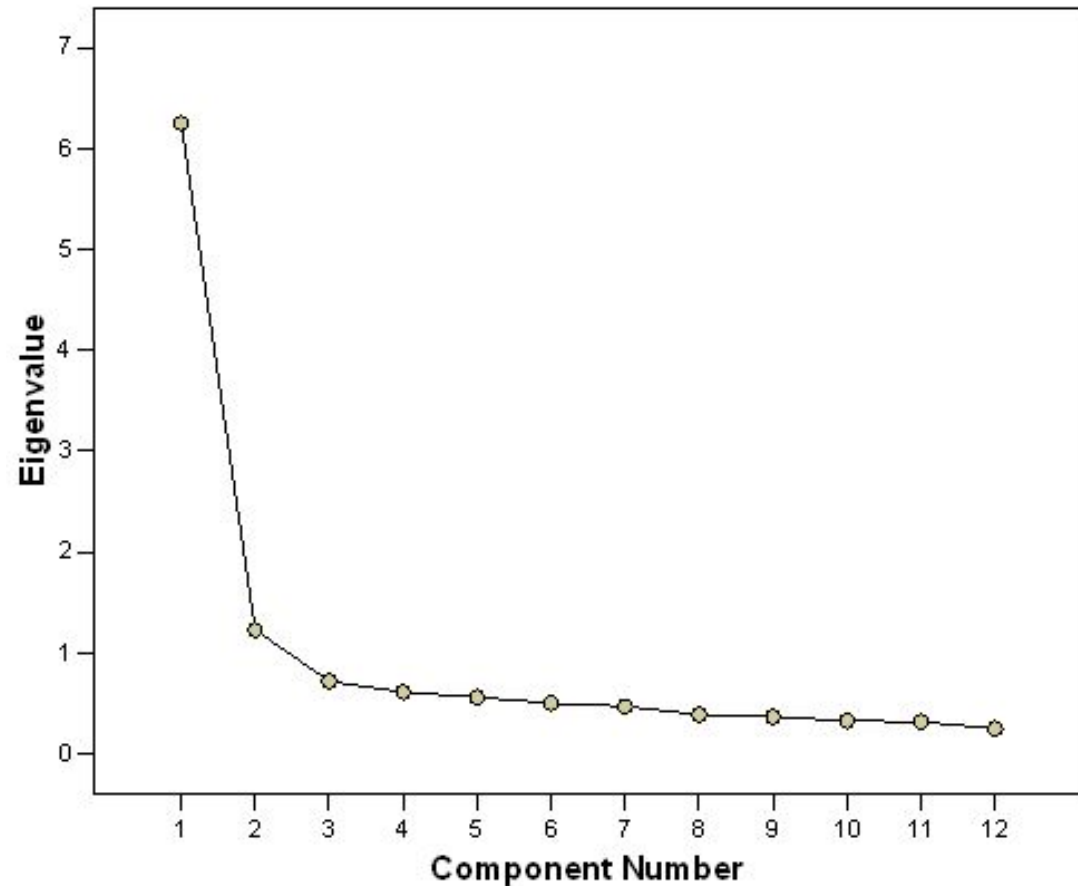


# Principal Component Analysis

- A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Works by identifying the dimensions (principal components) that explain the largest variance in the dataset.
- Reduces the complexity of the dataset by defining new variables that “capture” most of the variation in the dataset
- Starts by finding the dimension capturing the most variation, then finds the dimension orthogonal (meaning “at right angles” -actually the lines are perpendicular to each other in n-dimensional space) to the first dimension capturing the most of the remaining variation for all n-dimensions (n variables).
- n-Dimensional Space: the variable sample space. There are as many dimensions as there are variables (here SNPs), so in a data set with 10 variables (SNPs) the sample space is 10-dimensional.

# Scree plot - how many PCAs to use?

Scree Plot



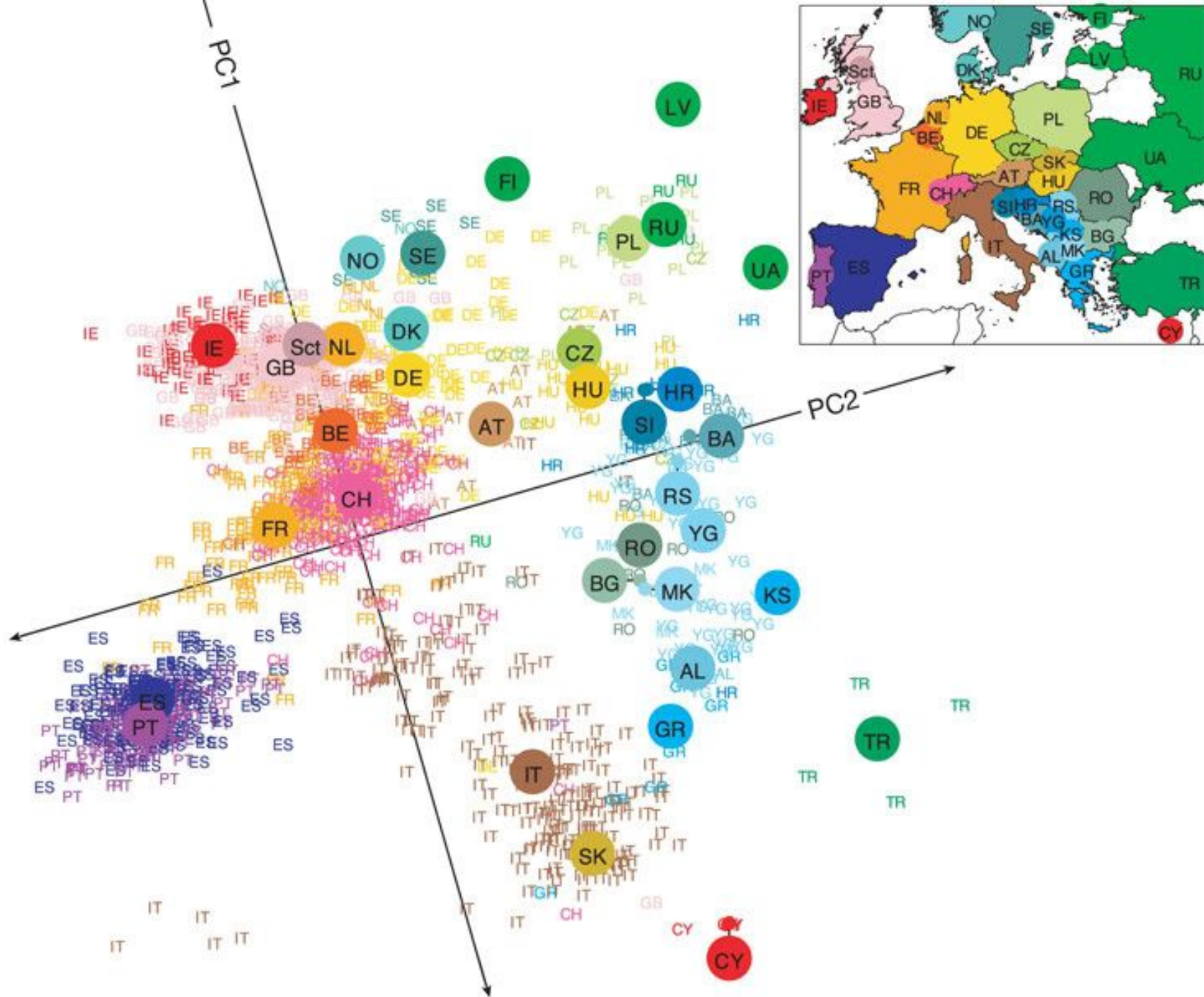
scree



# Principal Component Analysis

The assumptions of PCA:

1. Linearity: Assumes the data set to be linear combinations of the variables.
2. The importance of mean and covariance: There is no guarantee that the directions of maximum variance will contain good features for discrimination.
3. That large variances have important dynamics: Assumes that components with larger variance correspond to interesting dynamics and lower ones correspond to noise.



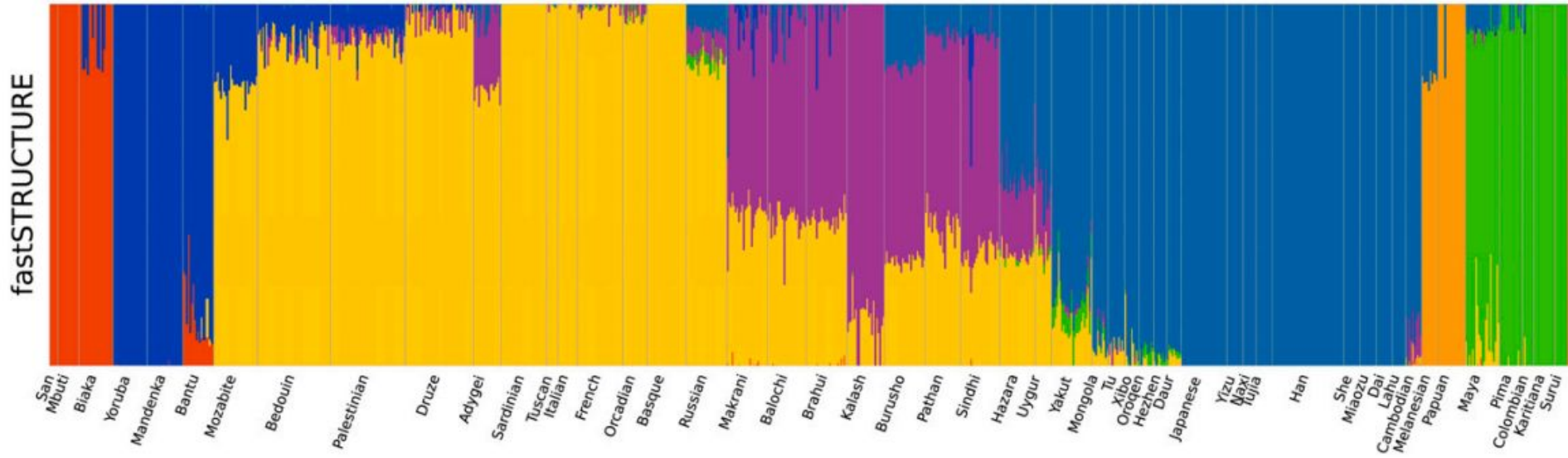
# EIGENSOFT/EIGENSTRAT

- One of many software solutions for Principal Component Analysis for GWAS
- Maintained by Nick Patterson and Alkes Price (++)
- Implements methods from
  - Population Structure and Eigenanalysis <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190>
  - Principal components analysis corrects for stratification in genome-wide association studies <http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html>
  - Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia [http://www.cell.com/ajhg/abstract/S0002-9297\(16\)00003-3](http://www.cell.com/ajhg/abstract/S0002-9297(16)00003-3)
- Package Eigensoft contains
  - Eigenstrat - Stratification of populations using PCA
  - Convertf - conversion tool from PLINK format to Eigenstrat format
- For the latest version visit Alkes Price's Harvard homepage:
  - <http://www.hsph.harvard.edu/alkes-price/software/>
  - latest version as of writing: 6.1.1 (May 2016) - supports multithreading!

# STRUCTURE/fastSTRUCTURE

- Popular software used to infer population structure in GWAS data.
- Uses include...
  - assigning individuals to populations (requires merging with known refs. eg. HapMap)
  - studying hybrid zones
  - identifying migrants and admixed individuals
  - estimating population allele frequencies in situations where many individuals are migrants or admixed
- Under the hood...
  - generates clusters based on Hardy–Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) caused by admixture between populations.
  - Works by clustering individuals in groups, where both linkage and HWD are minimized, and therefore, the presence of LD in the data improves clustering results
  - Requires an a priori knowledge about the number of populations
- fastStructure
  - Alternative implementation of Structure
  - Faster at the expense of accuracy (still adequate for most)
  - Includes a tool for choosing model complexity (“guessing” the number of populations in the dataset)

# fastStructure - identifying ethnic groups



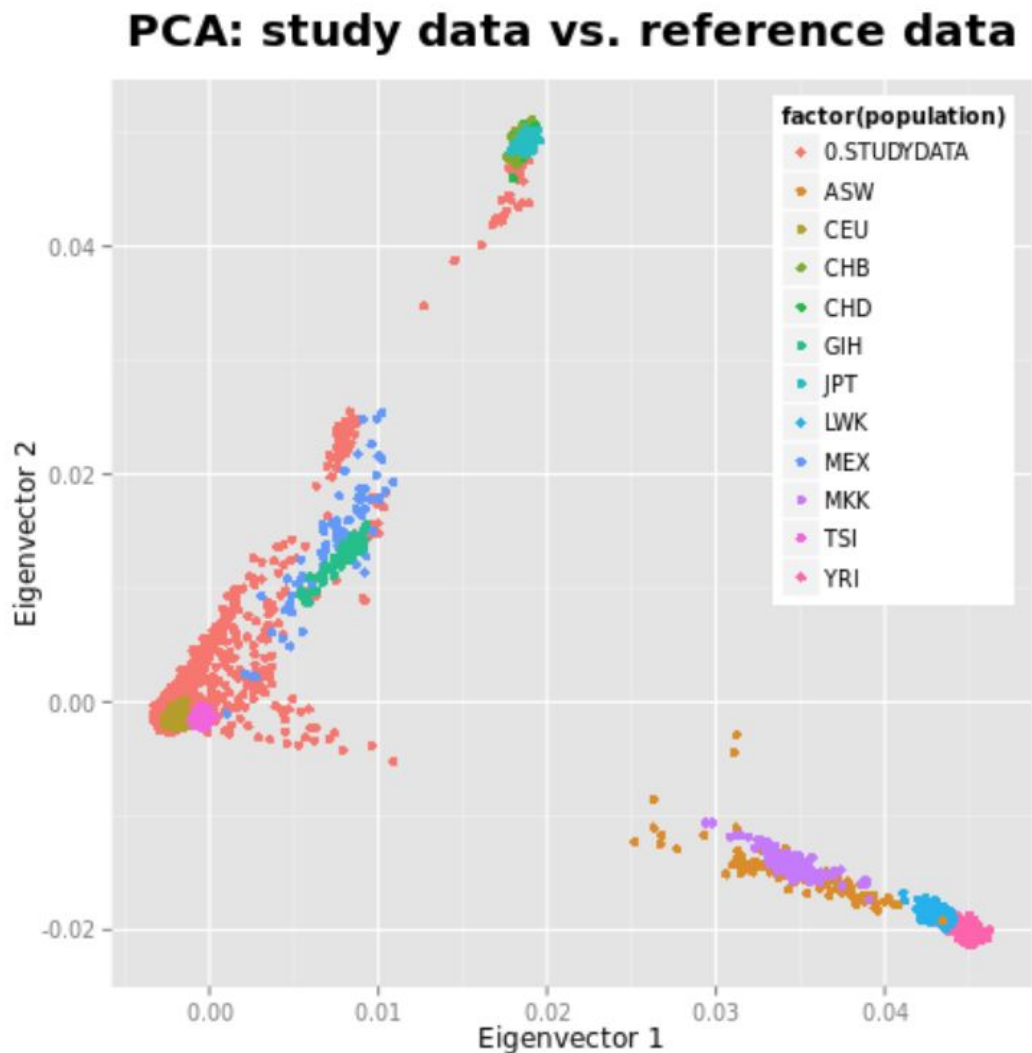
## 7. Check ethnicities - PCA/MDS plotting

PLINK: --genome, --mds

- The most common method for identifying (and subsequently removing) individuals with large-scale differences in ancestry is **principal components analysis** (PCA)
- An alternative to PCA is the **multidimensional scaling** (implemented in PLINK), requires a pair-wise IBD matrix to be constructed and is therefore more computationally complex (ie. running --genome)
- PCA is a multivariate statistical method used to produce a number of uncorrelated variables (or principal components) from a data matrix containing observations across a number of potentially correlated variables.
- The principal components are calculated so that the first principal component accounts for as much variation in the data as possible in a single component, followed then by the second component and so on.
- When using PCA to detect ancestry the **observations are the individuals** and the **potentially correlated variables are the markers**.
- A principal component model is built using **pruned genome-wide genotype data from populations of known ancestry** (eg. HapMap)



# PCA plot - MOBA merged with HapMap3



## **ASW**

African ancestry in Southwest USA

## **CEU**

Utah residents with Northern and Western European ancestry from the CEPH collection

## **CHB**

Han Chinese in Beijing, China

## **CHD**

Chinese in Metropolitan Denver, Colorado

## **GIH**

Gujarati Indians in Houston, Texas

## **JPT**

Japanese in Tokyo, Japan

## **LWK**

Luhya in Webuye, Kenya

## **MXL**

Mexican ancestry in Los Angeles, California

## **MKK**

Maasai in Kinyawa, Kenya

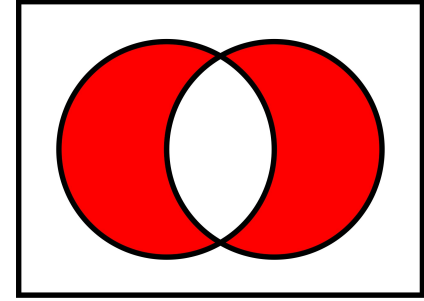
## **TSI**

Toscani in Italia

## **YRI**

Yoruba in Ibadan, Nigeria

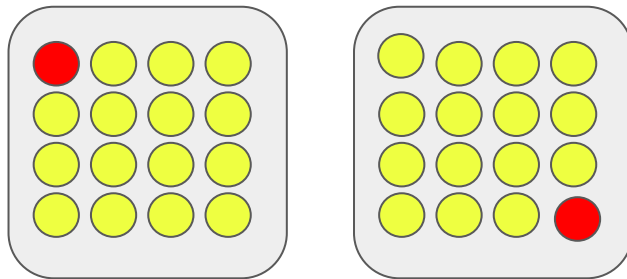
## 8. Duplicate concordance



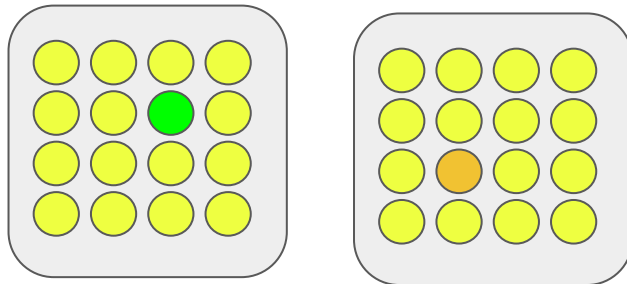
- Including duplicate samples is way of internal and external quality control of the genotyping accuracy post-QC
- Genotyping the same samples several times provides the possibility to check the concordance between the genotype calls
- HapMap samples often genotyped together with the rest of the samples in the study in large datasets as a “gold standard” since these have been thoroughly sequenced/genotyped
- A high rate of discordance post-QC should invoke further investigation
  - Too lenient QC parameters?
  - Bad calls in one of the two duplicates? Do you have triplicates to check? How many duplicates are available?
  - Is the discordance present in only a specific group of markers?
    - Only rare markers?
    - Only X-chr markers?

# Duplicate samples - configurations

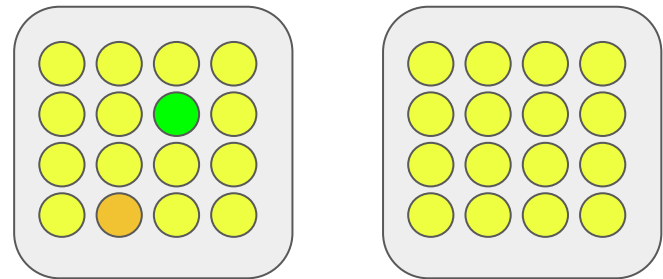
Batch 1



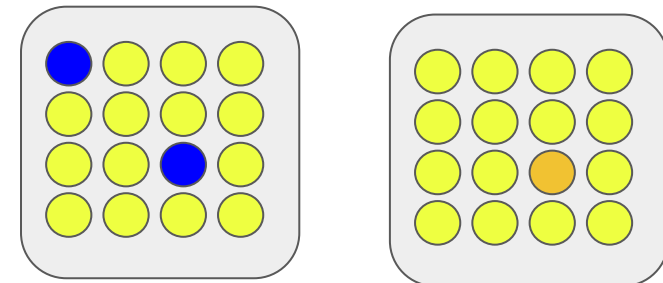
well plates







Batch 2



well plates



-  Duplicates across batch
-  Same plate
-  Same batch - different plates
-  Triplicates - across batch

# Duplicate concordance - examples from HARVEST

## moba12-b37-reclustered

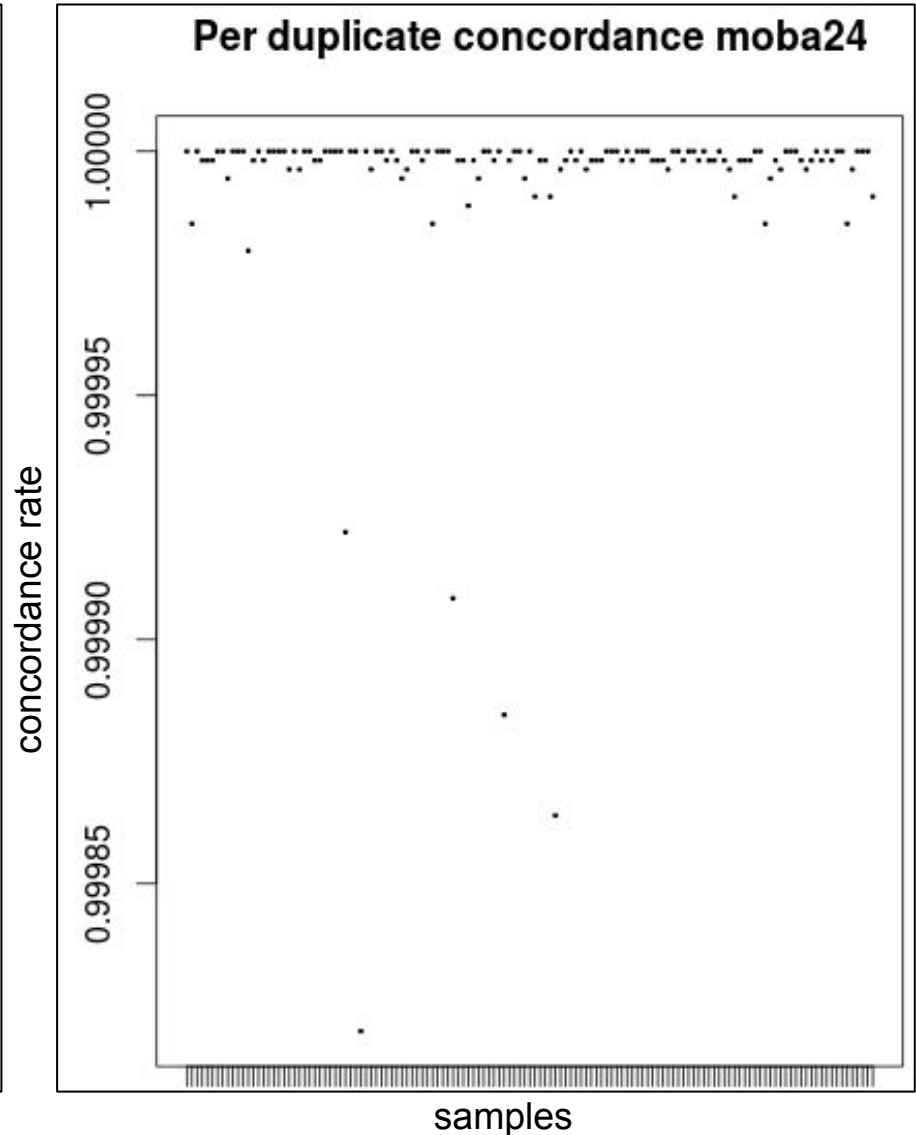
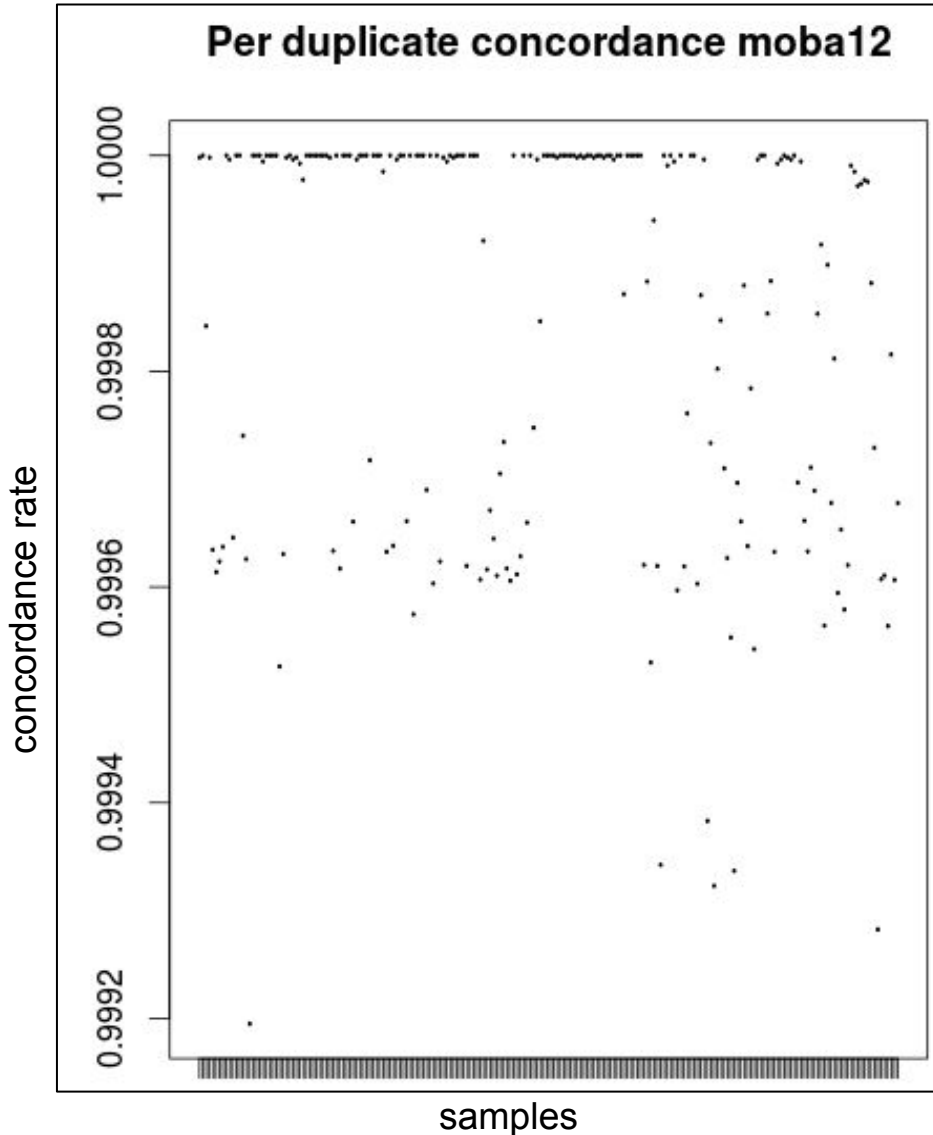
Data	Cat.	N	Halved	Markers	Correct	Miscall	One_NC	Both_NC	Conc
gencall-clean-atcg	All	210	0	532963	531422.8	78.77142857	728.7571429	732.6714286	0.9998517946
gencall-clean-atcg	~Rare	210	0	276453	275656.2571	38.85238095	373.6285714	384.2619048	0.9998590748
gencall-clean-atcg	~Common	210	0	256510	255766.5429	39.91904762	355.1285714	348.4095238	0.9998439482

## moba24-b37-reclustered

Data	Cat.	N	Halved	Markers	Correct	Miscall	One_NC	Both_NC	Conc
gencall-clean-atcg	All	135	0	540325	539242.3333	3.474074074	287.0962963	792.0962963	0.9999935575
gencall-clean-atcg	~Rare	135	0	277566	277023.0741	1.896296296	144.9333333	396.0962963	0.9999931548
gencall-clean-atcg	~Common	135	0	262759	262219.2593	1.577777778	142.162963	396	0.999993983

# Duplicate concordance - examples from HARVEST

Post-QC concordance rates



# Mendelian errors

PLINK: `--mendel`, `--me`, `--set-me-missing`

- A Mendelian error in the genetic analysis of a species, describes an allele in an individual which could not have been received from either of its biological parents by Mendelian inheritance.
  - Example:
    - Mother's genotype is AA
    - Father's genotype is AA
    - Offspring genotype is AB → mendelian error
- Requires trio data (duo data can also be used to some extent)
- Can be used in QC to check genotype quality
- PLINK `--mendel` outputs mendelian error rates
  - per individual in `.imendel`
  - per family in `.fmendel`
  - per marker `.lmendel`
  - all errors defined by codes in `.mendel`

# Mendelian errors

PLINK: --mendel

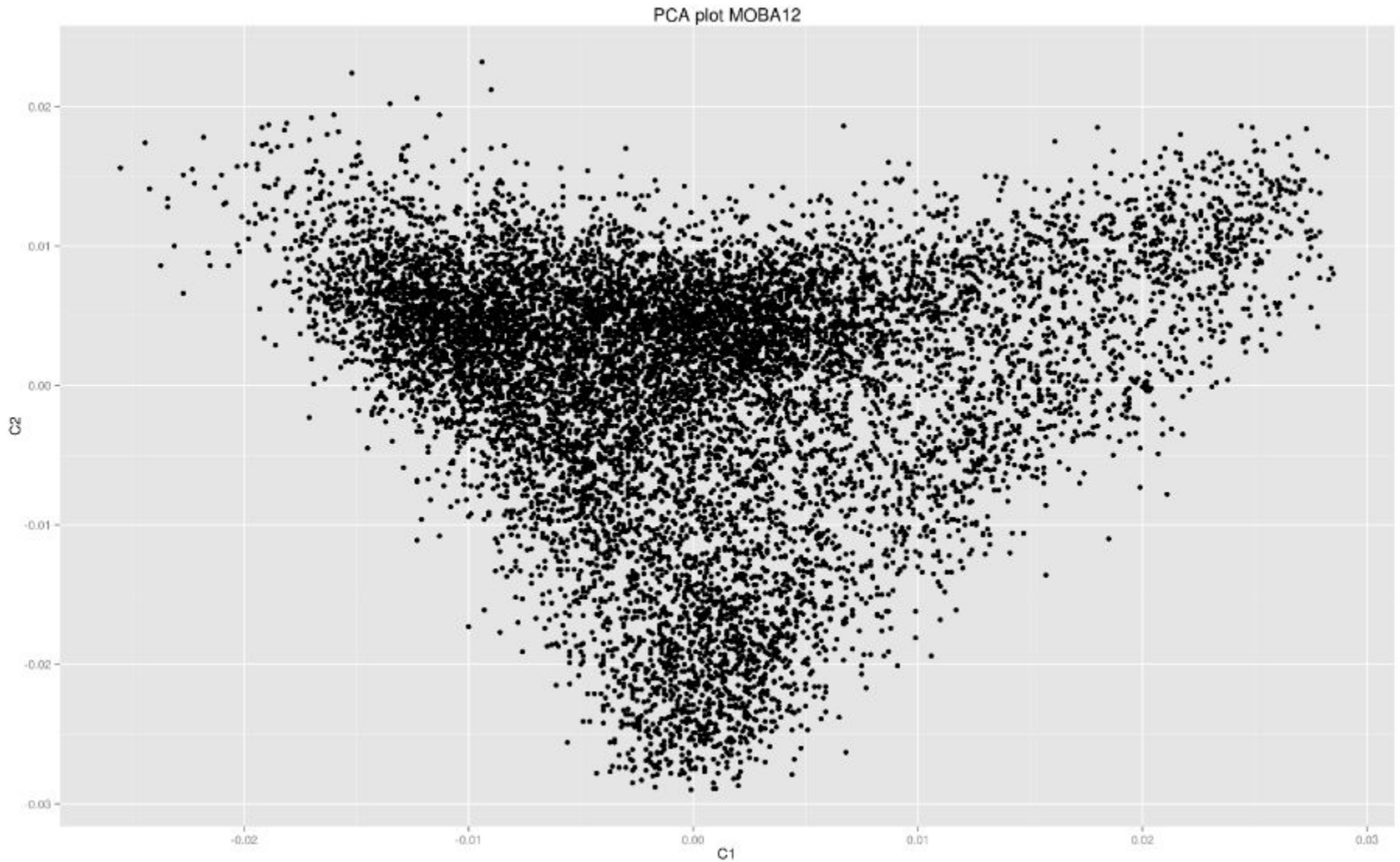
PLINK error codes implicates what sample is likely to be the “offending” sample in producing the mendelian error

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

## 9. Batch effects

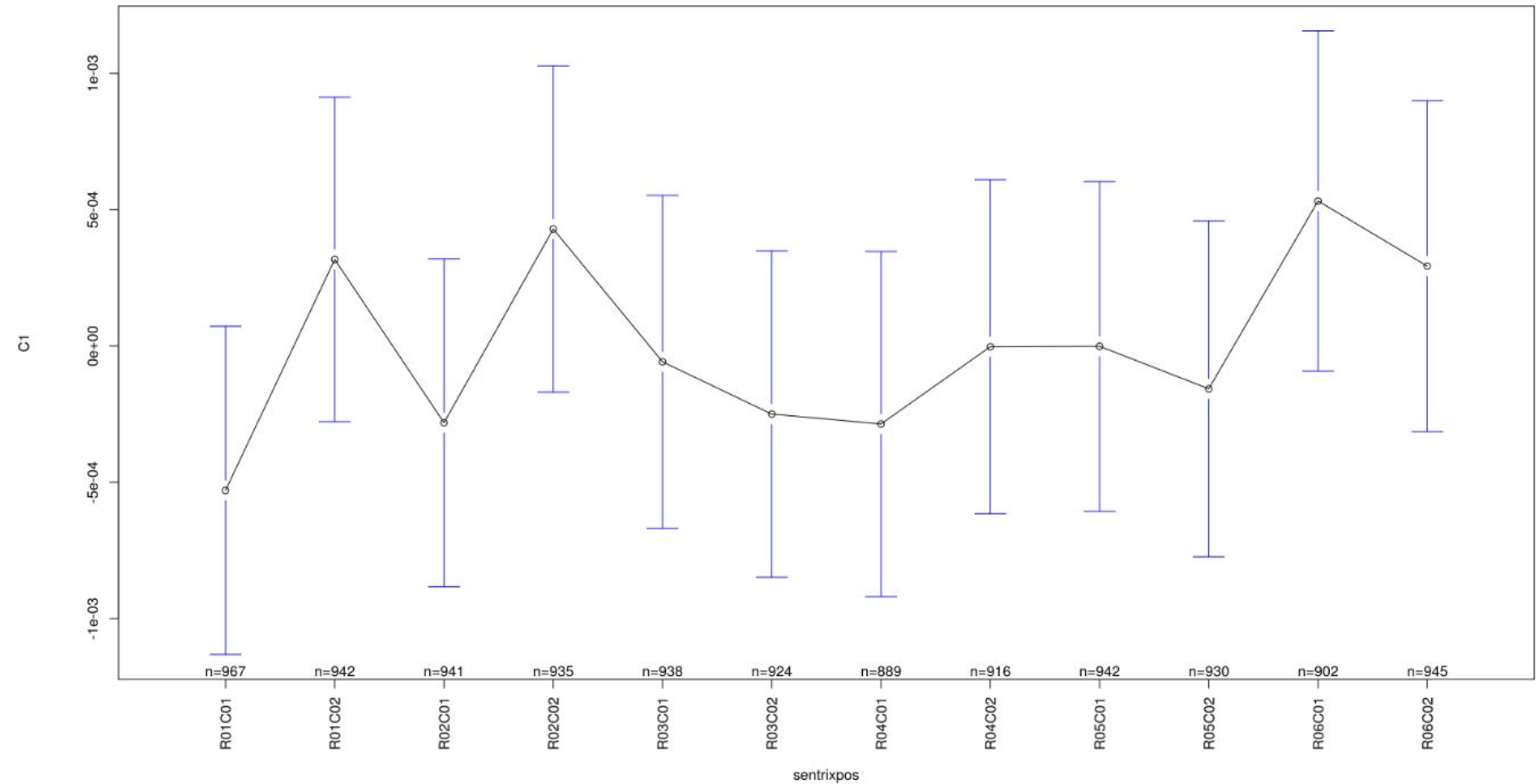
- Batch effects are technical sources of variation that have been added to the samples during handling
- It is important that this type of technical variation does not confound with the biology
- Example of sources of variation can be
  - Change of genotyping platform (different version → different bead pool)
  - New batch of reagents used during genotyping
  - Multiple scanners used during genotyping
  - Multiple individuals/facilities carrying out genotyping
  - Different biobanks handling the samples
  - Position on the sentrix
  - Different date of genotyping (unknown source of error)
- To identify batch effects one needs to look for systematic differences in subgroups of the data
- One way of doing this is by using ANOVA test of variances





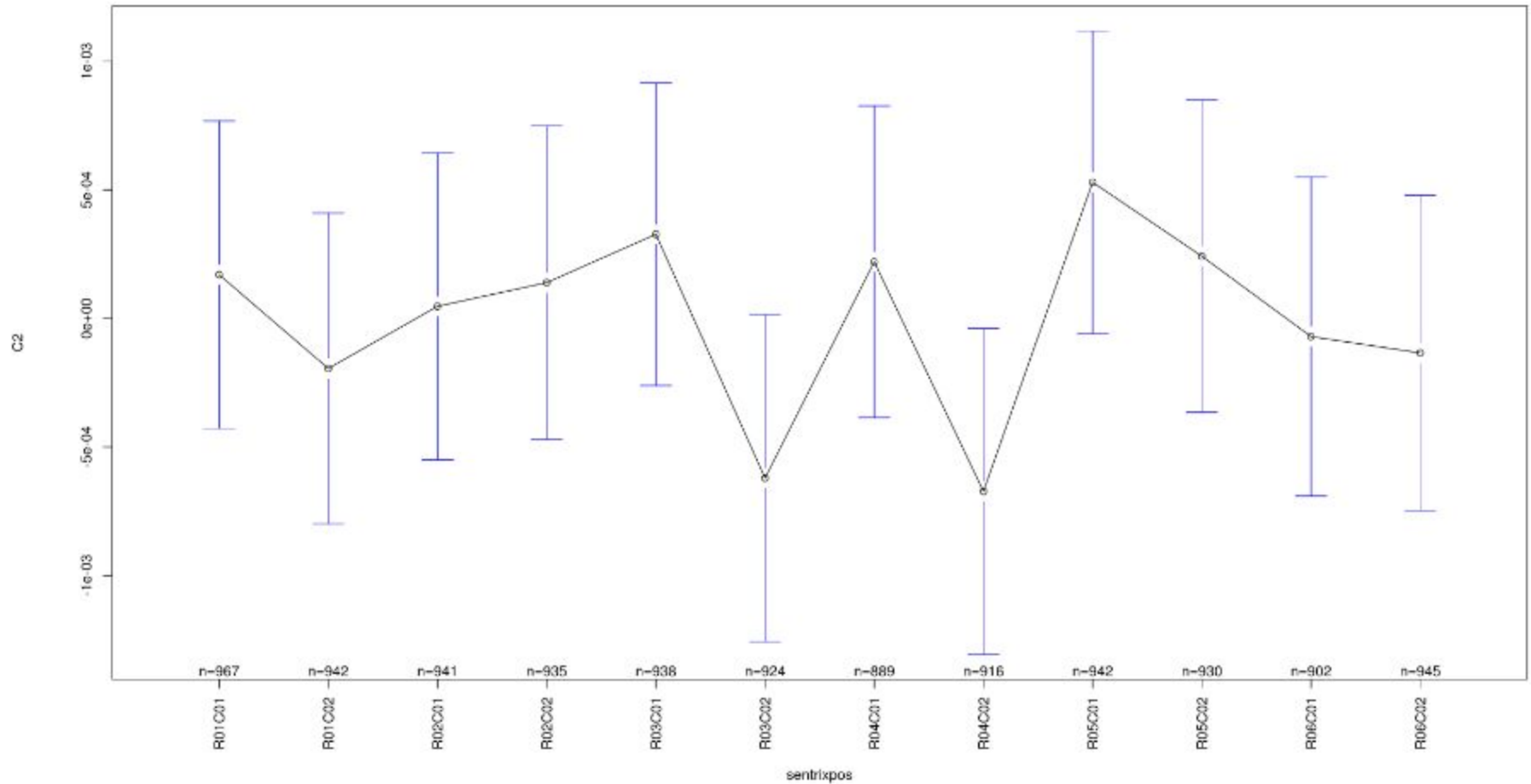
- Before checking for batch effects, ethnic outliers should be removed.
- Ethnic outliers clustering in groups tested in ANOVA could give batch effects because of ethnicity, not necessarily quality.

# MOBA12: PC1 ~ sentrixpos.



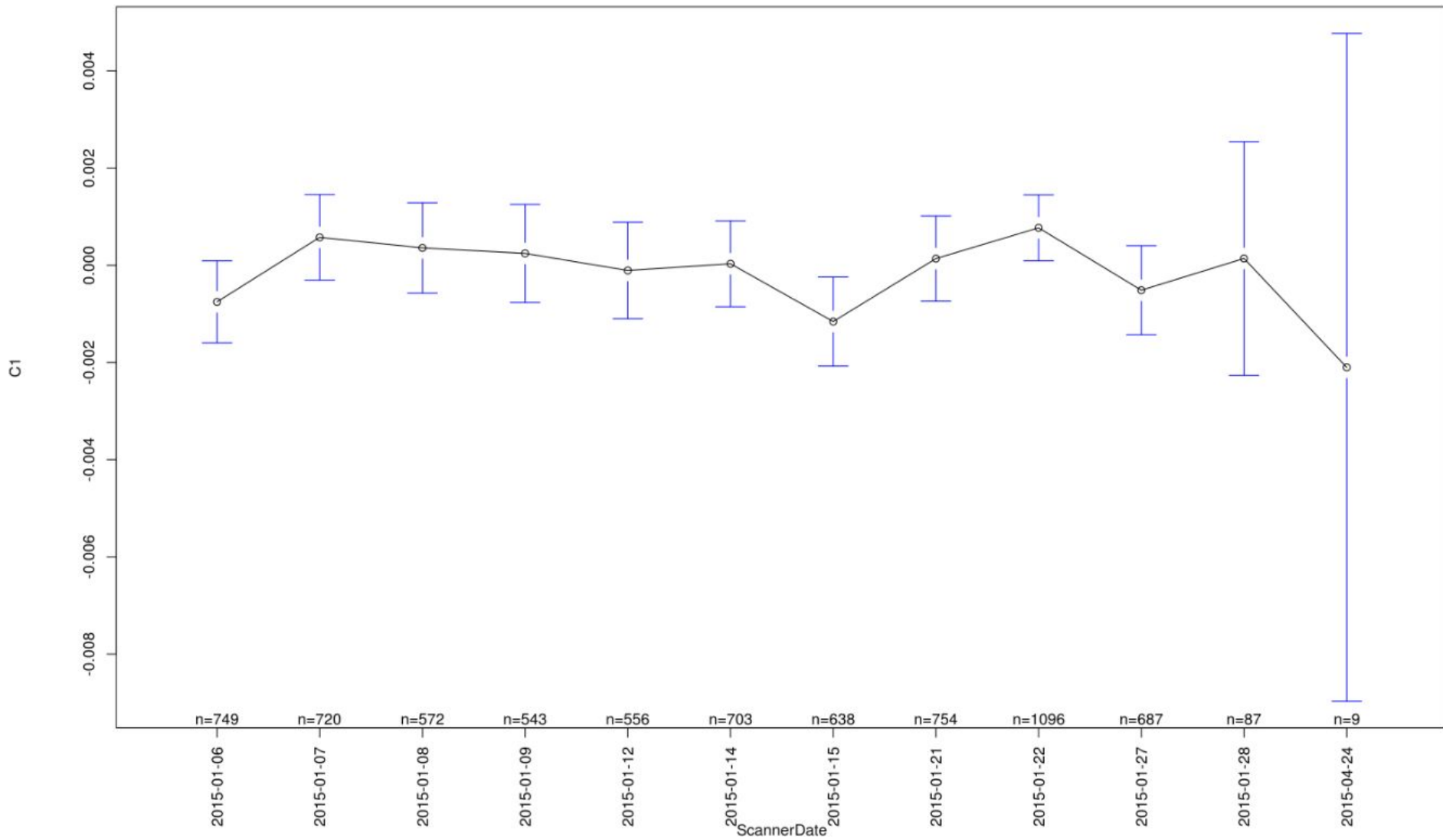
One-way ANOVA test of variances: **PCA1** ~ position on the sentrix in MOBA12

# MOBA12: PC2 ~ sentrixpos.



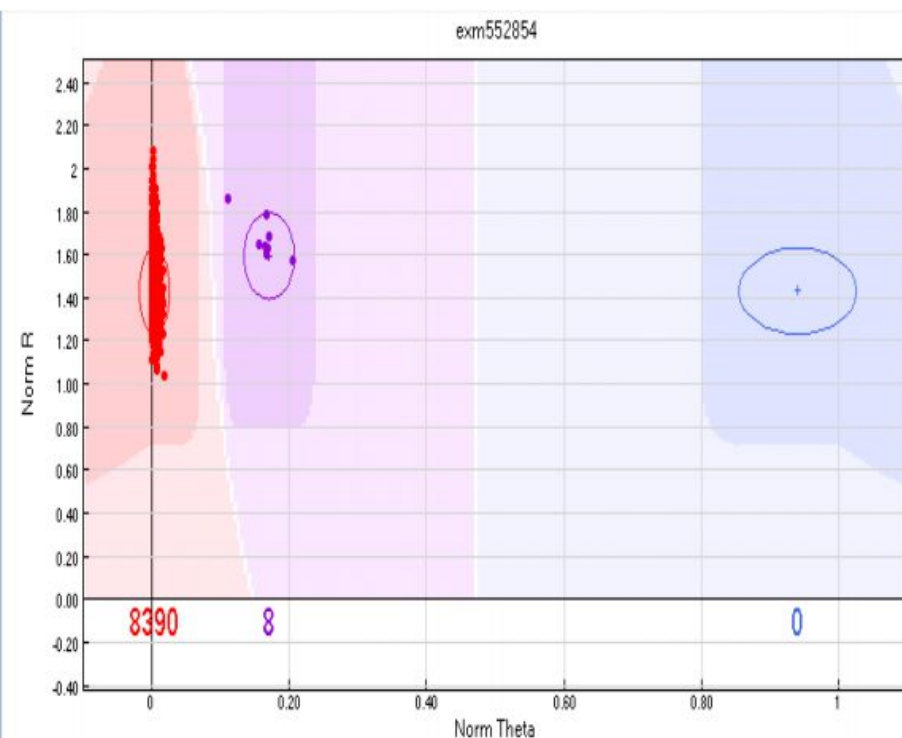
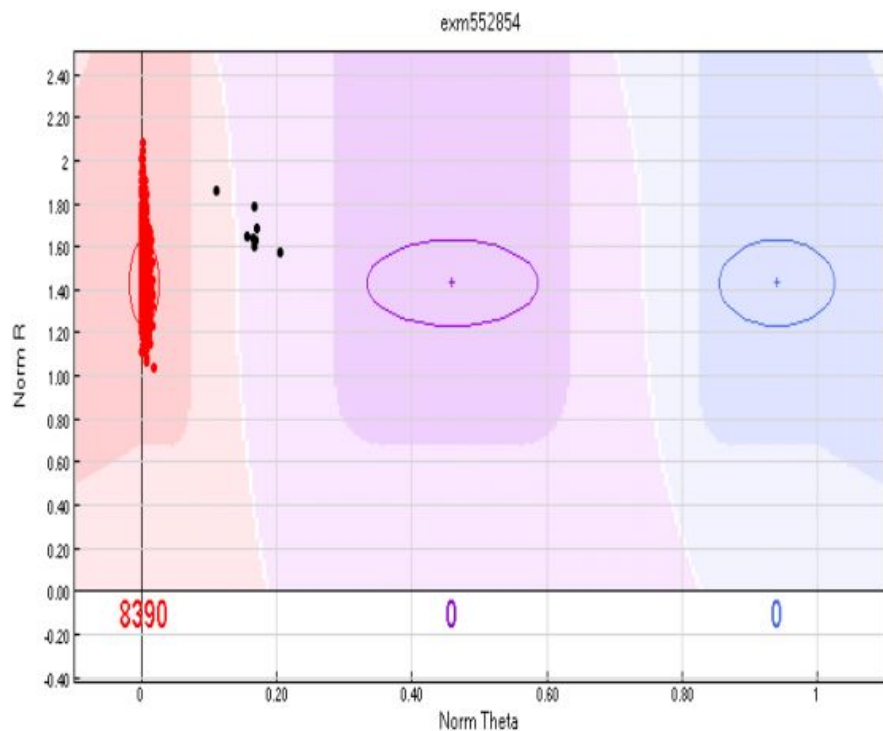
One-way ANOVA test of variances: PCA2 ~ position on the sentrix in MOBA12

# MOBA24: PC1 ~ scanner date



One way ANOVA test of variances: **PCA1** ~ **scanner date** in MOBA24

# Calling rare variants

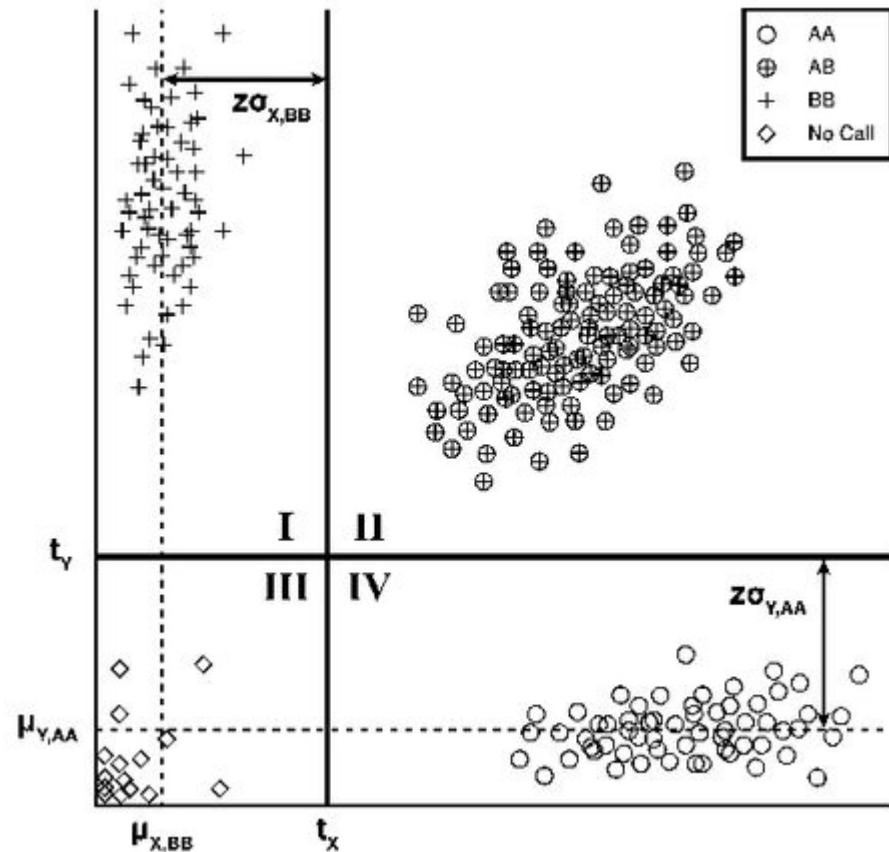


**Left side:** GenCall fails to identify the rare variants. Rares interpreted as outliers.

**Right side:** A program named zCall re-interprets the clusters and assigns correctly the rare variants to the heterozygous cluster

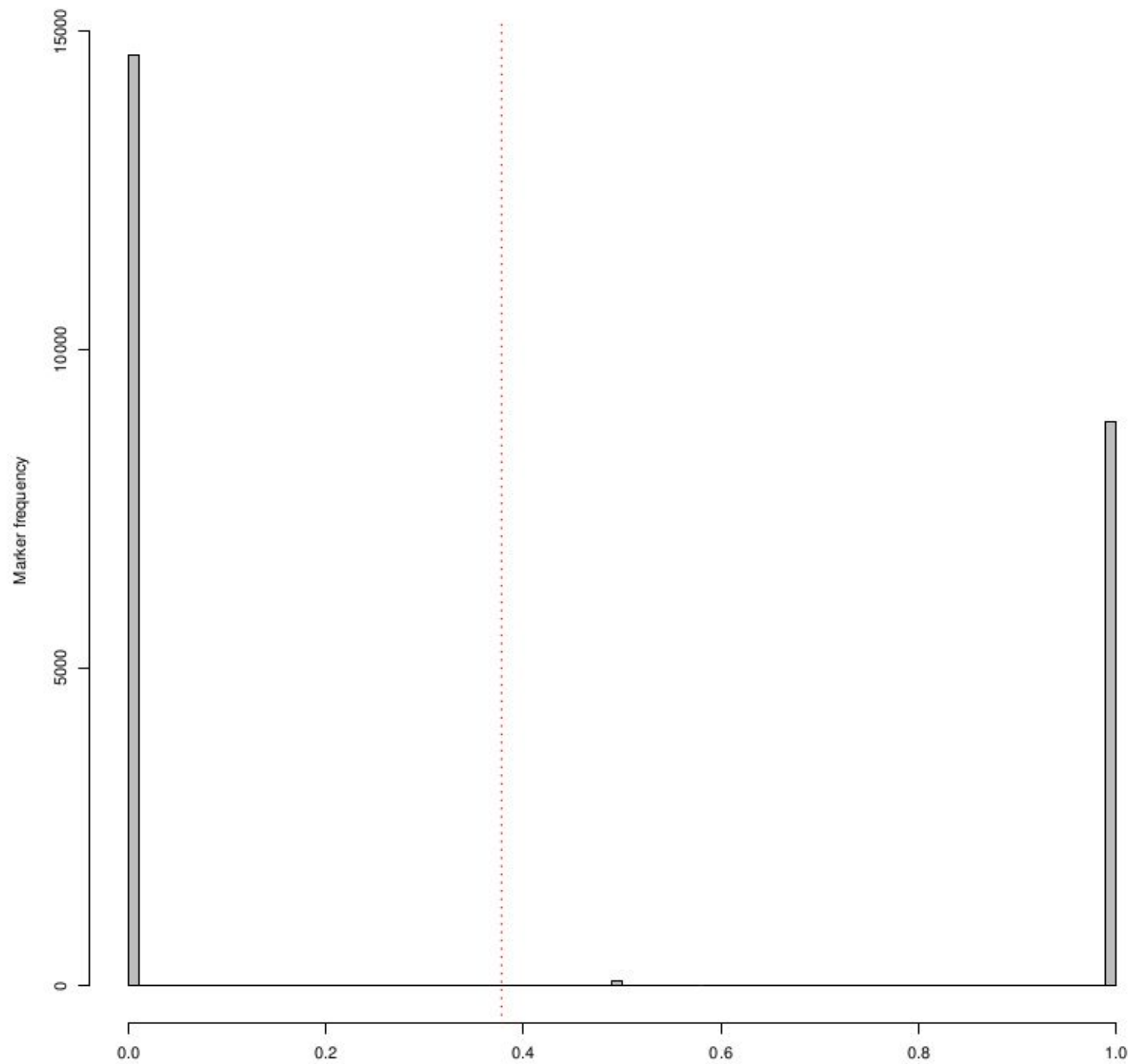
# zCall - a very short explanation

Schematic of how zCall assigns genotypes to points based on the normalized intensity distribution of the homozygote clusters. In normalized intensity space, the common allele homozygote clusters lie along the x and y axes (circles and pluses) and the heterozygote cluster lies along the line  $y = x$  (circled pluses). The three genotype clusters can be separated by a vertical ( $x = t_v$ ) and horizontal line ( $y = t_v$ ) that are derived by solving for the location of z standard deviations from the mean in the direction of the minor axis of the cluster. After  $t_v$  and  $t_v$  have been defined, points are assigned genotypes based on their position relative to the thresholds. Points in Quadrant I are classified as homozygotes (BB), points in Quadrant II are classified as heterozygotes (AB), points in Quadrant III are classified as No Calls, and points in Quadrant IV are classified as homozygotes (AA)

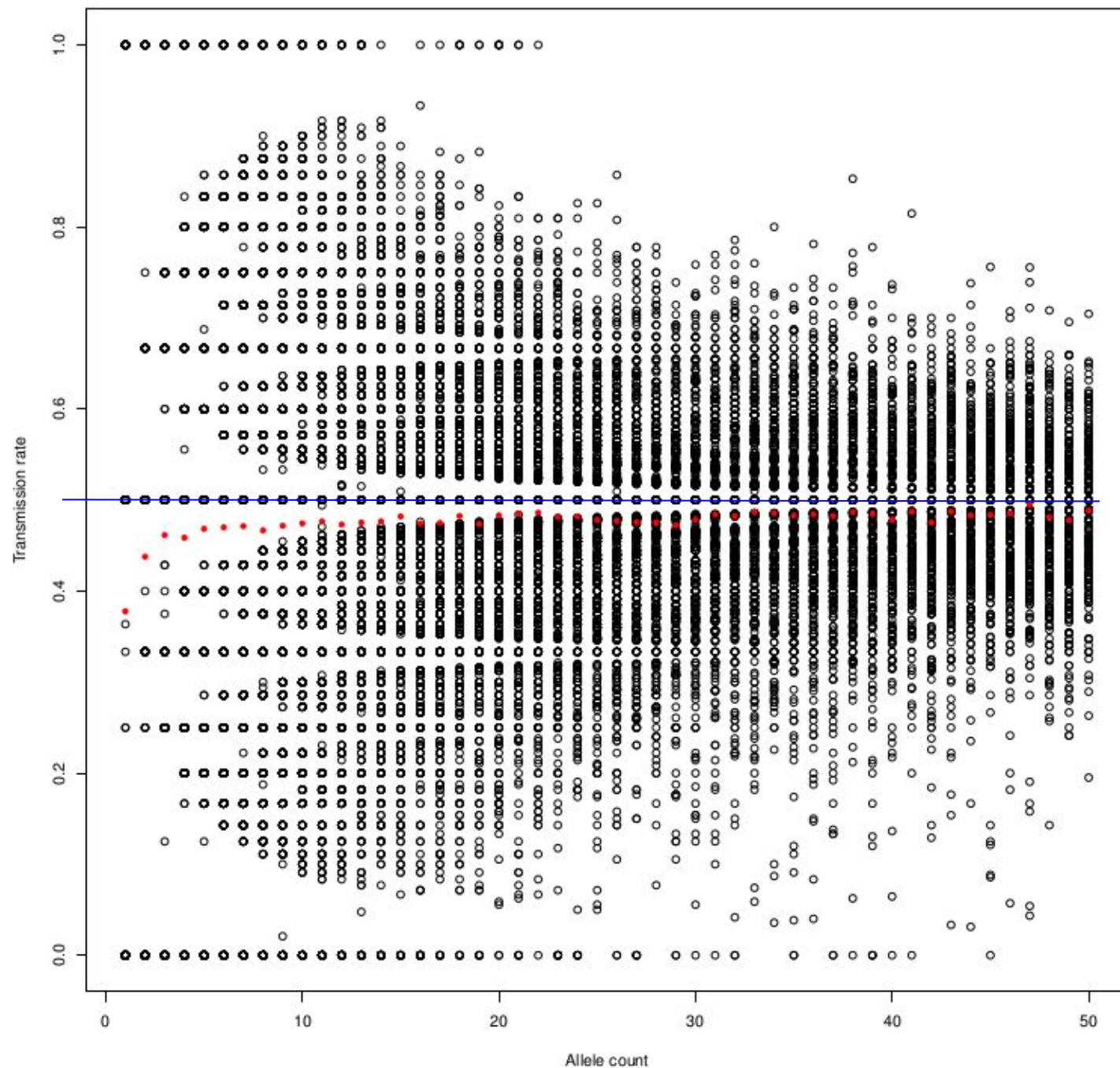


Example genotype intensity plot

Transmission rate distribution for allele count: 1 . Number of markers: 31175



Transmission rate per allele count for allele count 1–50. Red points: mean transmission rate per a.c.

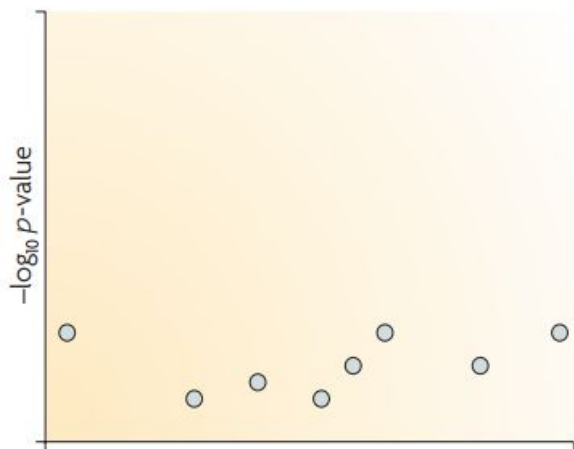




# Genotype imputation

# How imputation works...

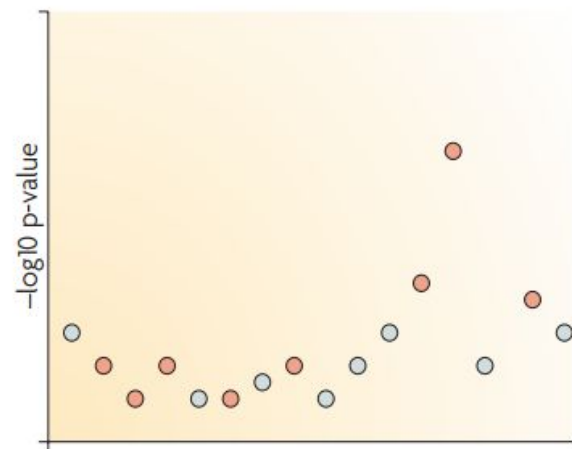
**b** Testing association at typed SNPs may not lead to a clear signal



**d** Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

**f** Testing association at imputed SNPs may boost the signal



**a** Genotype data with missing data at untyped SNPs (grey question marks)

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

**c** Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0	
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0	
⋮																
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0	
1	?	?	?	1	?	1	?	?	1	1	1	?	?	1	?	0
⋮																
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0	
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0	

**e** The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

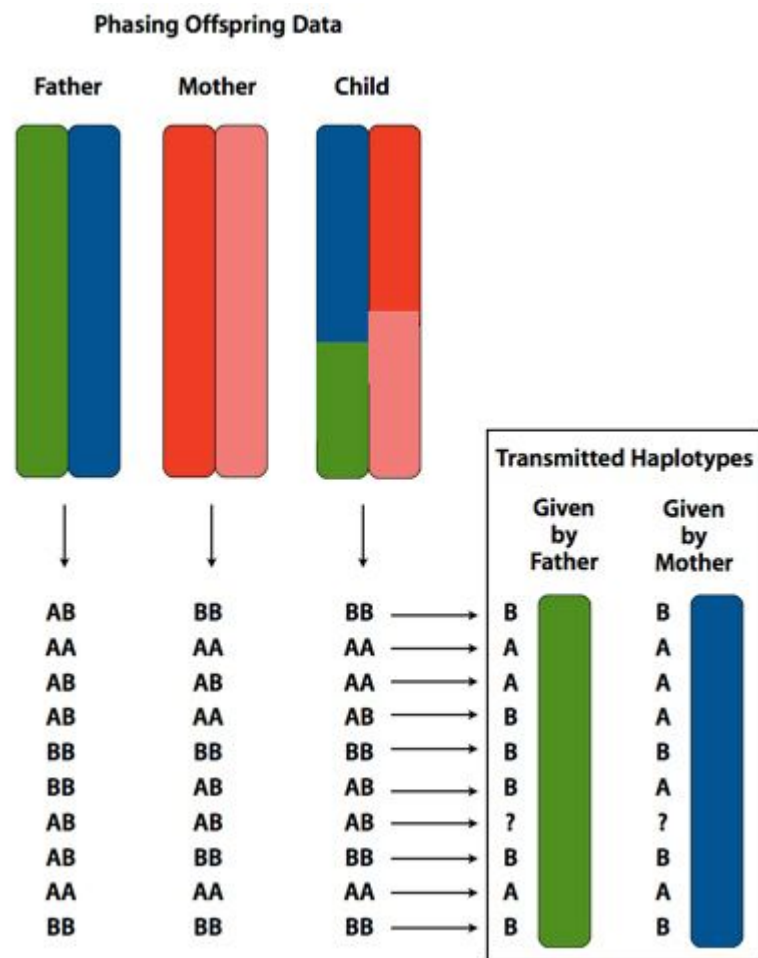
1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

# What is genotype imputation?

- Genotype imputation is the term used to describe the process of predicting genotypes that are not directly assayed in a sample of individuals
- Usually a reference panel of haplotypes at a dense set of SNPs is used to impute genotypes into a study sample that has only been genotyped at a subset of the SNPs
- The goal is to predict the genotypes at the SNPs that are not directly genotyped in the study sample
- These 'in silico' genotypes can be used to boost the number of SNPs that can be tested for association.
- Genotype imputation can be carried out across the whole genome as part of a genome-wide association (GWA) study or in a more focused region as part of a fine-mapping study
- **Increases the power of the study, the ability to resolve or fine-map the causal variant and facilitates meta-analysis**

# Pre(phasing)

- Phasing - often termed pre-phasing as phasing and imputation is separated into two processes
- Pre-phasing done separately from imputation since the phased files can be used to impute genotypes when new genotype reference sets are available (no need to rephase the data)
- Trio phasing (using data from a child and both parents) is the gold standard for phasing. It is possible to phase about 94% of the alleles in an autosomal dataset using a two parent/one child trio
- If large trio data → internal phasing (not using reference dataset) is more accurate than relying on the “resolution” in the reference data.
- External datasets with more samples and thus “higher resolution” of haplotypes yield better phasing results (captures more recombinations)



# Phasing and imputation

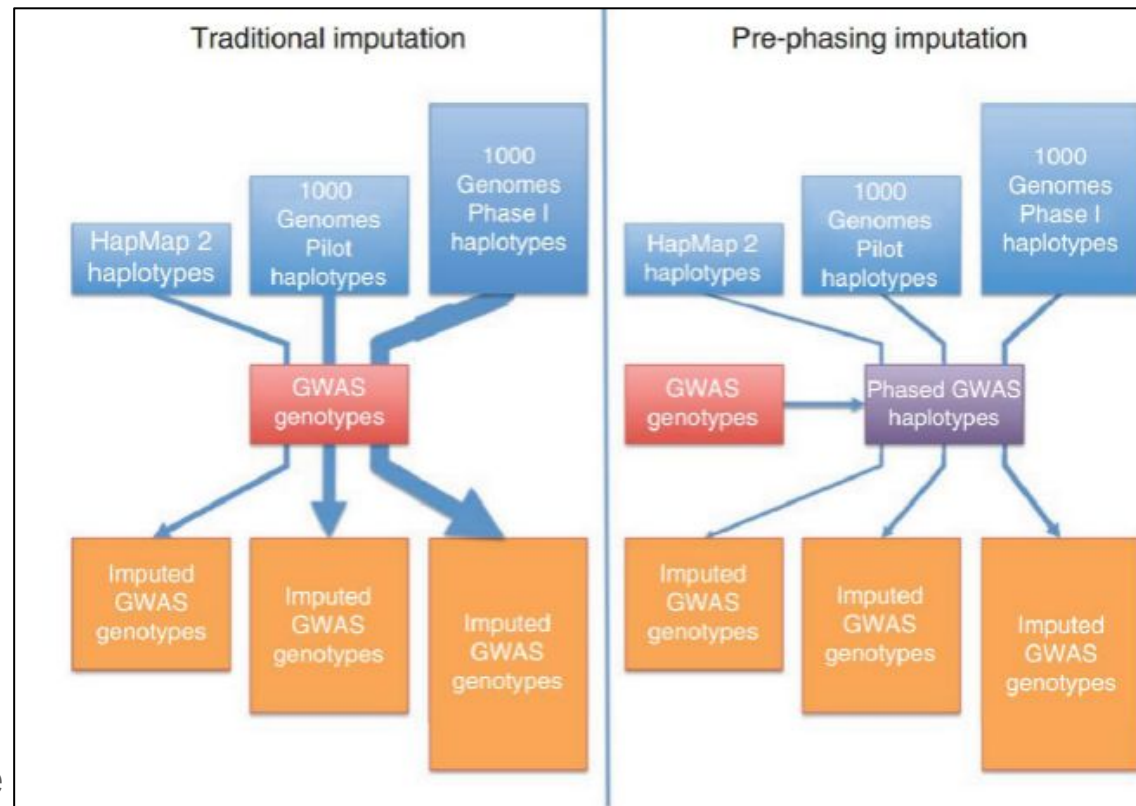
Usually divided into two steps:

## 1. Phasing/pre-phasing

Phasing is the process of assigning alleles (the As, Cs, Ts andGs) to the paternal and maternal chromosomes (i.e. identifying the haplotypes)

## 2. Imputation

The process of “filling in” the missing genotypes using a dense reference set of genotypes using the phased dataset



# Software

Several tools developed for genotype imputation. Previously phasing and imputation was done using the same tool, now most tools rely/recommend specific software for prephasing.

- Popular software for imputation
  - Impute2 - [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](https://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
  - Minimac3 - <http://genome.sph.umich.edu/wiki/Minimac3>
  - Beagle4.1 - <https://faculty.washington.edu/browning/beagle/beagle.html>
  - MaCH - <http://csg.sph.umich.edu//abecasis/MACH/tour/imputation.html>
- Popular software for pre-phasing
  - Shapeit - [https://mathgen.stats.ox.ac.uk/genetics\\_software/shapeit/shapeit.html](https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)
  - MaCH
- Combinations most commonly used today
  - Shapeit + Impute2 (Seems to be the most popular combination)
  - MaCH + Minimac3

Minimac3 vs. Impute2 vs. Beagle4.1?

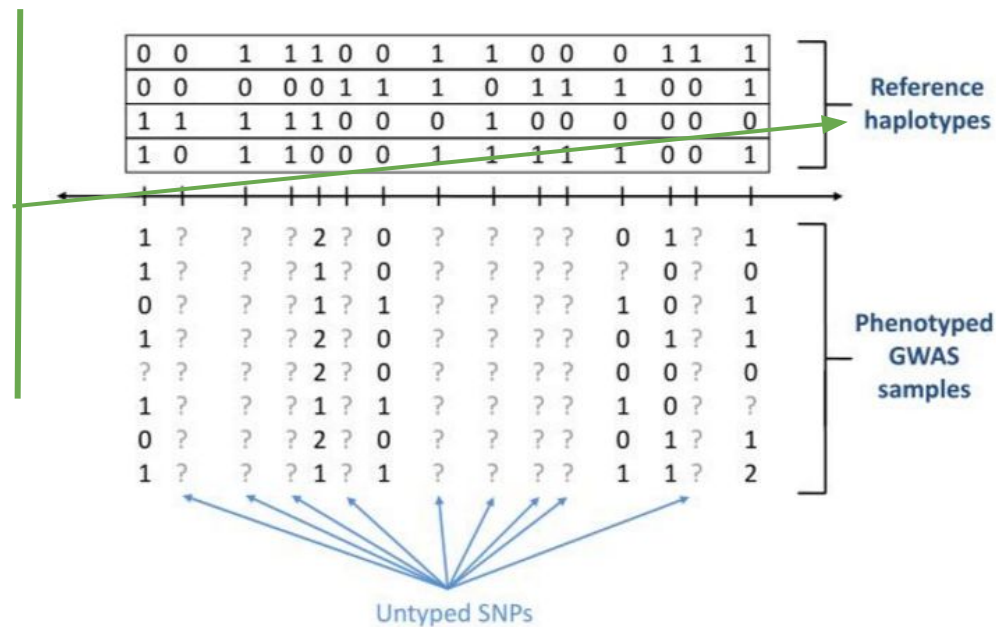
→ similar accuracy, similar computational costs, similar features

**And never use PLINK for imputation!**

# Reference data

- Gen. imputation relies on obtaining a reference dataset genotyped more densely than your own data.
- Several “publicly available” reference datasets exists:
  - **HapMap** - can be downloaded for local imputation
  - **1000 Genomes** - can be downloaded for local imputation
  - **DeCode** - imputation performed at DeCode, Reykjavik, Island
  - **UK10K** - Imputation performed at University of Michigan
- Imputation is most beneficial for rare variants (harder to tag in arrays)
- **Imputed genotypes are always associated with a degree of uncertainty**

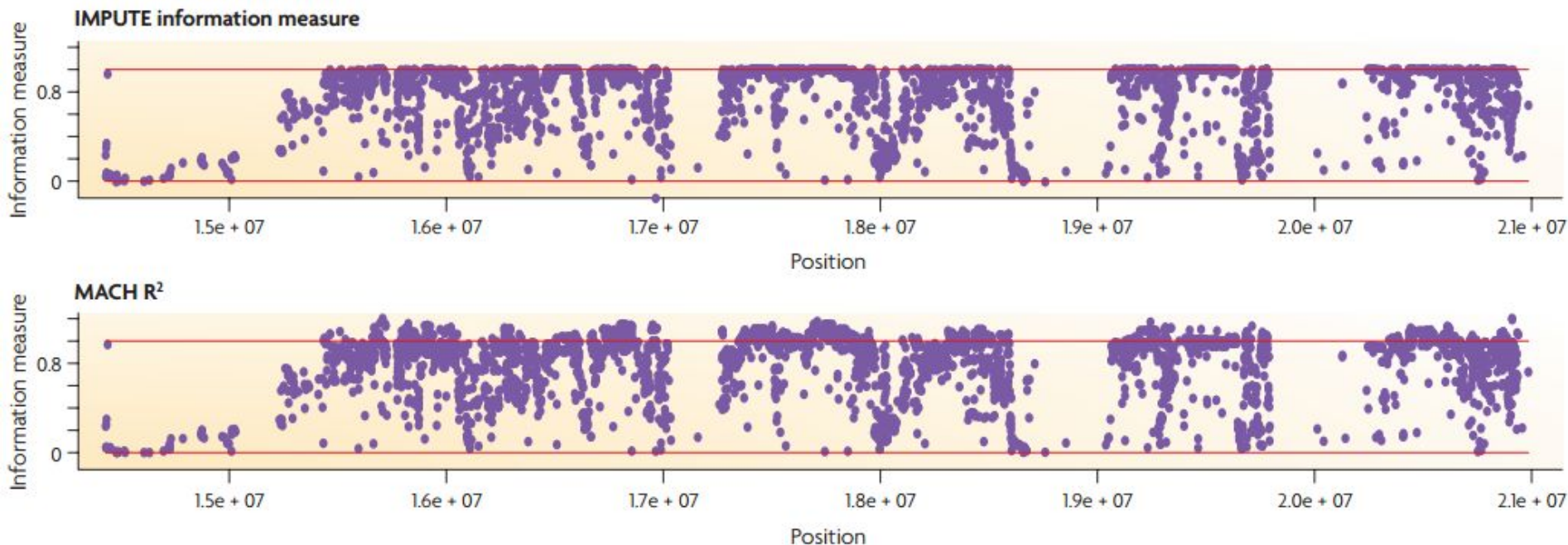
## Genotype imputation background





# Imputation quality scores

- Different software use different approaches to assess the quality/certainty of the imputed genotype
- Even though the approaches differ they have been shown to be comparable (e.g. very good correlation between Minimac3 and Impute2 score)
- Impute2 calls this *info-score* and Minimac3 *Rsquared*
- Most markers have too low quality (MOBA 48M  $\rightarrow$  9M)





# Quality and allele dosages

- After imputation, each genotype for each individual is associated with a level of uncertainty
- Instead of setting a fixed threshold for deciding whether the genotype is AA, AB or BB; allele dosages are created per genotype (per individual) as a number between 0 and 2
- The dosage is the “dose” of alternative allele present at that genotype taking this uncertainty into account - so if the uncertainty for a genotype is high it is “weighted” down by a lowered dosage
- Makes it possible to take imputation genotype uncertainty into account when running association analyses (by using the dosage files)



**BB - I am sure!**



**BB - but not that sure...**

# Quality and allele dosages - continued

## **But what if I want the genotypes?**

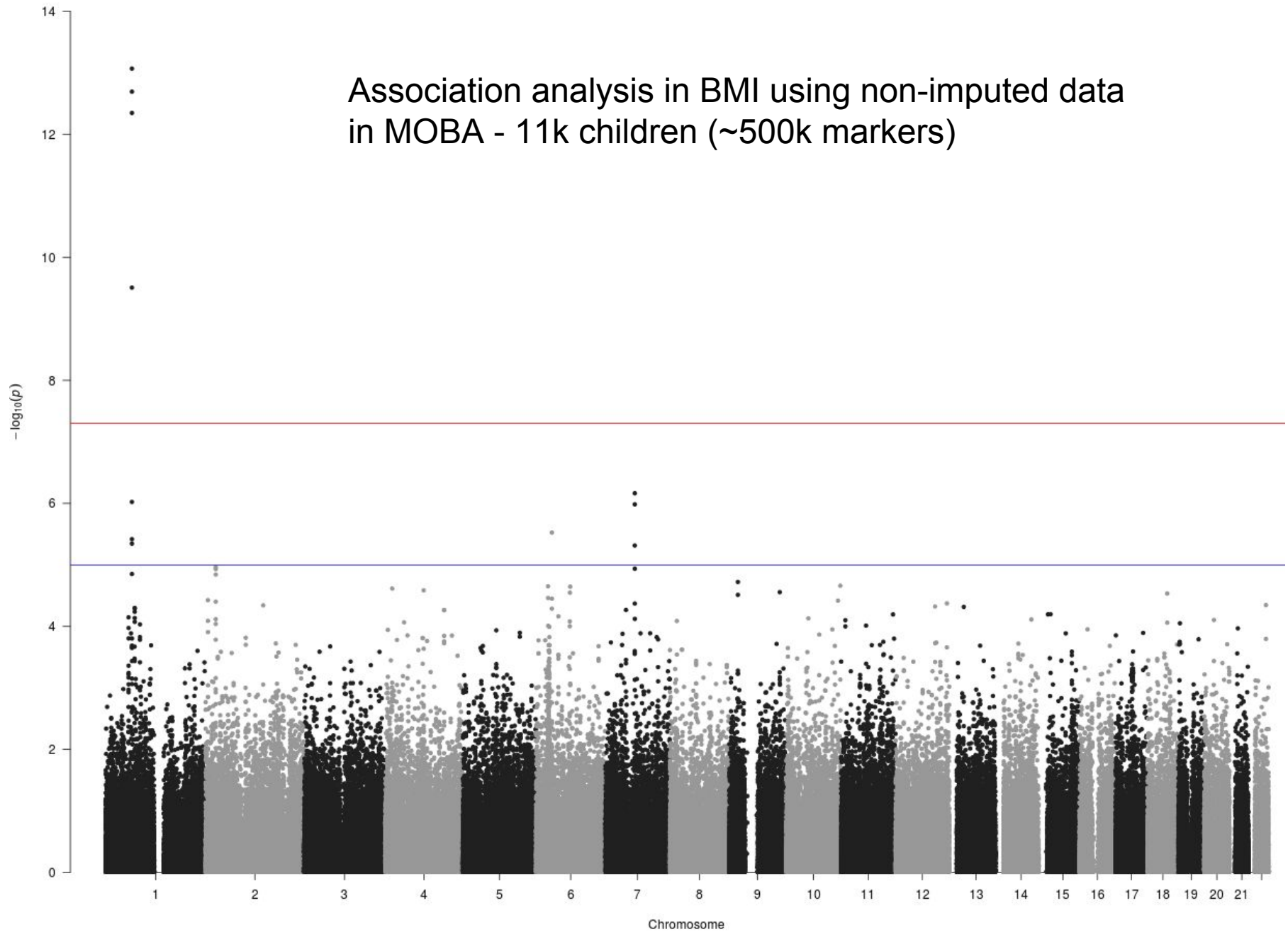
Then we have to convert the dosages to genotypes using the “buest guess” - approach - i.e. the most likely genotype based on the dosage

## **But what if the uncertainty of the “buest guess” genotype is high?**

When using buest guess genotypes we always have to use the “per marker” quality score.

Minimac3 compares the variance of the allele dosages to the expected binomial variance at Hardy-Weinberg equilibrium. Provides a quality score for each marker that can be used to filter the most reliably imputed markers.

Association analysis in BMI using non-imputed data  
in MOBA - 11k children (~500k markers)



Association analysis in BMI using imputed data in  
MOBA - 11k children (~16M markers)

