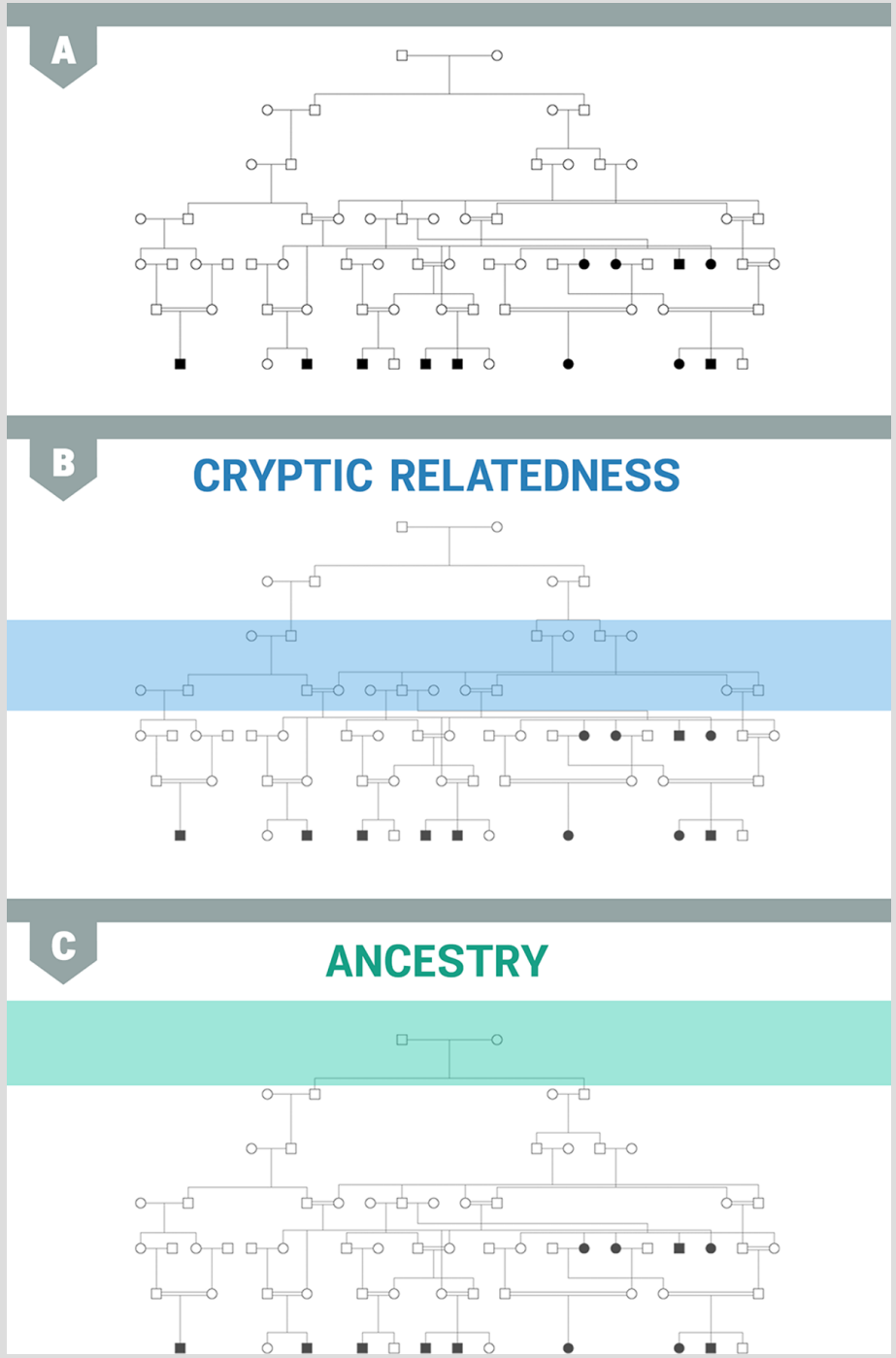
GWAS 5

Matti Pirinen
University of Helsinki
Week 3

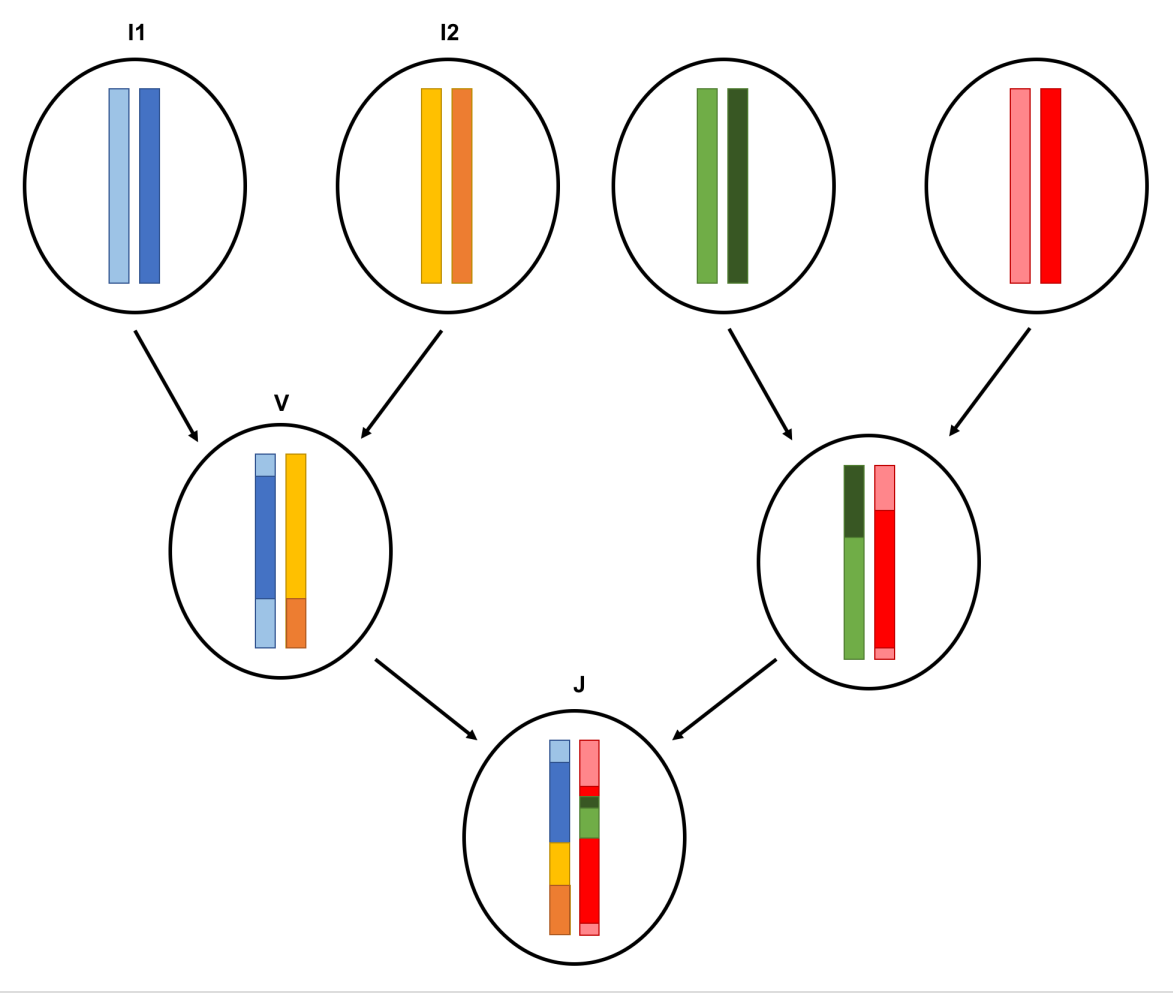


LEVELS OF RELATEDNESS

(A) All of the individuals in a genetic study are related through a large pedigree (family tree). Different parts of the tree induce different types of relatedness.
(B) **Cryptic relatedness** refers to relatively recent genetic relationships which are not otherwise reported in the data except being revealed by genetic analysis.
(C) Relatedness due to ancestry refers to relatedness caused by shared ancestors, possibly over longer periods of time in the past. This is also called genetic **population structure**.
The boxes in (B) and (C) represent the part of the pedigree that causes that type of relatedness in each case.

Sul et al. 2018 PLoS Gen https://doi.org/10.1371/journal.pgen.1007309

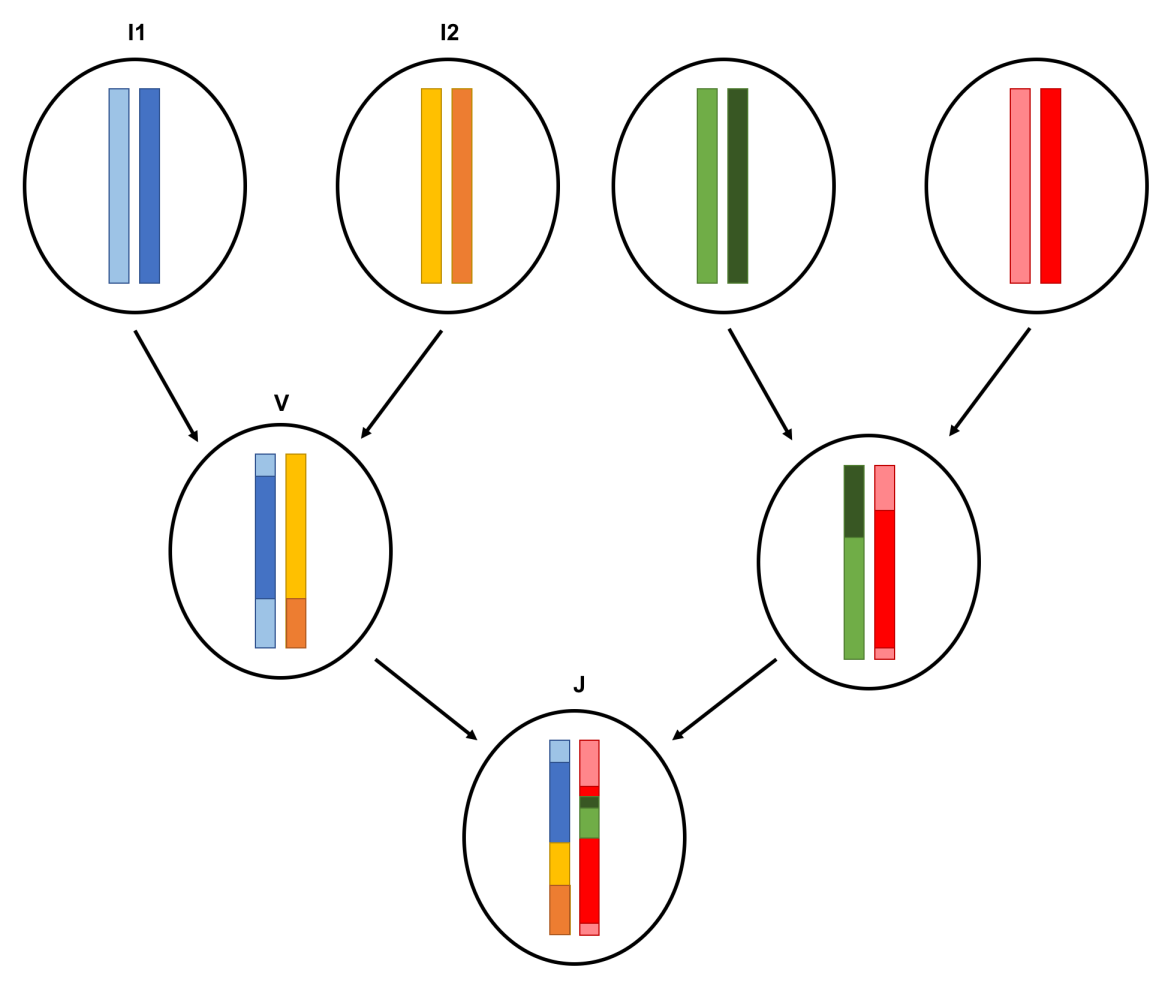
GENOMIC RECOMBINATION



Sini Kerminen; www.mv.helsinki.fi/home/mjxpirin/stamp/

Three generations are shown: 4 grandparents (top), 2 parents (middle) and an offspring (bottom).
The two genomes of each individual are shown for a particular genomic segment.
All 8 grandparental genomes are colored with separate colors.
The parents inherit recombined genomes from the grandparents, e.g., parent V inherits from grandparent I_1 a combination of light blue and dark blue genomes.
Similarly, as a result of a recombination, offspring J inherits from parent V a genome that contains segments from all 4 genomes of grandparents I_1 and I_2.
If we took the reference level of colored genomes further back in time to more distant ancestors, we would observe shorter intact segements, and more colorful genomes, in the offspring.

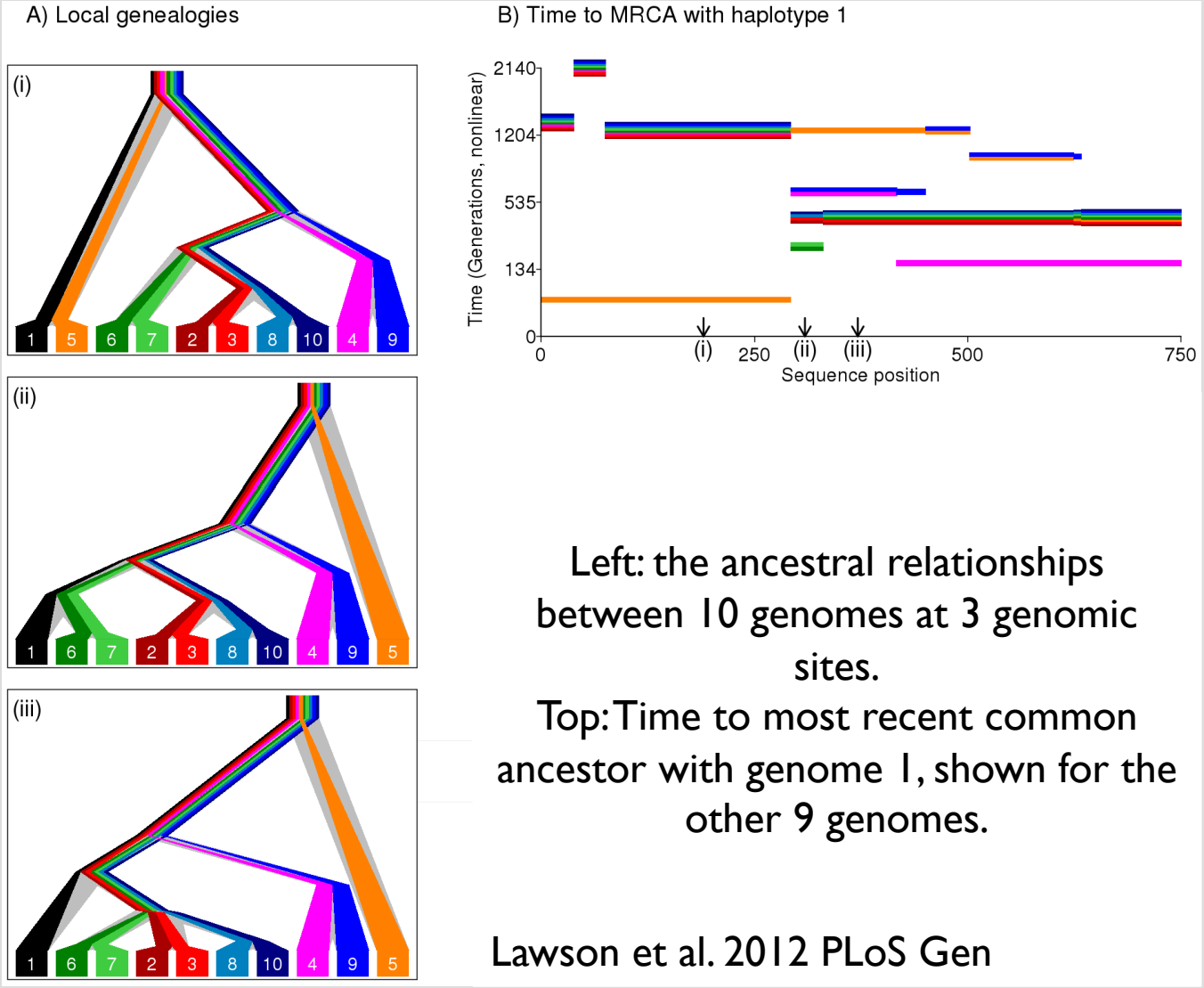
GENOMIC RECOMBINATION



Sini Kerminen; www.mv.helsinki.fi/home/mjxpirin/stamp/

- Offspring inherits genome as continuous segments of the parent's two genomes
- Segments from recent ancestors are longer than from more distant ancestors
- This process leads to correlations in genotype data
  - Genetic relatedness (today's topic)
    - Individual-by-individual correlation of genotypes
    - "closer relatives share more genome"
  - GWAS results at nearby SNPs are correlated (next week)
    - SNP-by-SNP correlation of genotypes

GENETIC RELATEDNESS IS ABOUT TIME TO MOST RECENT COMMON ANCESTOR

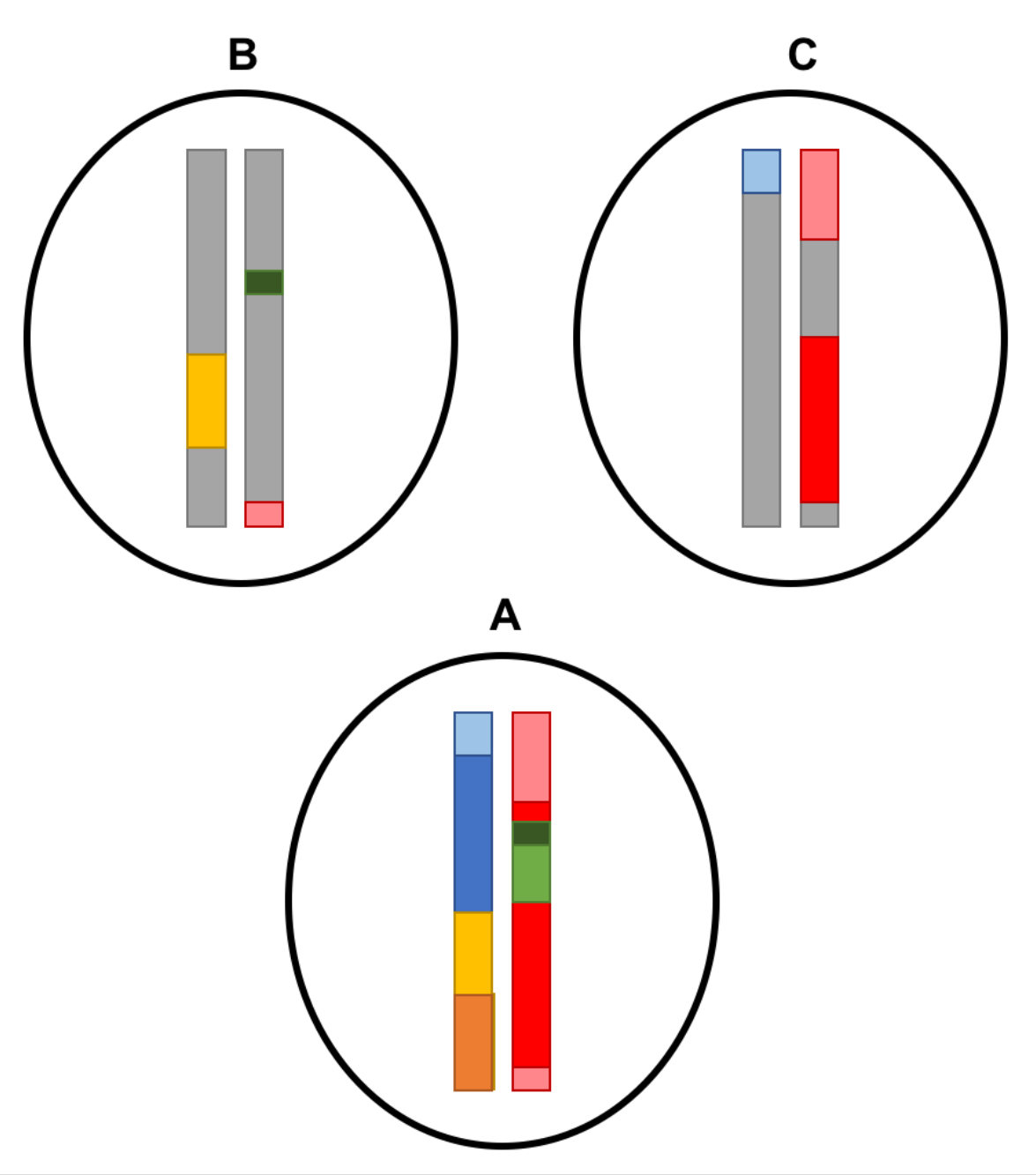


Left: the ancestral relationships between 10 genomes at 3 genomic sites.
Top: Time to most recent common ancestor with genome 1, shown for the other 9 genomes.

Lawson et al. 2012 PLoS Gen

Intuition:
Close relatives share more genetic ancestors in more recent past than distant relatives.

If we average such information across the genome we have a relatedness estimate for genomes,
and if we average over the two genomes of individuals we have a relatedness estimate for individuals.



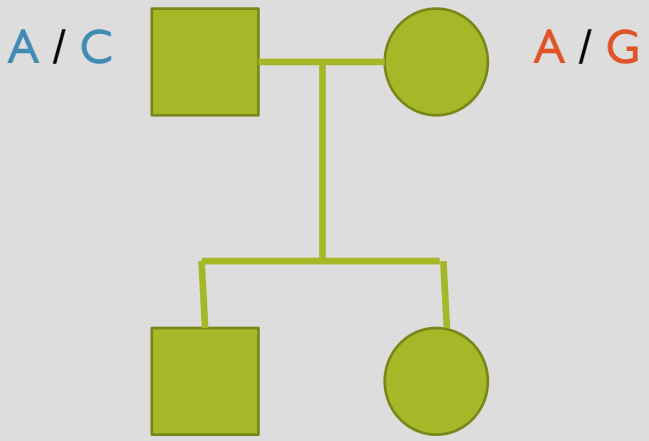
ESTIMATING RELATEDNESS

Two genomes of individual A are colored according to their ancestral origins with respect to some time point back in time (See Slide 2). The two additional individuals B and C belong to the present day generation (as does A) and by computational methods we have estimated in which parts of the genome B or C shares segments with A.
The most probable matches have been colored and the rest of the genomes have been left gray.

We estimate that C is a closer relative to A than B because C and A have more shared DNA from recent ancestors than B and A.

Sini Kerminen; www.helsinki.fi/~mjxpirin/stamp/

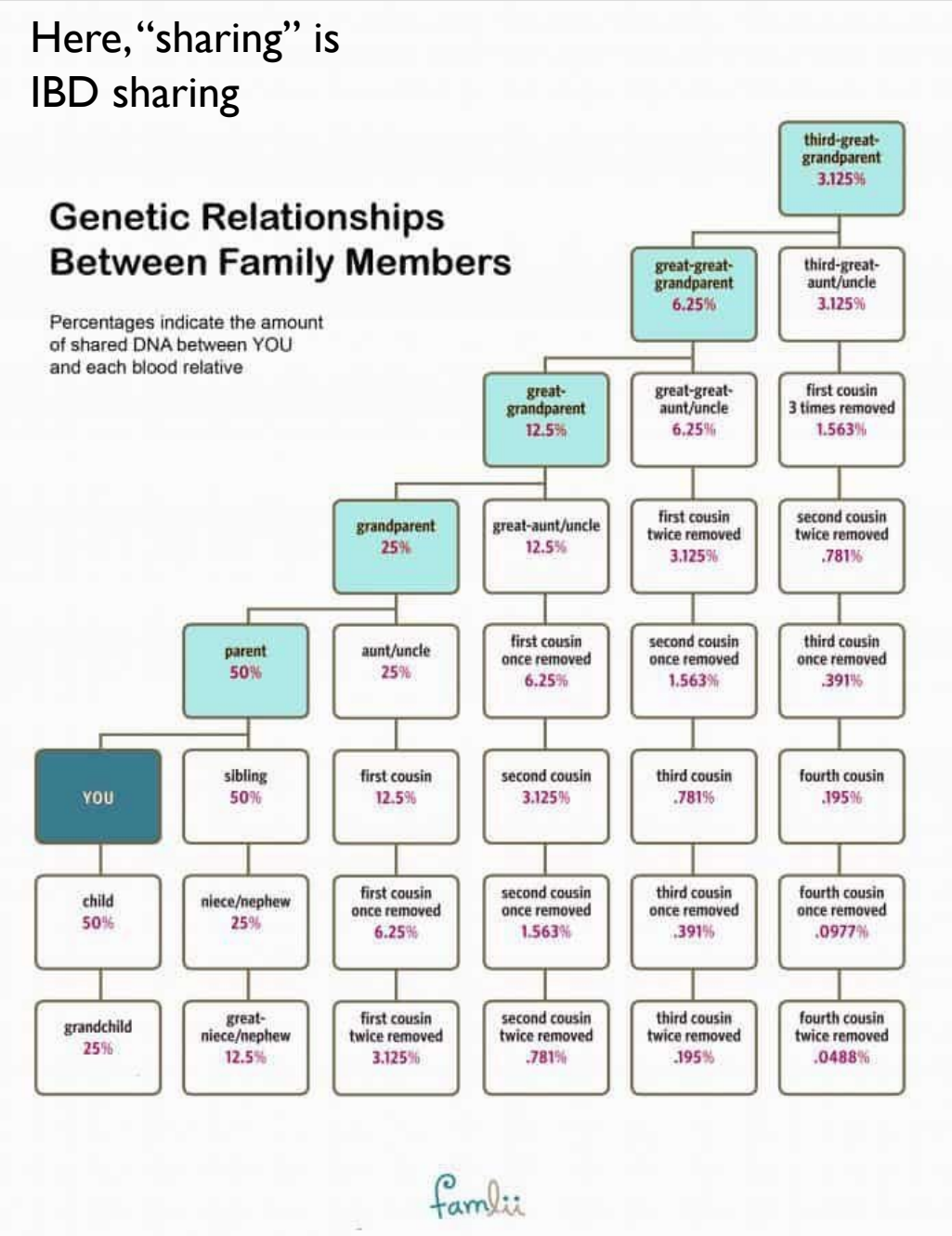
IBS VS IBD



	IBS	IBD	
A / G	1	1	A / A
C / A	1	0	A / G
A / A	2	2	A / A

- IBS (identical-by-state) means that DNA sequence is identical
- IBD (identical-by-descent) means that DNA sequence is inherited from a common ancestor (within a given timeframe)
  - For pedigrees the timeframe is the ”founder generation” (top level)
  - For population data, we don’t have an exact timeframe and then IBD is measured by how much more the pair shares IBS than would be expected from a random pair from the population
  - Most accurate IBD estimation methods look for sharing of longer segments, not just individual loci, but here we consider sharing at independent loci
- IBD implies IBS but not vice versa: there can be IBS sharing without there being a common ancestor **within a given timeframe,** such as within the known pedigree structure of, say, last 3 generations

Here, "sharing" is IBD sharing

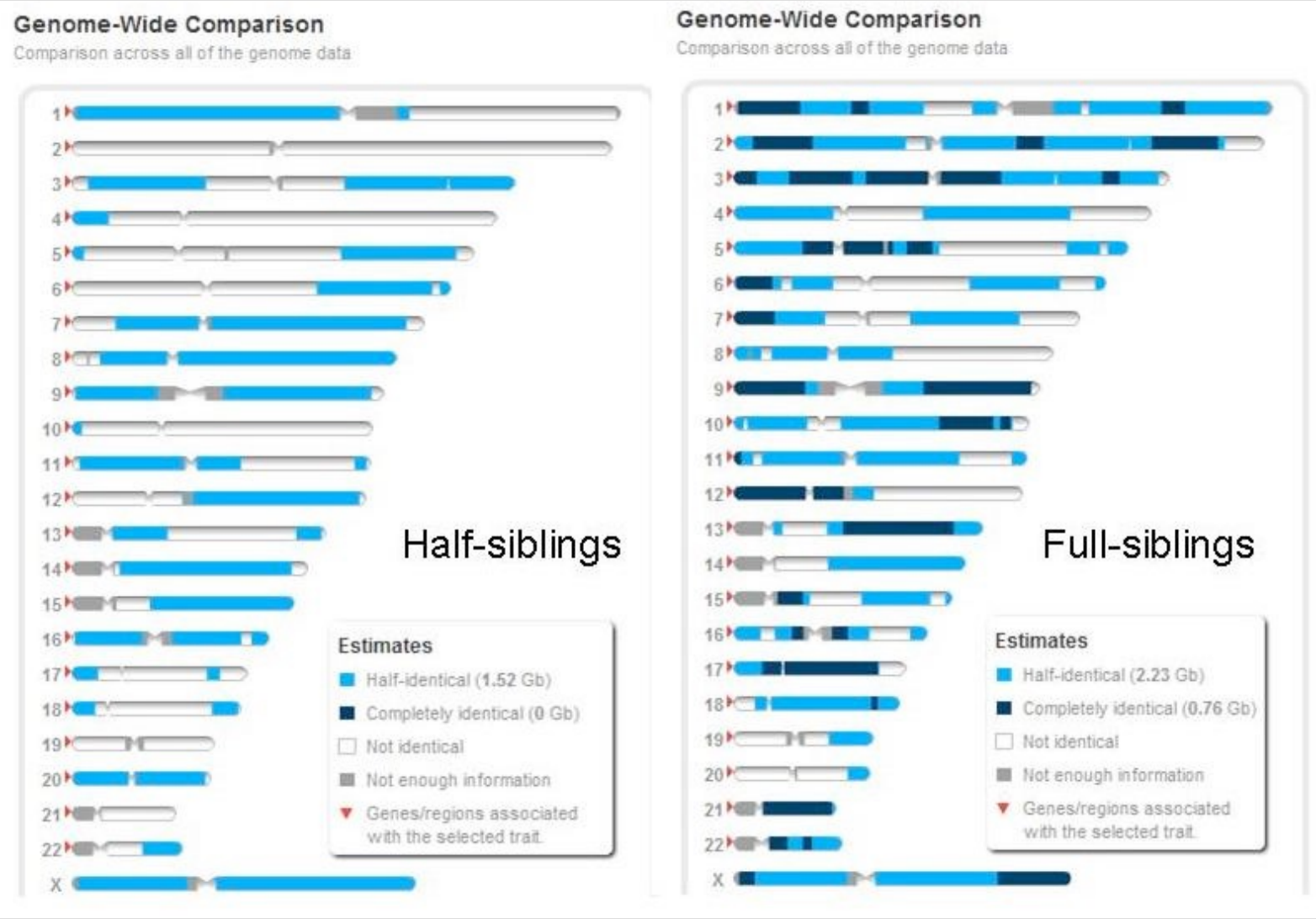


www.famlii.com

EXPECTED IBD SHARING

- 1st degree relatives: $r_{ij} \sim 50\%$
- 2nd degree: $r_{ij} \sim 25\%$
- 3rd degree $r_{ij} \sim 12.5\%$
- etc.
- Several categories have same expectation

RECENT IBD SEGEMENTS ARE INFORMATIVE



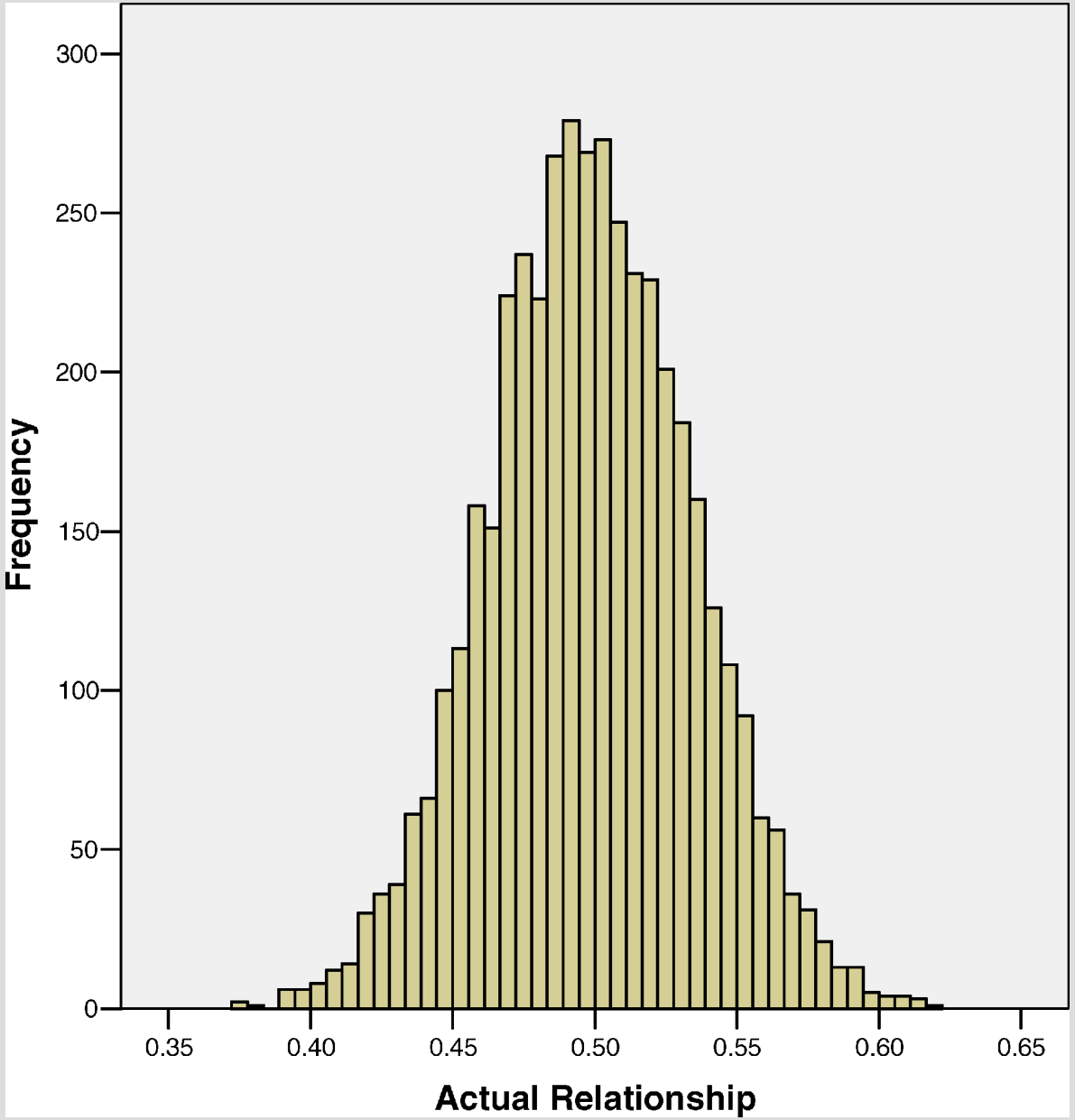
Each panel describes estimated IBD sharing between one pair of individuals.

Left pair seem like half siblings as they share half of the genome IBD1

Right pair seem full sibs as they share 25% IBD2, 50% IBD1 and 25% IBD0.

Relatedness coefficient r_{ij} for half-sibs is ~25% and for full sibs is ~50%.

http://ongenes.blogspot.fi/2011/02/half-siblings-vs-full-siblings.html



FULL SIBLINGS' IBD SHARING

Empirical distribution of actual genetic relationships r_{ij} of 4,401 pairs of full sibs estimated using up to 2,000 variants across the genome.

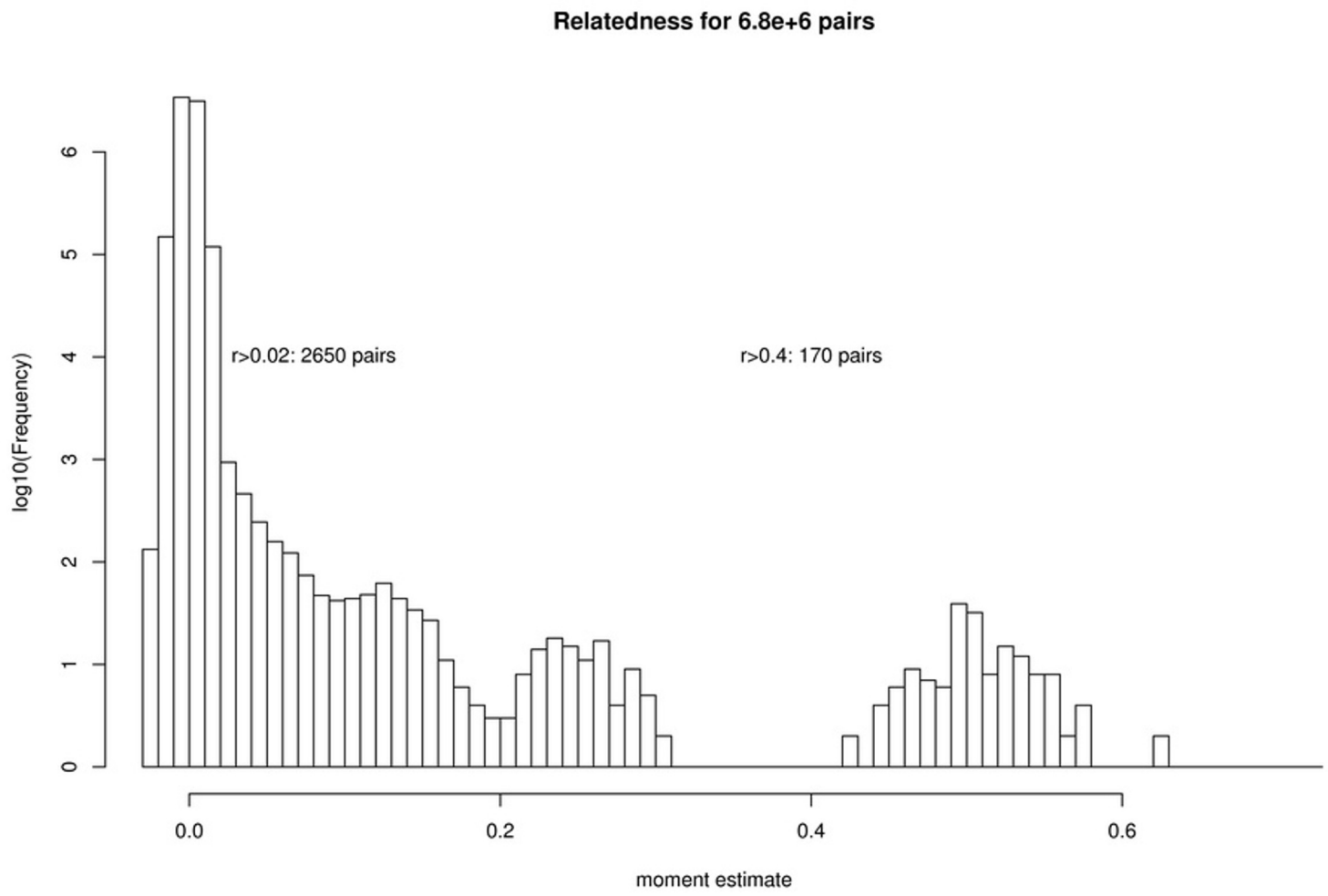
Values range between 0.374–0.617

The exact amount of IBD sharing varies considerably within a relative catrgory (here full sibs).

Exception is the parent-offspring relationship that has IBD1 throughout the genome.

Visscher et al. https://doi.org/10.1371/journal.pgen.0020041

EXAMPLE RELATEDNESS FROM A GWAS STUDY



Correlation estimator was used as an estimate of r_{ij}

We detect groups of
50% (e.g. Full sibs)
25% (e.g. Half sibs)
12.5% (e.g. First cousins)

And almost all pairs are around 0%, ”unrelated”

NOTE y-axis is on log-scale.

KING APPLIED TO UK BIOBANK

N ~ 500,000
1.2e+11 pairs

From: The UK Biobank resource with deep phenotyping and genomic data

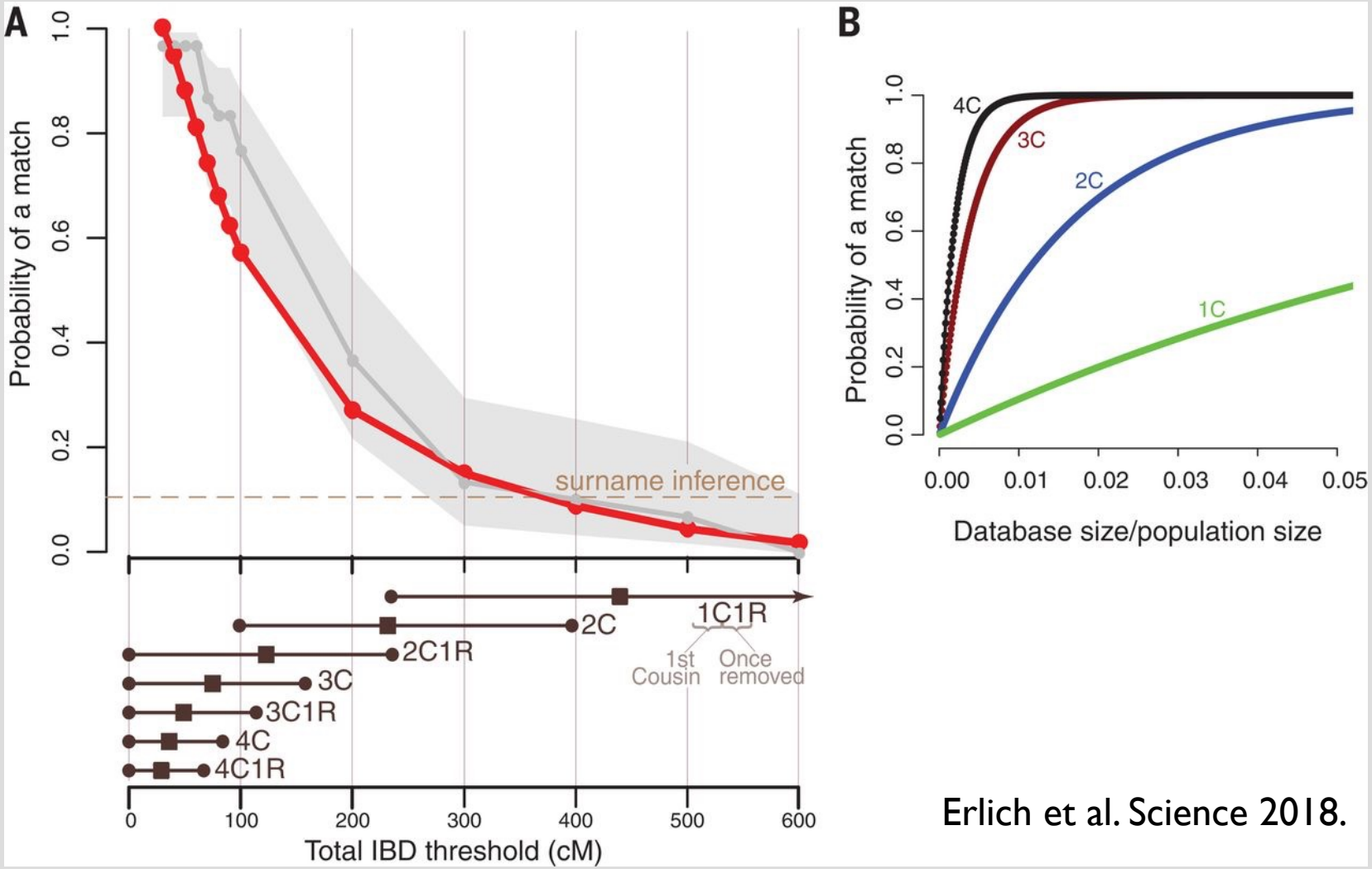
	Monozygotic twins	Parent-offspring	Full siblings	2nd degree	3rd degree	Total
Number of pairs	179	6,276	22,666	11,113	66,928	107,162

Counts are derived from the kinship coefficients (see Methods). The count of monozygotic twins is after excluding samples identified as duplicates (Supplementary Information).

ARREST OF GOLDEN STATE KILLER

- "GSK" did tens of murders, assaults and rapes in 1970s and 1980s
- In 2018, police uploaded the DNA of the killer to GEDmatch
  - GEDmatch contains > million genome-wide data sets uploaded by volunteers in order to identify their own relatives
- The search returned a 3rd cousin of the killer (sharing a pair of great-great-grandparents)
  - With manual work, police traced down the suspect based on the info about the 3rd cousin
- Police took a DNA sample of the suspect – which matched to the killer
- In 2020, the suspect confessed and was sentenced for life

FINDING RELATIVES FROM DATABASE

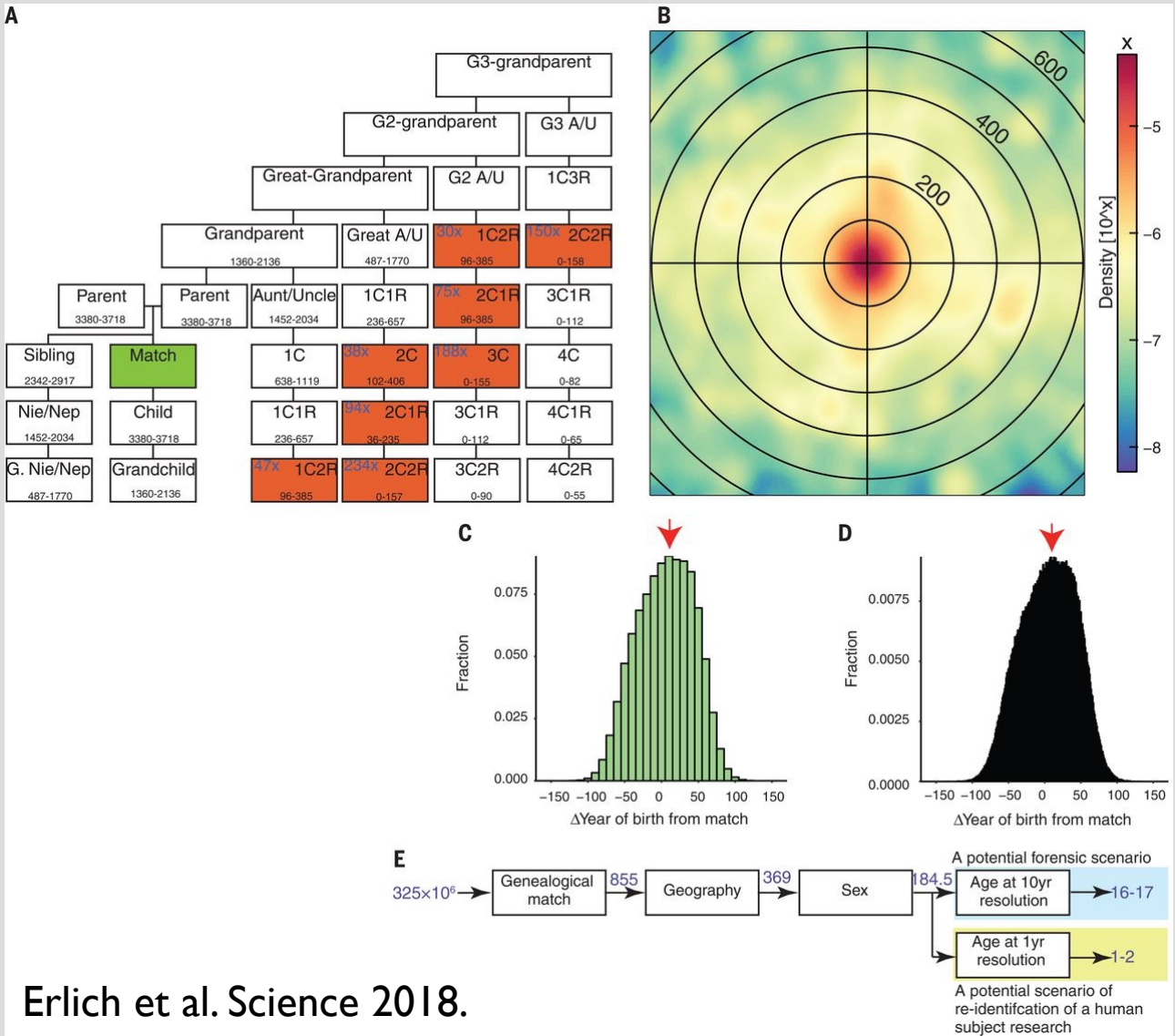


Erlich et al. Science 2018.

A) The probability of finding at least one relative for various IBD thresholds (top) with 1.28 million searches of DTC-tested individuals (red) and 30 random GEDmatch searches (gray). Light gray shading indicates the 95% CI for the GEDmatch estimates. The dashed line indicates the probability of a surname inference from Y chromosome data. The bottom panel shows the 95% CIs (circles) and average total IBD length (squares) for a first cousin once removed (1C1R) to a fourth cousin once removed (4C1R).

(**B**) A population-genetic theoretical model for the probability of finding relatives up to a certain type of cousinship as a function of the database coverage of the population. 1C to 4C indicate first to fourth cousins.

FROM RELATIVE TO MATCH

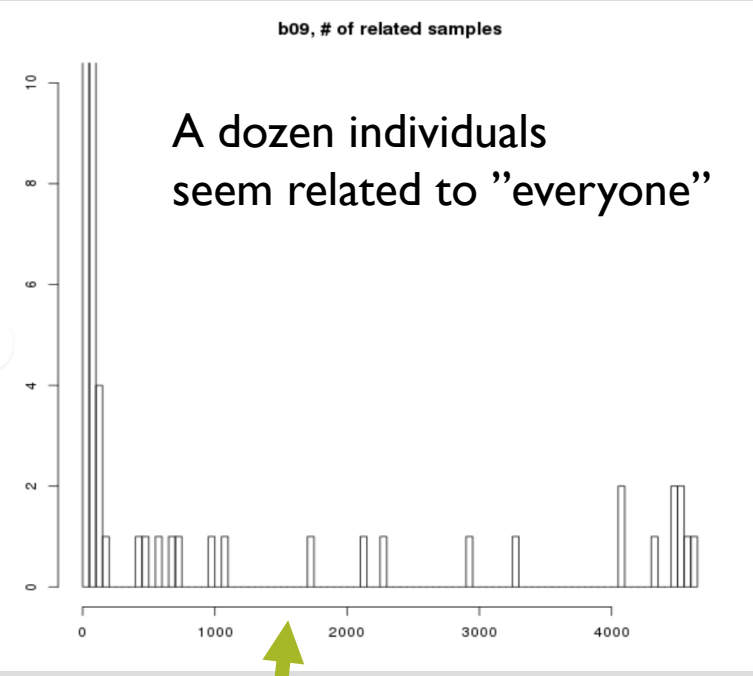


Erlich et al. Science 2018.

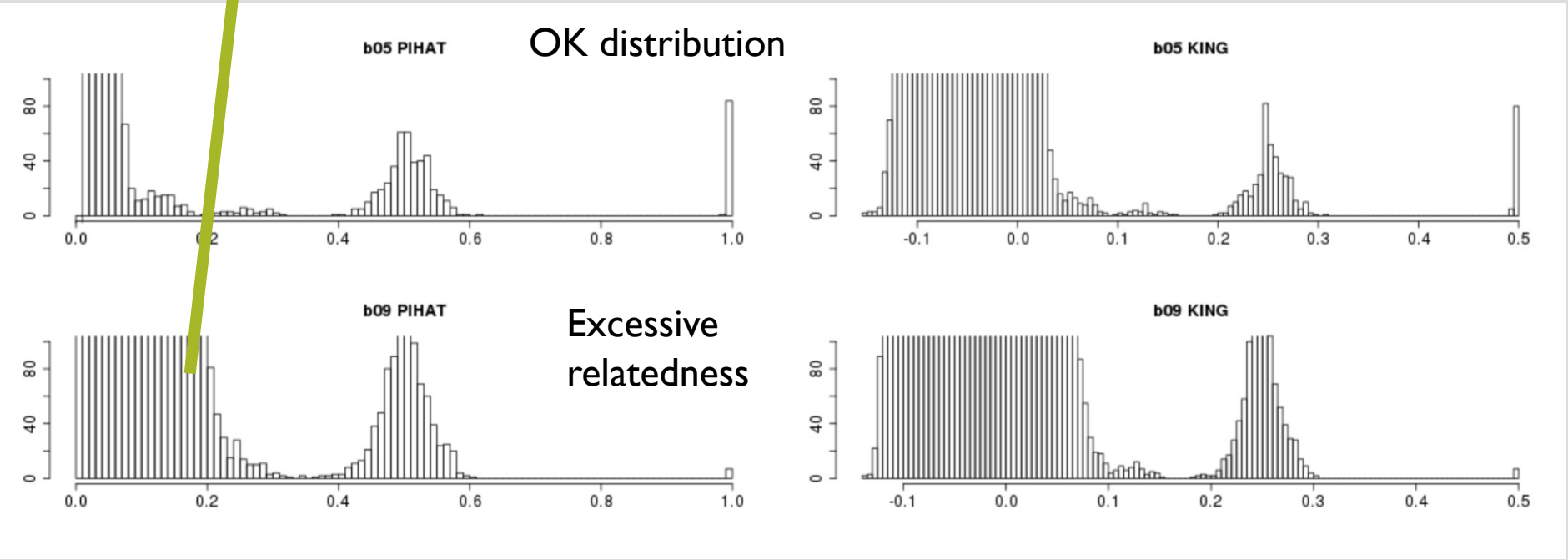
Tracing a person of interest from a distant match using demographic identifiers. (**A**) The possible relatives of a match (green) in a database. Each square represents a potential degree of relatedness. The range corresponds to the 5th to 95th percentile of shared IBD in centimorgans from (*16*). Red indicates relatives that could fit a bona fide 3C match (~100 cM). The average number of relatives is indicated in the top-left corner of each square on the basis of a fertility rate of 2.5 children per couple. Only genealogical relationships that are within 100-cM range include the average number of relatives. (**B**) An example of the geographical dispersion of 3C or 2C1R around the matched relative. Every circle indicates 100 km.

(**C-D**) The distribution of the expected age differences between matches and their potential relatives with a genetic distance of third cousins. The age distribution is shown at a 10-year resolution (C) and at a 1-year resolution (D). (**E**) The entire pipeline of using demographic identifiers along with a long-range familial match to identify a U.S. person (blue type indicates the average number of people after incorporating each piece of information.).

RELATEDNESS AS QC TOOL



- High relatedness to many individuals can imply contamination
  - When DNA from many samples are mixed the result has excessive heterozygozity



FINNGEN

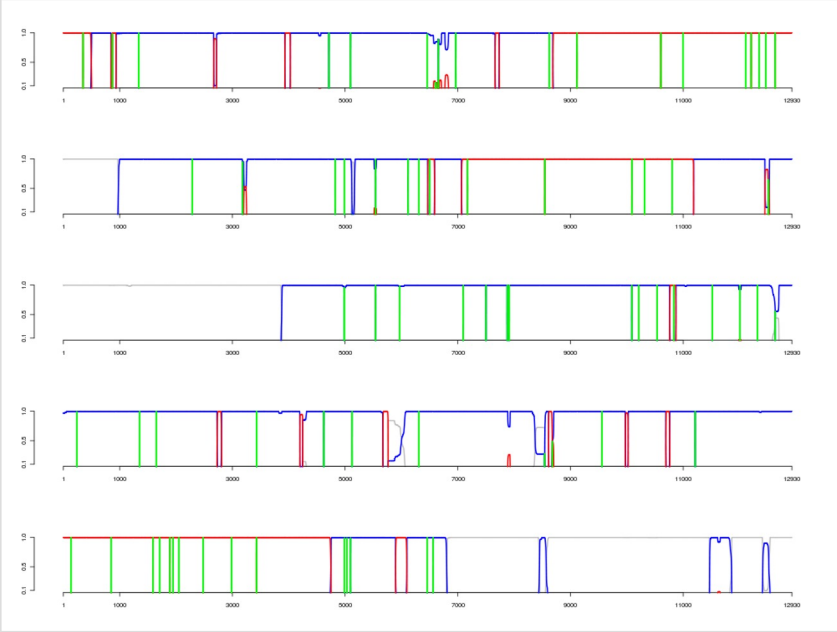
Priit Palta &
Sequencing
Informatics
team

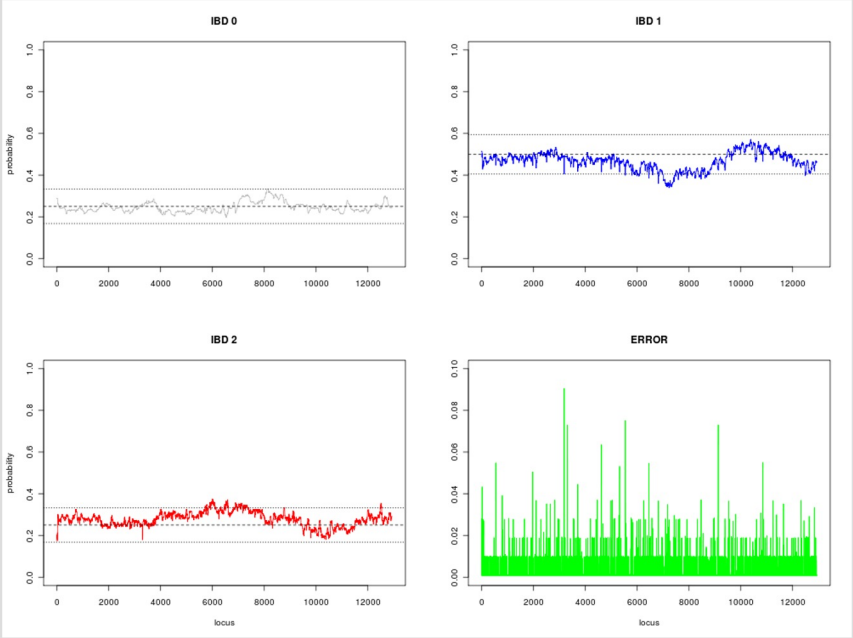
USING IBD IN FULL SIBS TO IDENTIFY PROBLEMATIC SNPS

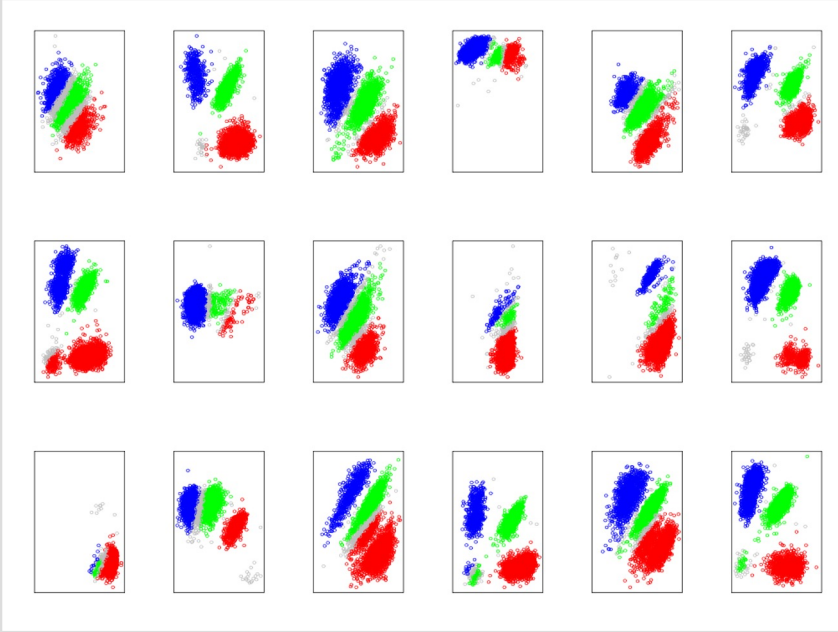
1. For each pair use a hidden Markov Model to model IBD state along the genome.
Allows 4 states: IBD0, IBD1, IBD2 and ERROR.

2. Check the average of each state along the genome.
the ones with exceptionally many ERRORs are suspicious SNPs.

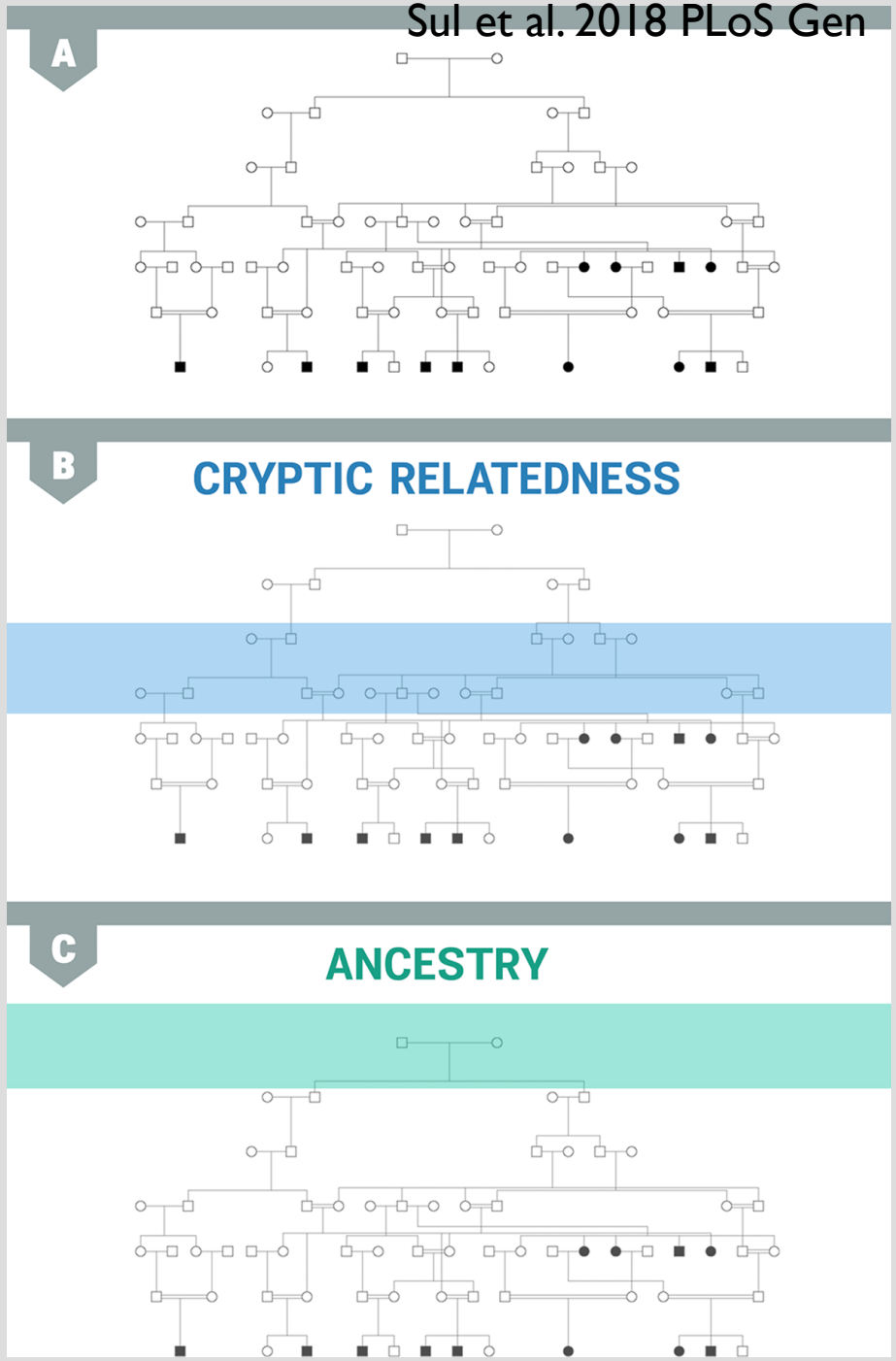
3. Manually check (some) of the suspicious SNPs to see whether there are obvious calling problems.
(Yes there are!)







Sul et al. 2018 PLoS Gen



POPULATION STRUCTURE

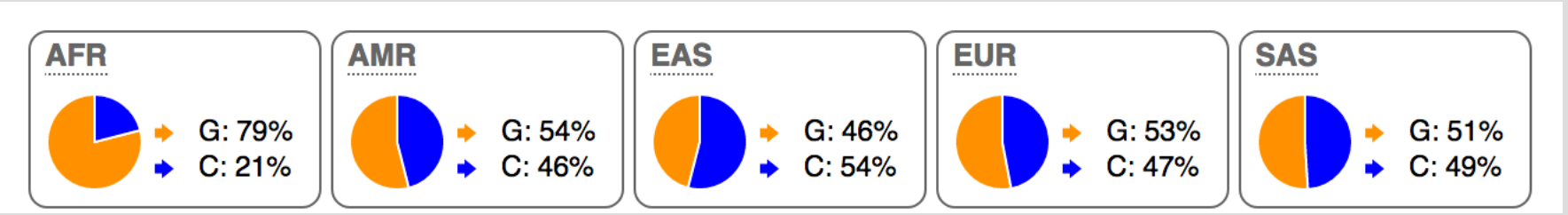
So far we have talked about close relatives that share Relatively long segment(s) of DNA.
Now we move to more distant relatedness that we call ”population structure”.

Populations = ”groups of individuals who **on average** are more closely related to the other members of the same group than to the members of another group”

Asian population vs European population
Finnish population vs Swedish population
Eastern Finns vs Western Finns

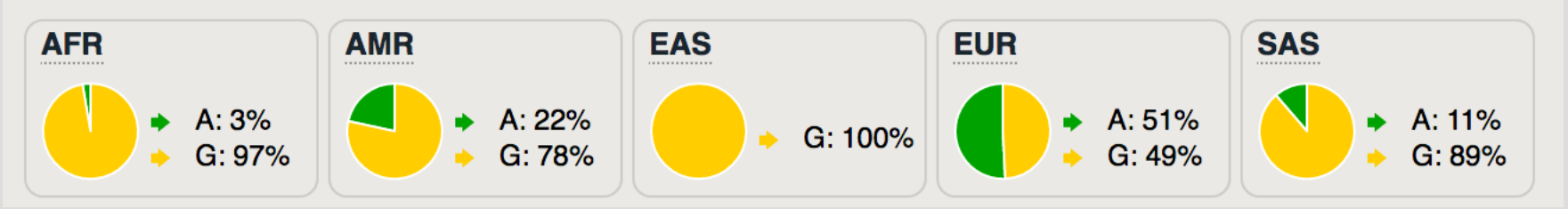
Population membership is not an objectively defined characteristic but depends on the level of detail.

GLOBAL ALLELE FREQUENCY DIFFERENCES

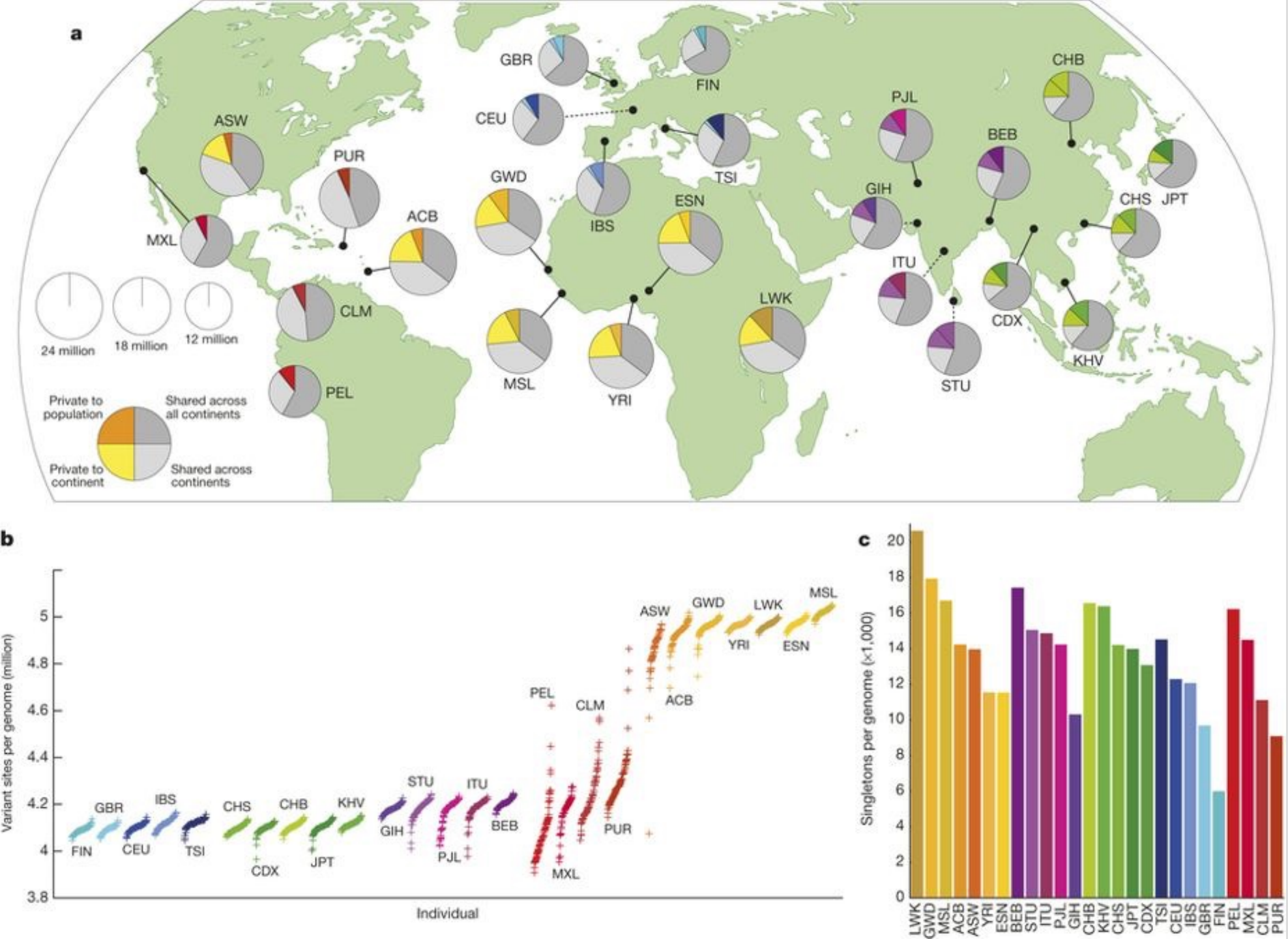


A ”random” SNP

Frequencies can change due to genetic drift, selection, admixture or all of them.



rs4988235 (Lactase tolerance in Europeans, selection has had effect)



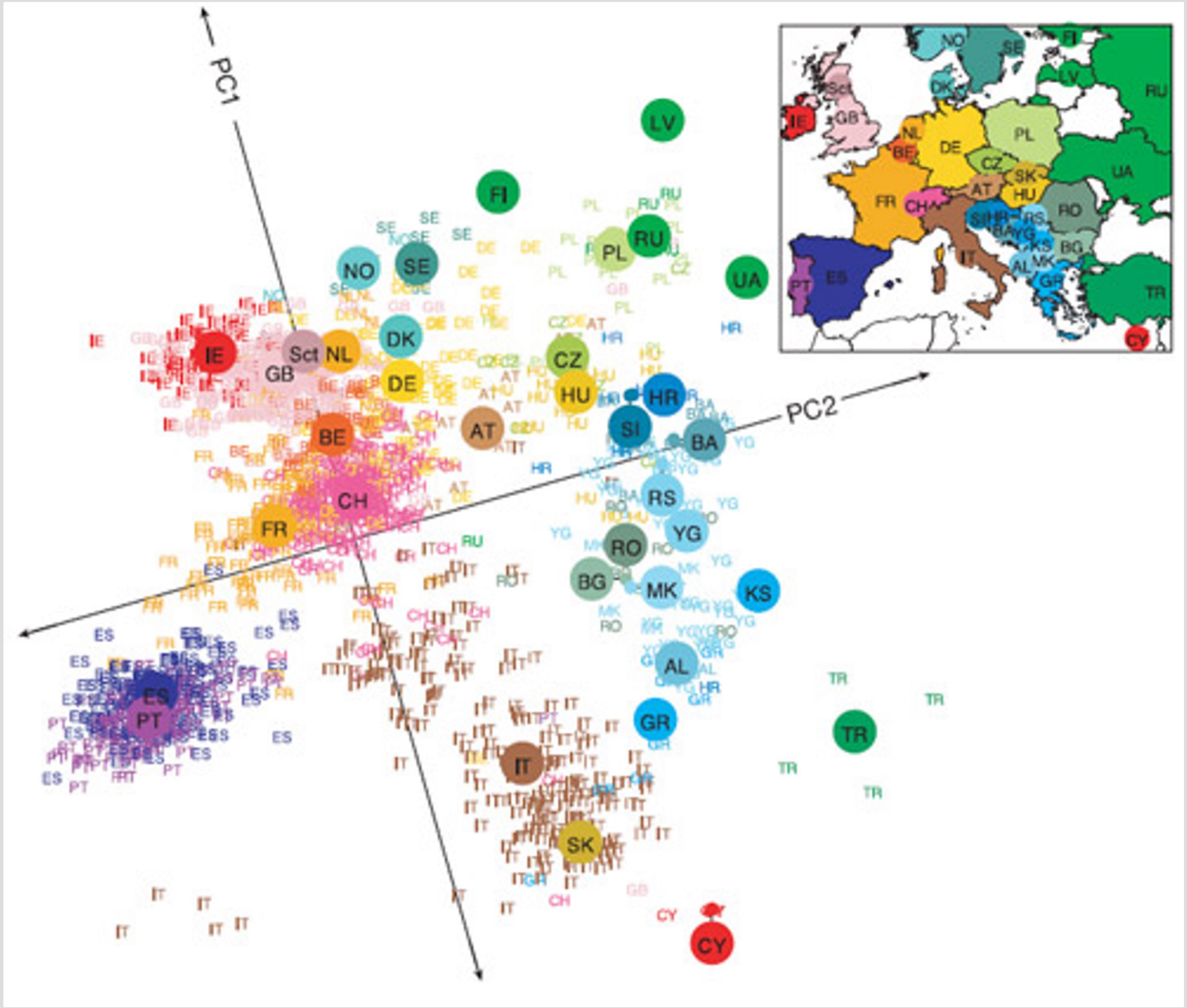
1000 Genomes project
Nature 2015
Figure 1.

a,
Polymorphic variants.
The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, according to sharing pattern across globe.

b,
The number of variant sites per genome.

c,
The average number of singletons per genome.

PCA OF EUROPEANS

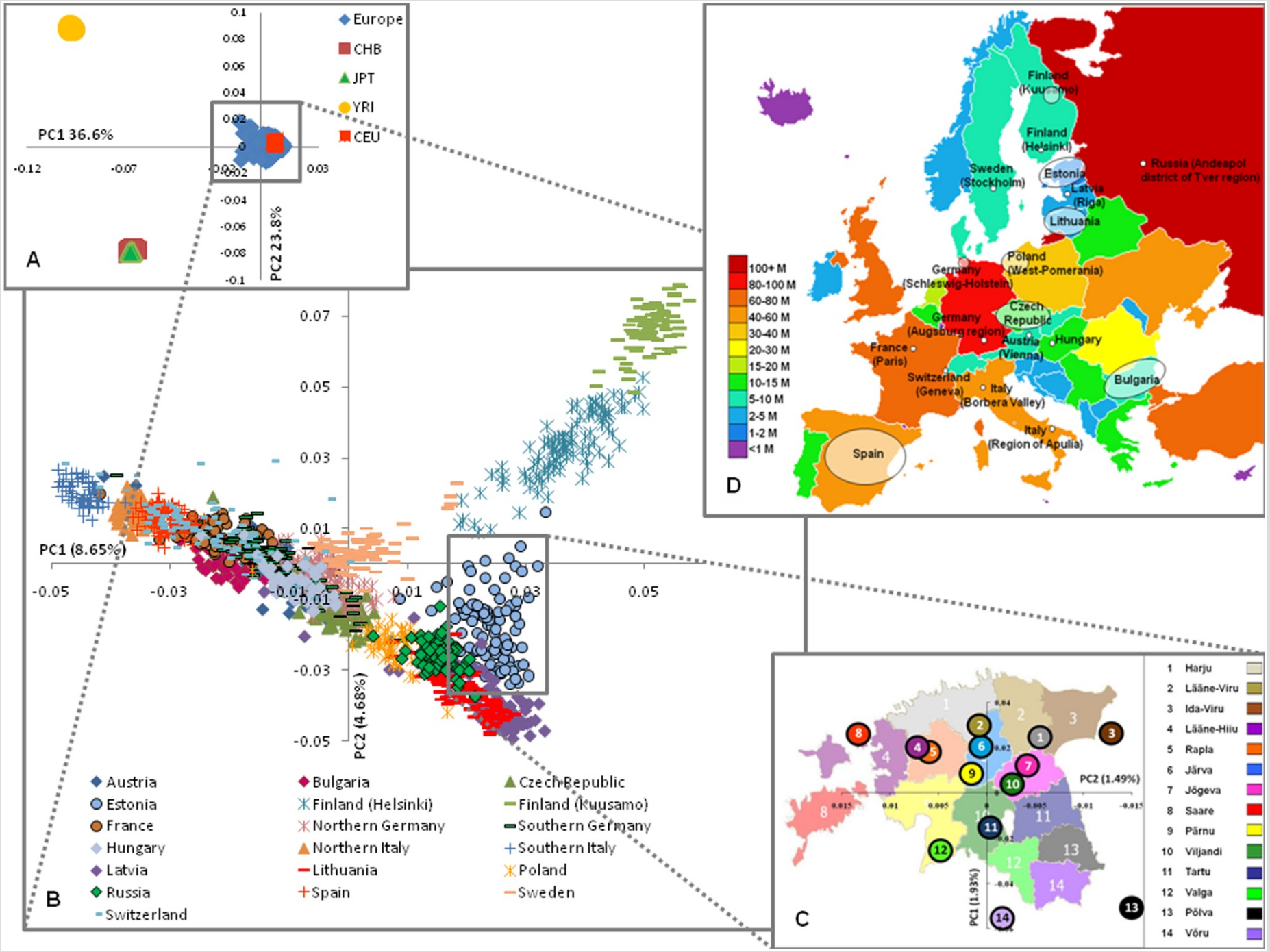


"Genes mirror geography within Europe"
Novembre et al. 2008, Nat Gen

1400 individuals from Europe
200,000 SNPs

In Fig., PC1 and PC2 rotated to match best the geographic map of Europe

NELIS ET AL. PLOS ONE 2009



Three levels of structure as revealed by PC analysis:

A) inter-continental;

B) intra-continental; and

C) inside Estonia,

where median values of the PC1&2 are shown.

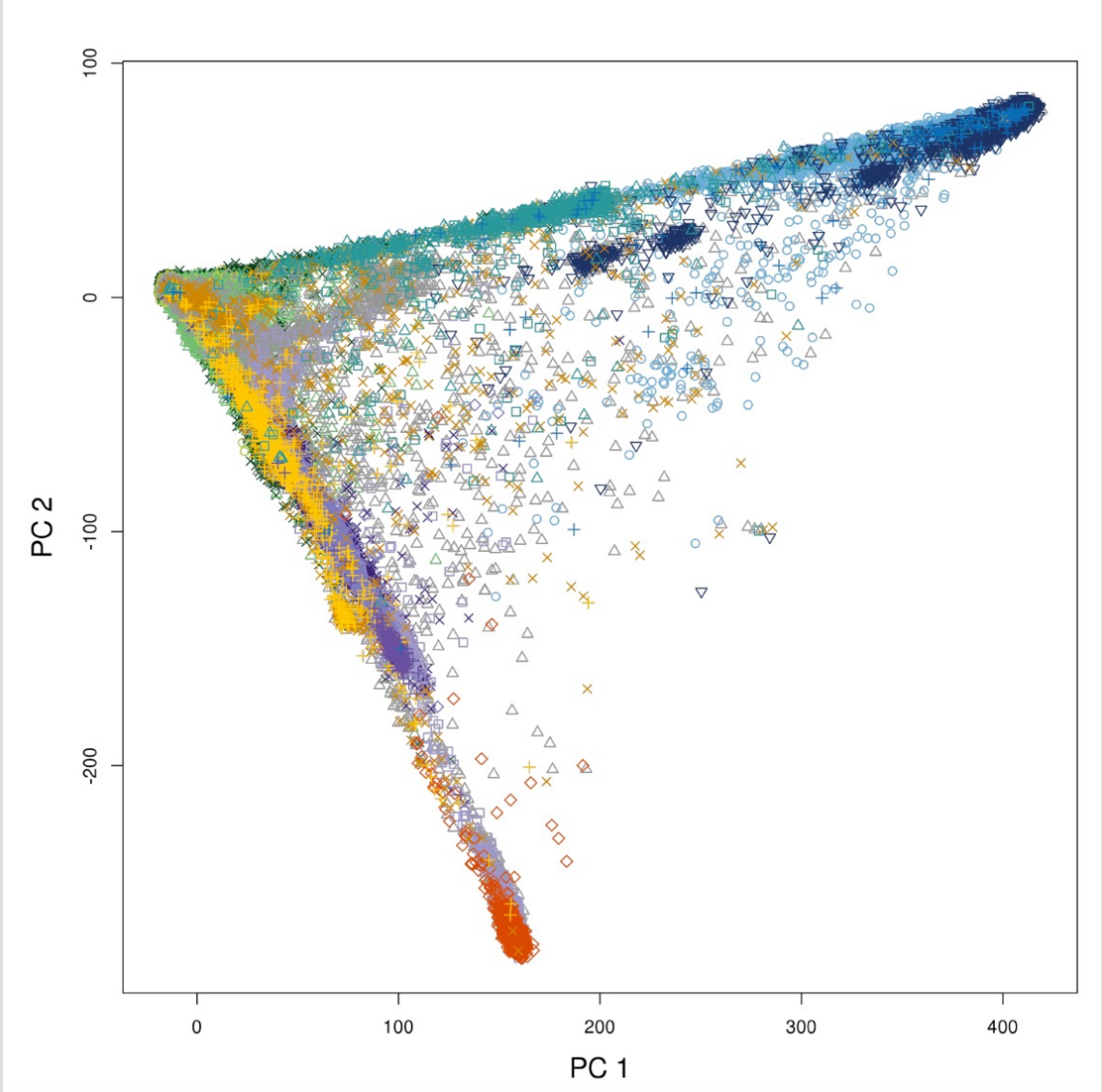
D) European map illustrating the origin of sample and population size.

CEU - Utah residents with ancestry from Northern and Western Europe,

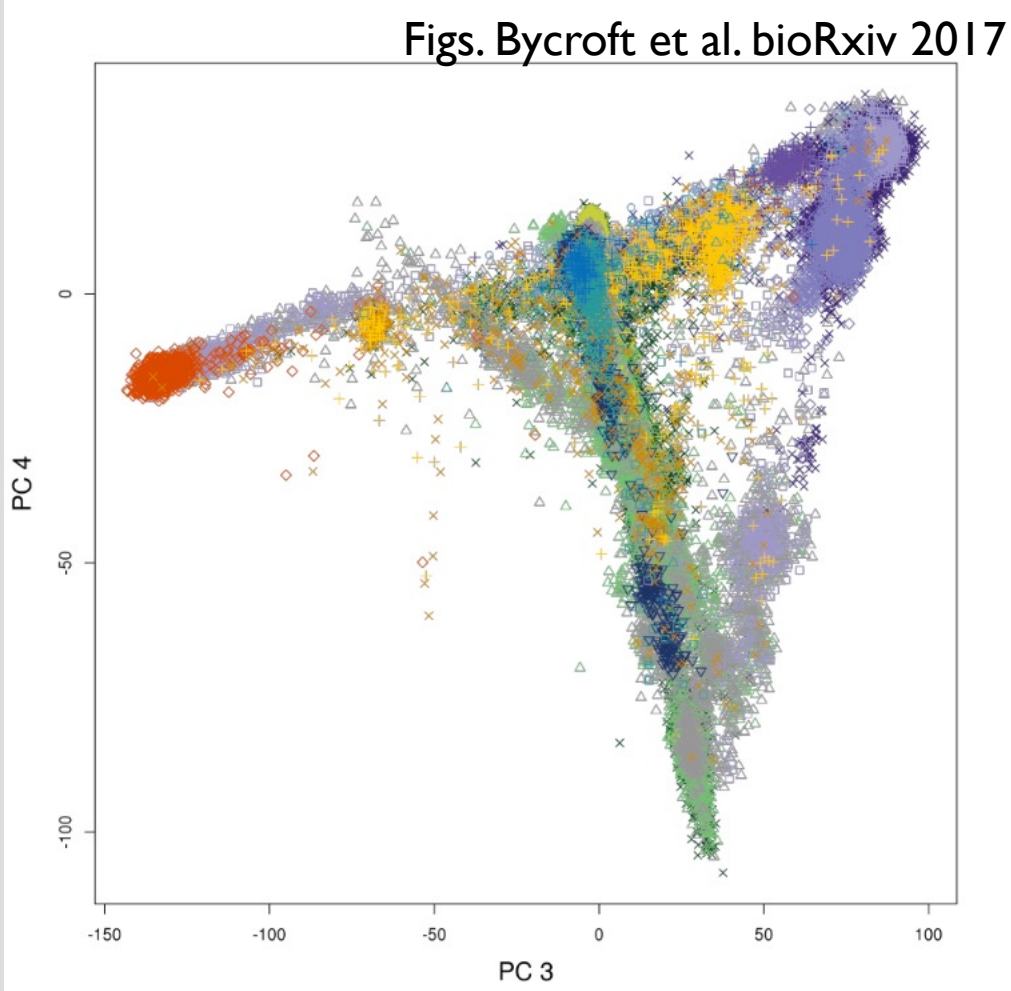
CHB – Han Chinese from Beijing,

JPT - Japanese from Tokyo, and

YRI - Yoruba from Ibadan, Nigeria.



Figs. Bycroft et al. bioRxiv 2017

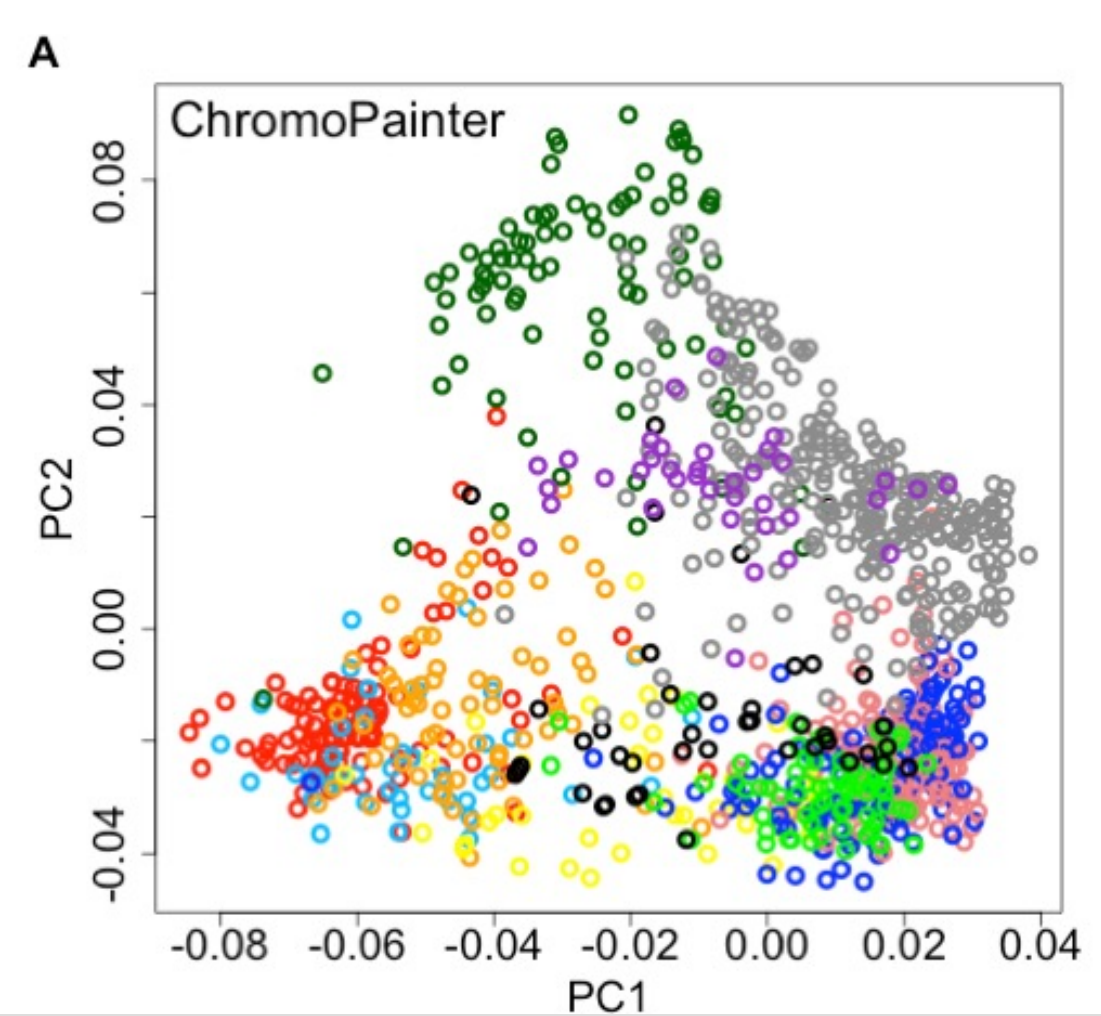


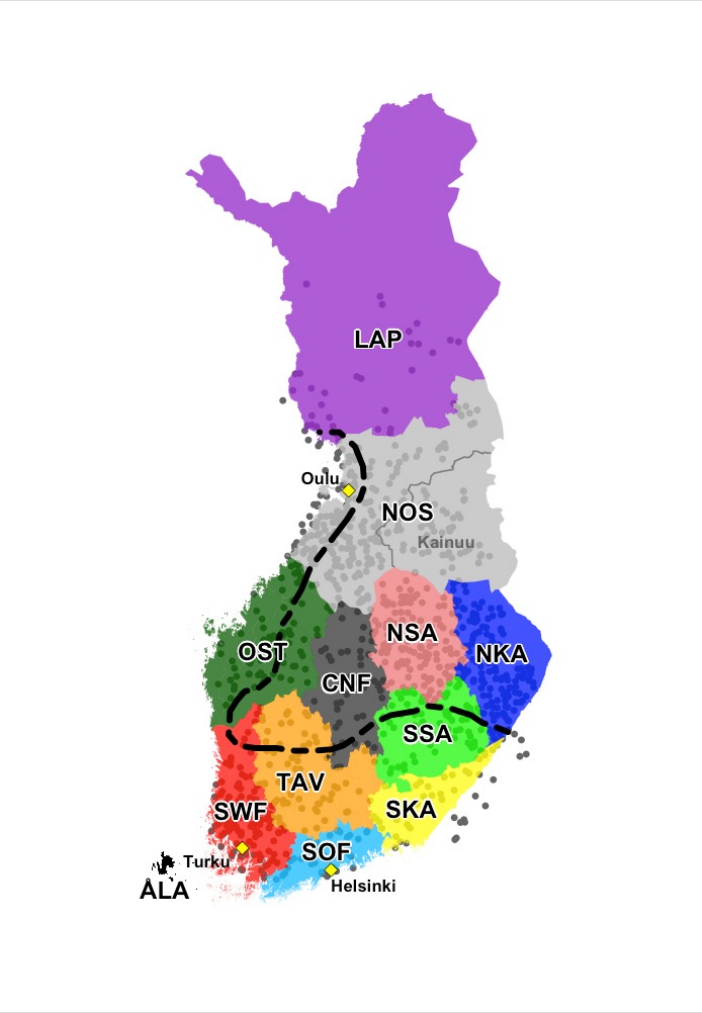
Self-reported ethnic background

- British
- Irish
- Any other white background
- African
- Caribbean
- Any other Black background
- Indian
- Pakistani
- Bangladeshi
- Any other Asian background
- White and Asian
- White and Black African
- White and Black Caribbean
- Any other mixed background
- Chinese
- Other/Unknown

- UK Biobank genetic data with n=**407,000** and p=150,000,
- Done with FastPCA (Galinsky et al 2015) that implements *blanczos* algorithm, a stable version of power iteration

PCA WITHIN FINLAND

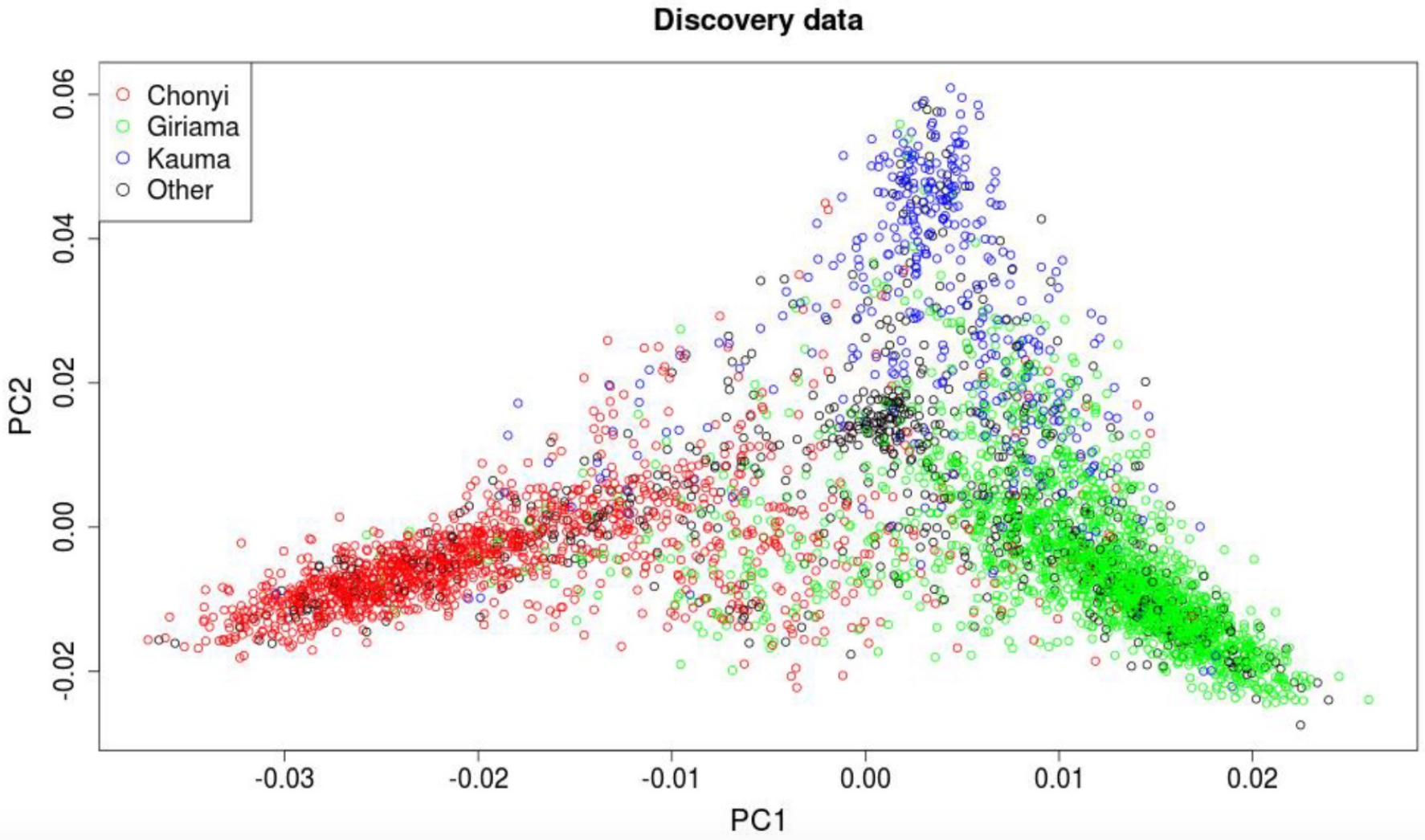




Kerminen et al. 2017 G3. 1042 Finns with both parents born < 80km. PCA on left colored according to regions on right.

Figure S5. The first and second axes of the principal components analysis in the discovery dataset using 168,217 SNPs. Color coding represents A) reported ethnicity

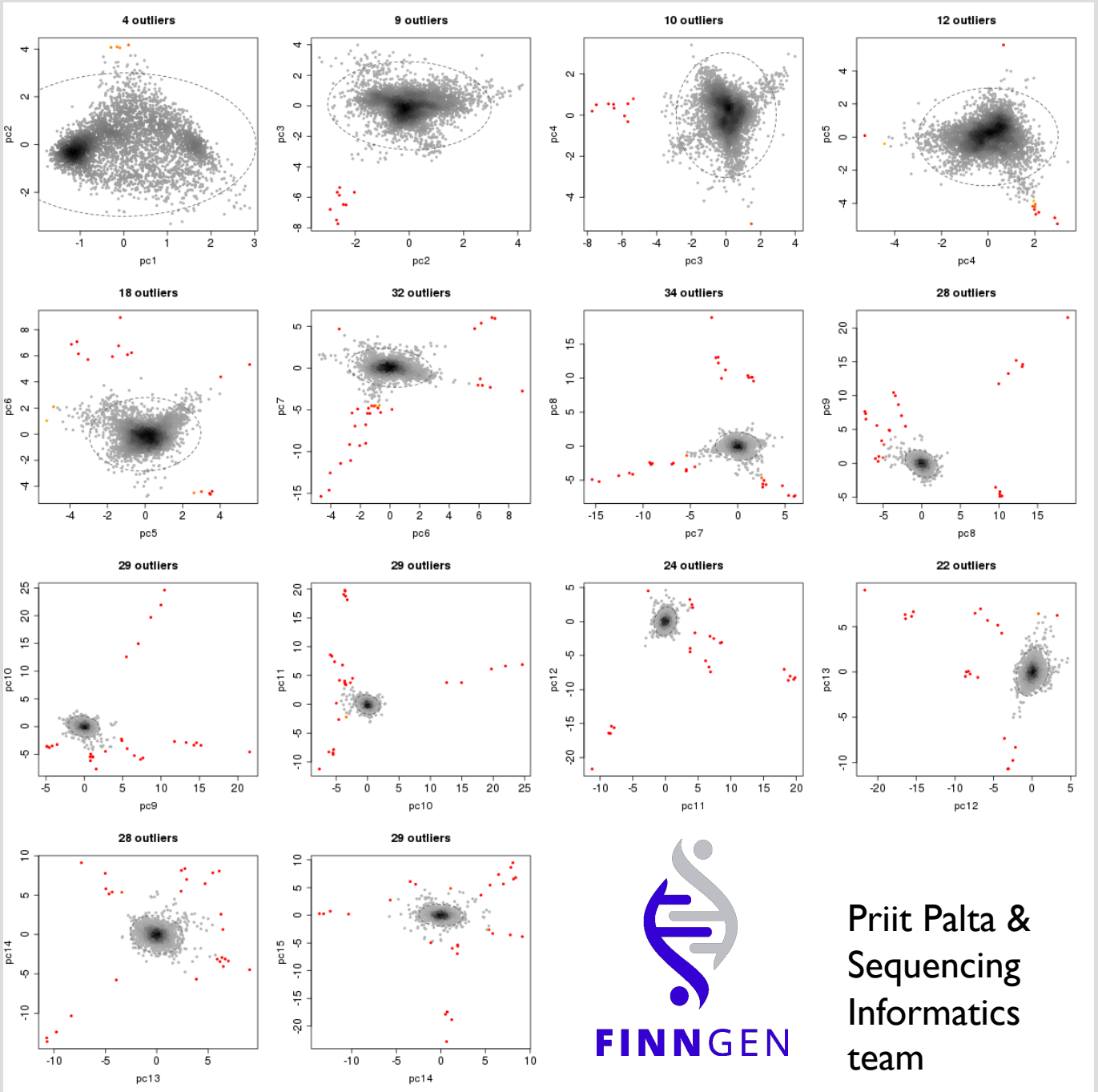
A)



From a study of genetics of Bacteraemia in Kenyan Children

First 2 PCs associate with ethnic group.

(Rautanen et al. 2016 AJHG)



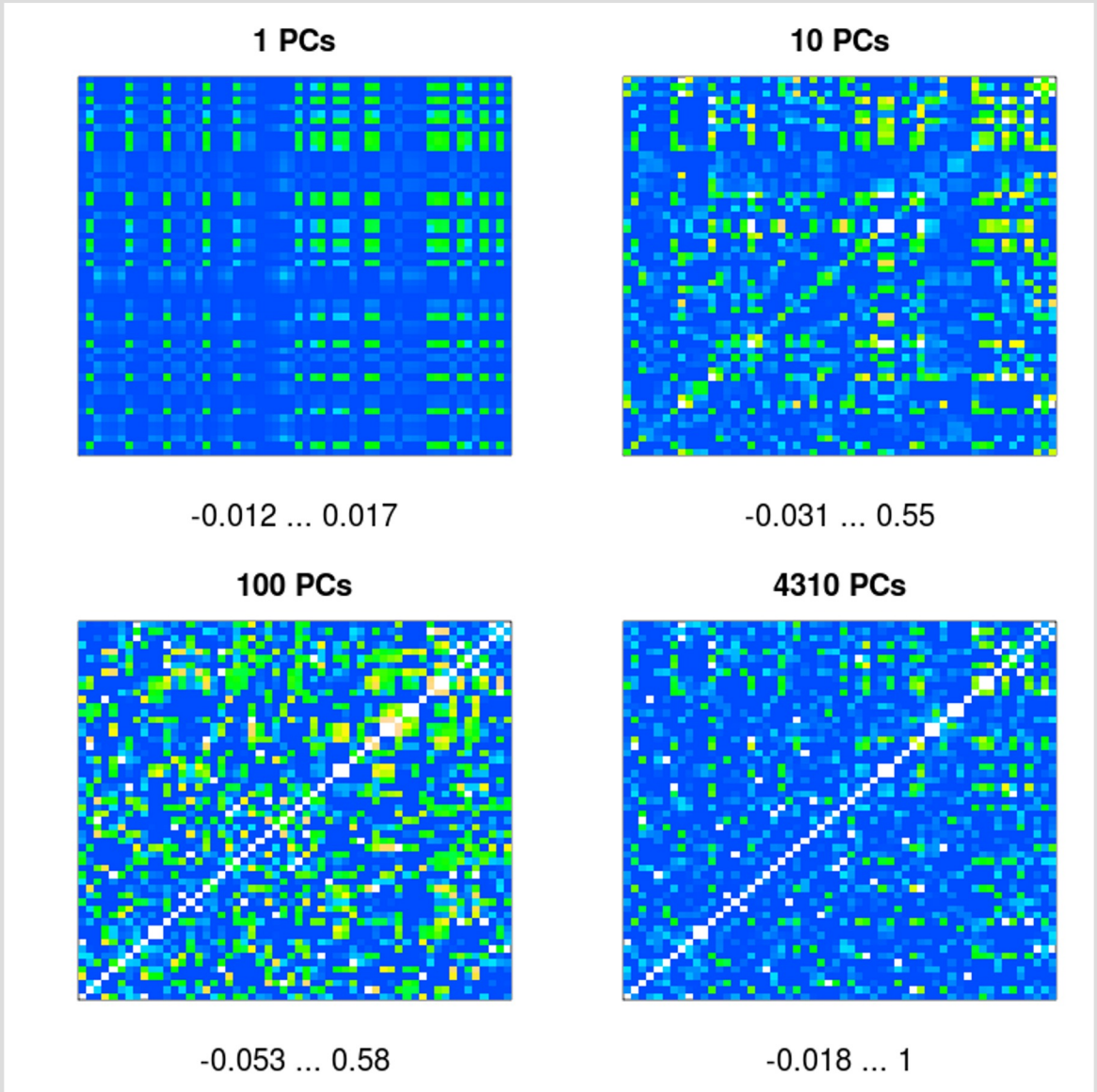
PCA AS QUALITY CONTROL METHOD

- PCA can be used to inform if data have samples with
  - relative groups
  - different ancestry
  - technical problems (e.g. contamination)
- These are typically removed from further analyses

FINNGEN

Priit Palta & Sequencing Informatics team

GRM-COR IS BUILT FROM PCS



$$R = UDU^T = \sum_{i=1}^{n} d_i U_{\cdot i} U_{\cdot i}^T$$

R is genetic relatedness matrix (GRM-cor)
U has eigenvectors of R as columns
D has eigenvalues on diagonal
$U_{\cdot i}$ is the i^{th} column, i.e., scores of individuals on PC i
$U_{\cdot i}\ U_{\cdot i}^T$ is a rank-1 matrix

Any subset of PCs corresponds to a truncated GRM