From Bologna to Melbourne: a research experience in modern statistical genetics

Anna Quaglieri
University of Bologna
22nd January 2016

Walter and Eliza Hall Institute of Medical Research

15 different scientific divisions

Wehi divisions

Walter+Eliza Hall
Institute of Medical Research

BIOINFORMATIC RESOURCES

Welcome to the Walter and Eliza Hall Institute website for local bioinformatic resources.

WEHI bioinformatic resources

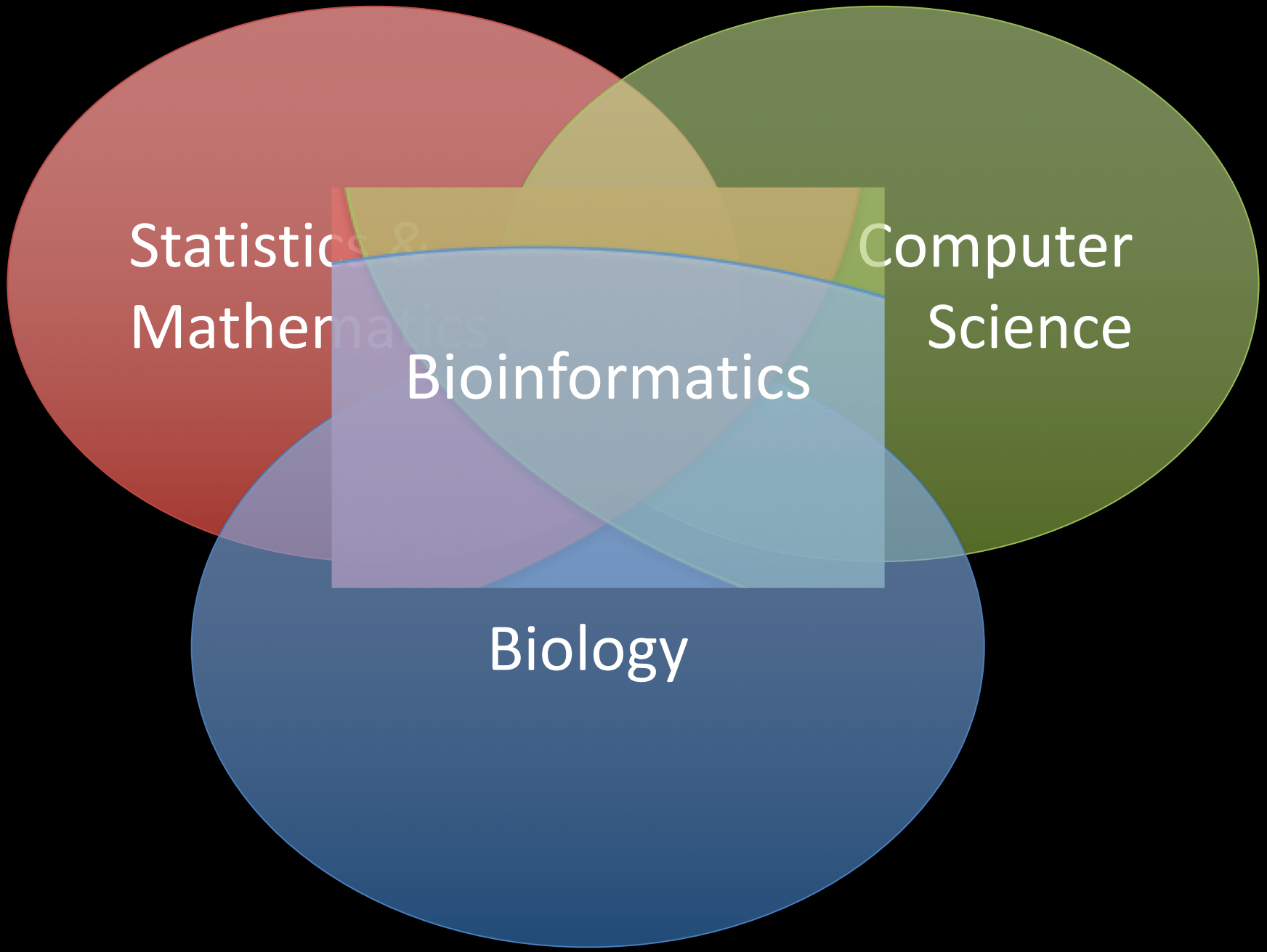
- Software projects
- Supplementary materials for published articles

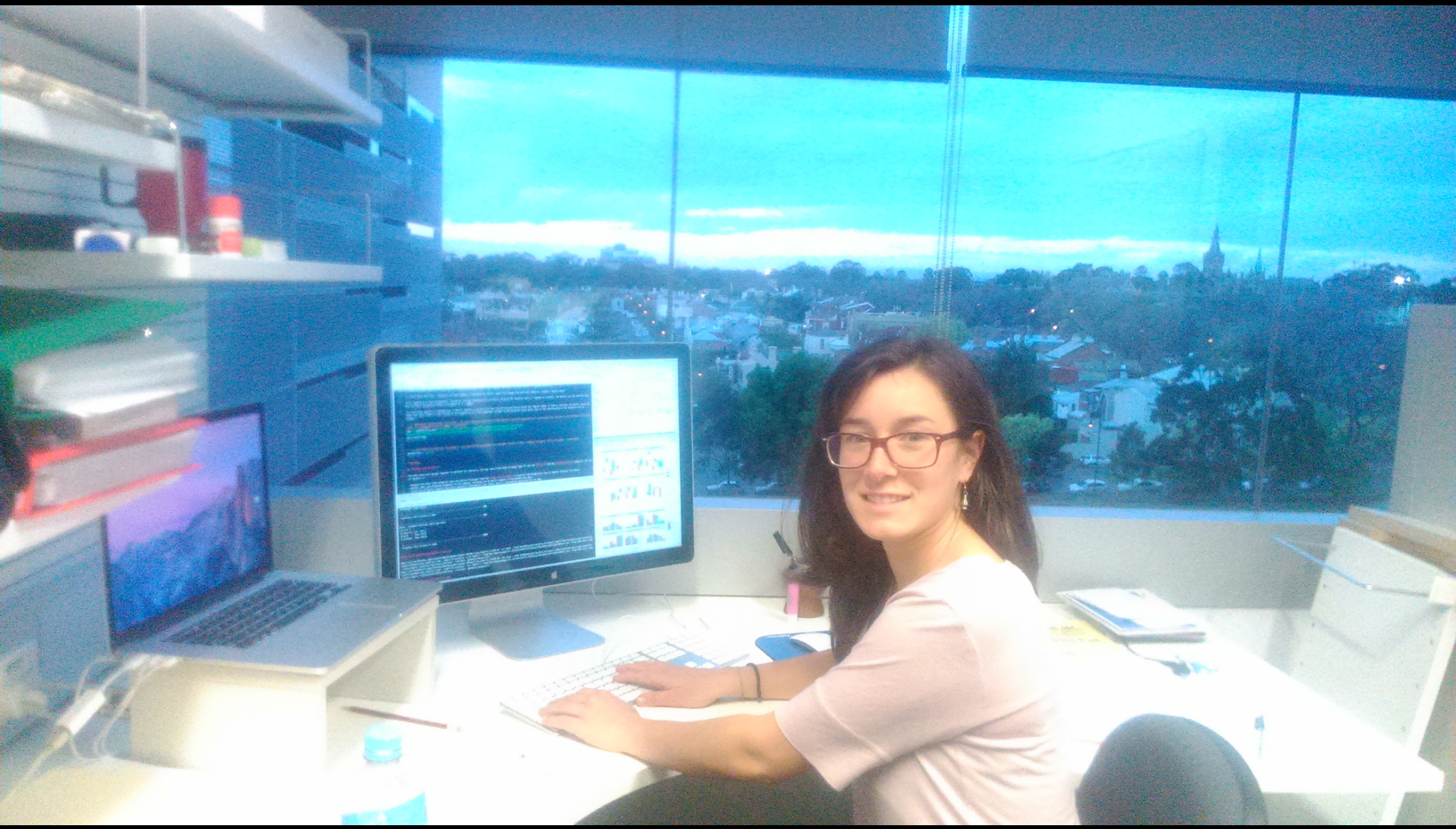
Bioinformatics division seminars

- Current seminar
- All seminars for this year
- Search past seminars

Bioinformatics-orientated research laboratories

- Bahlo Lab
- Papenfuss Lab
- Ritchie Lab
- Smyth Lab
- Speed Lab


Statistics &
Mathematics
Computer
Science
Bioinformatics
Biology



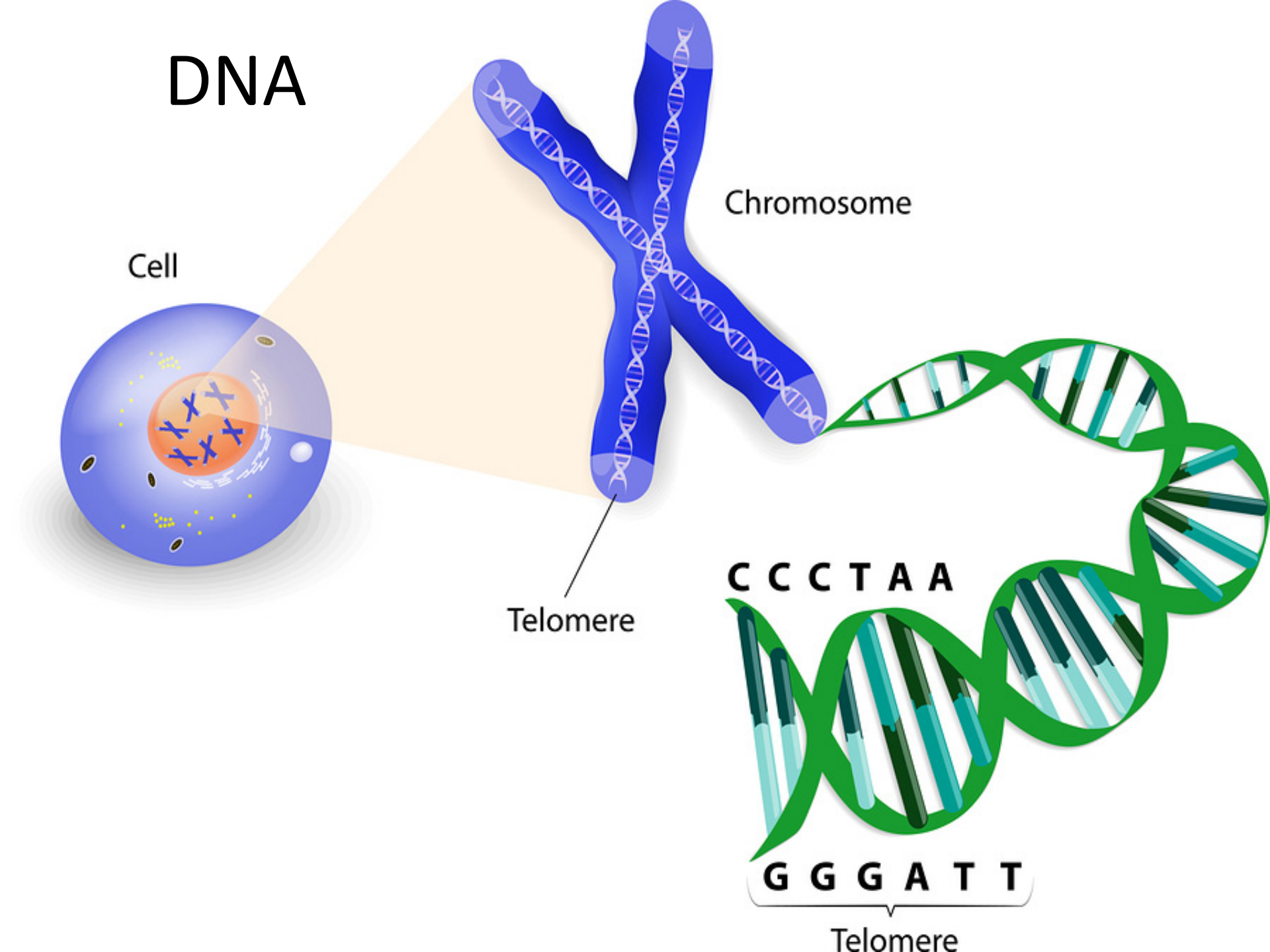

PubMed


R

My main projects

- DNA – GWAS on MacTel – an eye disease
- Protein-DNA interaction data: ChIP-Seq

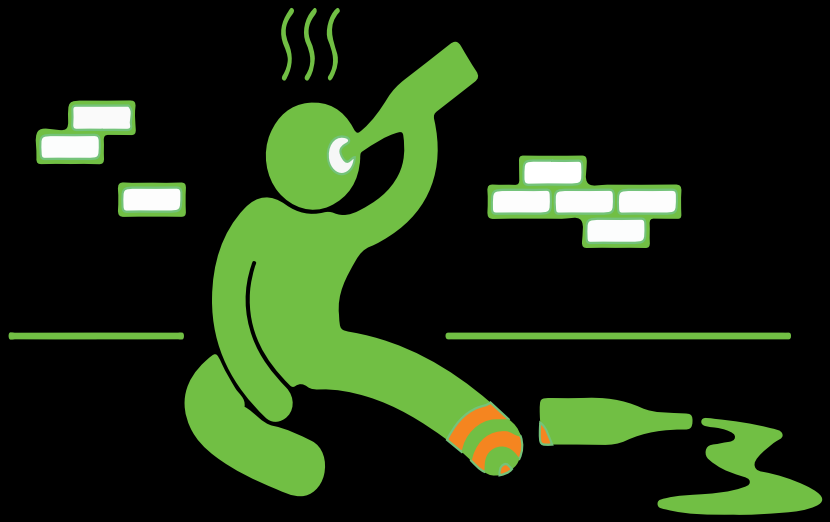
DNA



DNA analysis

- Comparing *people's code*
- 3 billion letters each
- Extremely intensive/expensive task

Select representative letters



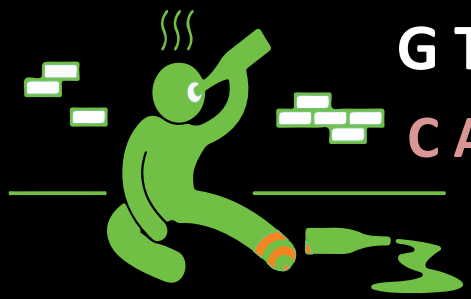
99% the same

GTACTGGTCATTGAG T TCTTTGCA AC

CATG ACCAG TAACTCA A GAAACGTTGAC

GGTCATCGTATGGCAT G ACAAATGCCAT

CCA GTAGATAC CG TA C TGTTTA CGGTA



GTACTGGTCAT TGAG T T CTTTGCA AC

CATTGACCAGTAACTCA A GAAACGTTGAC

GGTCATCGTATGGCAT T ACAAATGCCAT

CCA GTAGATAC CG TA A TGTTTA CGGTA


SNP T/G
T 90%
G 10%
© david mchugh

SNPs as representatives of people’s genome

Associate a SNP to the occurrence of a disease

Vanilla GWAS

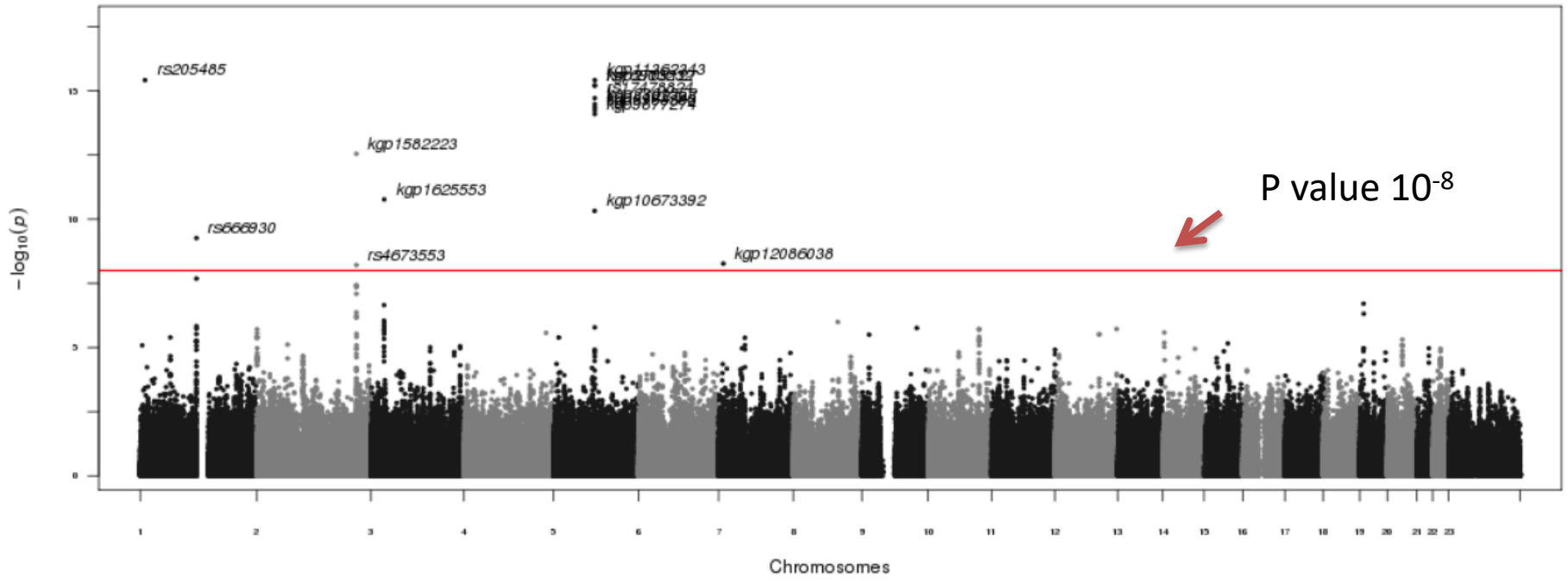
$SNP_{A/B}$

Allele	A	B	Total
Case	n_{11}	n_{12}	N_{10}
Control	n_{21}	n_{22}	N_{20}
Total	N_{01}	N_{02}	N

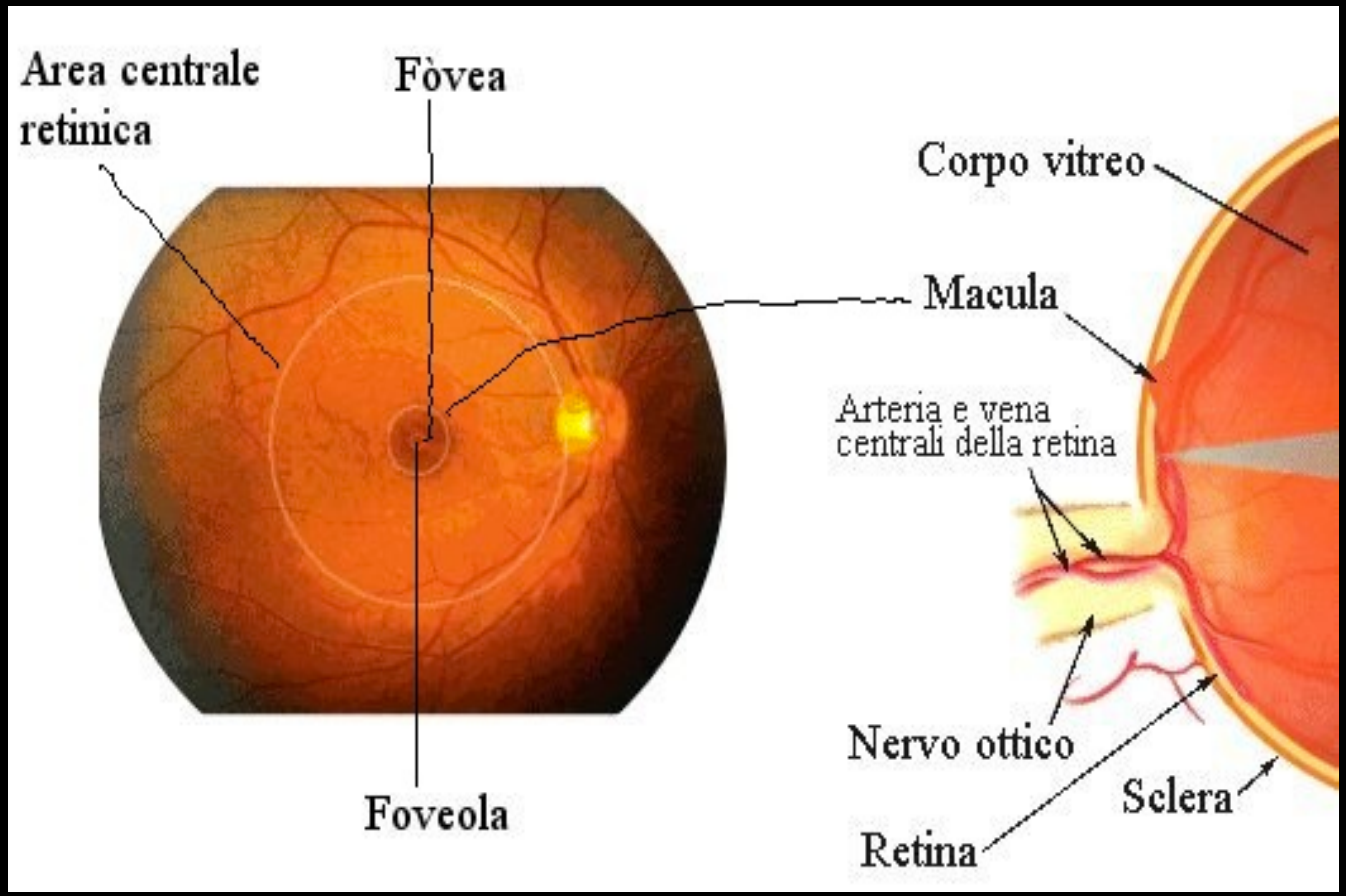
Odds ratios of association SNP by SNP

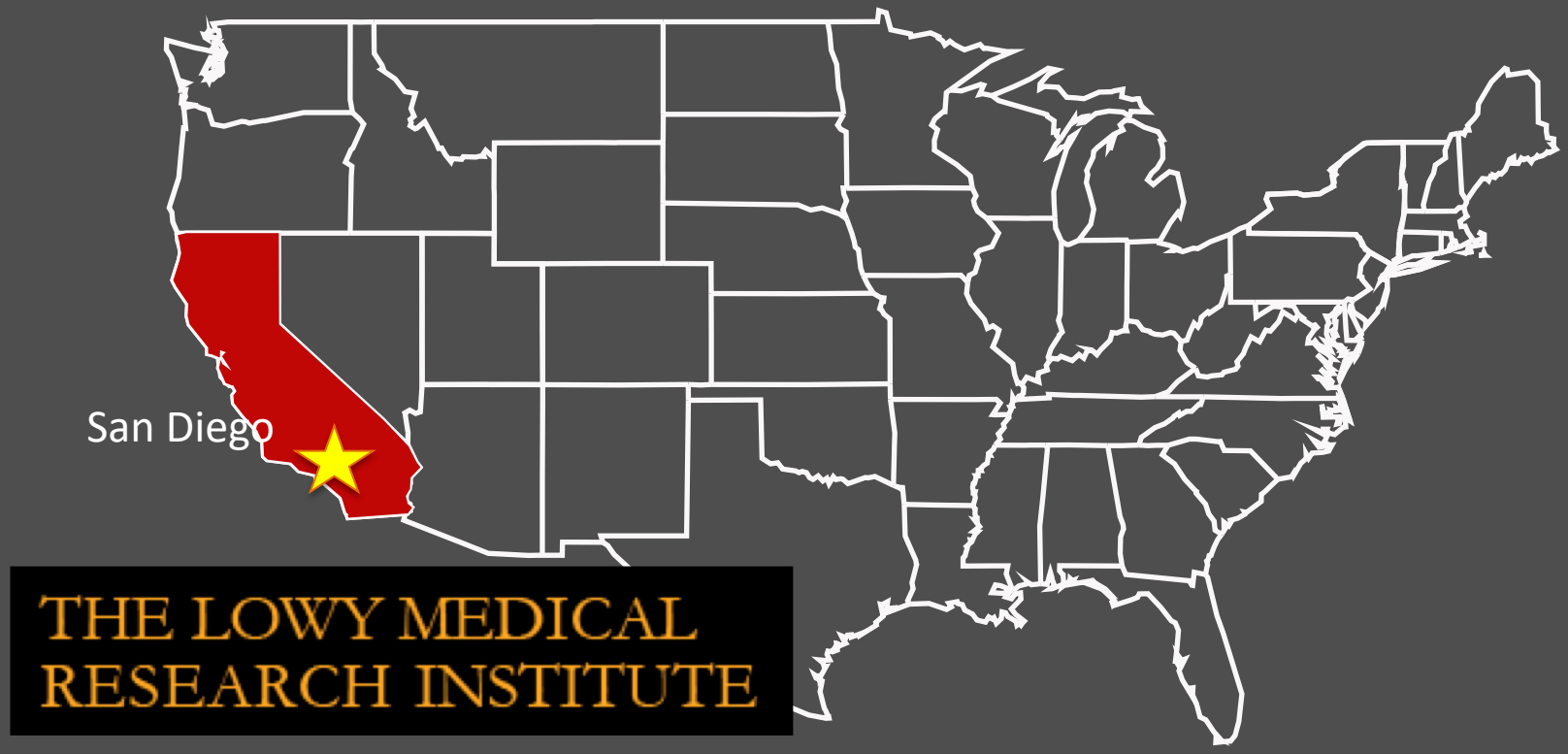
GWAS are association studies. Any discovered SNPs **not** be thought of as **causative**, until investigated further.

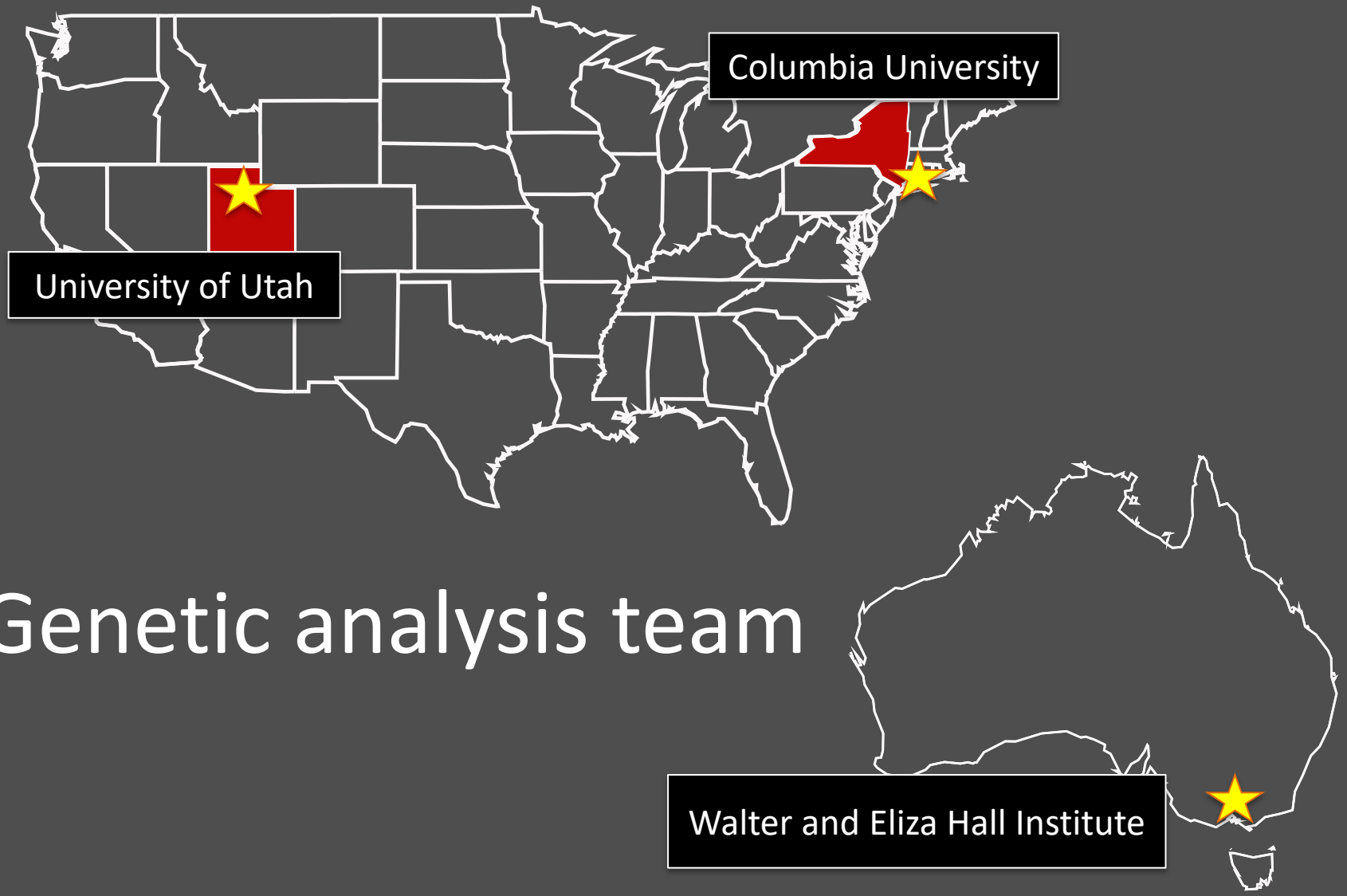
~1million SNPs



Case Study: MacTel



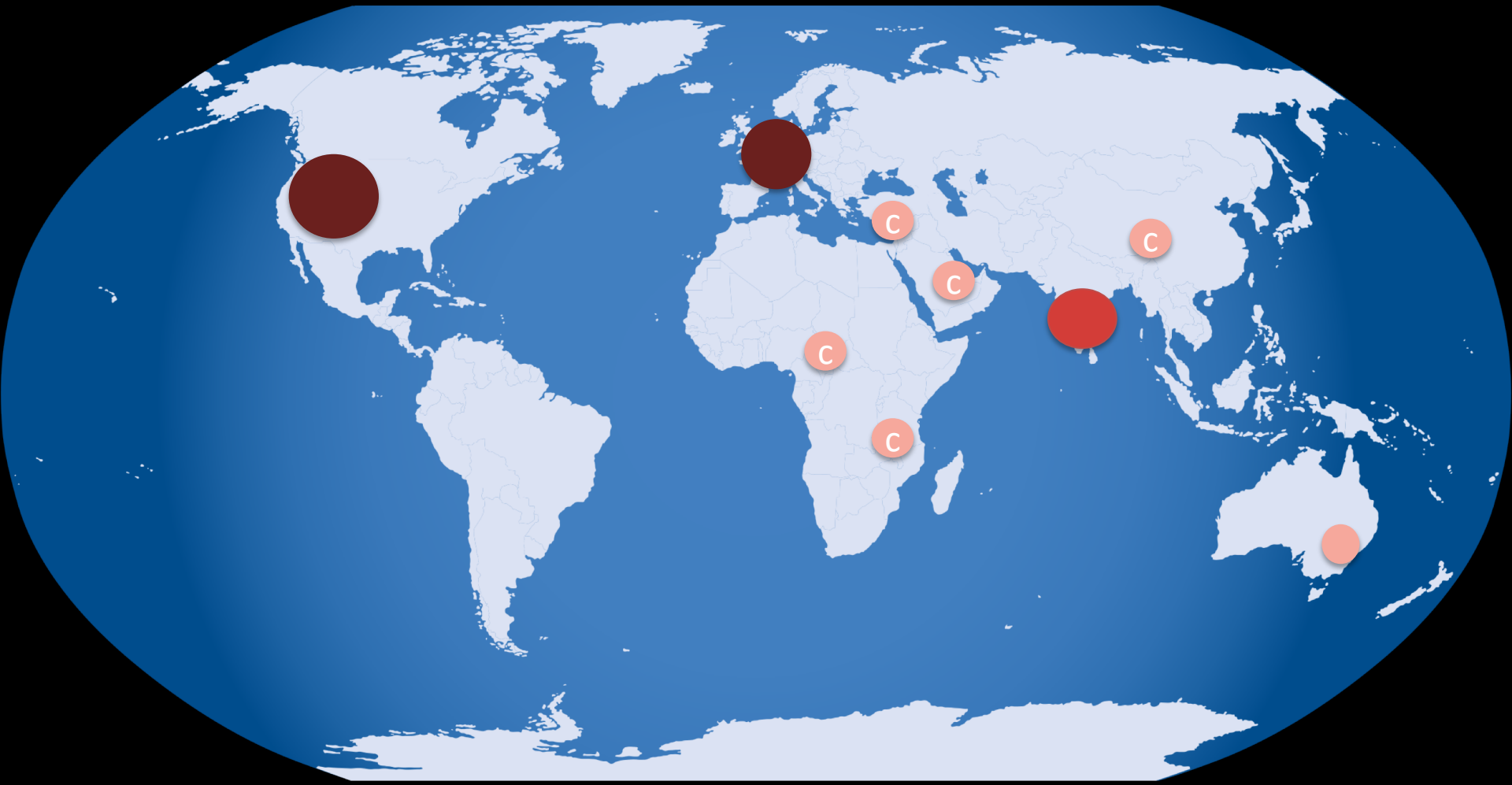

San Diego
THE LOWY MEDICAL
RESEARCH INSTITUTE



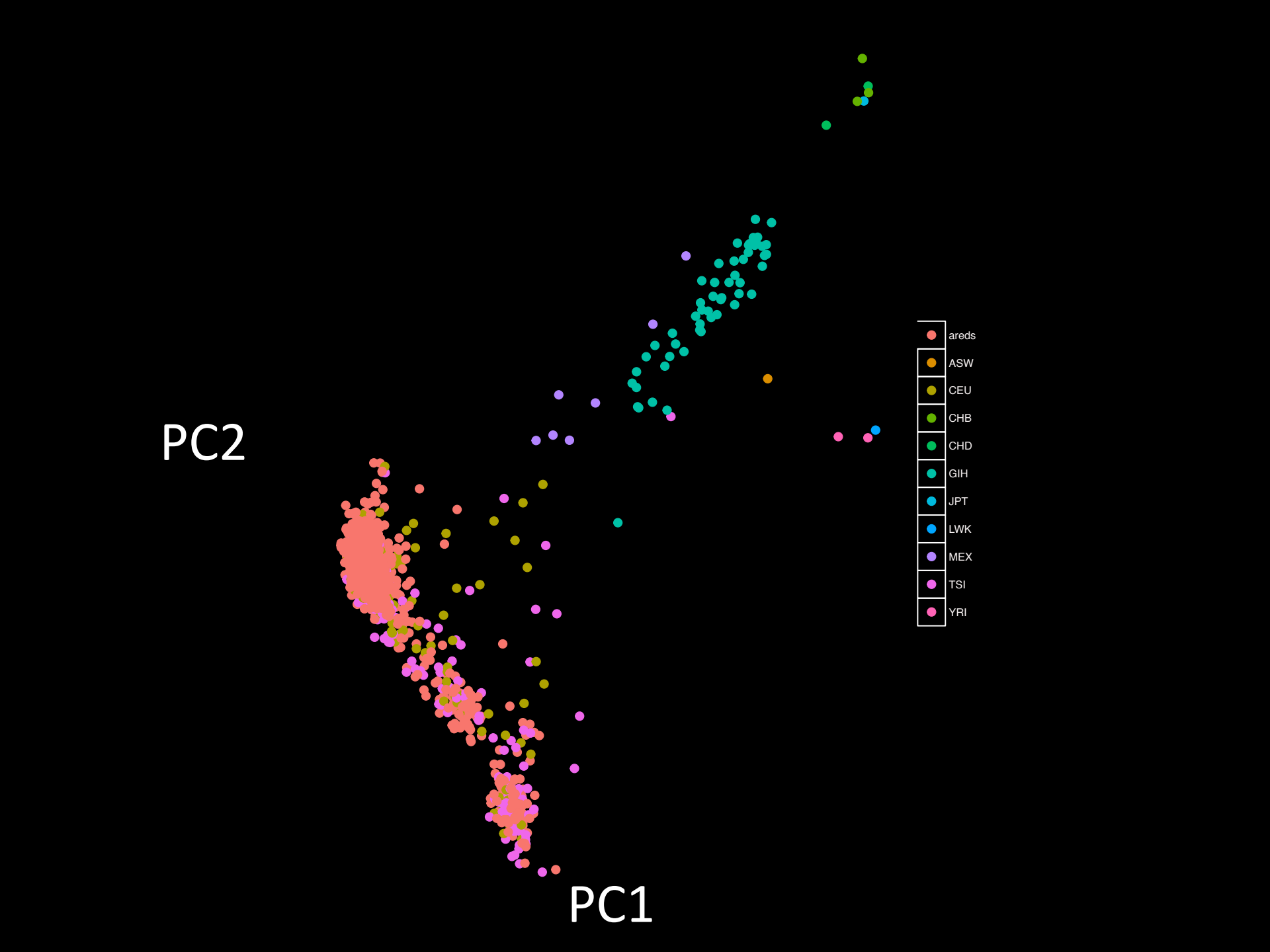
Genetic analysis team

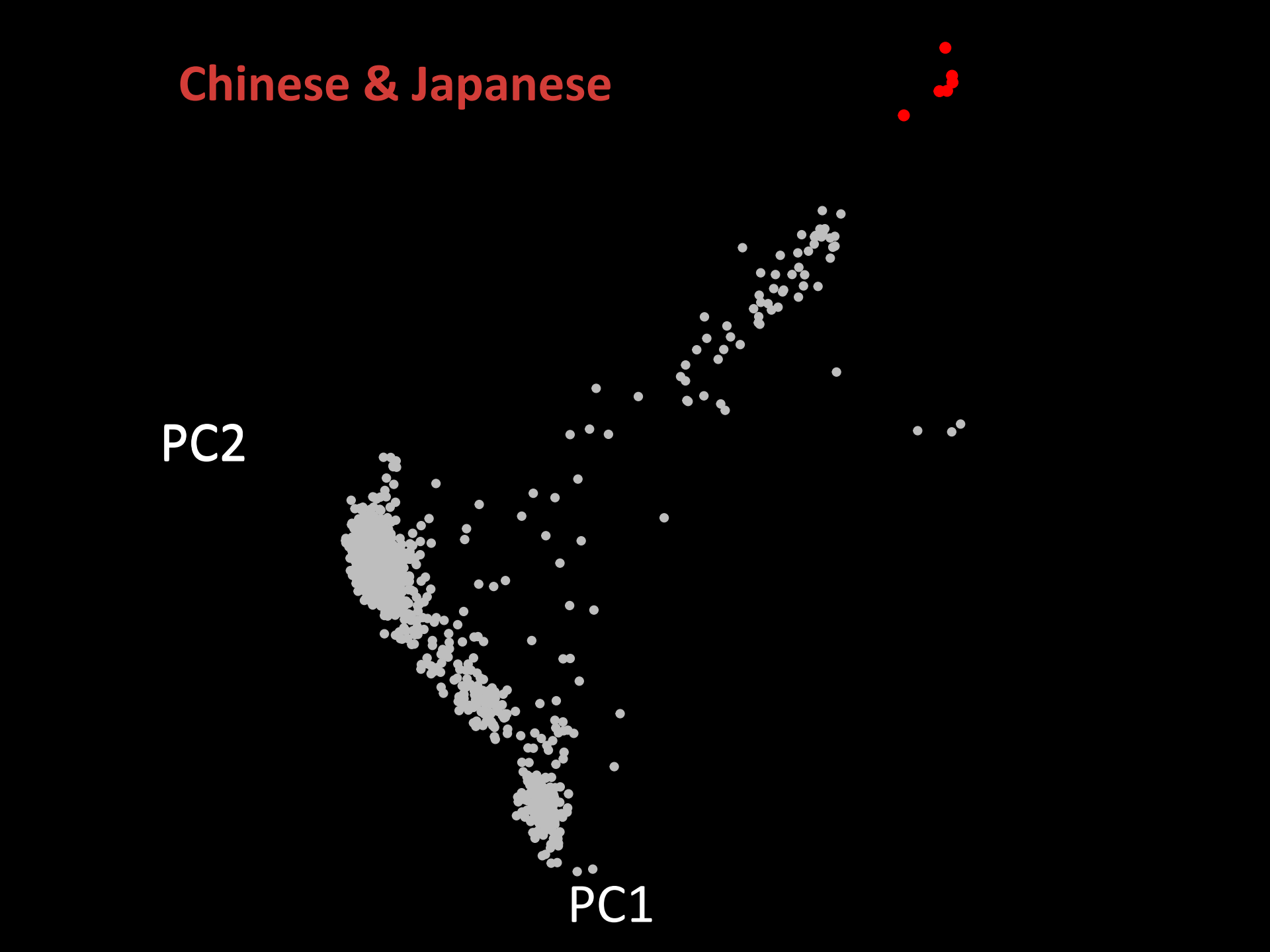
WEHI - Statistical genetic group

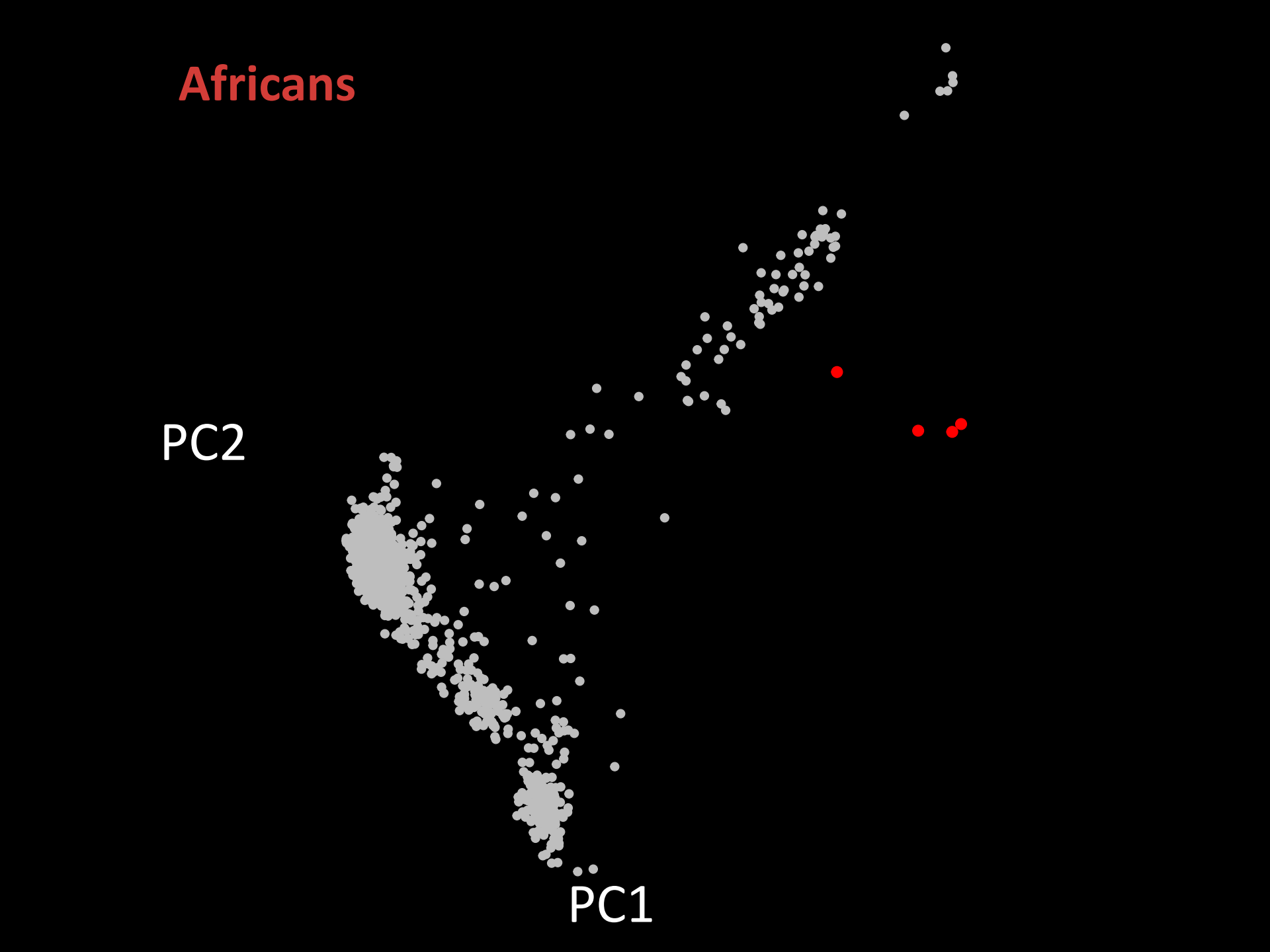
Affected people's DNA

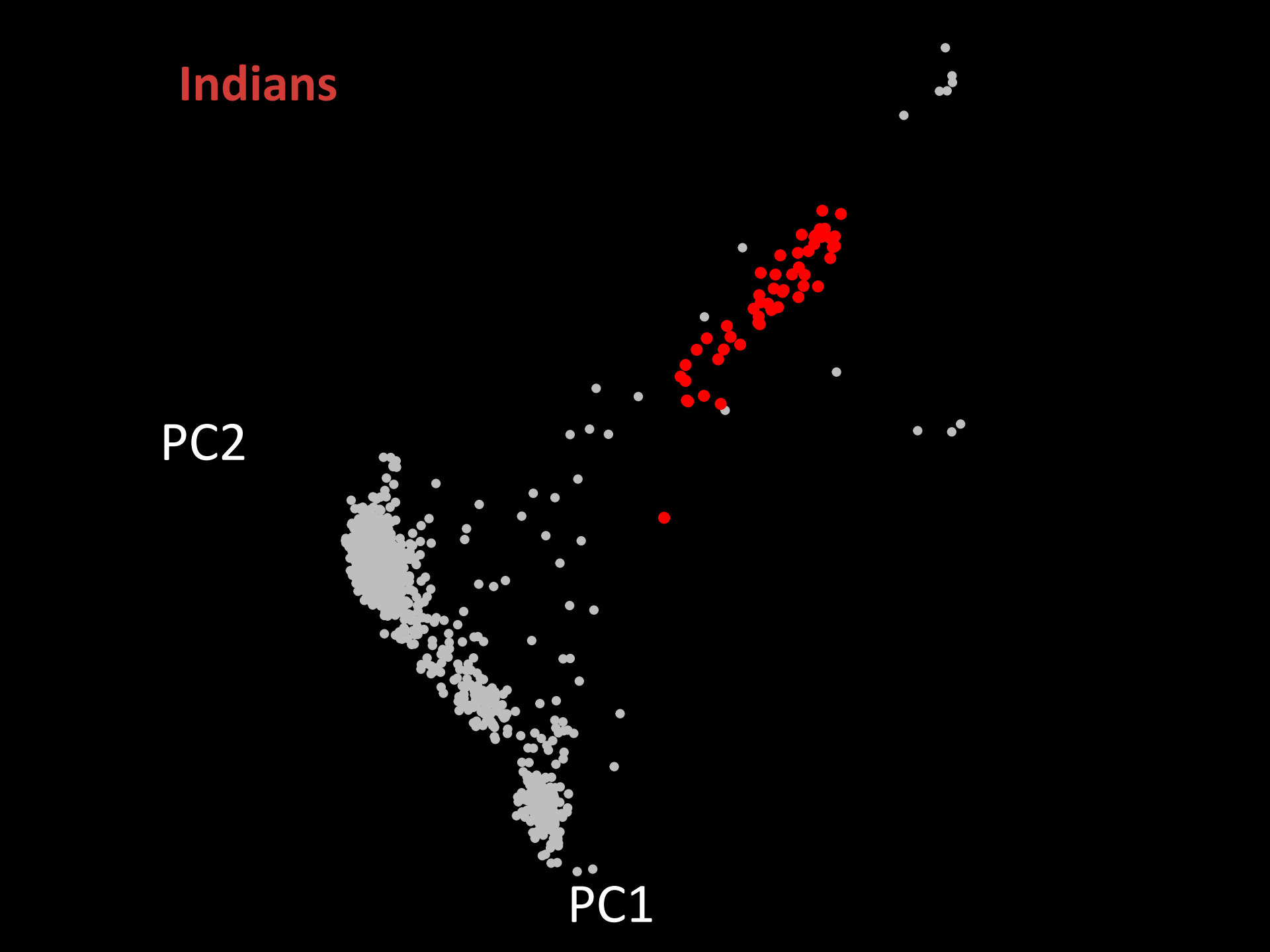


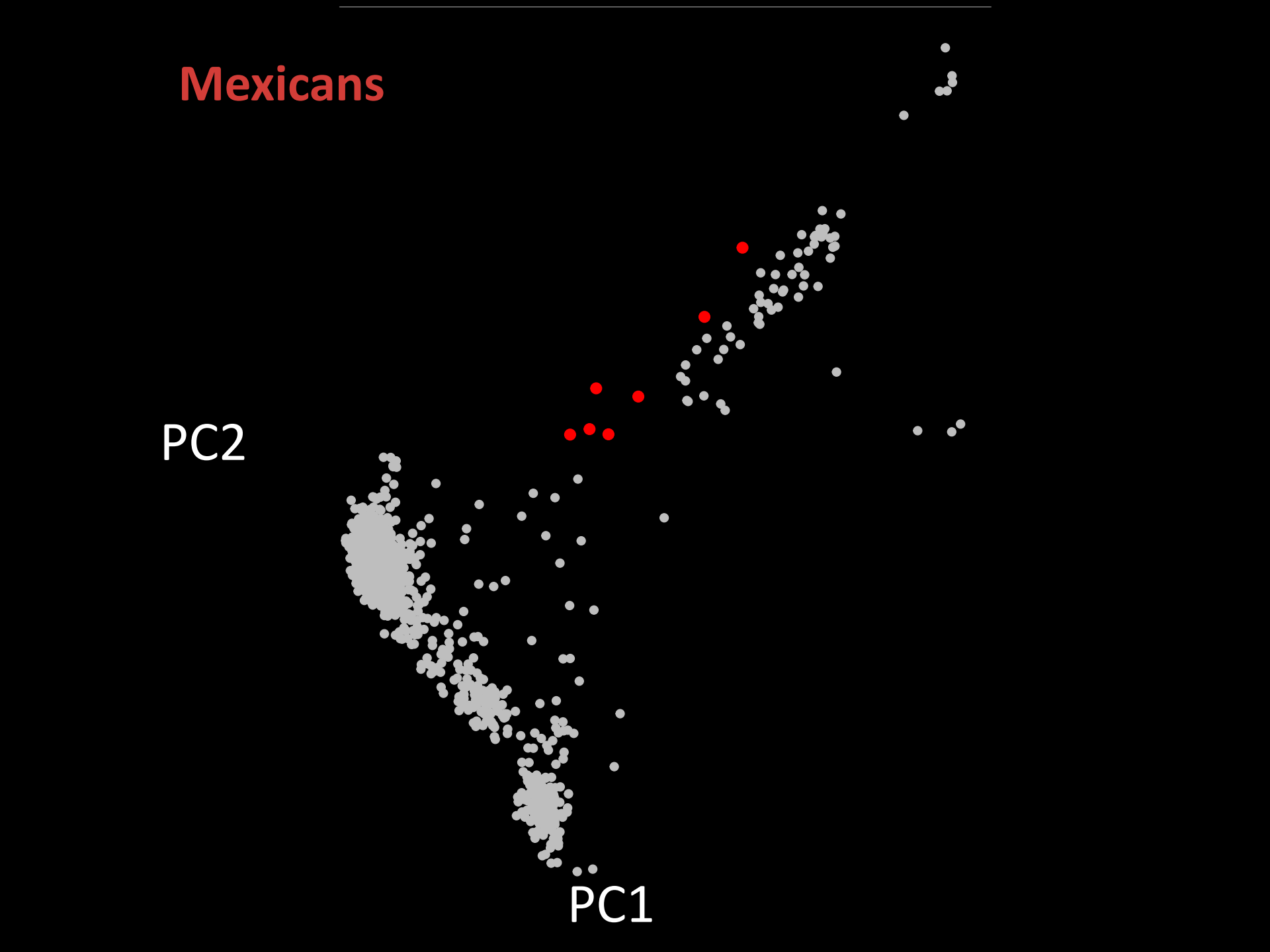

Samples QC of the MacTel dataset (Omni5 chip)
- 704 samples
- 4636551 SNPs
~2.5M common SNPs
~2M rare SNPs (MAF < 0.01)
Concordance between replicates (proportion of concordant genotypes)
Call rate (cr), sex chromosome anomalies, and population stratification
Relatedness analysis
11 duplicate samples within Omni5m (22 samples involved):
- 3 pairs of repeats due to low call rate
- 8 pairs comprising: technical replicates, repeats deduced from relatedness analysis, or from infos in the database
30 samples genotyped on both the Illumina1m and the Omni5m chip
Analysis with PLINK and Linkdatagen (Bahlo M, Bromhead CJ (2009))
- To confirm the known relationhips
- To detect wrongly defined or hidden relatedness
Illumina call rate < 0.95
- 12 samples to be dropped
692 left in the analysis
- 15982 non population specific selected markers
- 691 samples (also dropped one duplicate
Samples with minimum lower call rate showed also lower
Just to clean the data!!
discrepancy to run PCA for outliers detection
Classified as correct expected relationship
Classified as wrong expected relationship
8 such relationships detected
PCA analysis (EIGENSOFT 6.0BETA) to detect outliers:
- keep only ACTG SNPs on autosomes
- Keep one SNP for each pair with r2 > 0.7 in a window of 0.3cM
PCA 1 - Population confirmation (663 samples, 451623 SNPs) Confirmed the population structure estimated from
PCA2 - PCA on the matrix with 1 (missing geno.) and 0 (non missing geno.) (683 samples, 451623 SNPs)
PCA3 - Outliers detection
- Keep only the founder for each family
- Keep only samples close to Caucasian ancestry (CEU, TSI, MEX)
- Drop samples pop. discrepancy
Most of the expected relationships are confirmed
- 3 wrong expected relationships between A877 and 3 members of the same family (F204m). A877 is a sample swap (already highlighted in the concordance analysis)
- A686-A1296 (F156m) more likely uncle-nephew than parent-offspring
- 4 distant relationships detected (close to second double cousins) among expected independent samples
PCA3 dataset
- 569 samples
- 431297 SNPs
Summary of the exclusion list
- Illumina call rate < 0.98 and duplicates: 28 samples (12 with cr < 0.95 (including 3 duplicates), 11 duplicates, 7 with cr < 0.98)
- Estimated with a population either than CEU, TSI, or MEX: 79 samples (include samples with call rate)
- 6 for population discrepancy (BestPopTest vs database)
- Related individuals other than founders: 46 samples
- Outliers from PCA: 18 samples
Total of 153 samples to drop
18 detected outliers (mainly MEX) to be dropped from further analysis
Conclusions
- A877 is dropped from the analysis
- The relationship in F156m is corrected in the pedigree
- The detected distant relationships are left in the dataset
- Only the founders for each family are left in the study for the PCA3 and the GWAS analysis
Final cleaned datasets contains:
- ~4m SNPs
- 551 samples

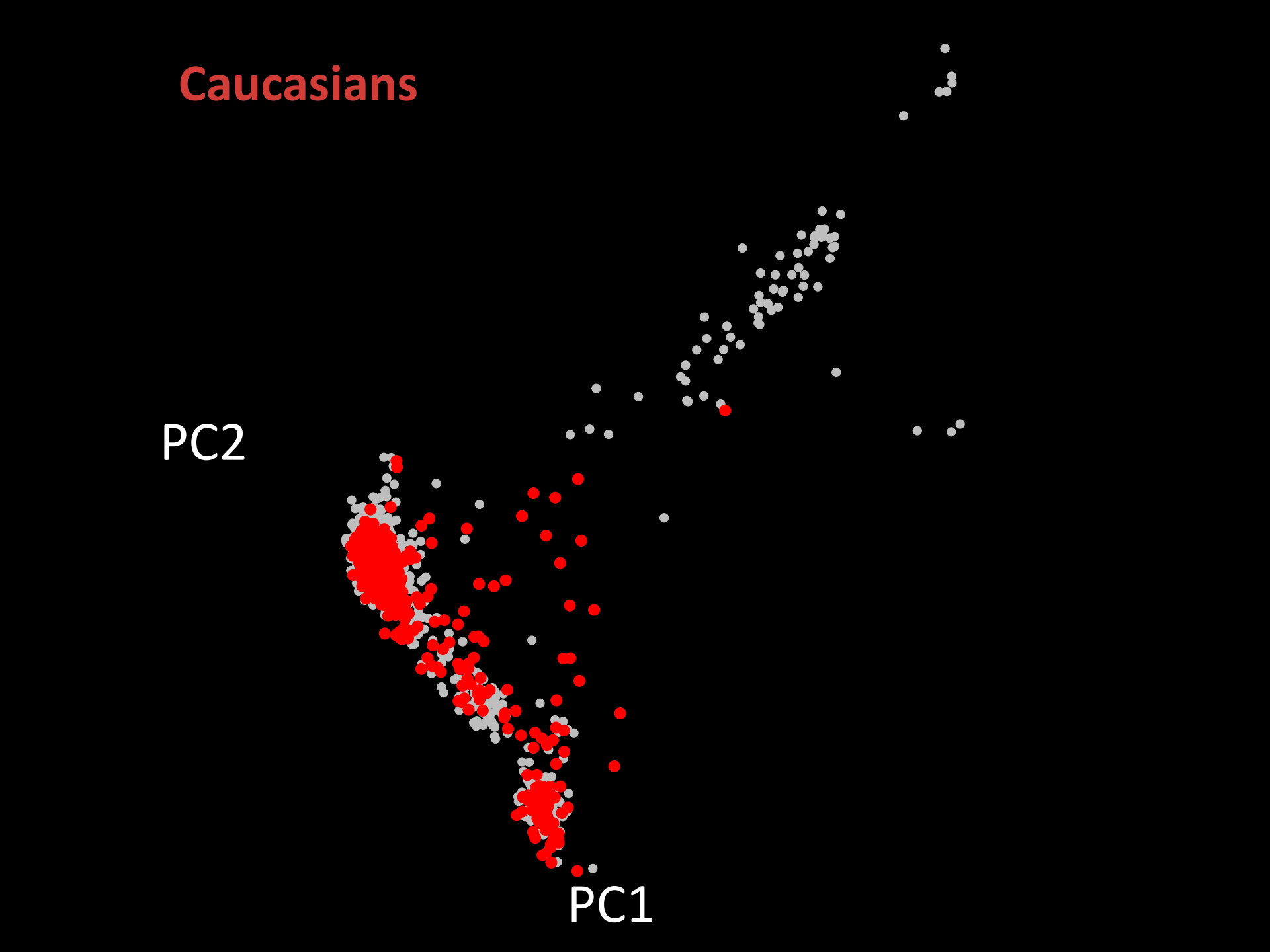

PC2
PC1
areds
ASW
CEU
CHB
CHD
GIH
JPT
LWK
MEX
TSI
YRI

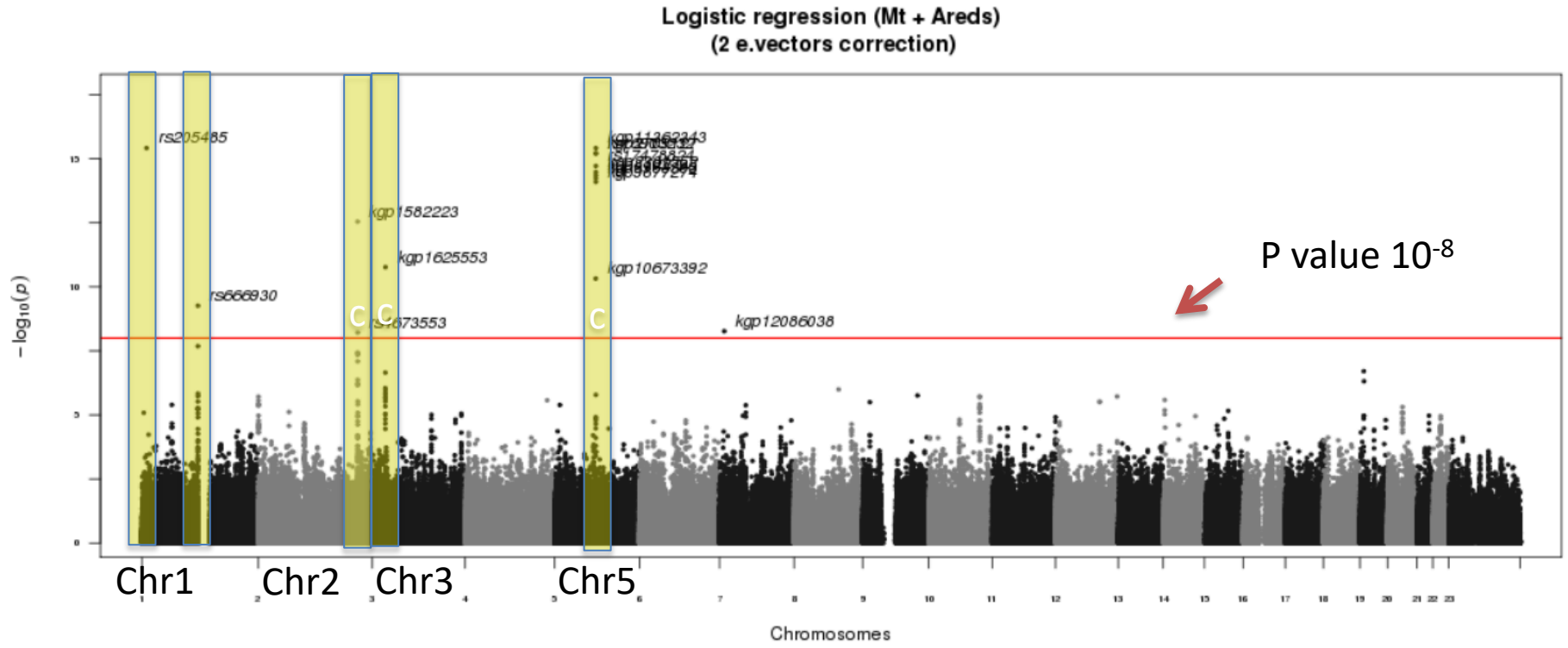

Chinese & Japanese
PC2
PC1

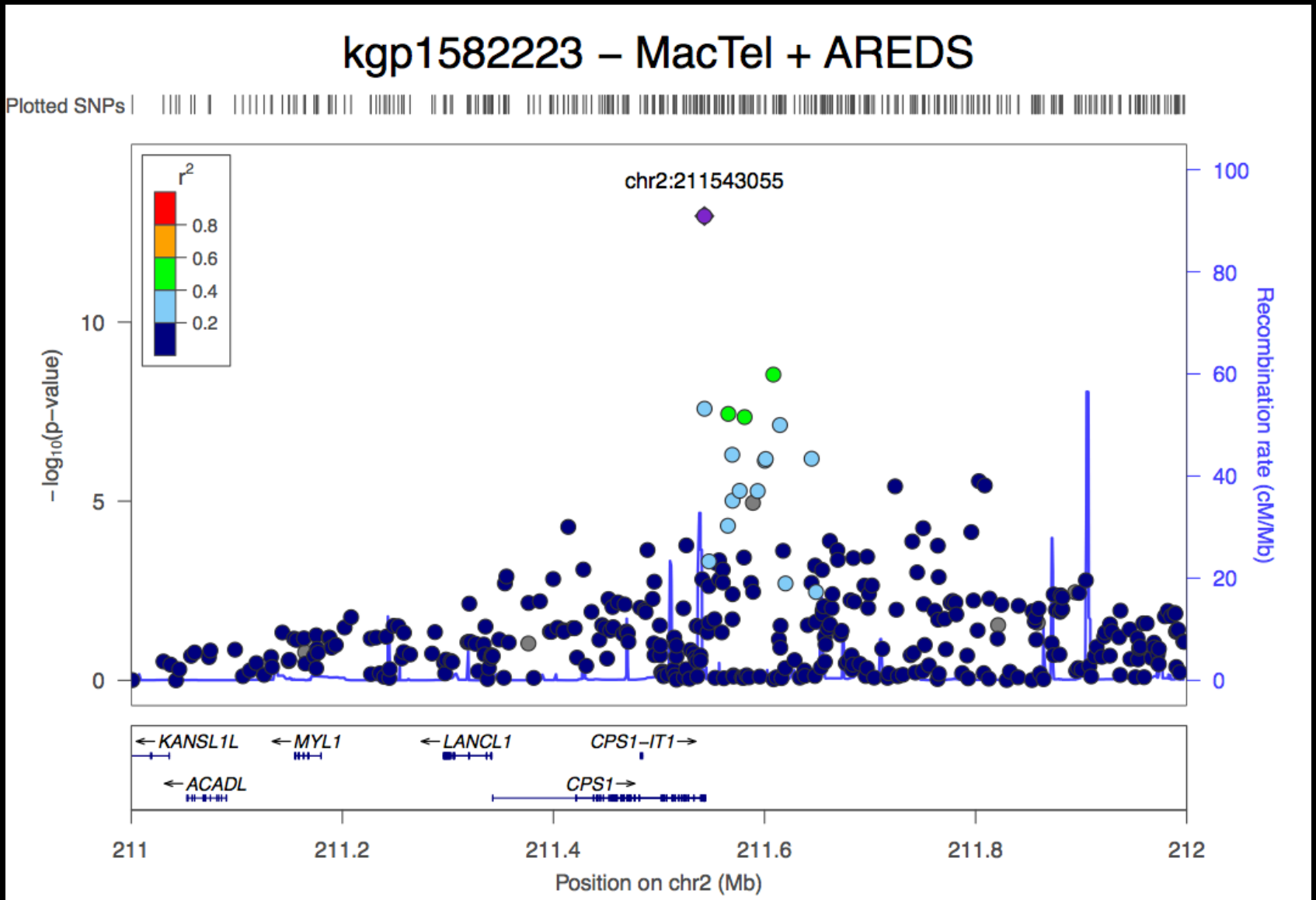

Africans
PC2
PC1


Indians
PC2
PC1


Mexicans
PC2
PC1


Caucasians
PC2
PC1

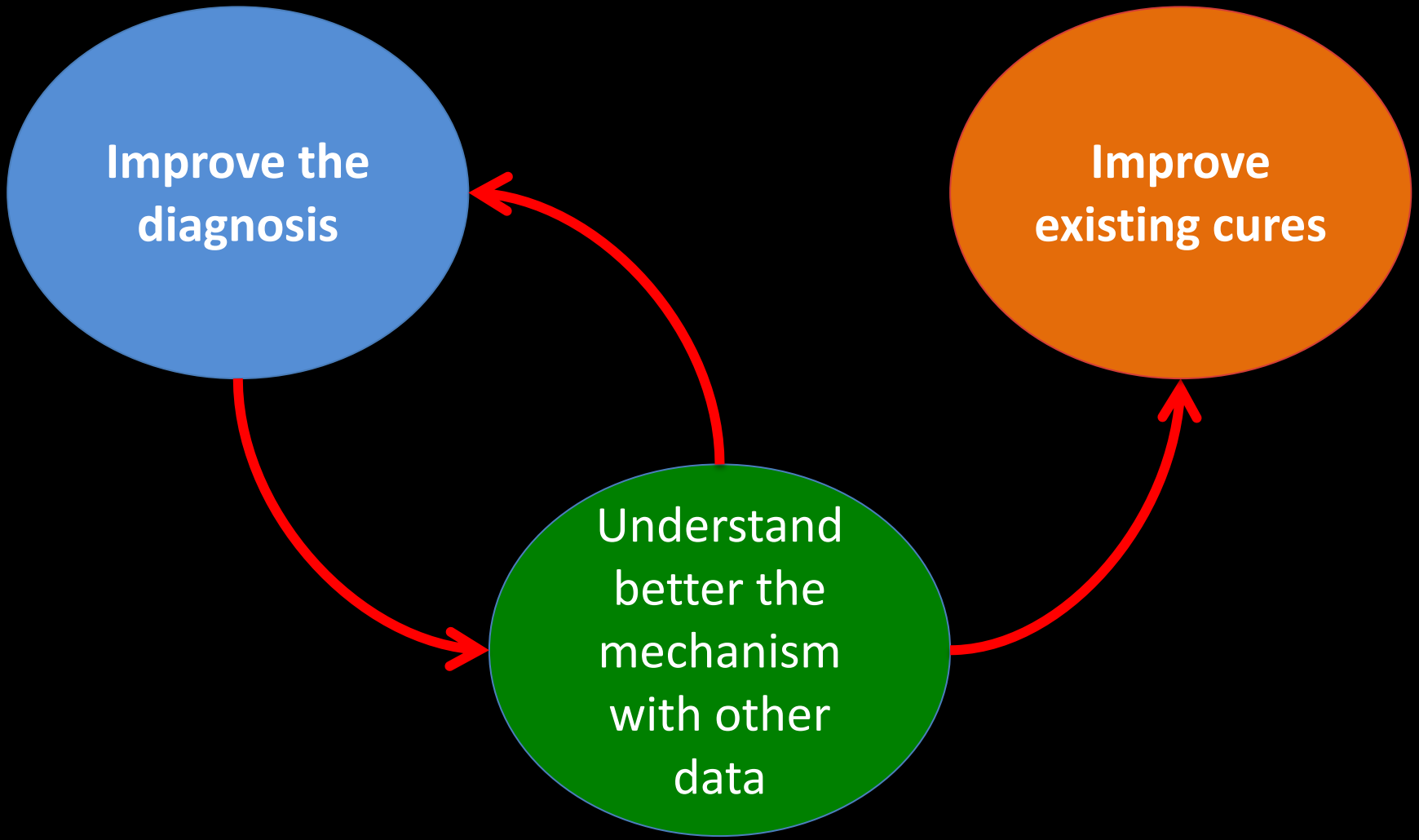

Logistic regression (Mt + Areds)
(2 e.vectors correction)
rs205485
rs666930
kgp1582223
kgp1625553
kgp10673392
kgp12086038
P value 10-8
Chr1
Chr2
Chr3
Chr5
Chromosomes
−log10(p)


kgp1582223 – MacTel + AREDS
Plotted SNPs
r²
0.8
0.6
0.4
0.2
chr2:211543055
10
5
0
− log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
←KANSL1L
←MYL1
←LANCL1
CPS1−IT1→
←ACADL
CPS1→
211
211.2
211.4
211.6
211.8
212
Position on chr2 (Mb)

Interpretation to understand causation

- Search through literature
- More experiments related to the specific genes highlighted by GWAS
- Understand the function of the genes involved

Next important aims



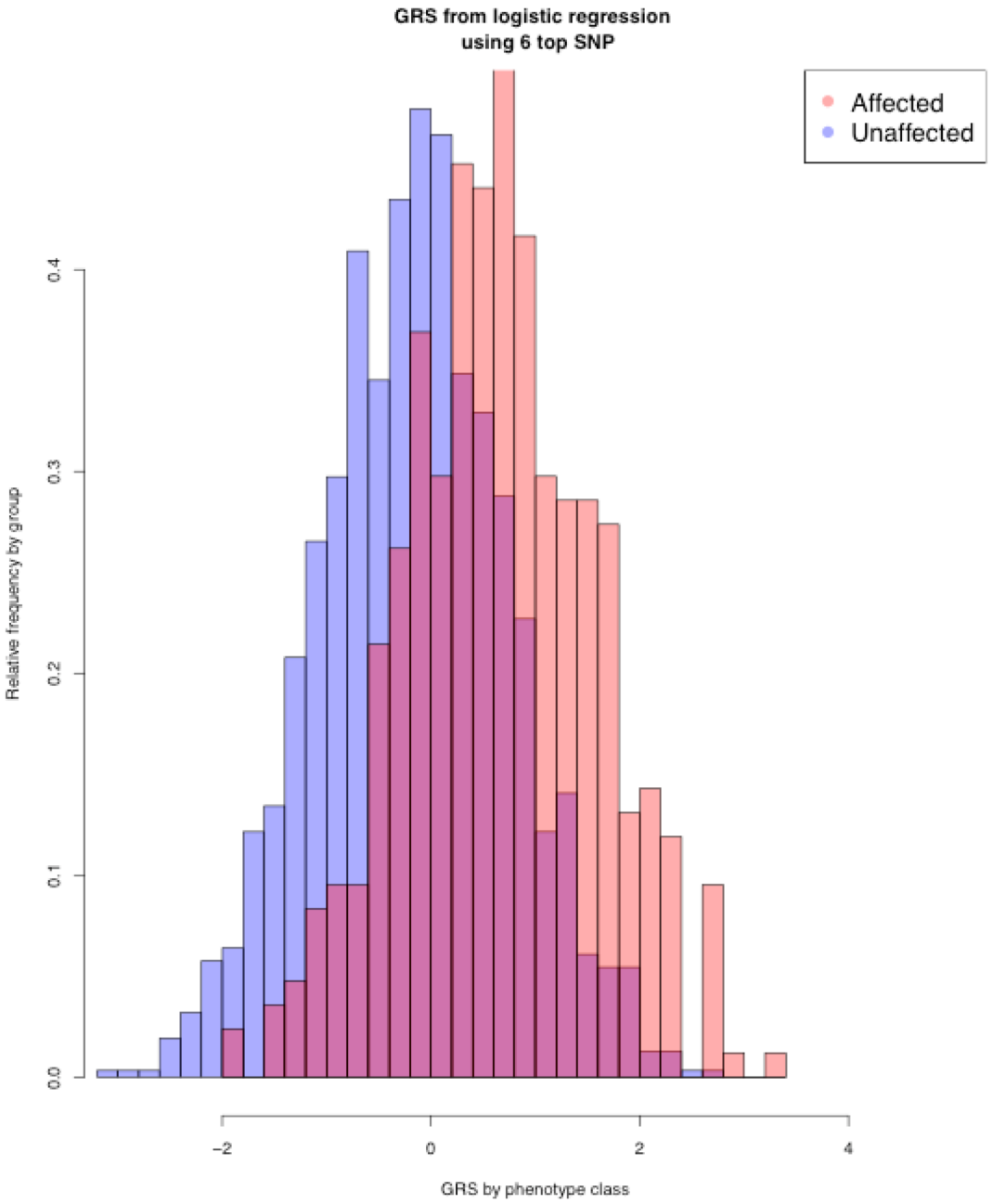
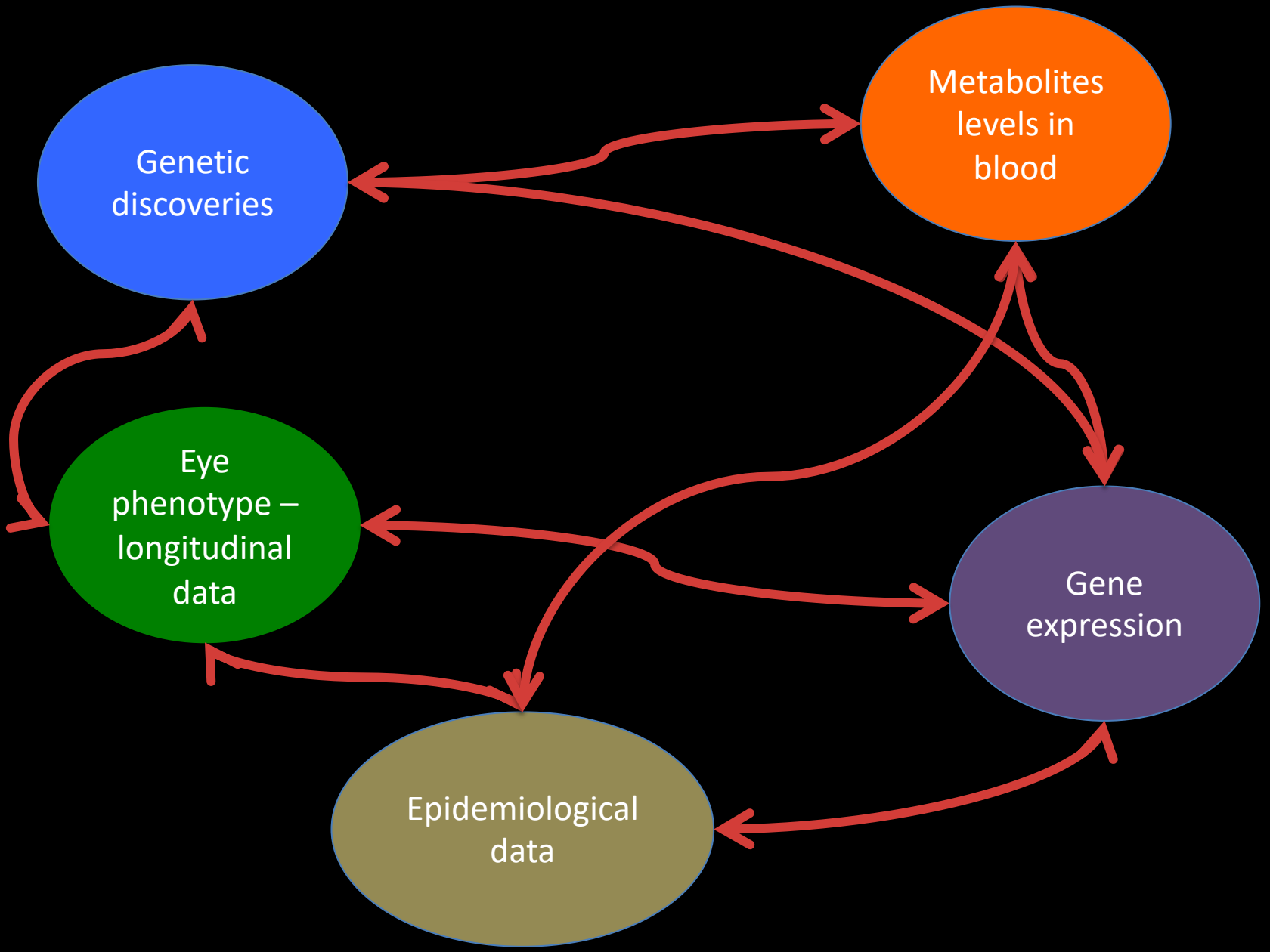


Figure 12: Relative frequency by affection group.


Genetic discoveries
Metabolites levels in blood
Eye phenotype – longitudinal data
Gene expression
Epidemiological data

Conclusions

Conclusions

- Never be afraid of asking for something

Conclusions

- Never be afraid of asking for something
- Study as better as you can

Conclusions

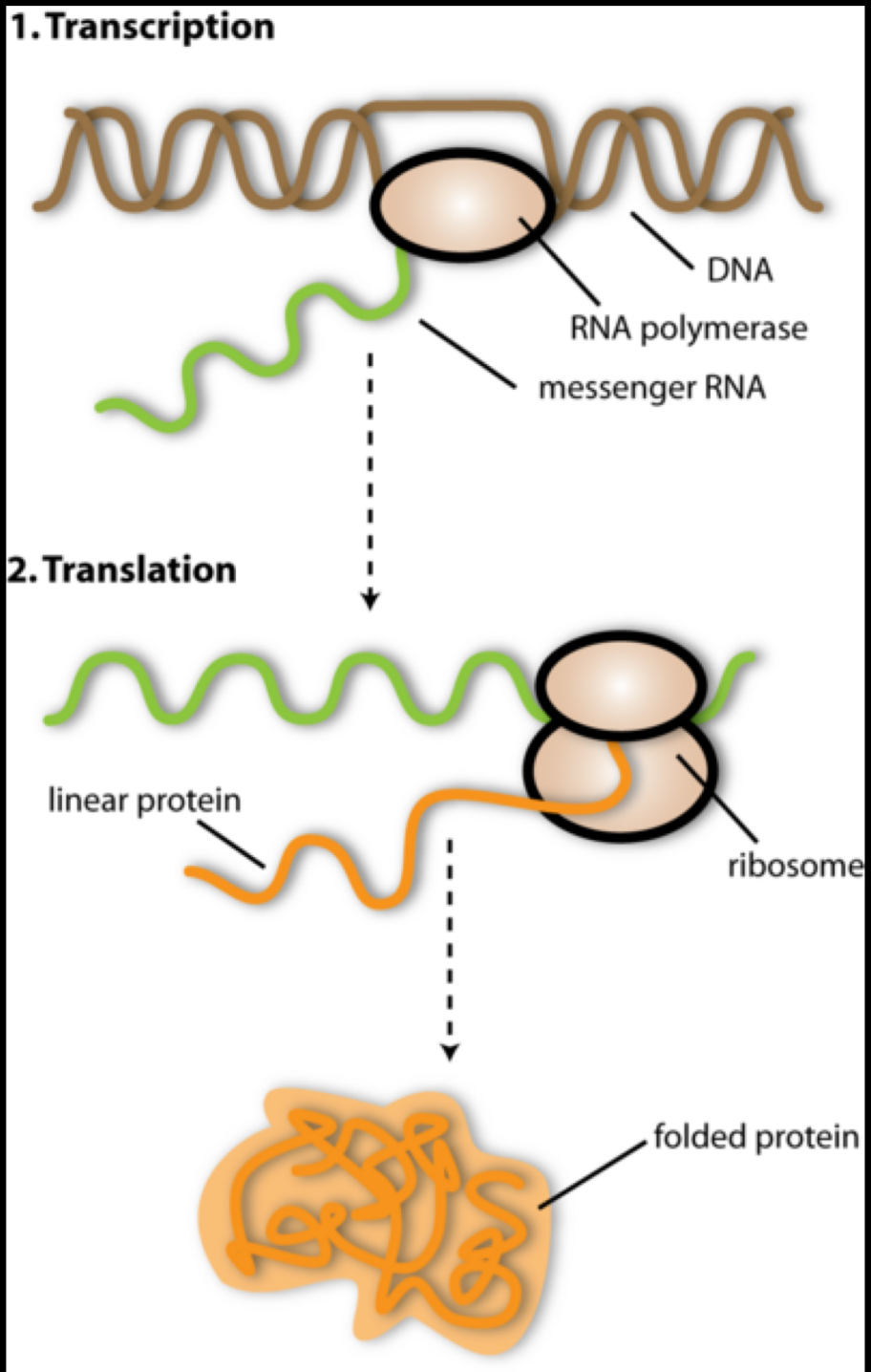
- Never be afraid of asking for something
- Study as better as you can
- Learn R!

Conclusions

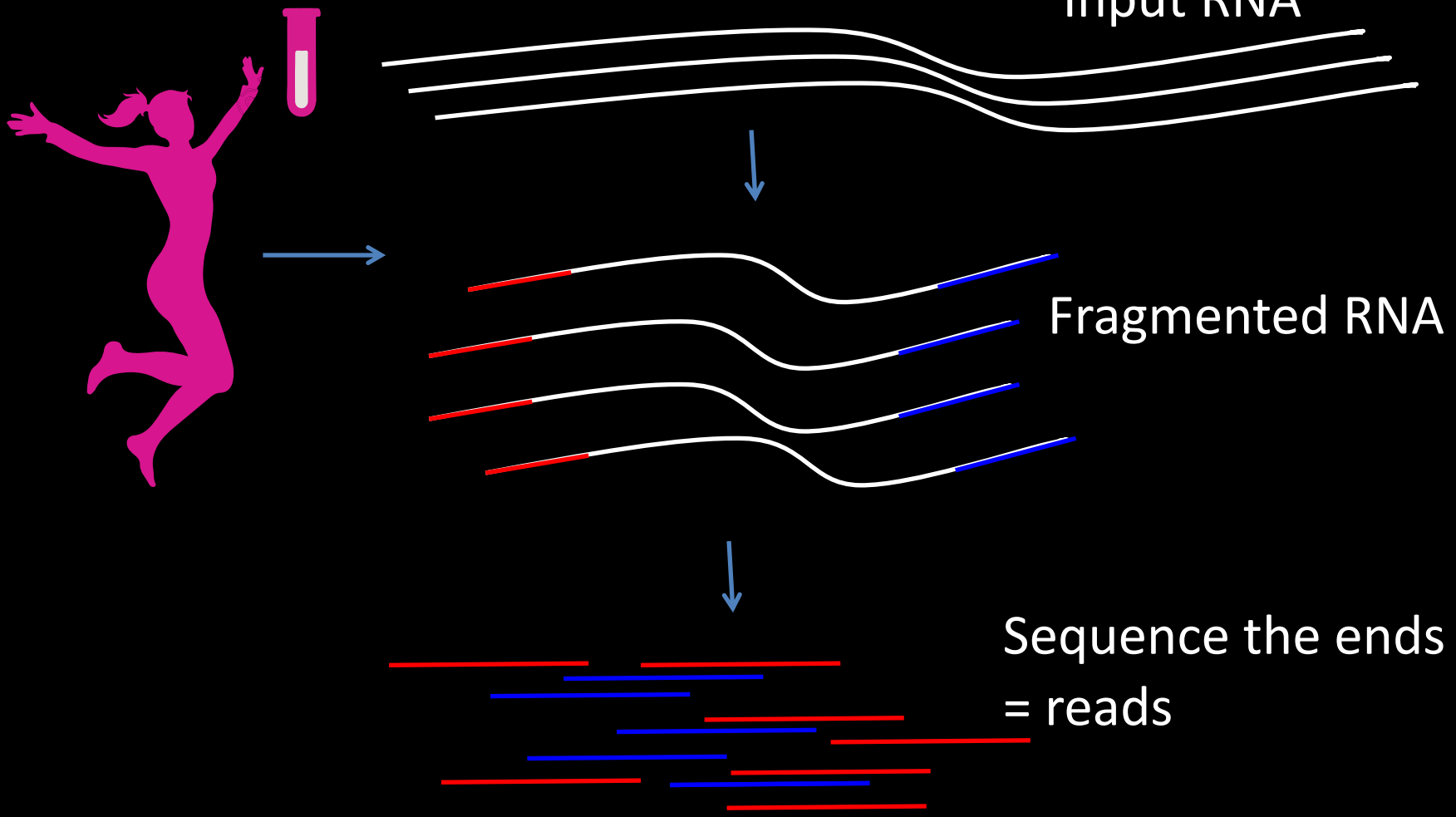
- Never be afraid of asking for something
- Study as better as you can
- Learn R!
- Be curious

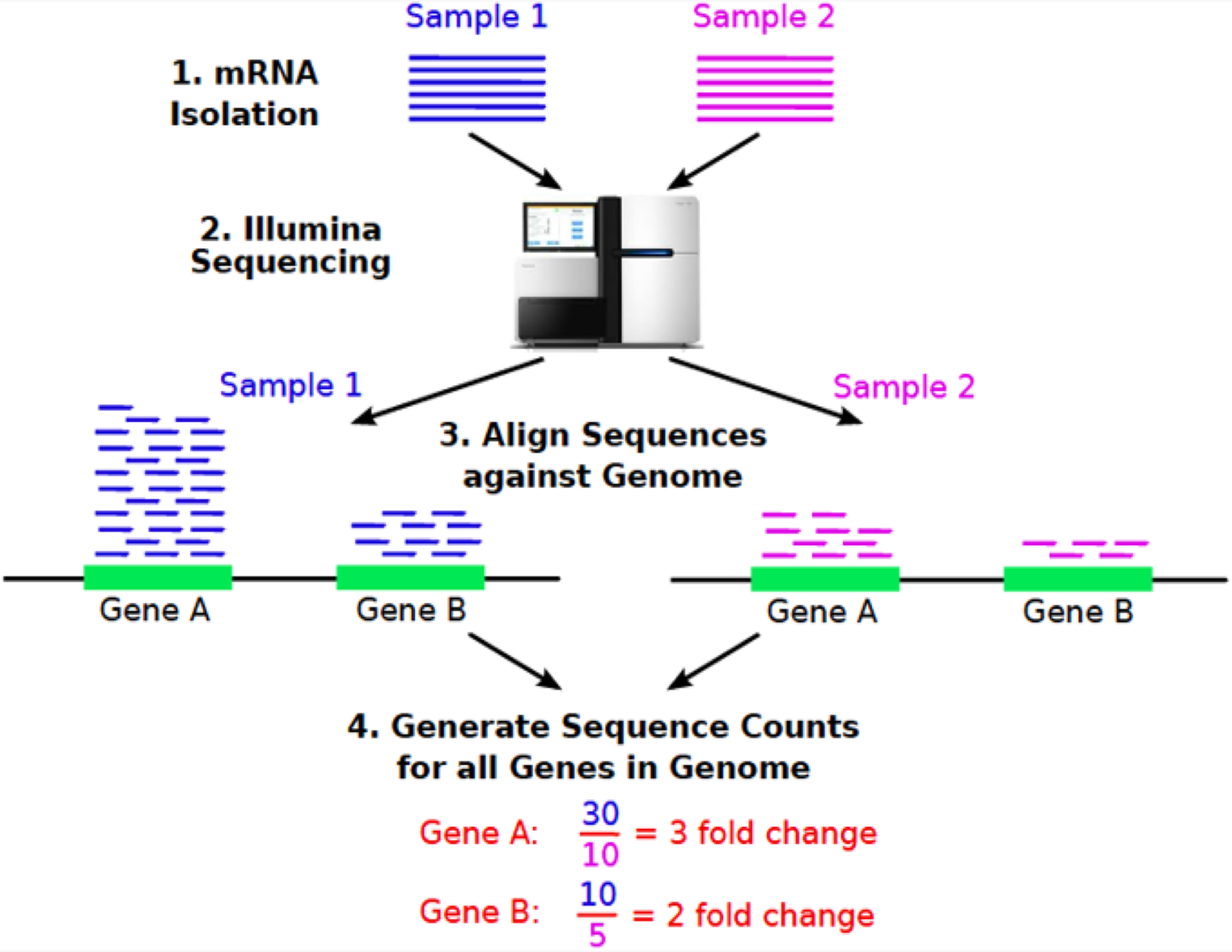
Thanks for listening!

Any question about food and kangaroos?

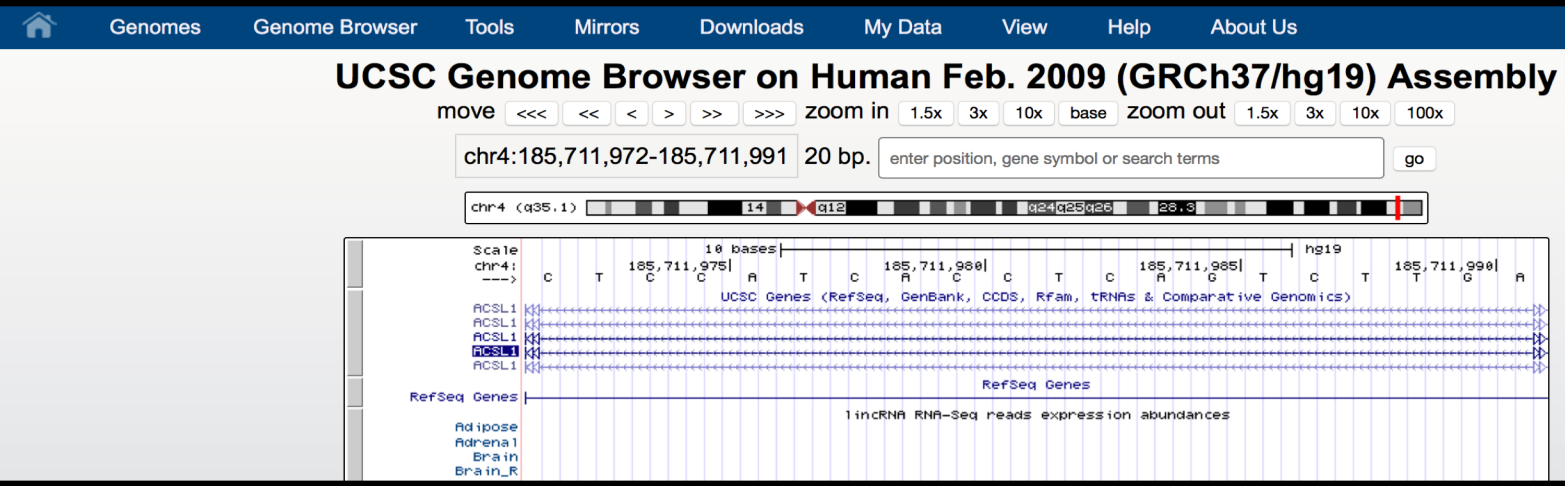

1. Transcription
DNA
RNA polymerase
messenger RNA
2. Translation
linear protein
ribosome
folded protein

RNA-sequencing




Sample 1
Sample 2
1. mRNA Isolation
2. Illumina Sequencing
Sample 1
Sample 2
3. Align Sequences against Genome
Gene A
Gene B
Gene A
Gene B
4. Generate Sequence Counts for all Genes in Genome
Gene A: 30/10 = 3 fold change
Gene B: 10/5 = 2 fold change

Reference genomes



[genome browser]

RNA analysis

	Treat1	Treat2	Control1	Control2
Gene1	100	120	20	15
Gene2	150	80	15	15
Gene3	0	15	80	100
Gene4	0	0	0	0
Gene5	180	200	250	200

Table of analysis

	Cases1	Cases2	Control1	Control2
Gene1	100	120	20	15
Gene2	150	80	15	15
Gene3	0	15	80	100
Gene4	0	0	0	0
Gene5	180	200	250	200

Table of analysis

	Treat1	Treat2	Control1	Control2
Gene1	100	120	20	15
Gene2	150	80	15	15
Gene3	0	15	80	100
Gene4	0	0	0	0
Gene5	180	200	250	200

Table of analysis

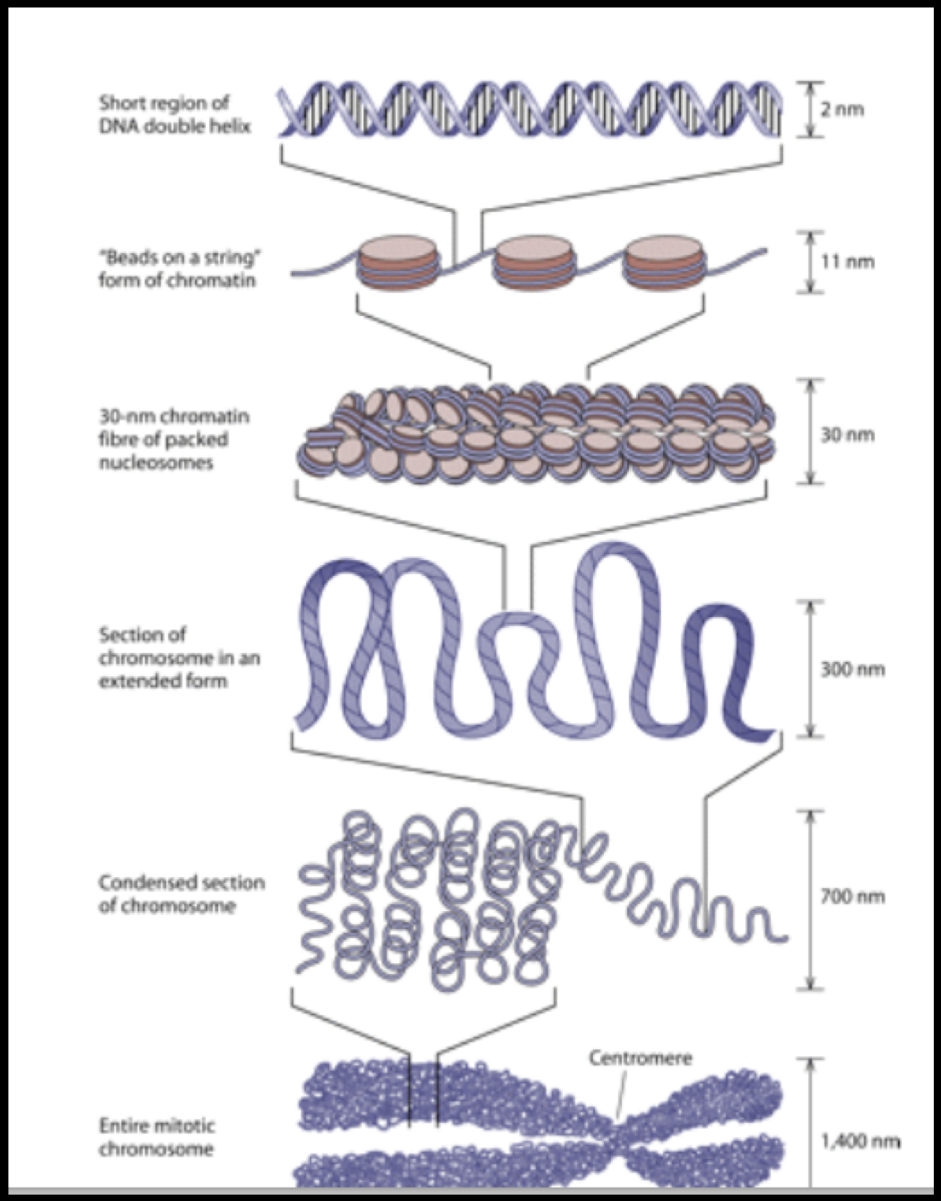
	Treat1	Treat2	Control1	Control2
Gene1	100	120	20	15
Gene2	150	80	15	15
Gene3	0	15	80	100
Gene4	0	0	0	0
Gene5	180	200	250	200

Protein-DNA interaction

Epigenetics - Ted Talk1

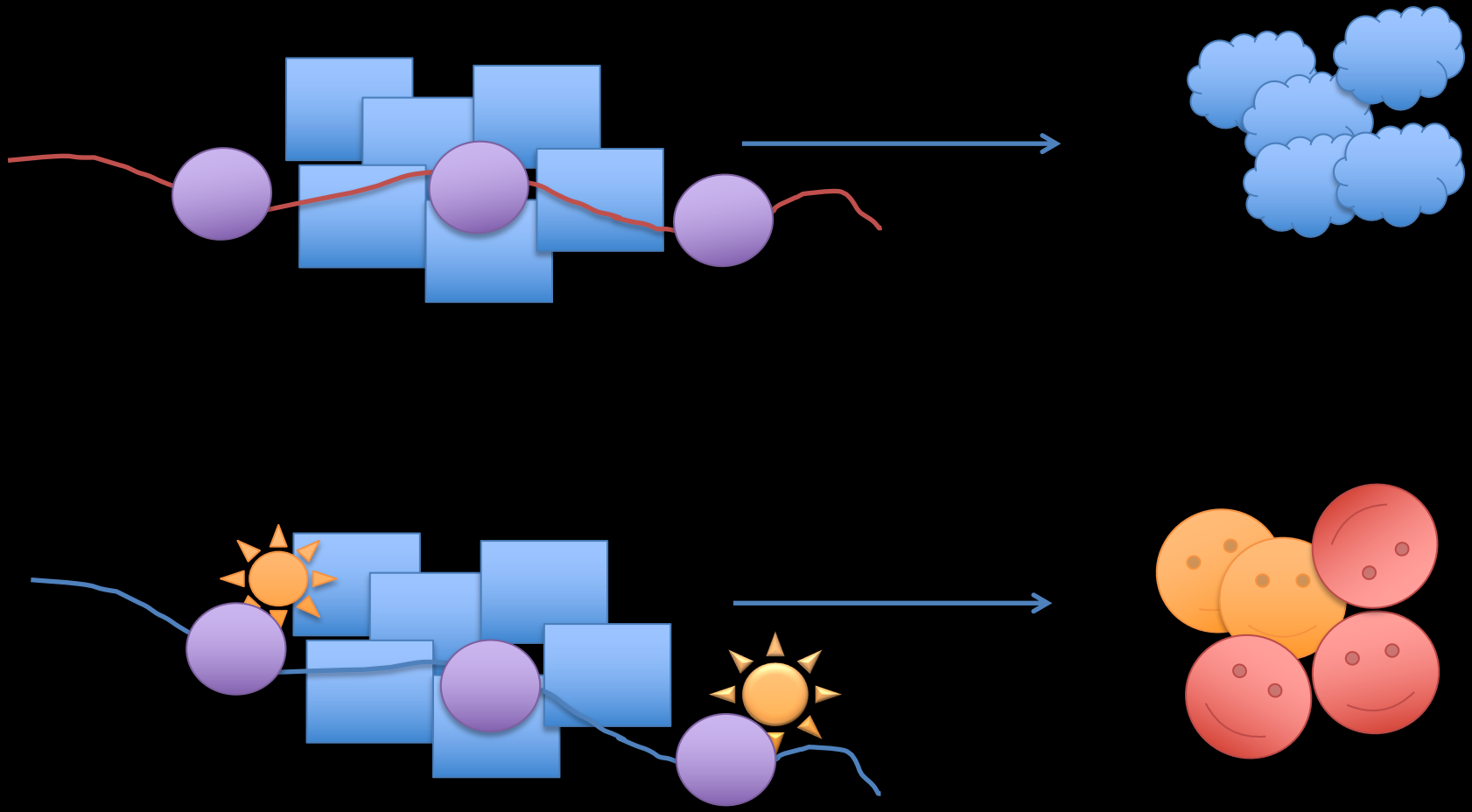
Epigenetic - Ted Talk2

Protein-DNA interaction

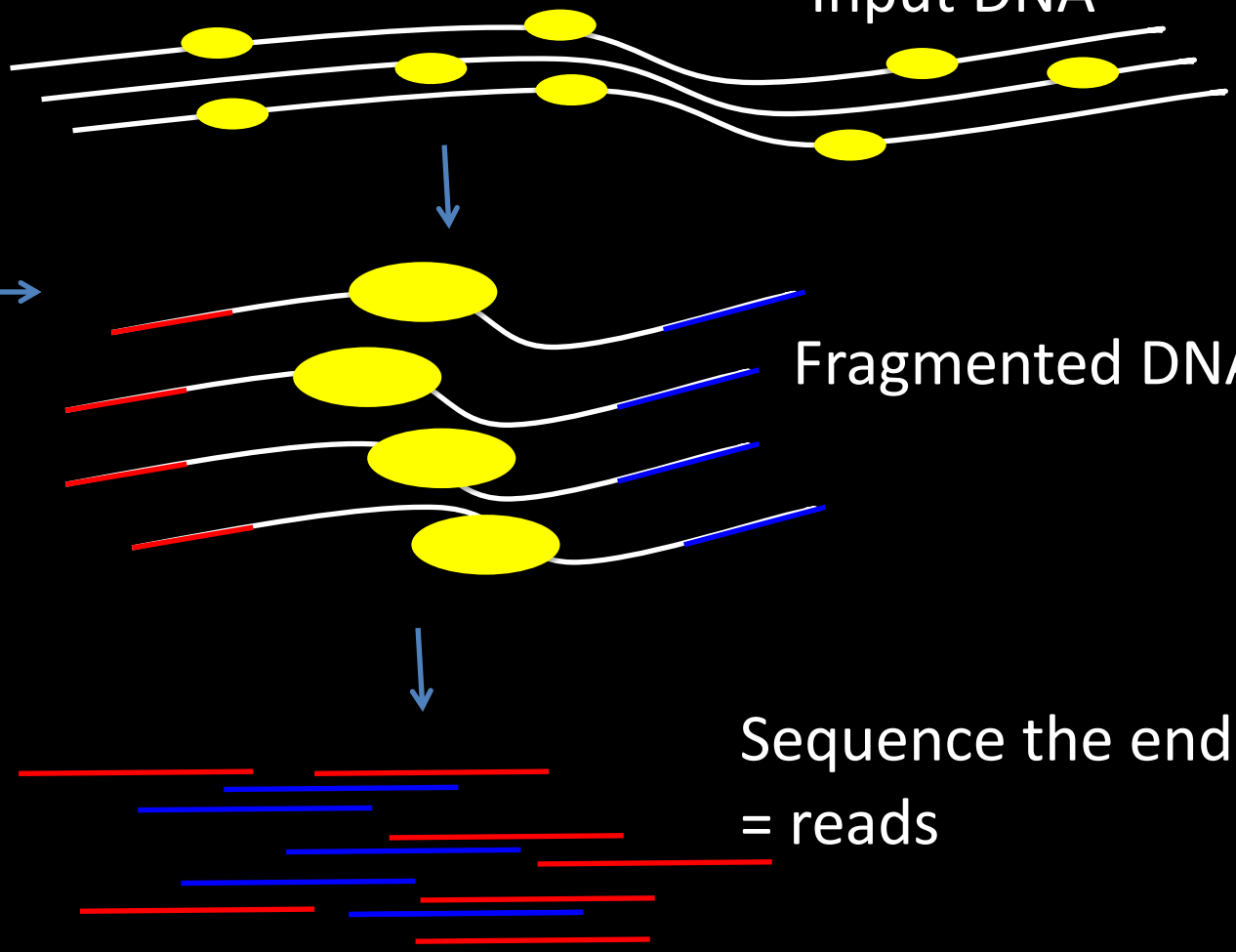


Initial cells

Differentiated cells

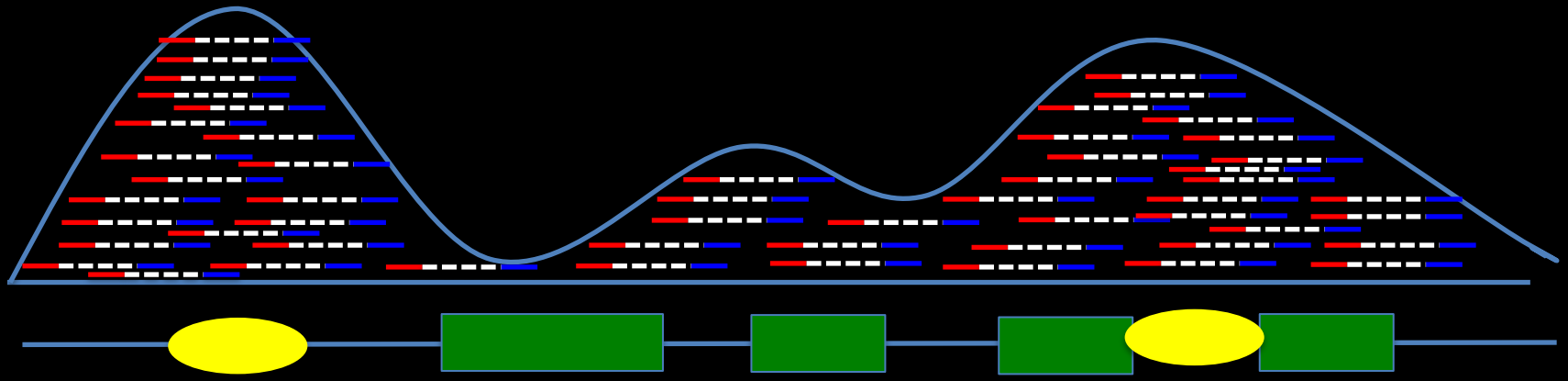


ChIP-sequencing



ChIP-Seq data

Objective: separate noise from true signal

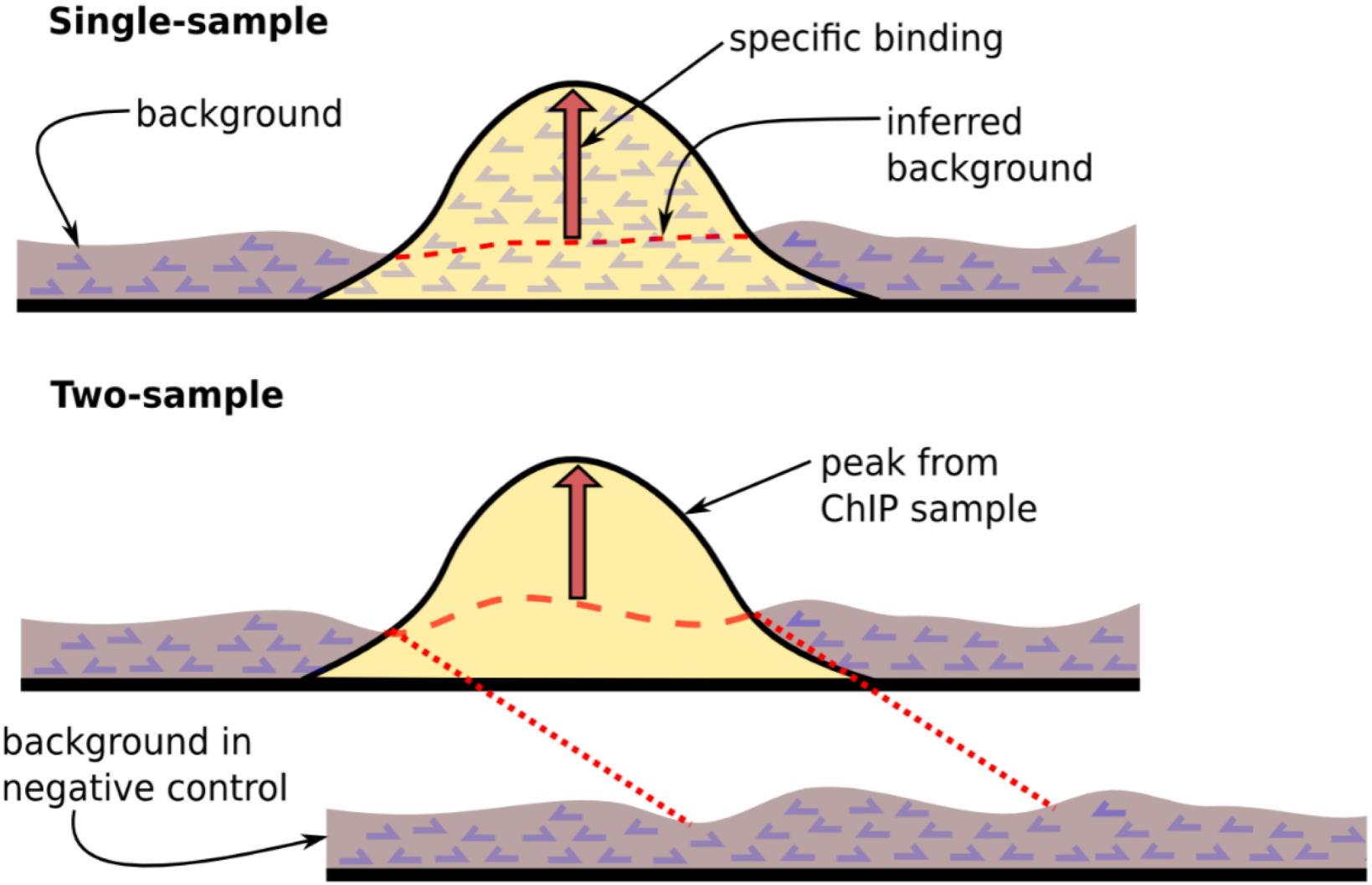


? ? ?

Noise also induced by the biological steps for preparation

Conventional analyses for ChIP-seq

Detect absolute binding by peak calling:



ChIP-Seq most used software

- MACS2 – Liu lab

New from WEHI Bioinformatics

- csaw– Aaron Lun (Smyth lab)

Model counts

- Wide use of statistical model for counts
  - Poisson
  - Negative binomial
  - Binomial
- Multivariate analysis to study the noise
  - PCs
  - SVD

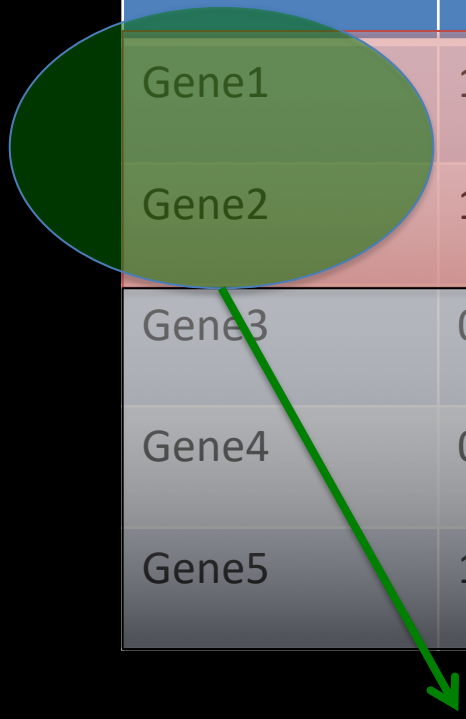
- Remove unwated variation

Computational & statistical challenges

- Multivariate analysis
- Data mining
- R and C as support
- Communication with biologist with user-friendly form

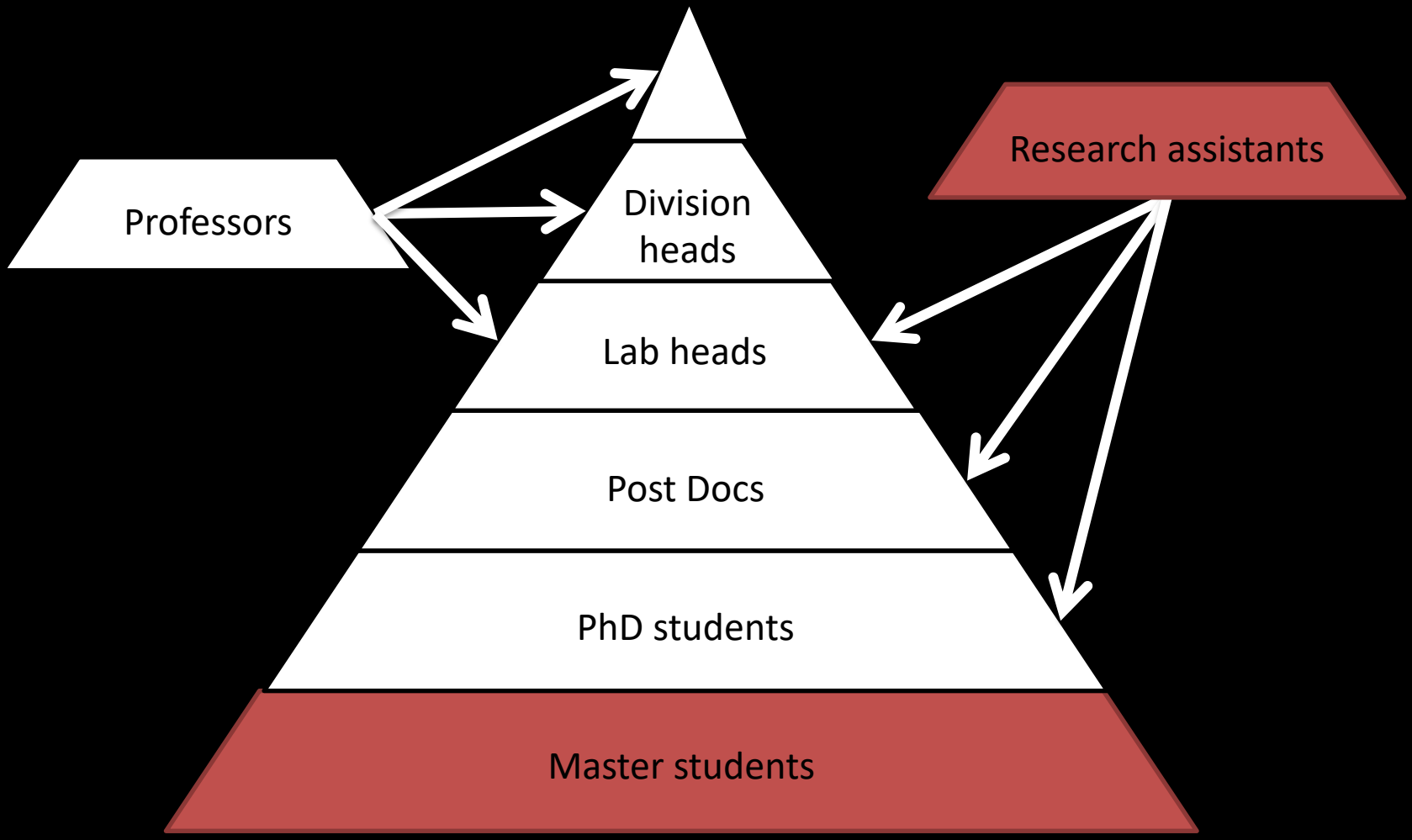
The dream

	Cases1	Cases2	Control1	Control2
Gene1	100	120	20	15
Gene2	150	80	15	15
Gene3	0	15	80	100
Gene4	0	0	0	0
Gene5	180	200	250	200



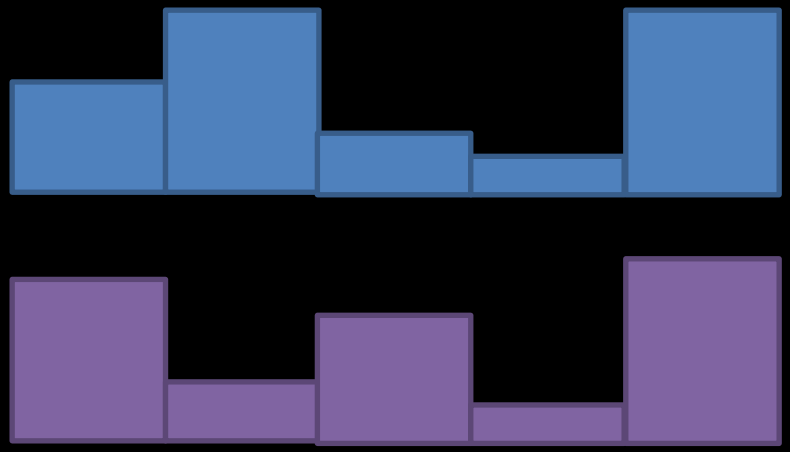
Genes with significant SNPs highlighted in the GWAS

To work in a research institute



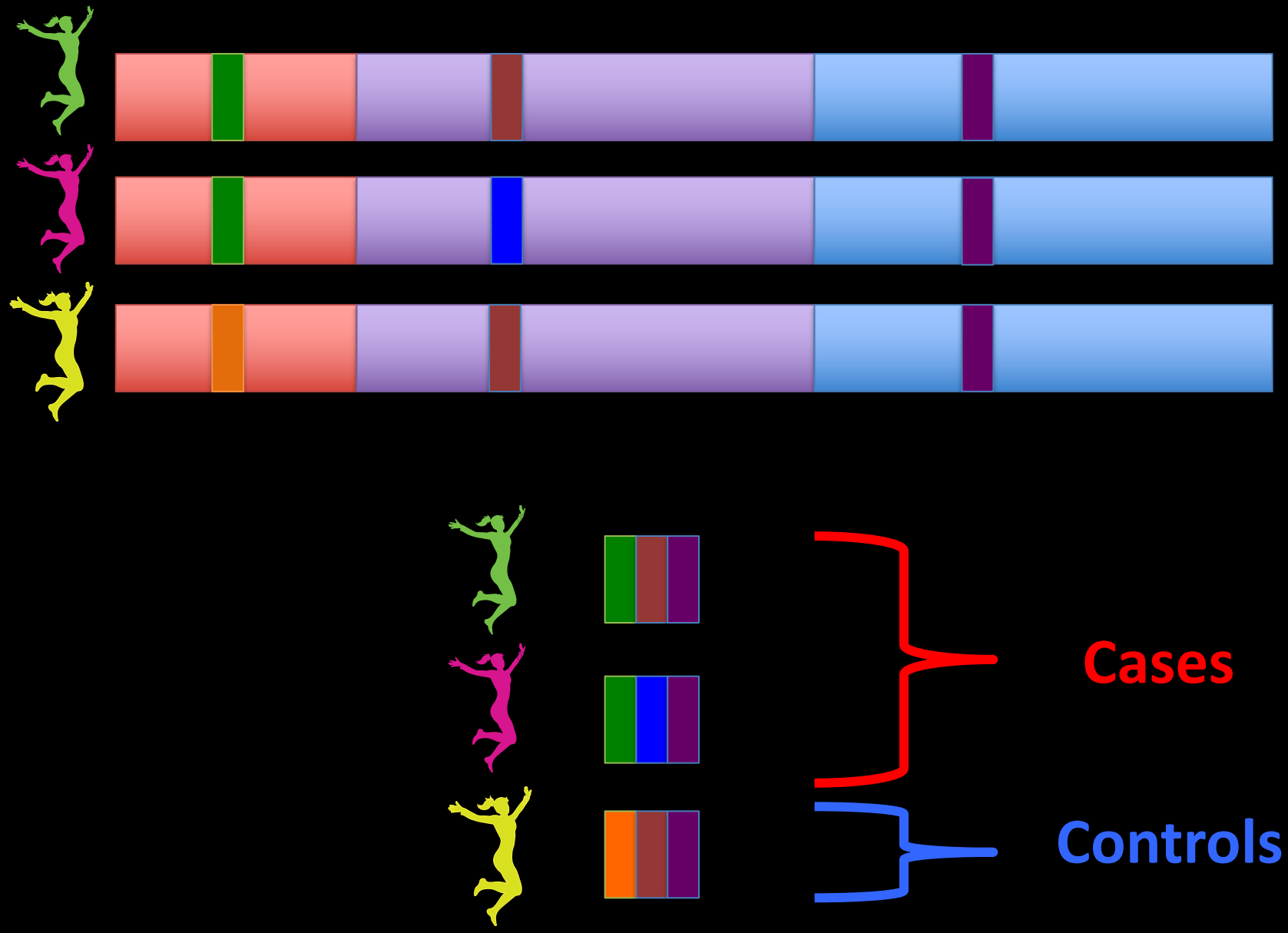
Classic ways of analysis

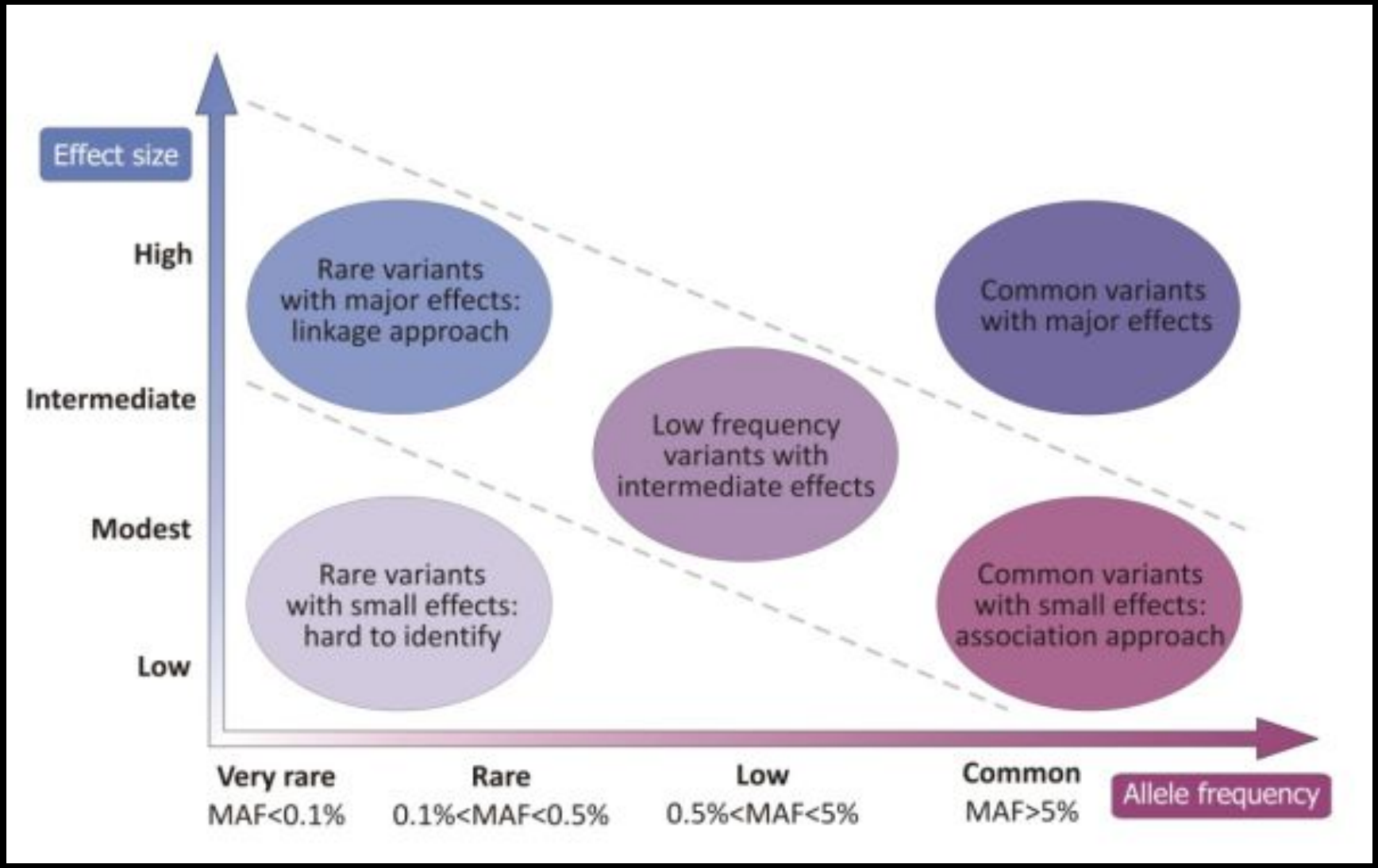
Compare
binned counts



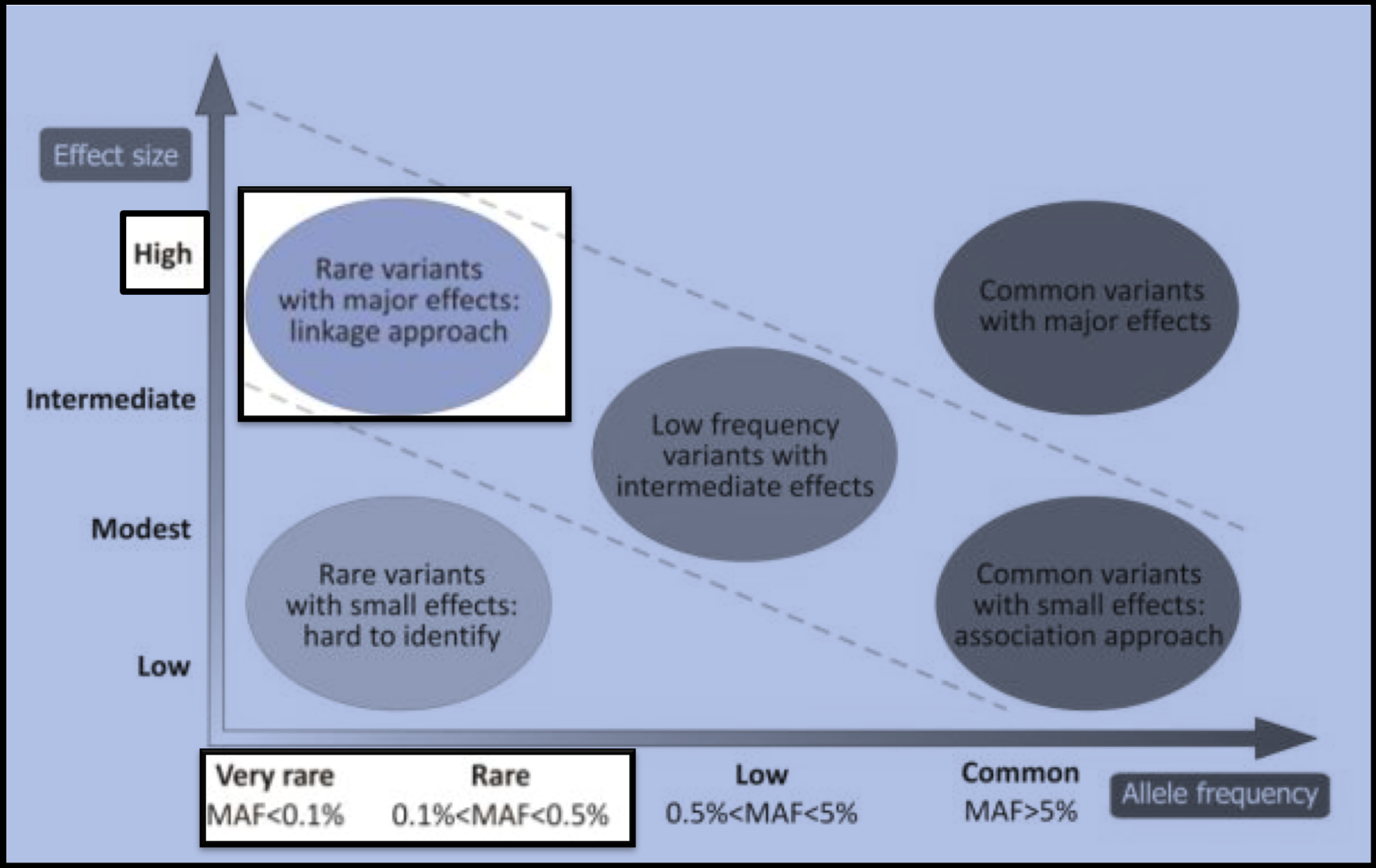
Lib1 ChIP-Seq

Lib2 ChIP-Seq - control


Cases
Controls


Effect size
High
Intermediate
Modest
Low
Rare variants with major effects: linkage approach
Common variants with major effects
Low frequency variants with intermediate effects
Rare variants with small effects: hard to identify
Common variants with small effects: association approach
Very rare
MAF<0.1%
Rare
0.1%<MAF<0.5%
Low
0.5%<MAF<5%
Common
MAF>5%
Allele frequency

Old approach to MacTel: poor results



New approach to MacTel: GWAS