Population Genetics

Section 4

Learning Objectives

- Understand the importance of Hardy Weinberg equilibrium and how to calculate deviance from HWE.
- Describe population substructure and how it can confound results. Also understand methods for accounting for it in analysis.
- Leverage linkage disequilibrium to identify genomic regions associated with phenotypes.

Revisiting linkage disequilibrium

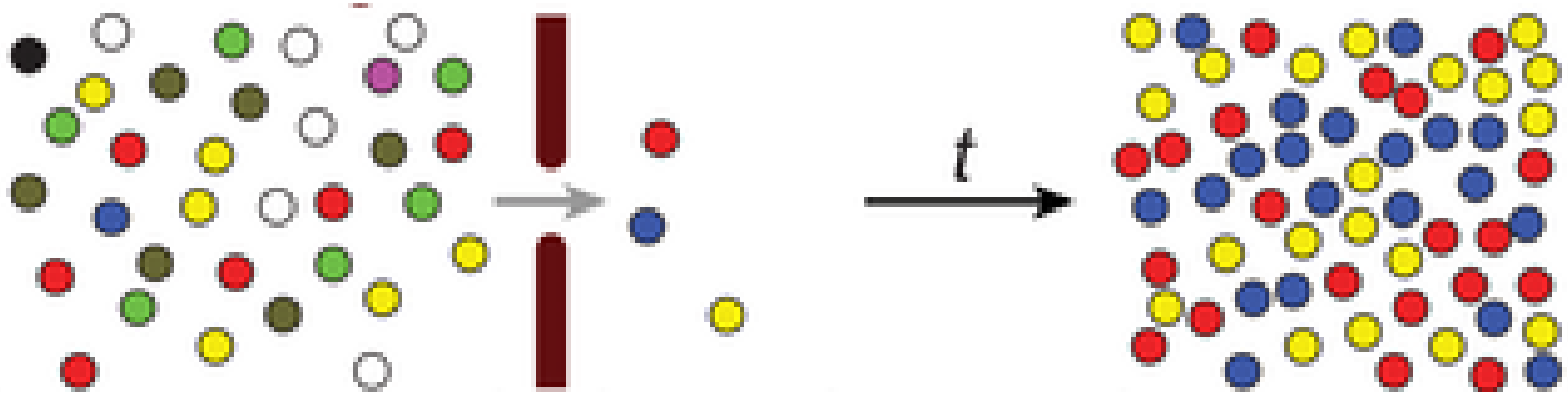
With what certainty can you know what variant is at position B/b if you know what is at A/a?

Linkage disequilibrium weakens over generations

Ardlie, Nat Rev Genetics 2002

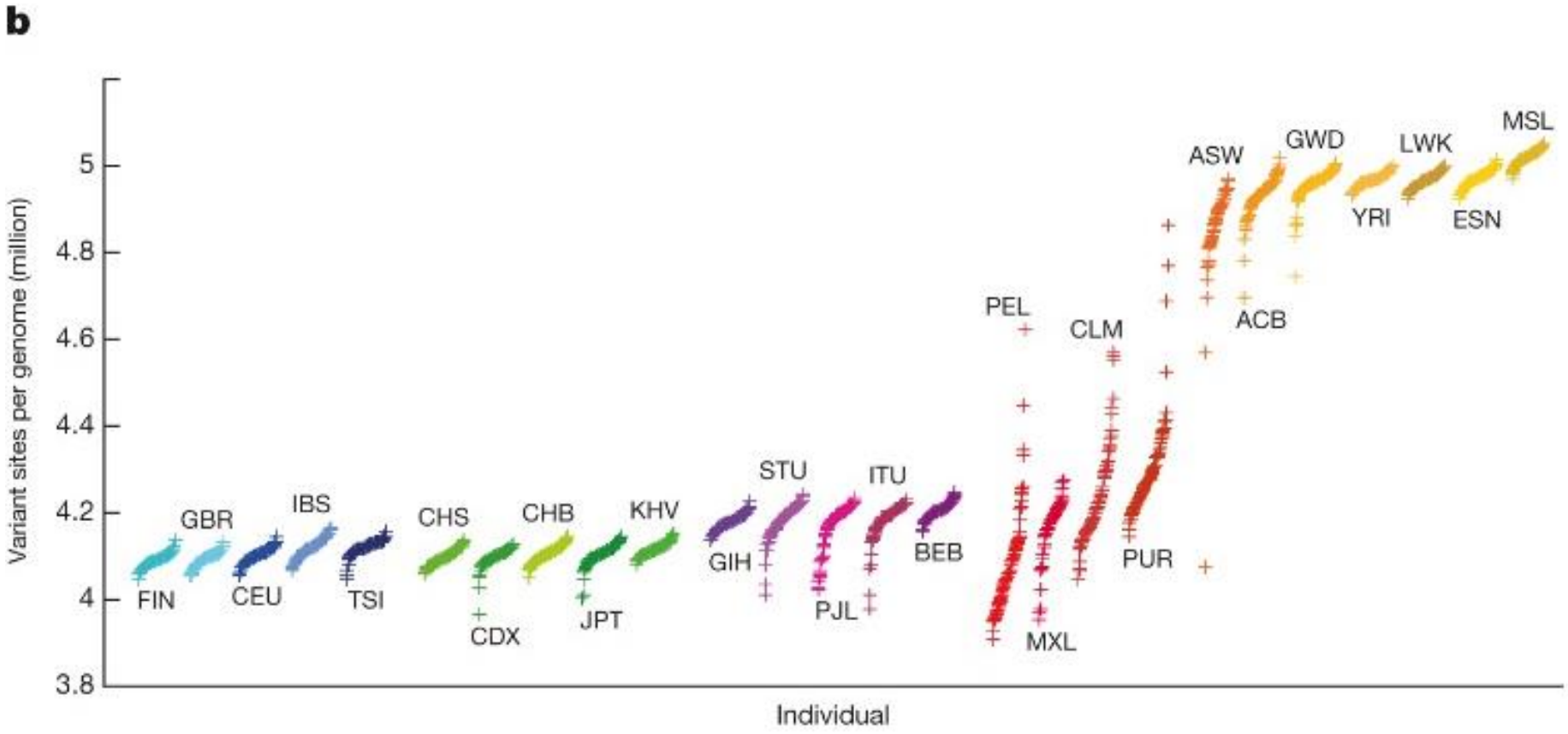
deMenocal & Stringer, Nature 2016

The “Out-of-Africa” migration is an example of a Population Bottleneck



Abel, PLoS Pathogenics 2015

Genetic diversity is greatest in Africans



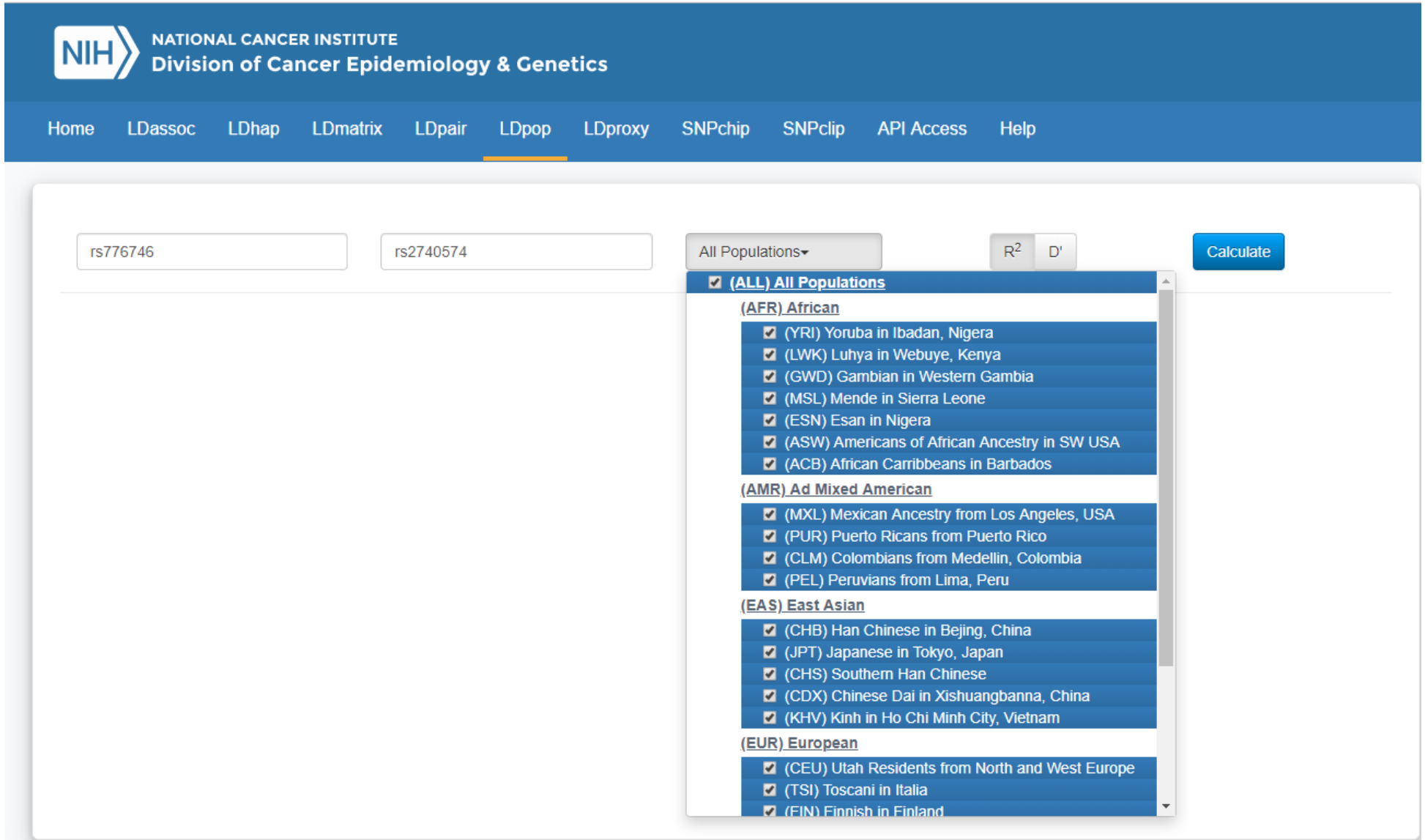
Population Color	Continental Group Color	Analysis Panel	
		AFR	Africa
		AFR	
		AFR	
		AFR	
		AFR	
		AFR/AMR	
		AFR/AMR	
		AMR	Americas
		AMR	
		AMR	
		AMR	
		EAS	East Asian
		EAS	
		EAS	
		EAS	
		EAS	
		EUR	European
		EUR	
		EUR	
		EUR	
		EUR	
		SAS	South Asian
		SAS	
		SAS	
		SAS	
		SAS	

1000 Genomes project, Nature 2015

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency

https://ldlink.nci.nih.gov/



rs776746 rs2740574 LD | rs776746 Allele Frequency | rs2740574 Allele Frequency

rs776746 rs2740574 LD

Export LD Map

Showing 1 to 32 of 32 entries **Search:**

Download Table

Population	N	rs776746 Allele Freq	rs2740574 Allele Freq	R^2	D'	LDpair
ACB	96	C: 25.0%, T: 75.0%	C: 66.15%, T: 33.85%	0.1406	0.4646	link
AFR	661	C: 18.0%, T: 82.0%	C: 76.55%, T: 23.45%	0.0566	0.281	link

Population genetics principles

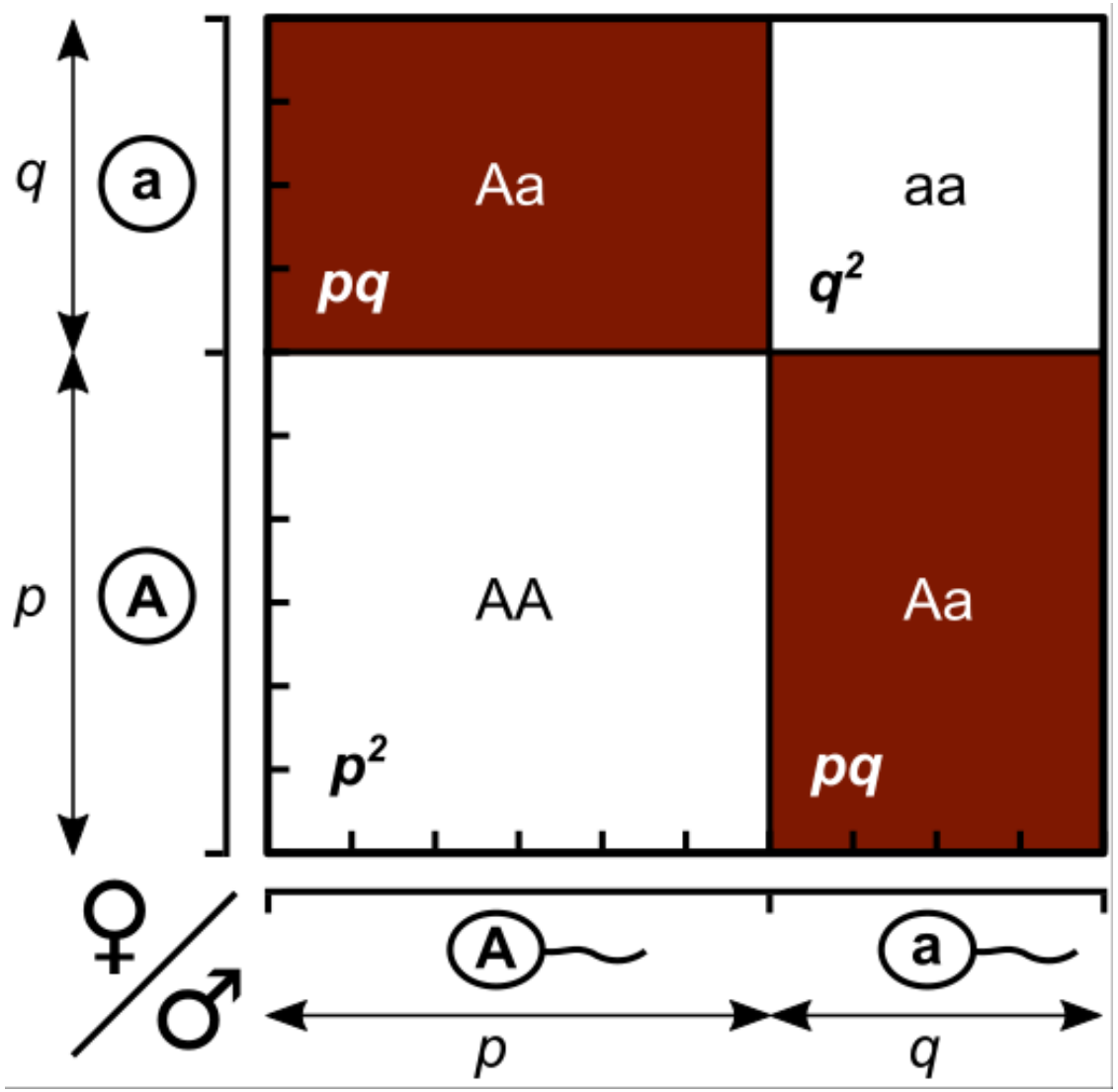
- Overall patterns of genetic variants within and between populations.
- Discipline originally developed to study evolution.
- Reflects interplay between genetic variation, phenotypes, and environmental pressures.
- Subject to mutation, mating and migration.

Yesterday: single mating pair and offspring

	A	a
A	**AA**	**Aa**
a	**Aa**	**aa**

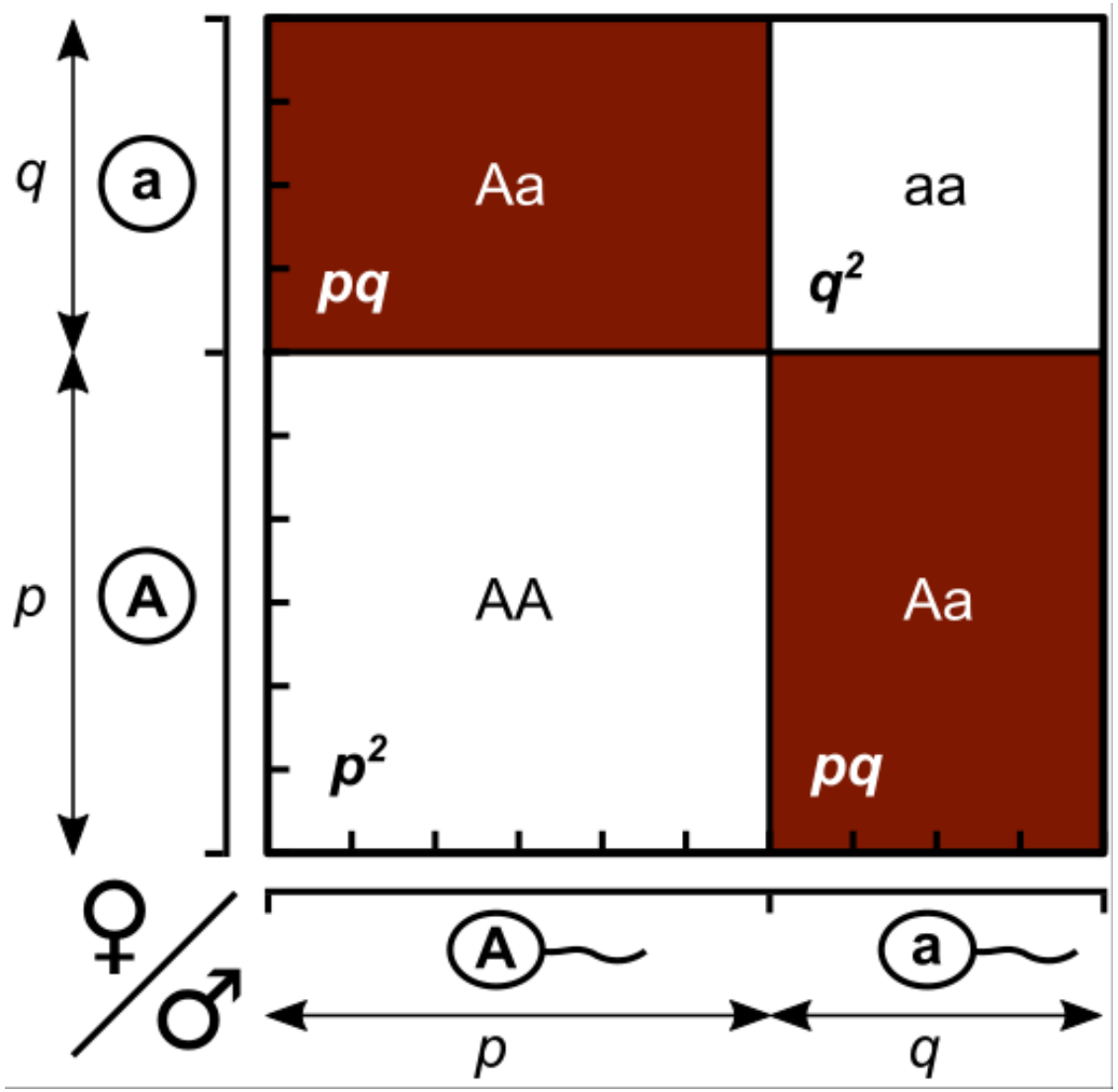
¼ (AA) + $^{2}/_{4}$ (Aa) + ¼ (aa)

Population scale expected genotype combinations



Probabilistic relationship between ALLELE frequencies and GENOTYPE frequencies

Population scale expected genotype combinations



Based on random mating:
Probability grab an “a” from the female is *q*
Probability grab an “a” from the male is *q*

So, probability grab an “a” from the female and an “a” from the male is *q*q*

The Hardy-Weinberg principle

- Assume that…
  - Population is large (coin flip likelihoods)
  - Mating is random (selective genotype matches)
  - No immigration or emigration
  - Natural selection is not occurring (all genotypes have an equal chance of surviving and reproducing)
  - No mutations

- If these assumptions are true, we say that a population is not evolving (allele frequencies stay the same) and in **Hardy-Weinberg Equilibrium**

The Hardy-Weinberg principle

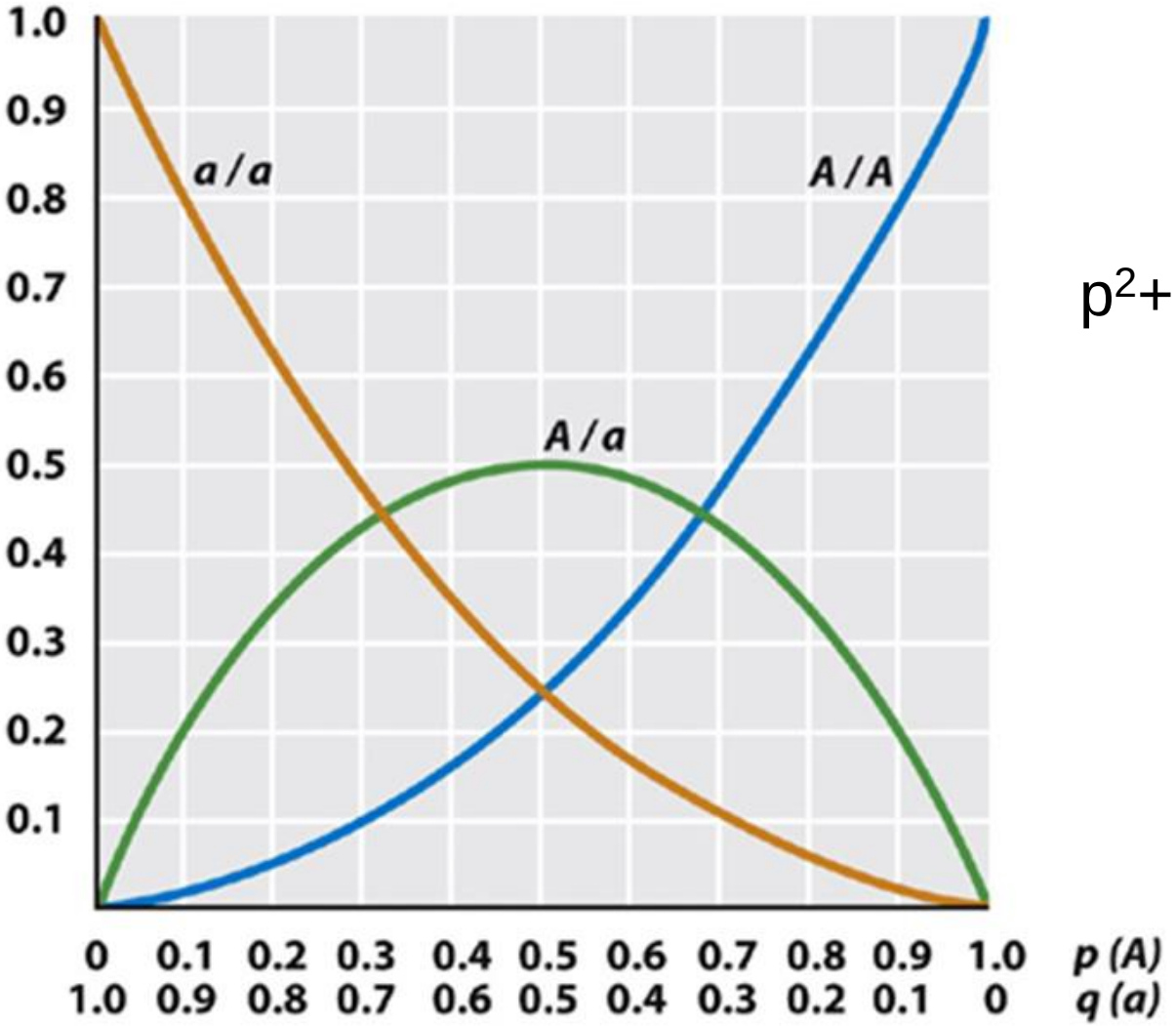


Figure 17-5
Introduction to Genetic Analysis, Ninth Edition

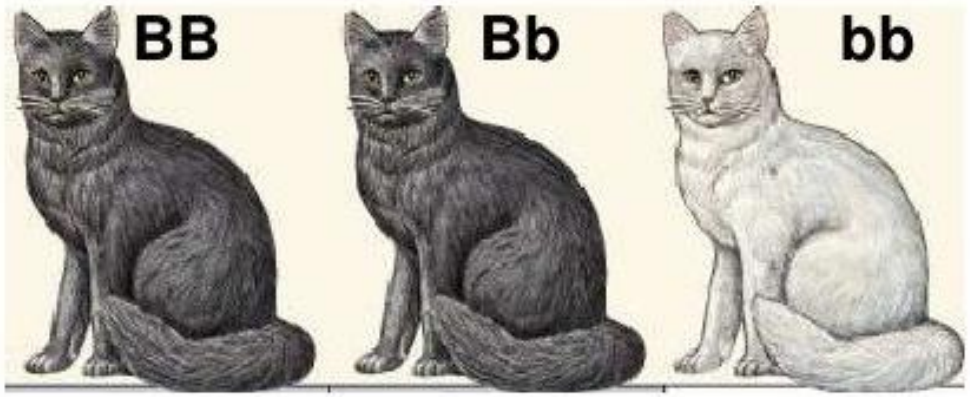
$p+q=1$ (allele frequencies)

$p^2+2qp+q^2=1$ (genotype frequencies)

HWE example

- Assume 100 cats (200 alleles) with alleles B and b. B allele is dominant and results in black coloring. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?

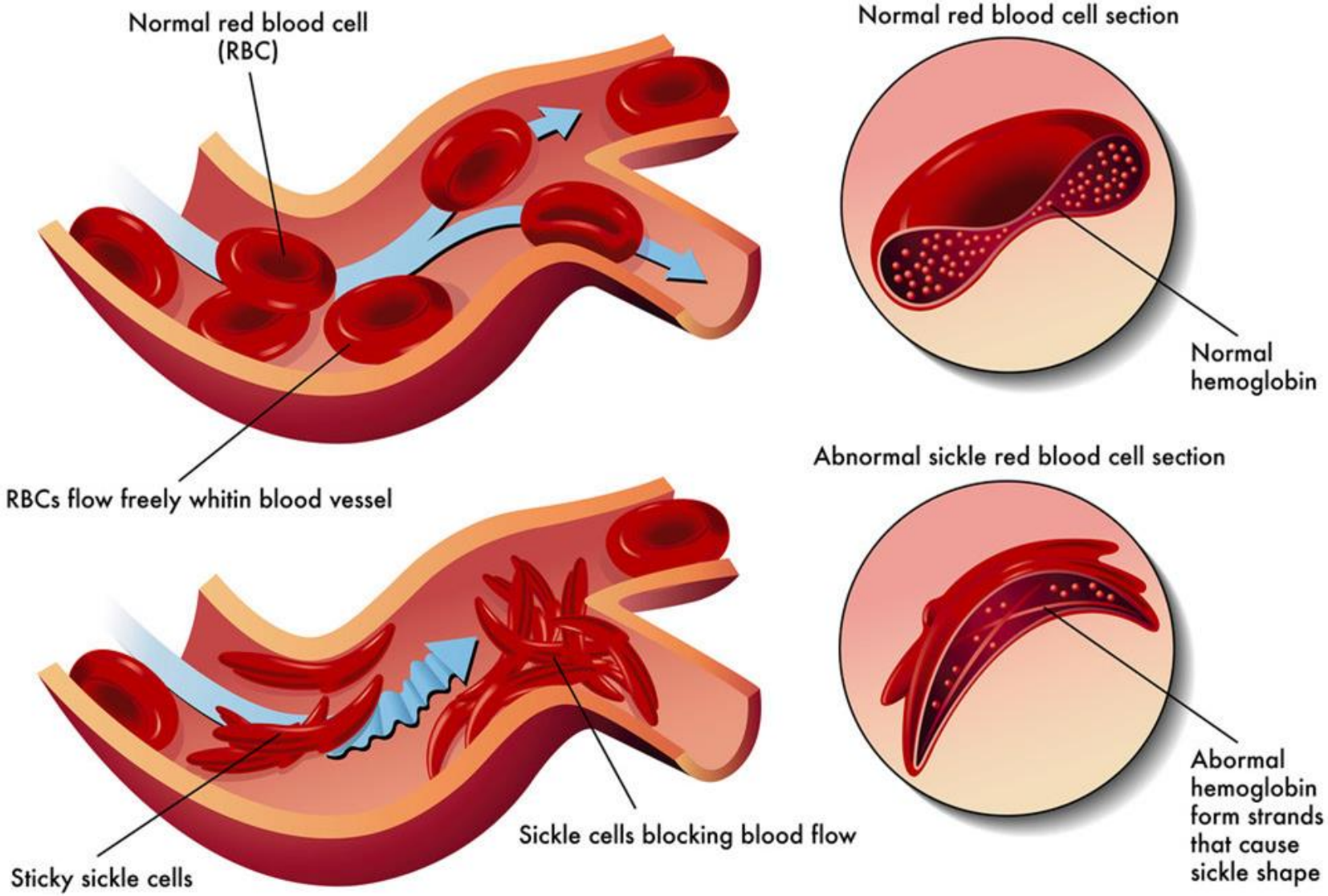
- $p+q=1$
- $p^2+2qp+q^2=1$



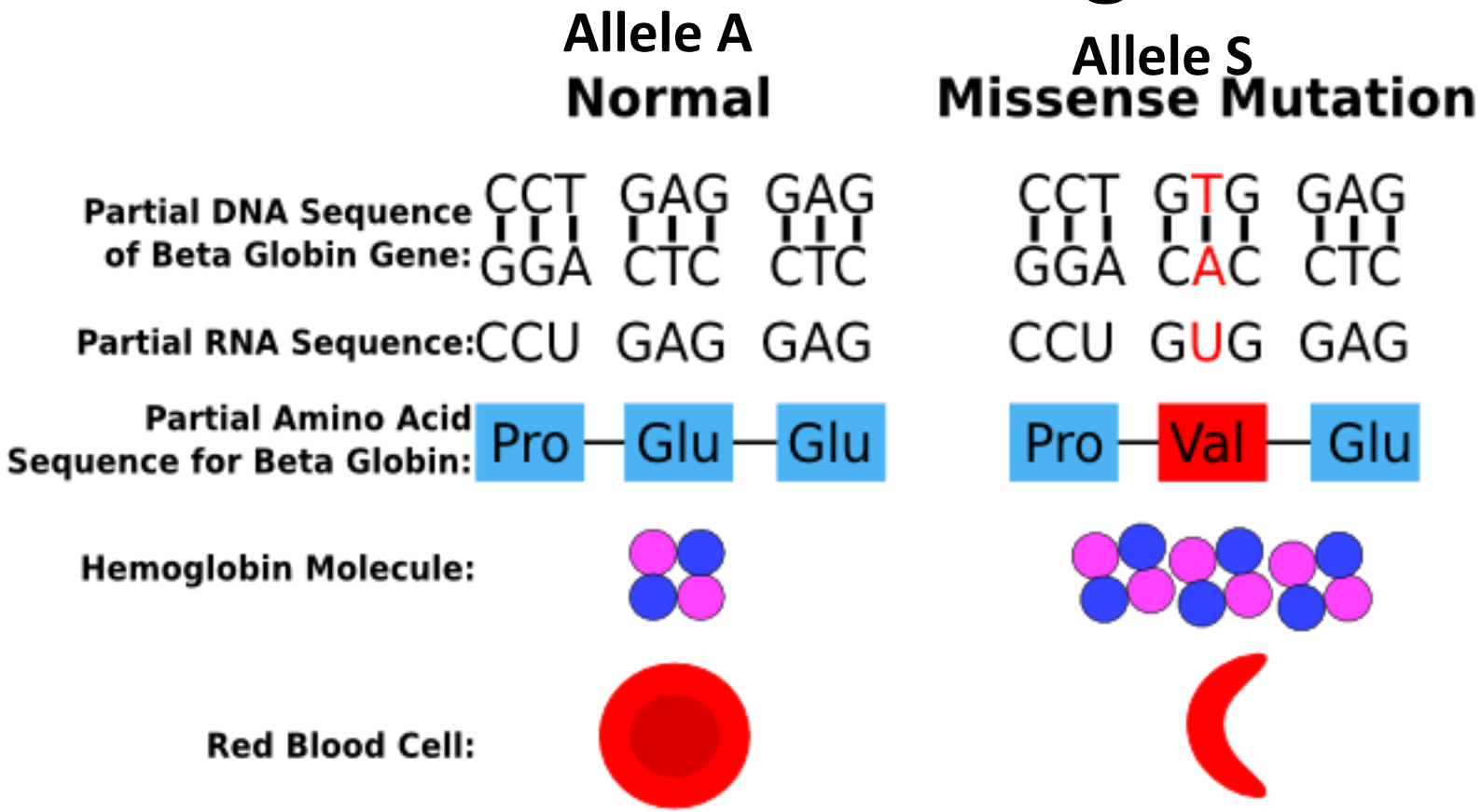
Sickle Cell Anemia Symptoms

1. Fatigue
2. Pain
3. Arthritis
4. Frequent bacterial infections
5. Sudden pooling of blood in internal organs
6. Lung and heart failure, tissue death, eye damage

Sickle Cell Anemia



Sickle Cell Anemia -- single amino acid change



Hemoglobin B subunit

Quite prevalent!

80,000 people in the US

200,000 people in Africa (9% of children have sickle cell disease)

120,000 people in India (in 1988!)

S allele vs. AS and SS genotypes

S allele frequency: 0.20

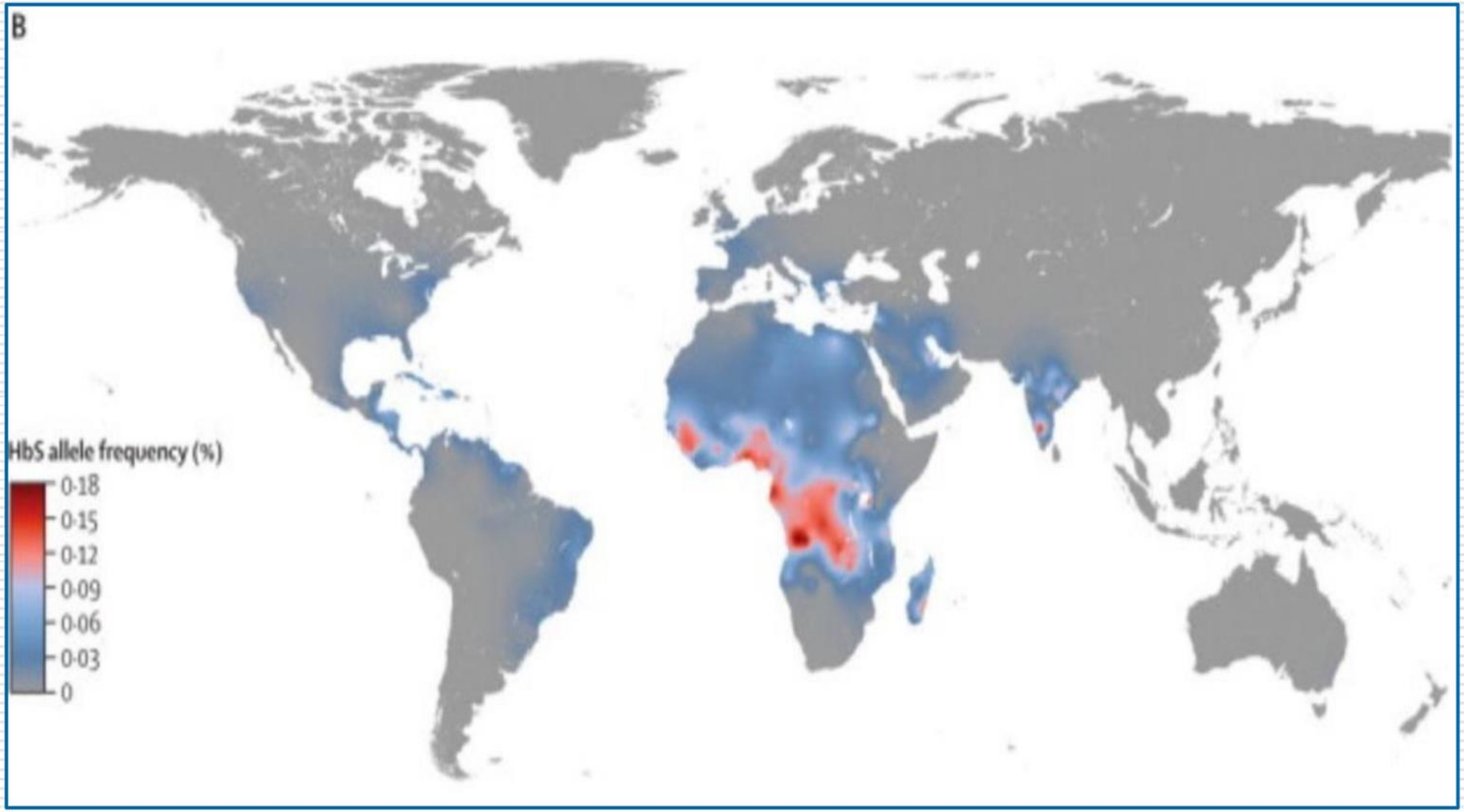
(Among adults -- each with two alleles -- the S allele comprises 20% of the alleles)

Use our Hardy Weinberg Equilibrium equations to calculate how many people out of 1000 would have each of your expected genotypes:

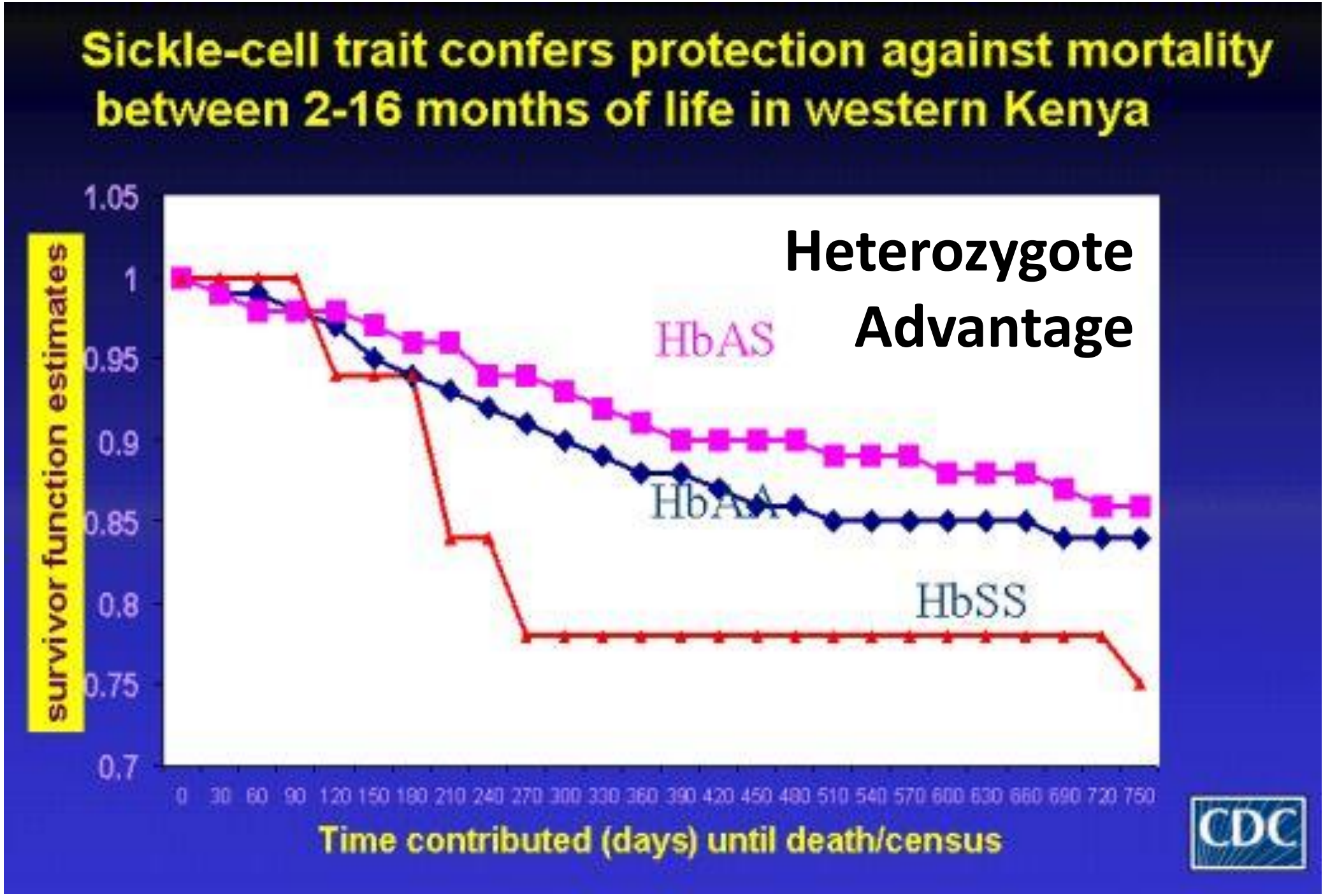
$$1=p+q$$

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

Frequencies of S allele of *HBB*



Piel et al. 2013. Lancet 381:142–51


Sickle-cell trait confers protection against mortality between 2-16 months of life in western Kenya
Heterozygote Advantage
HbAS
HbAA
HbSS
survivor function estimates
1.05
1
0.95
0.9
0.85
0.8
0.75
0.7
0 30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570 600 630 660 690 720 750
Time contributed (days) until death/census
CDC

Are the frequencies really that off?

χ2-goodness-of-fit (GOF) tests with 1 degree of freedom

Sum of observed minus expected

- *O* = observed counts, *E* = expected counts, sum across genotypes

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

Compare to chi-square distribution to determine whether the deviance is significant.

Deviation from Hardy Weinberg?

Check chi-square distribution with 1-degree of freedom:

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086

Deviation from Hardy Weinberg?

Check chi-square distribution with 1-degree of freedom:

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086

Reasons to defy Hardy Weinberg equilibrium

- True selective pressures
- Genotyping error! (most common reason)
- Undetected population substructure (differences in ancestry)
- Non-random procreation

Many statistical tests rely on SNPs being in hardy weinberg equilibrium, so we test this chi-square test on every SNP in a study.

Hardy-Weinberg and LD are useful tools to detect evolutionary forces acting on a population such as population bottlenecks

deMenocal & Stringer, Nature 2016

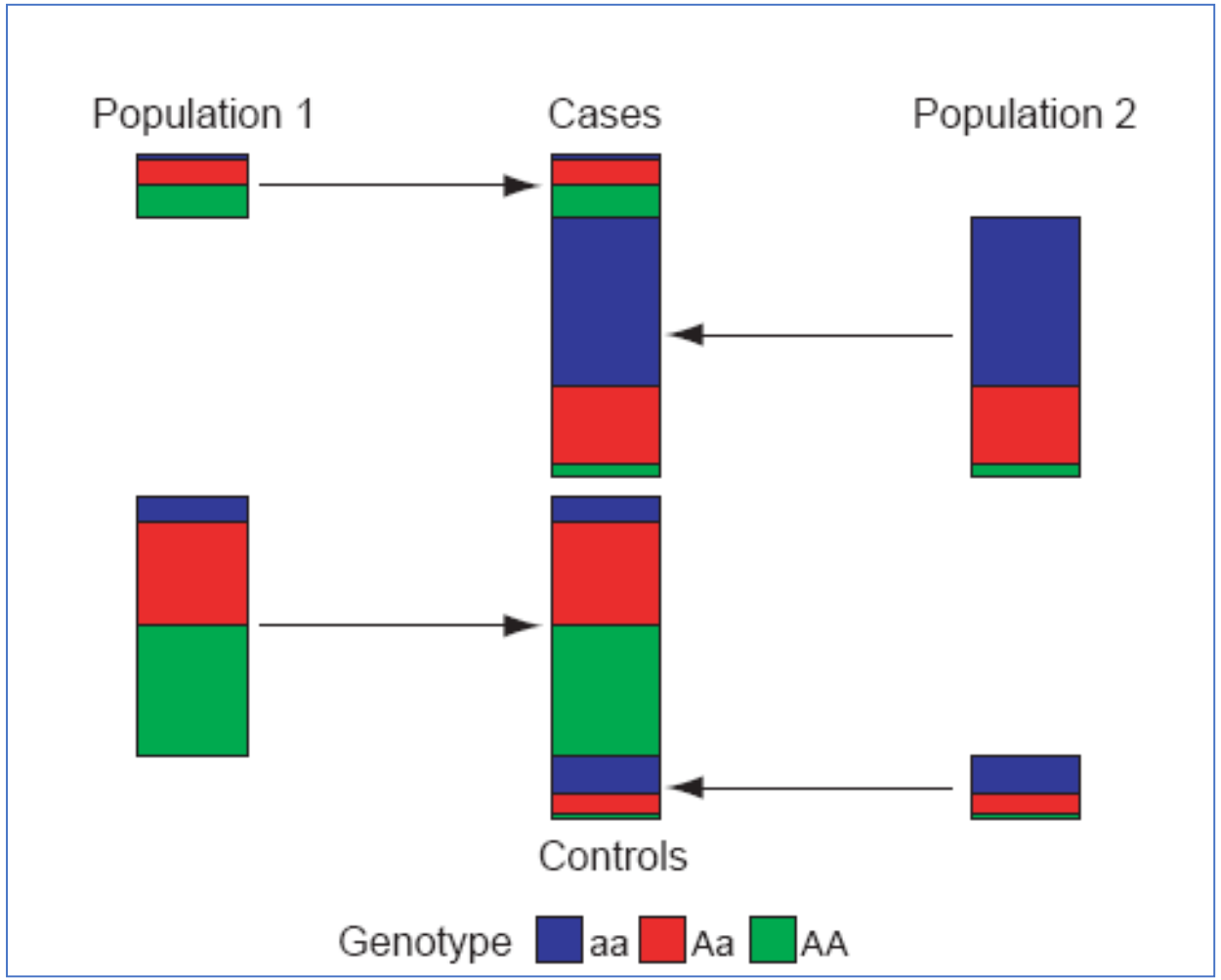
Ancestry in genetic data

Assume we conduct a case-control GWAS…

- Our cases were collected in Africa
- Our controls were collected in Asia
- If we find multiple SNPs that are significantly more/less common in cases than controls, do we believe that these results are due to association with disease or population differences?

Population Stratification - Confounding by ancestry

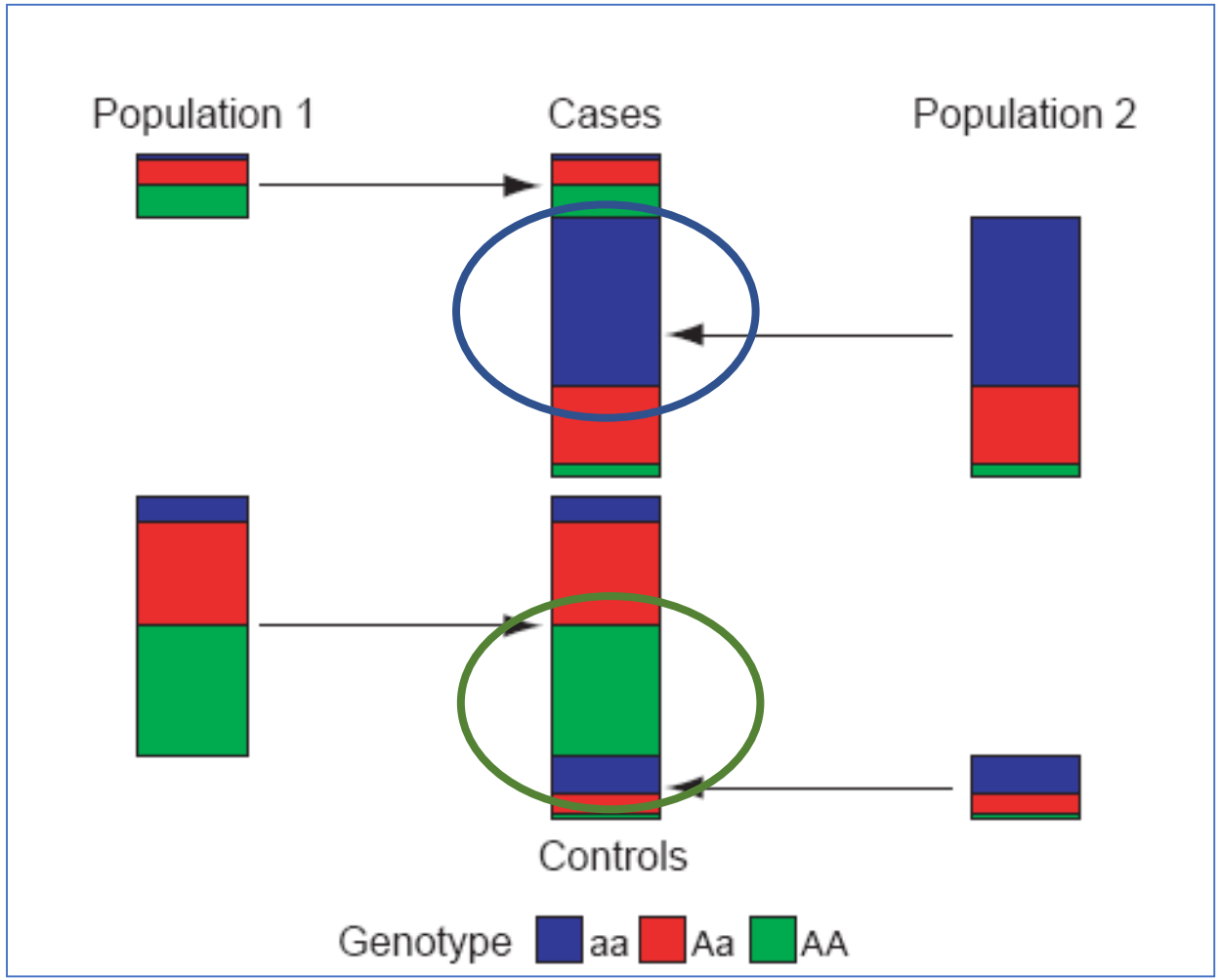
Group differences in ancestry
AND outcome



Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

Population Stratification - Confounding by ancestry

Group differences in ancestry
AND outcome

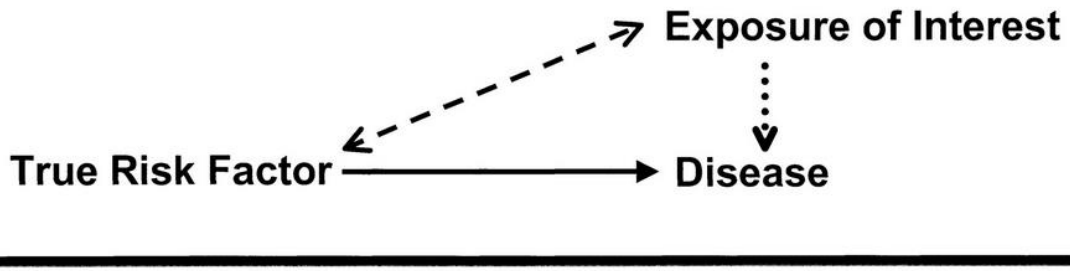


Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

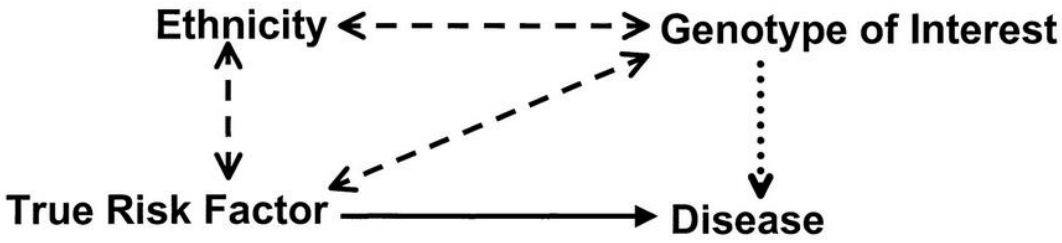
Population Substructure

The presence of a systematic difference in allele frequencies between subpopulations due to different ancestry

Confounding



Population Stratification



Assume we conduct a case-control GWAS…

- Our cases were collected in Africa
- Our controls were collected in Asia
- If we find multiple SNPs that are significantly more/less common in cases than controls, do we believe that these results are due to association with disease or population differences?

Assume we conduct a case-control GWAS…

- Our cases were collected in Africa
- Our controls were collected in Asia
- If we find multiple SNPs that are significantly more/less common in cases than controls, do we believe that these results are due to association with disease or population differences?

This is the extreme case, what about more subtle differences?

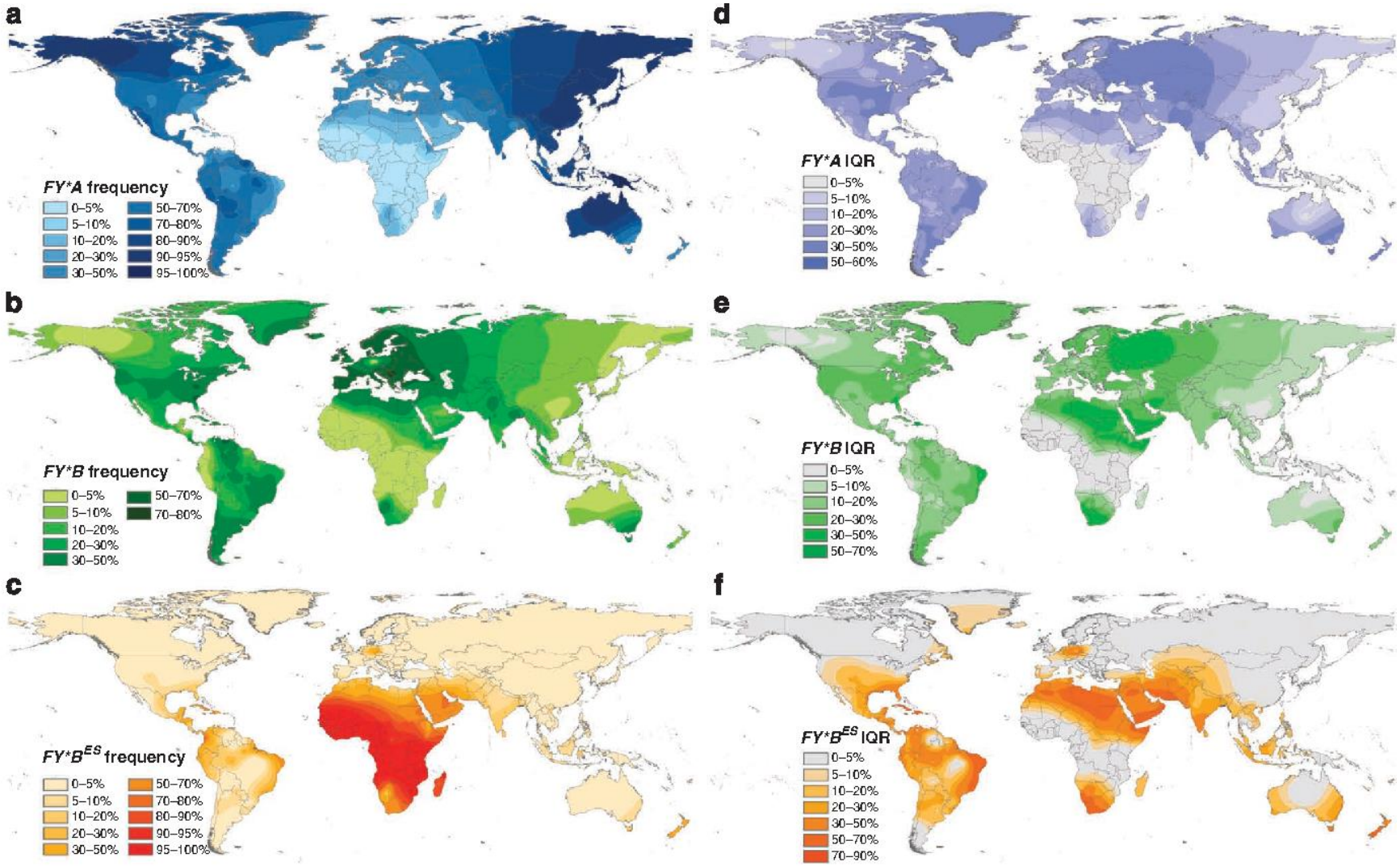
We can use genetic data to determine ancestry and to adjust for ancestry in association studies.

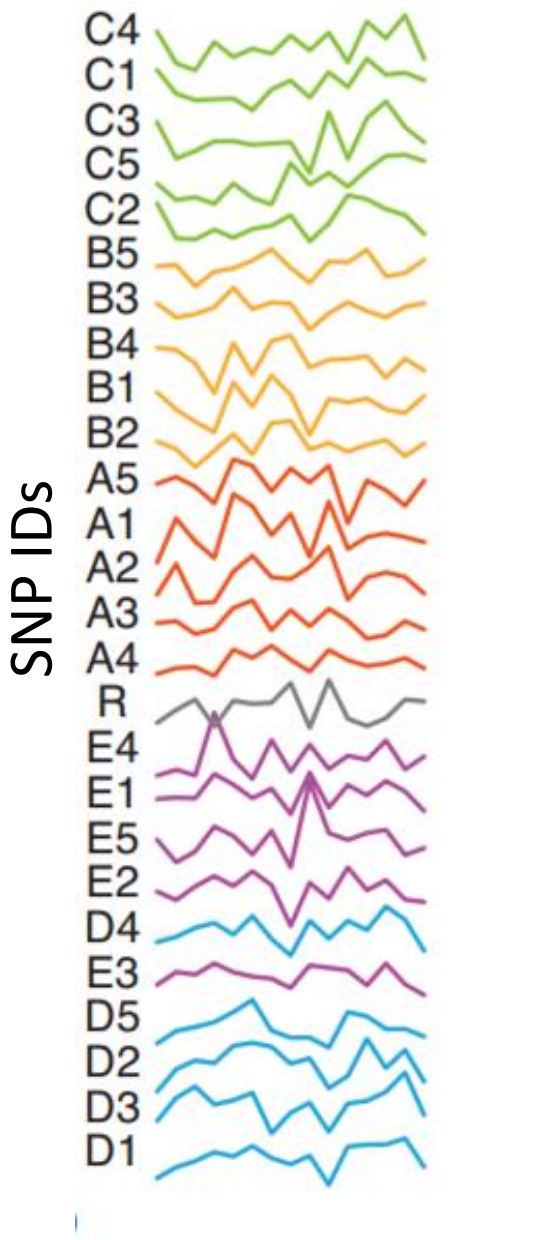
But these are very obviously different populations…

What about more subtle differences?

Relics of human history are present across the genomes, making some genetic variants more/less common in different populations, even if the variants don’t have any impact on human traits or health.

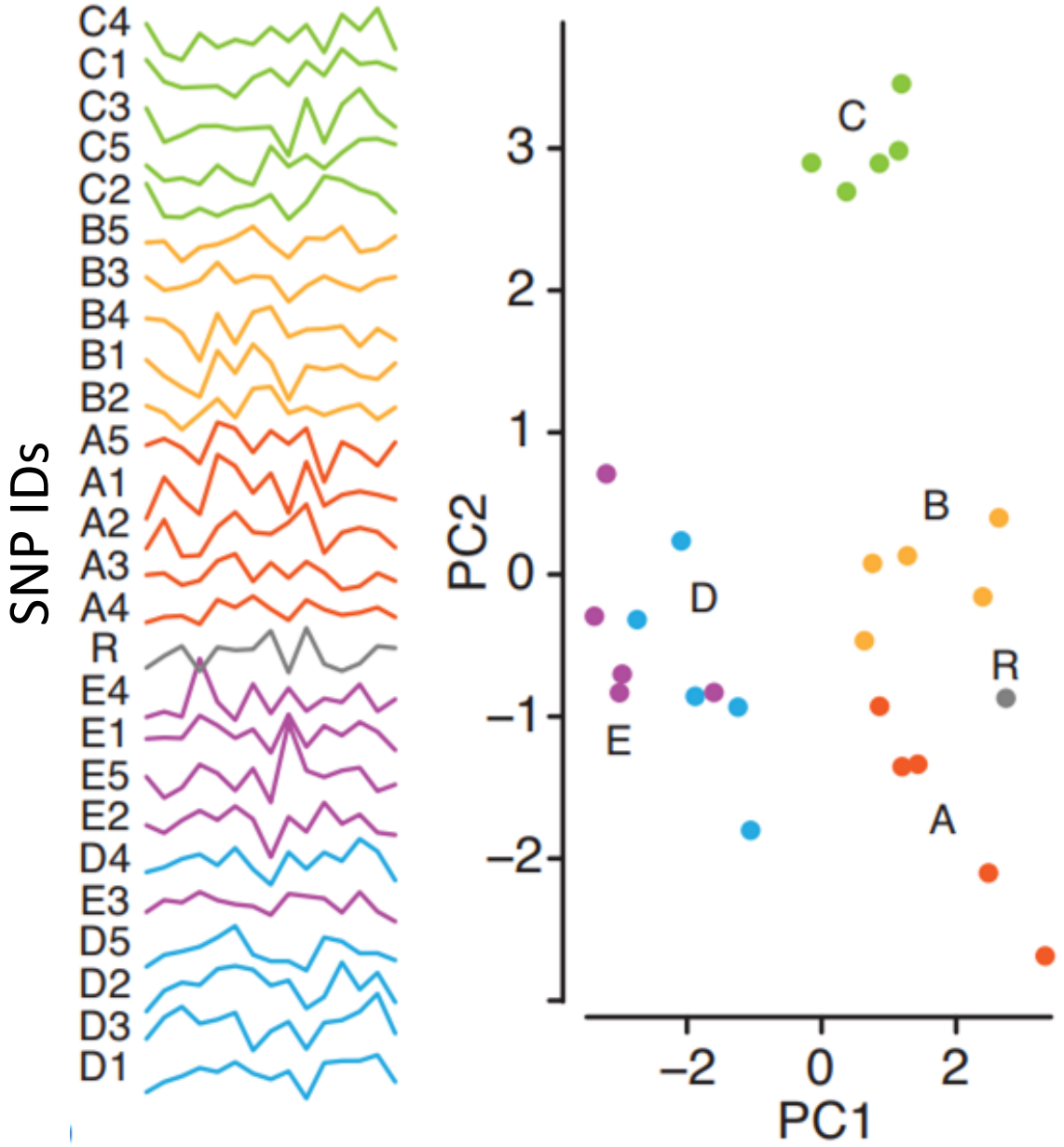
Consider all gradients together…





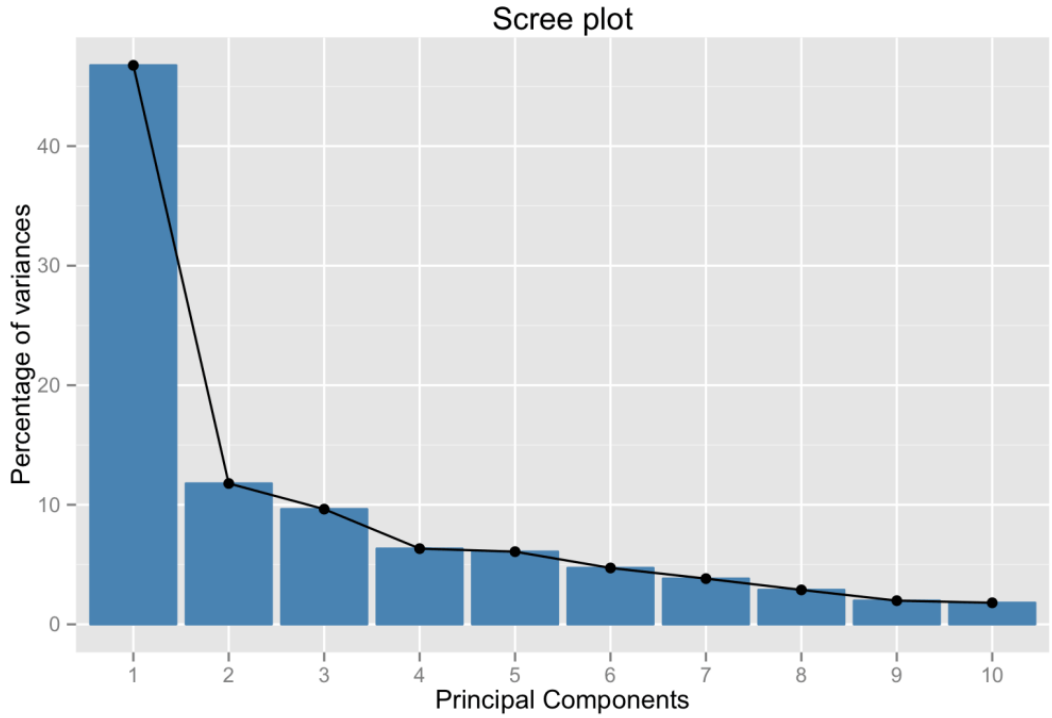
Principal component analysis (PCA)

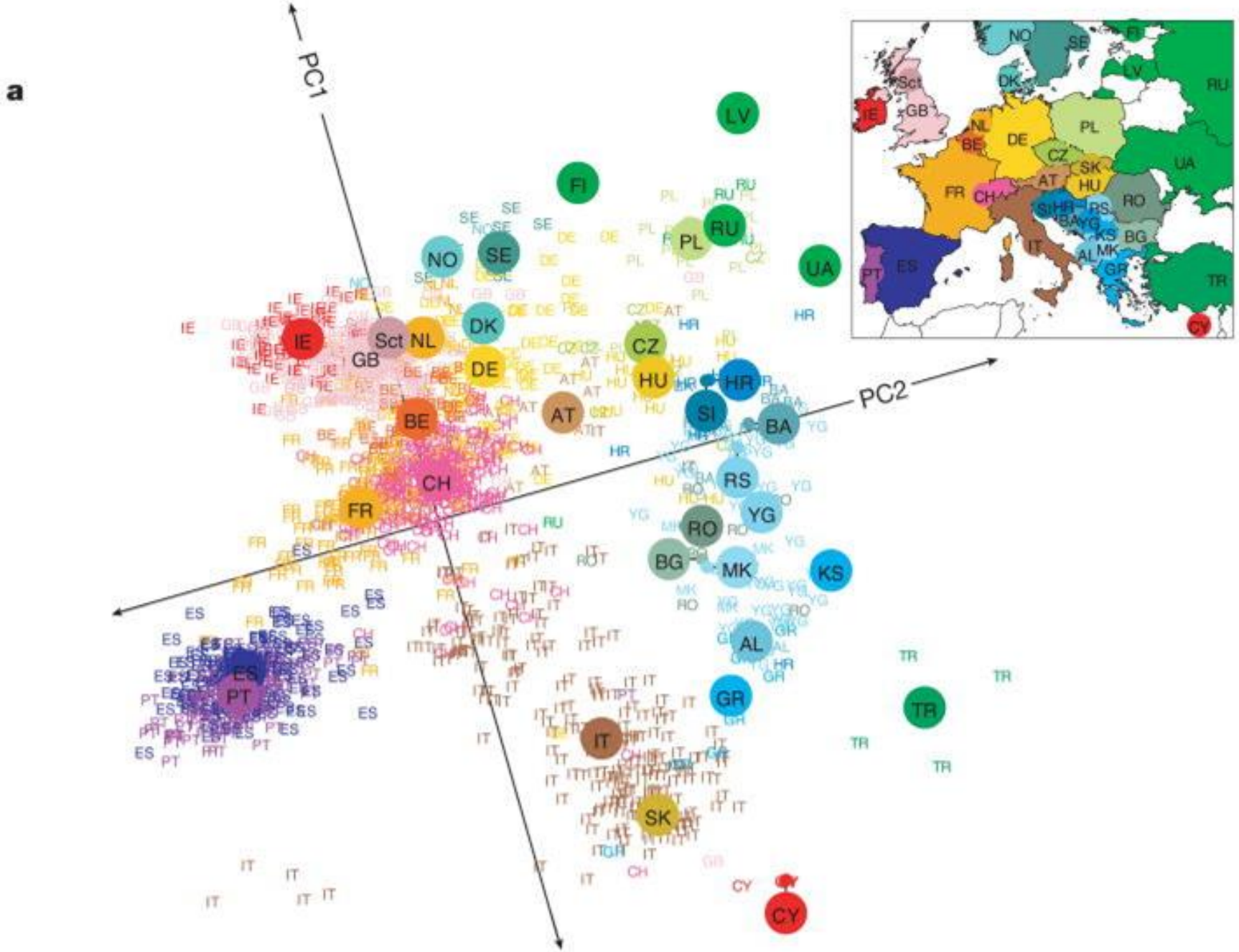
Principal component analysis (PCA)



Principal Component Analysis (PCA)

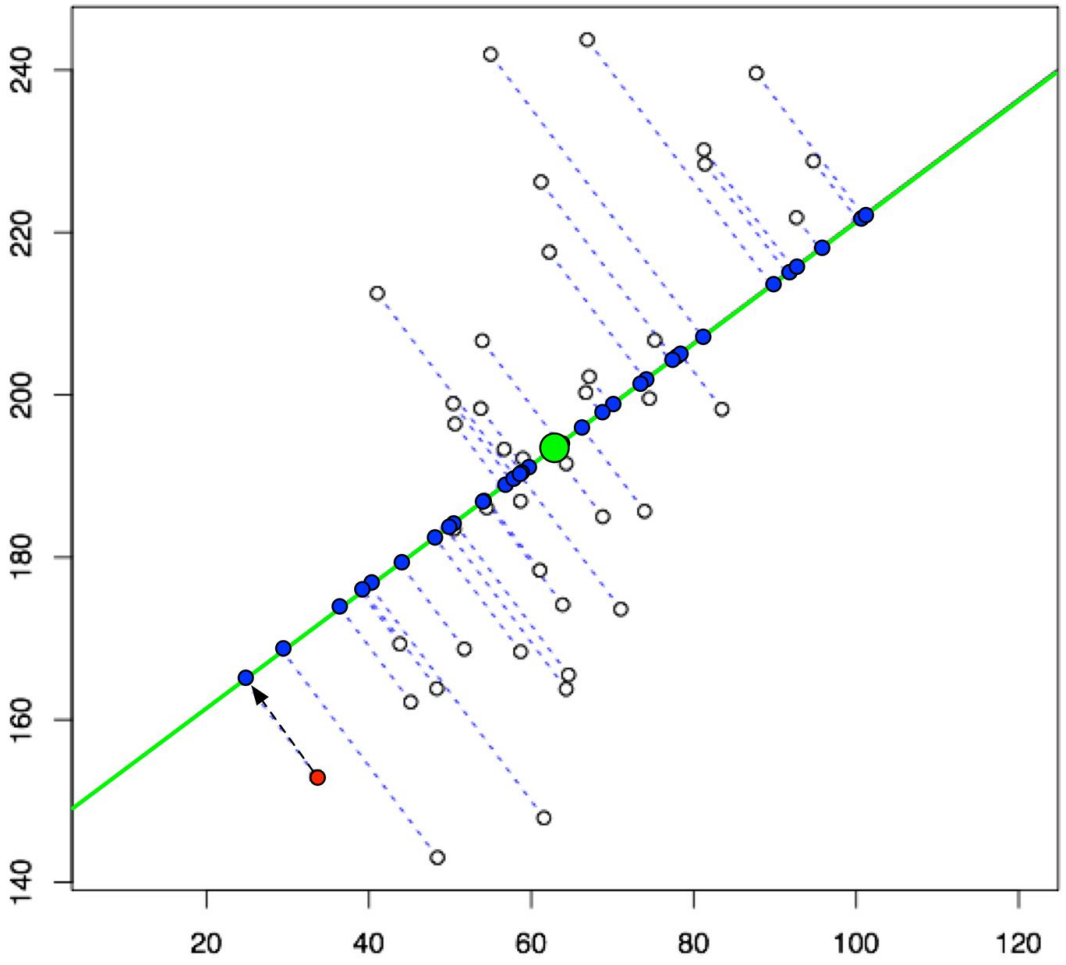
- Reduces the dimension of the data from many, many variables to a small set (“principal components ” or “PCs”– eigenvectors) that still explain the majority of variation seen in the data.
- The first PC (PC1) is constructed to explain as much of the variation as possible, the second (PC2) is constructed to explain as much of the remaining variation as possible….
- The more correlation in the data (i.e. between SNPs), the fewer PCs are needed to explain most of the variation.
- Each PC is a linear combination of the original variables (SNPs)
- PCs are independent of each other.





Novembre, 2008 Nature

How does PCA work?

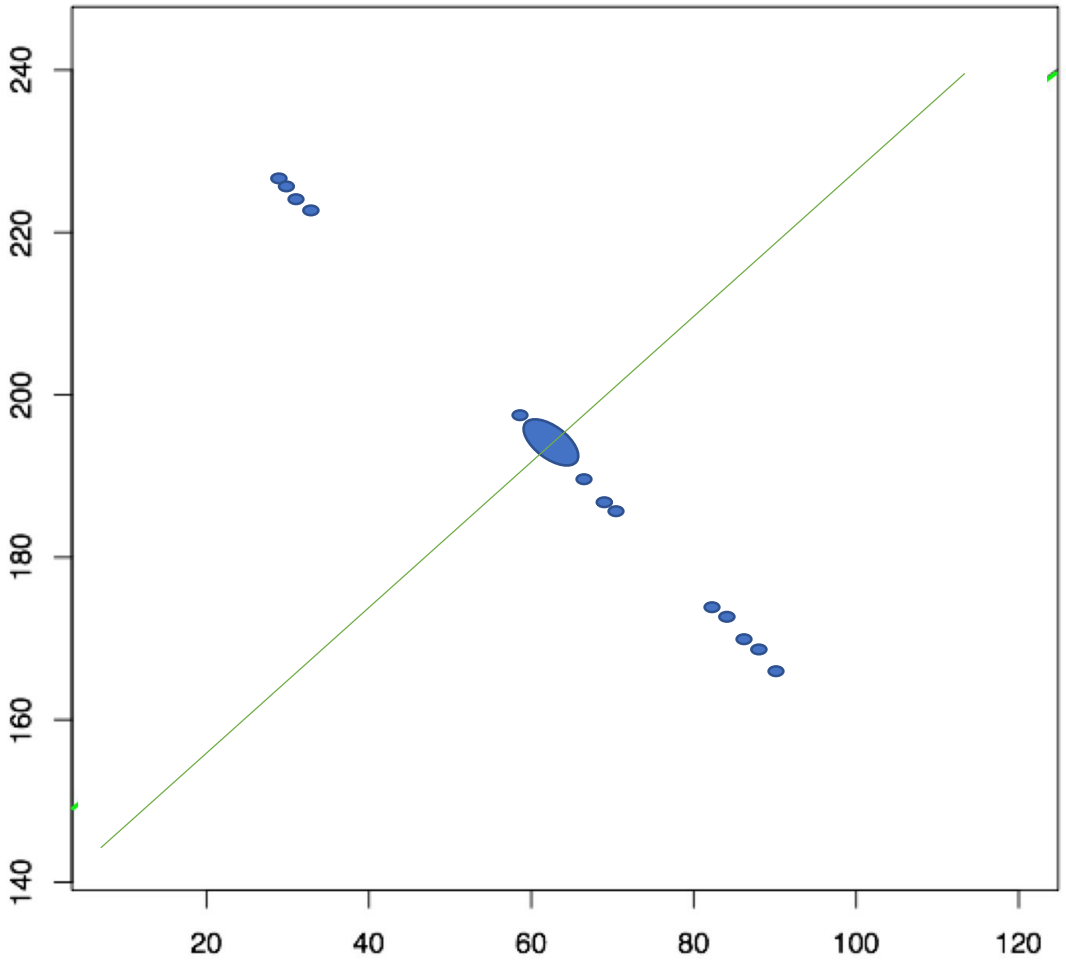


Each PCA maximizes variance and minimizes error

Basically to “absorb any systematic differences.

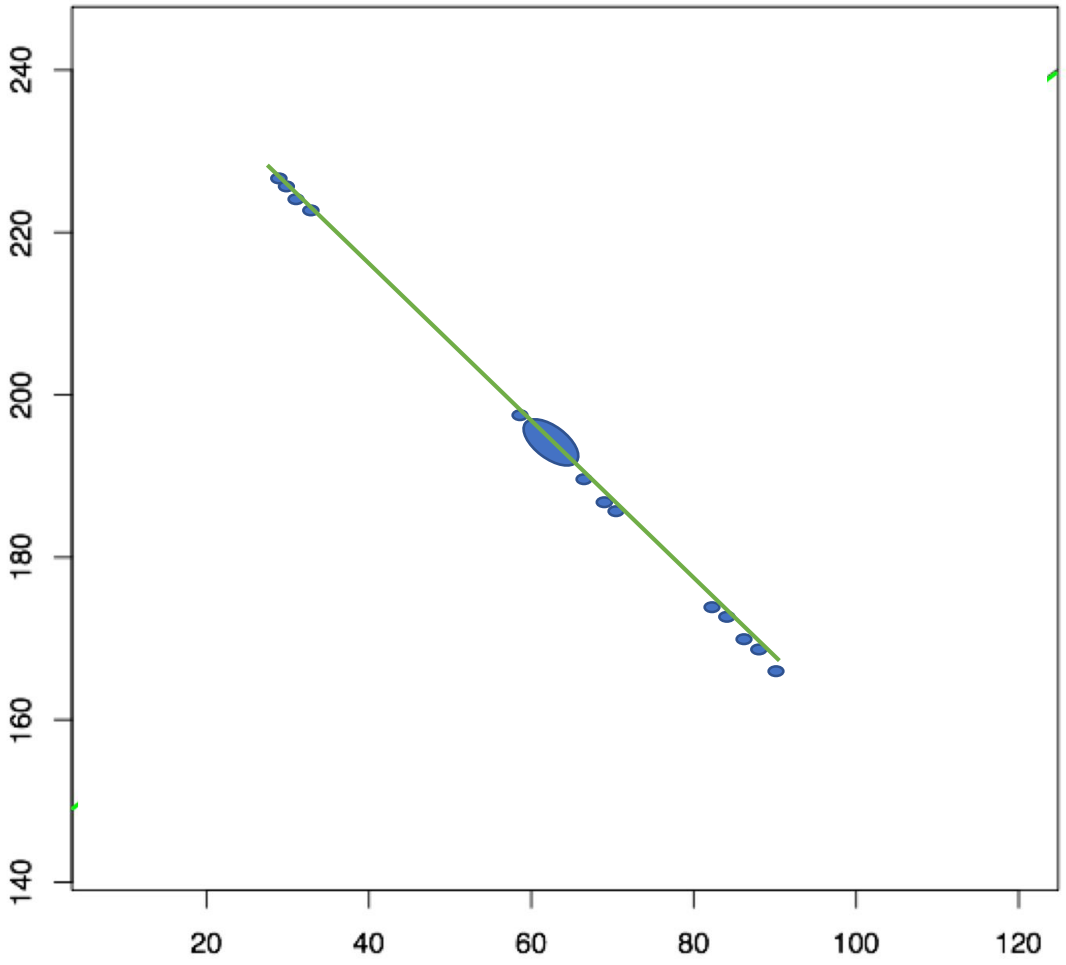
Reduces data dimensions.

Second PCA “soaks up” leftover variance



Remove the dimension from PC1 so that every point is squished together with zero variance along PC1 axis

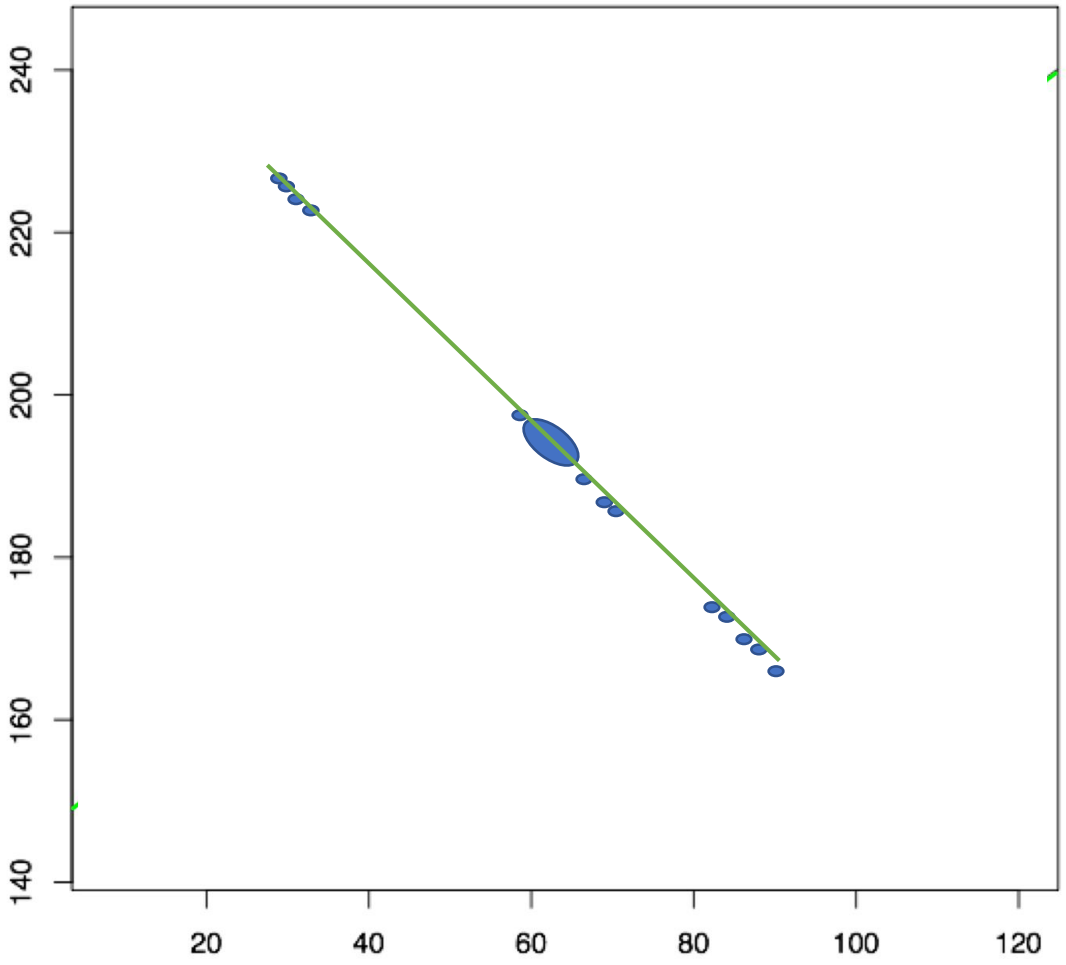
Second PCA “soaks up” leftover variance



Remove the dimension from PC1 so that every point is squished together with zero variance along PC1 axis

Now, PC2 absorbs the most variance from whatever is left after PC1 dimension is removed

Second PCA “soaks up” leftover variance

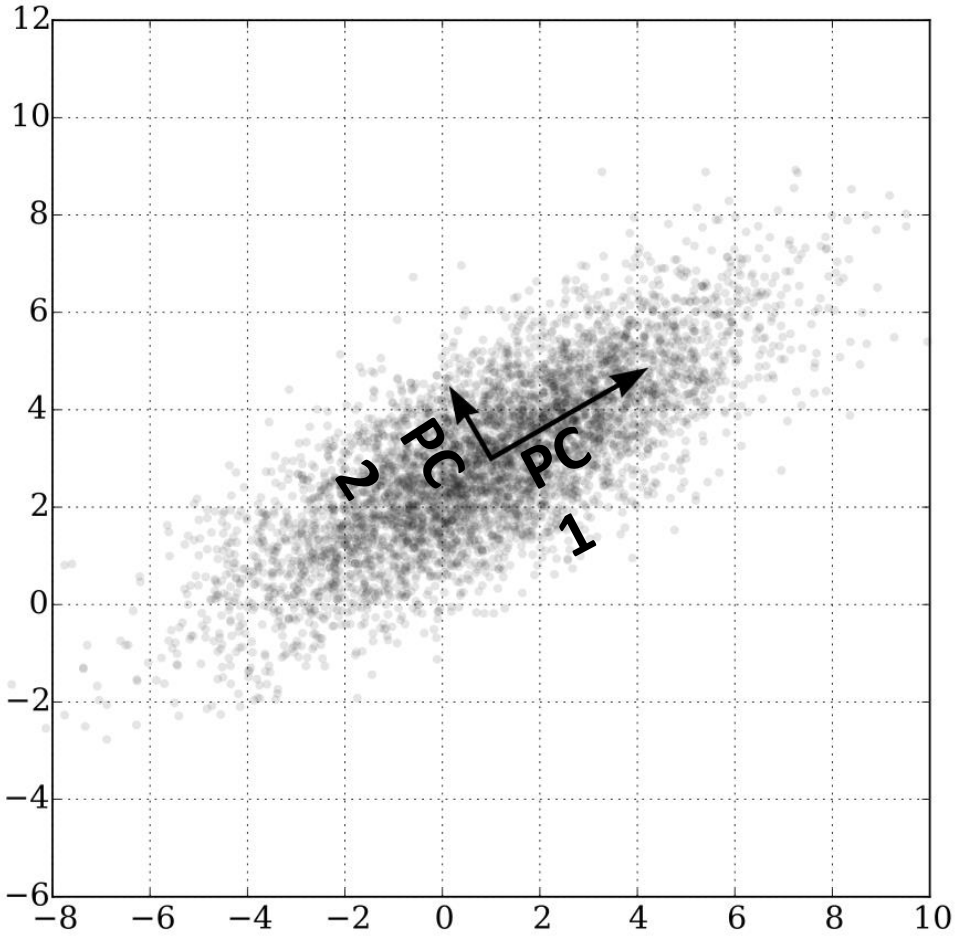


Remove the dimension from PC1 so that every point is squished together with zero variance along PC1 axis

Now, PC2 absorbs the most variance from whatever is left after PC1 dimension is removed

This is 2D, but potentially the number of dimensions as the number of variants.

PC1 vs PC2 vs PC3…



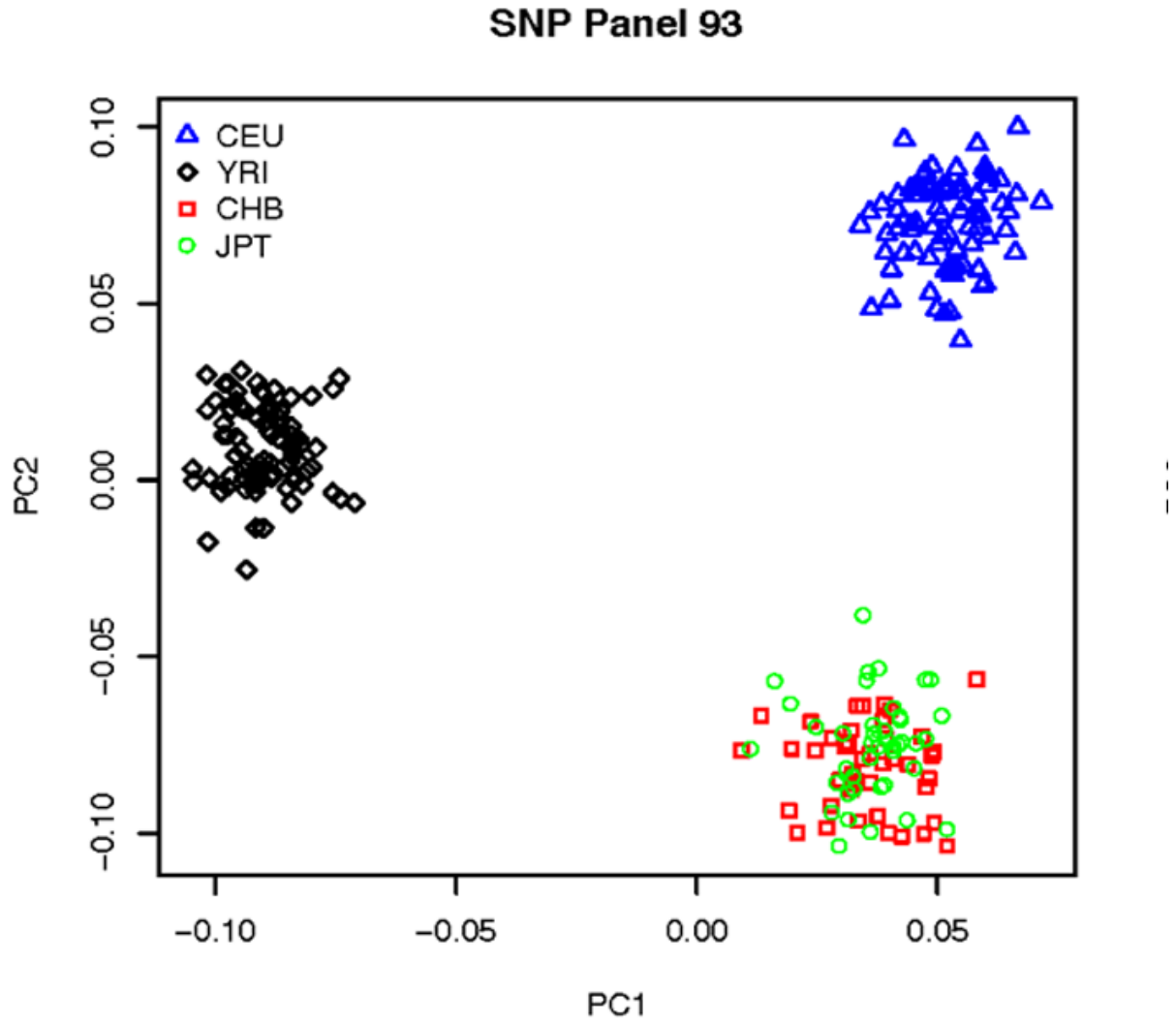
Each PC “absorbs” as much variance as possible in a new direction compared to all of the PCs before it.

The more correlation in the data (i.e. between SNPs), the fewer PCs are needed to explain most of the variation.

Each PC is a linear combination of the original variables (SNPs)

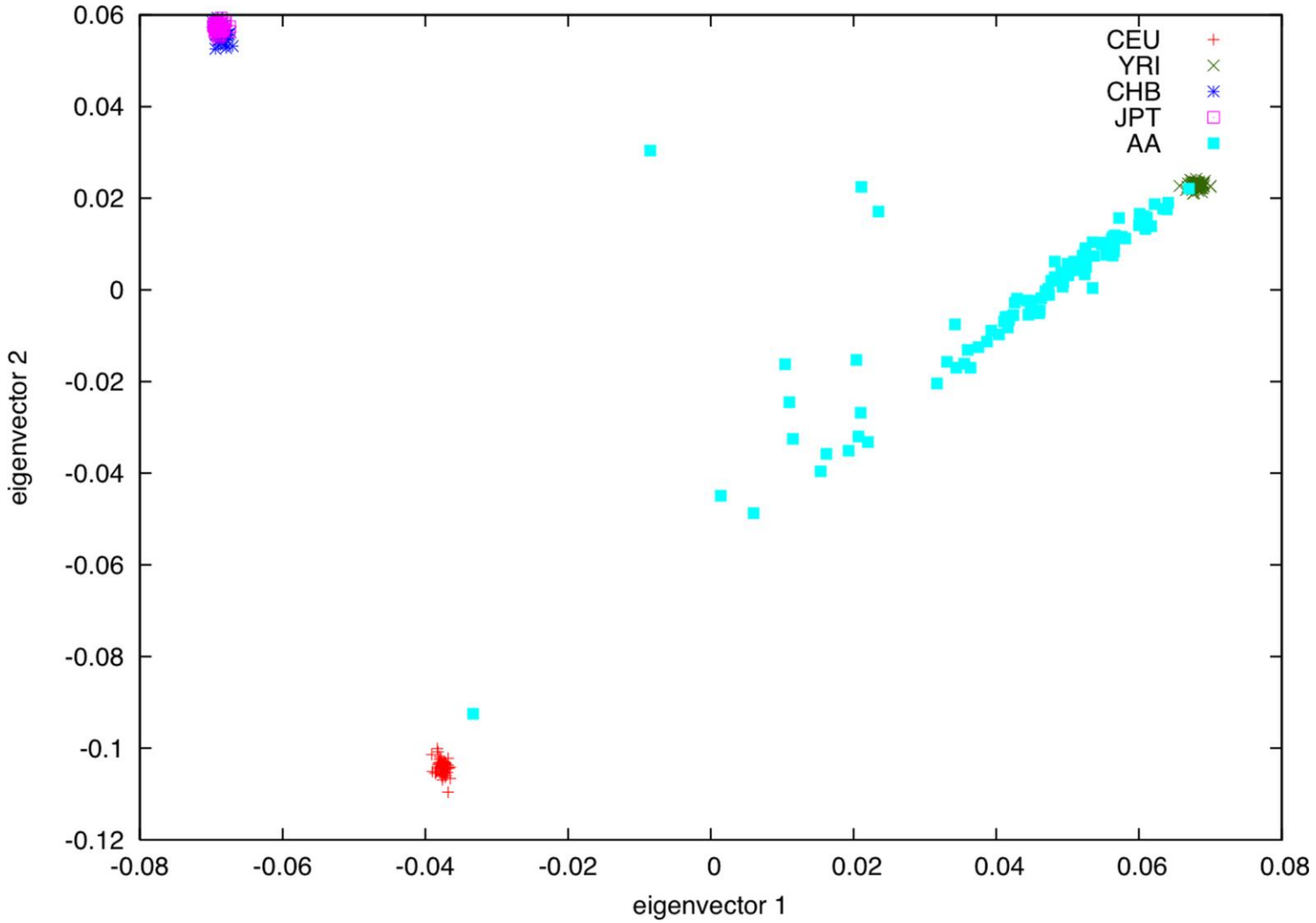
PCs are independent of each other.

The first two PCs can help distinguish ancestral populations



Hou, PLoS One 2011

The first two PCs can help distinguish ancestral populations



Price, PLoS Genetics 2008

Include top PCs in genetic association study

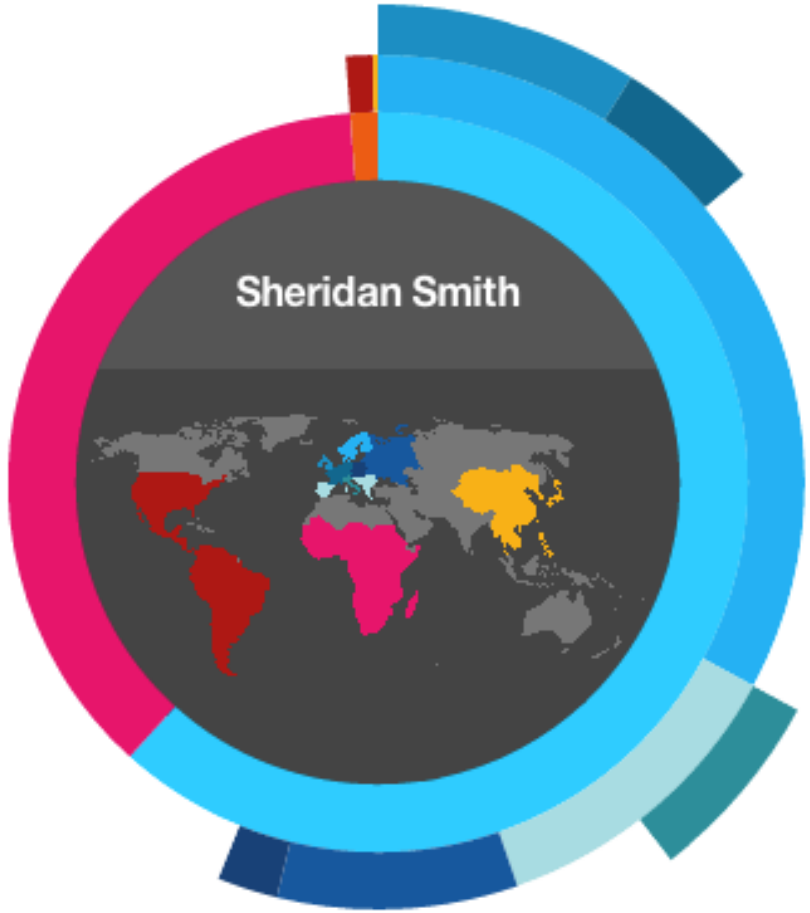
Phenotype = m*genotype + aPC1 +bPC2 + cPC3 + dPC4 + ePC5 + f

Accounts for underlying gradient patterns that aren’t truly associated with a phenotype, but may appear so due to allele frequency differences.

Ancestry pattern as a benefit in genetic epidemiology studies

Map View

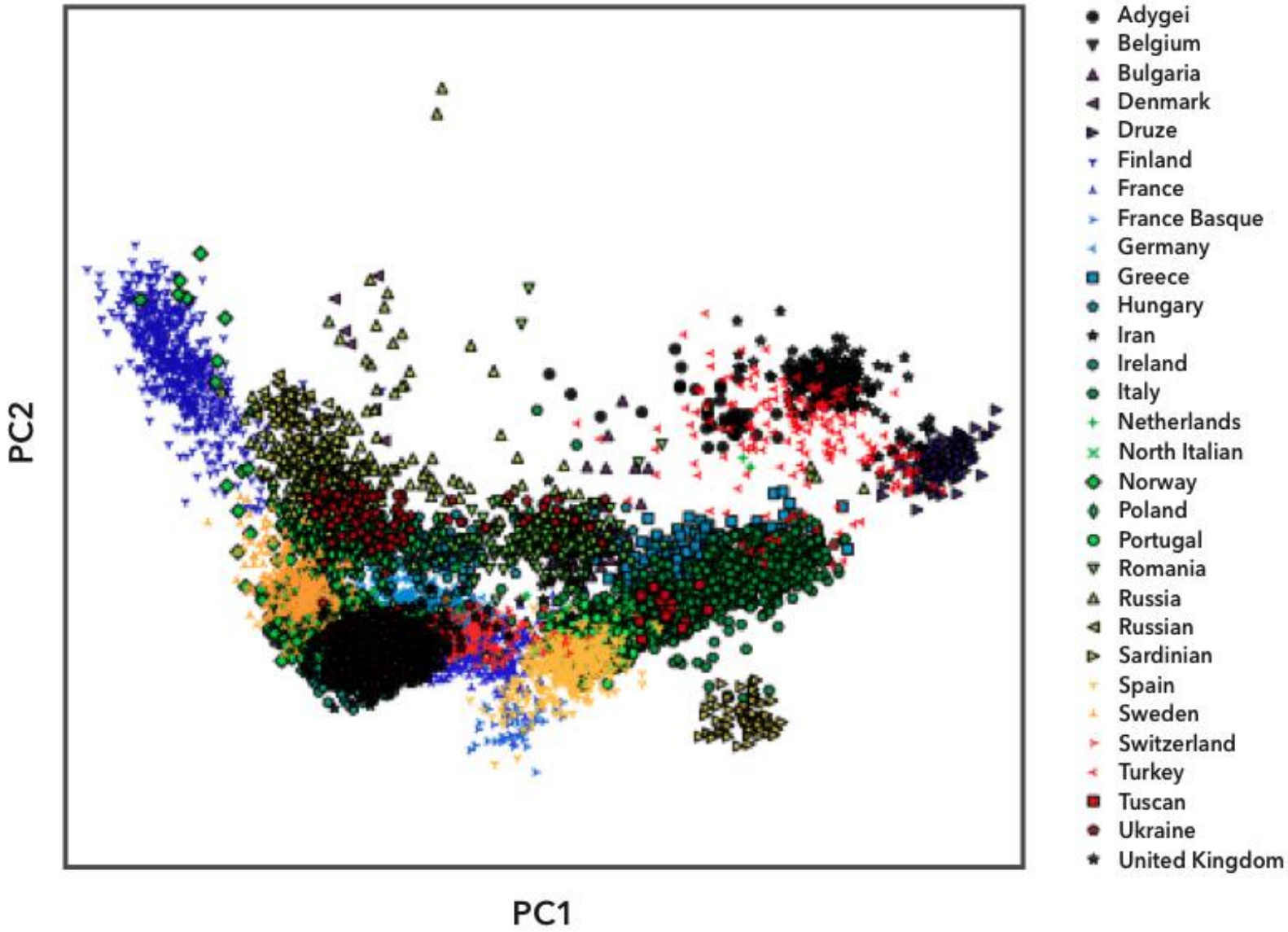
Sub-regional Resolution



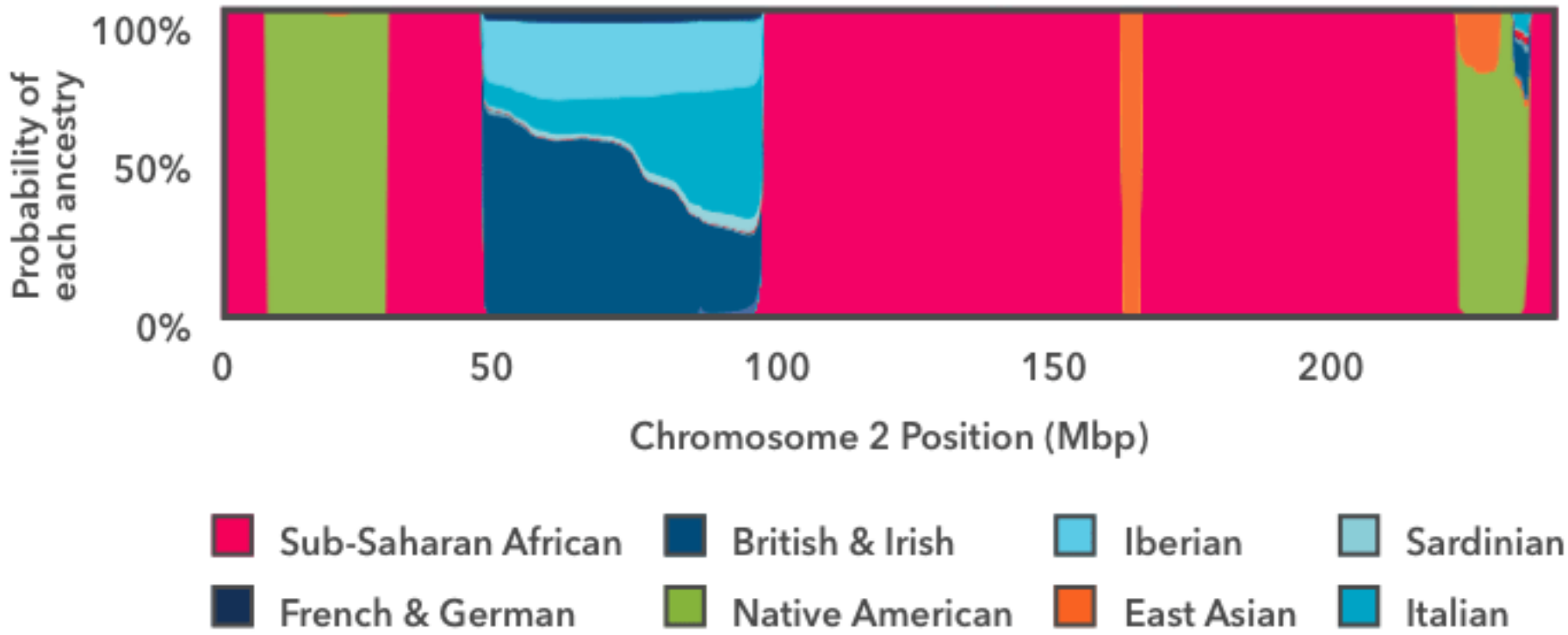
Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.

61.7%	European
	Northern European
8.9%	British and Irish
5.0%	French and German
19.0%	Nonspecific Northern Eur…
	Southern European
6.6%	Italian
5.2%	Nonspecific Southern Eur…
9.1%	Eastern European
2.3%	Ashkenazi
5.6%	Nonspecific European
37.1%	Sub-Saharan African
1.2%	East Asian & Native American
1.0%	Native American
0.2%	East Asian
< 0.1%	Unassigned
100.0%	**Sheridan Smith**

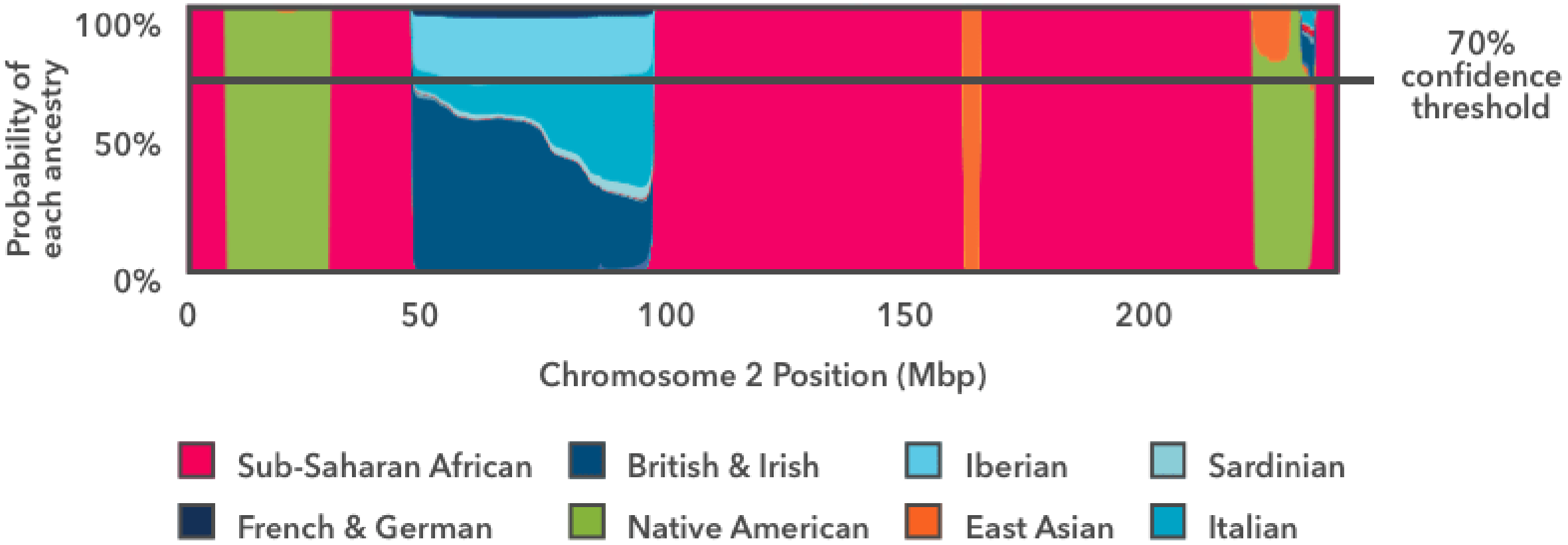
PC matching at 300bp genomic windows- 23andMe



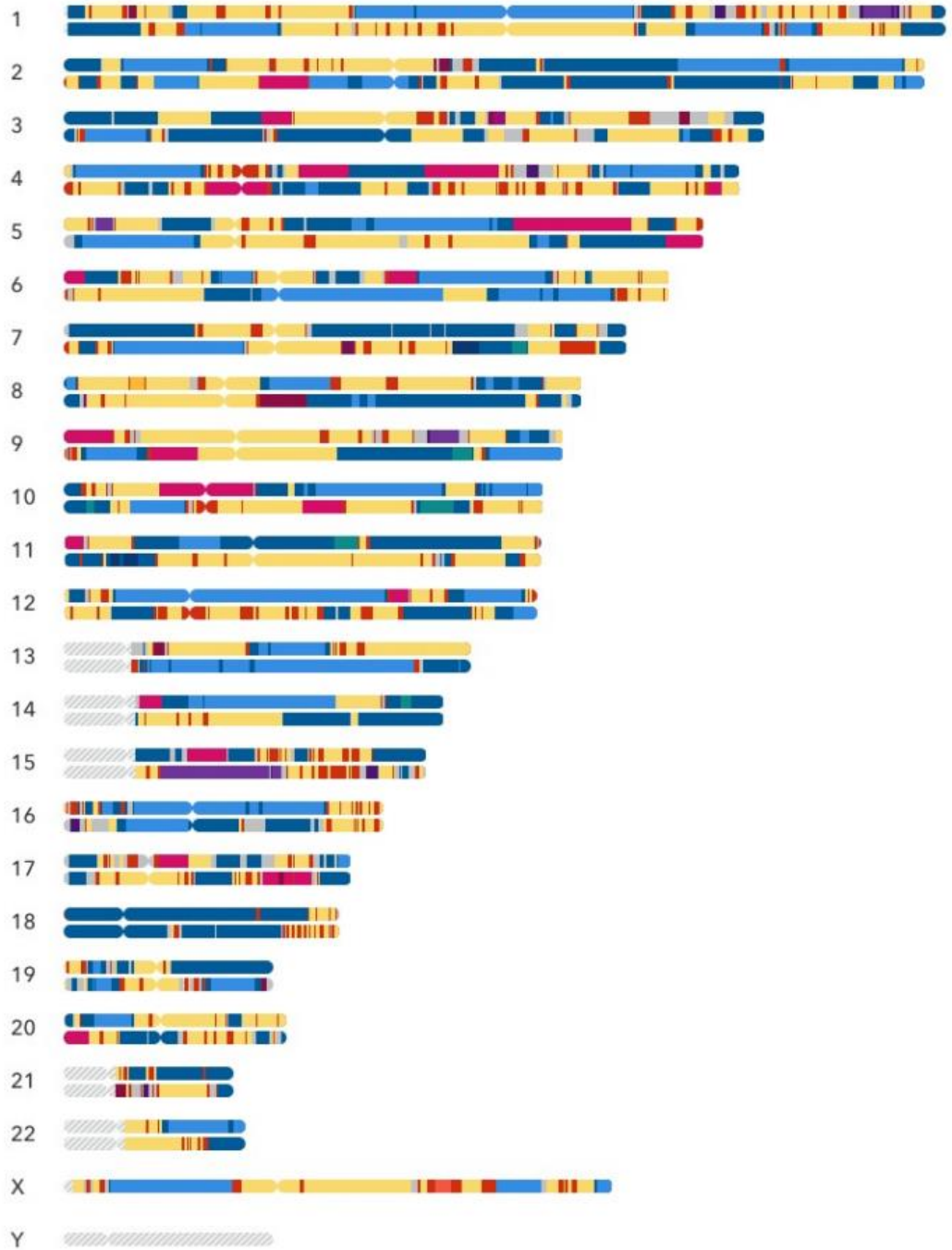
DNA segment and ancestry probability



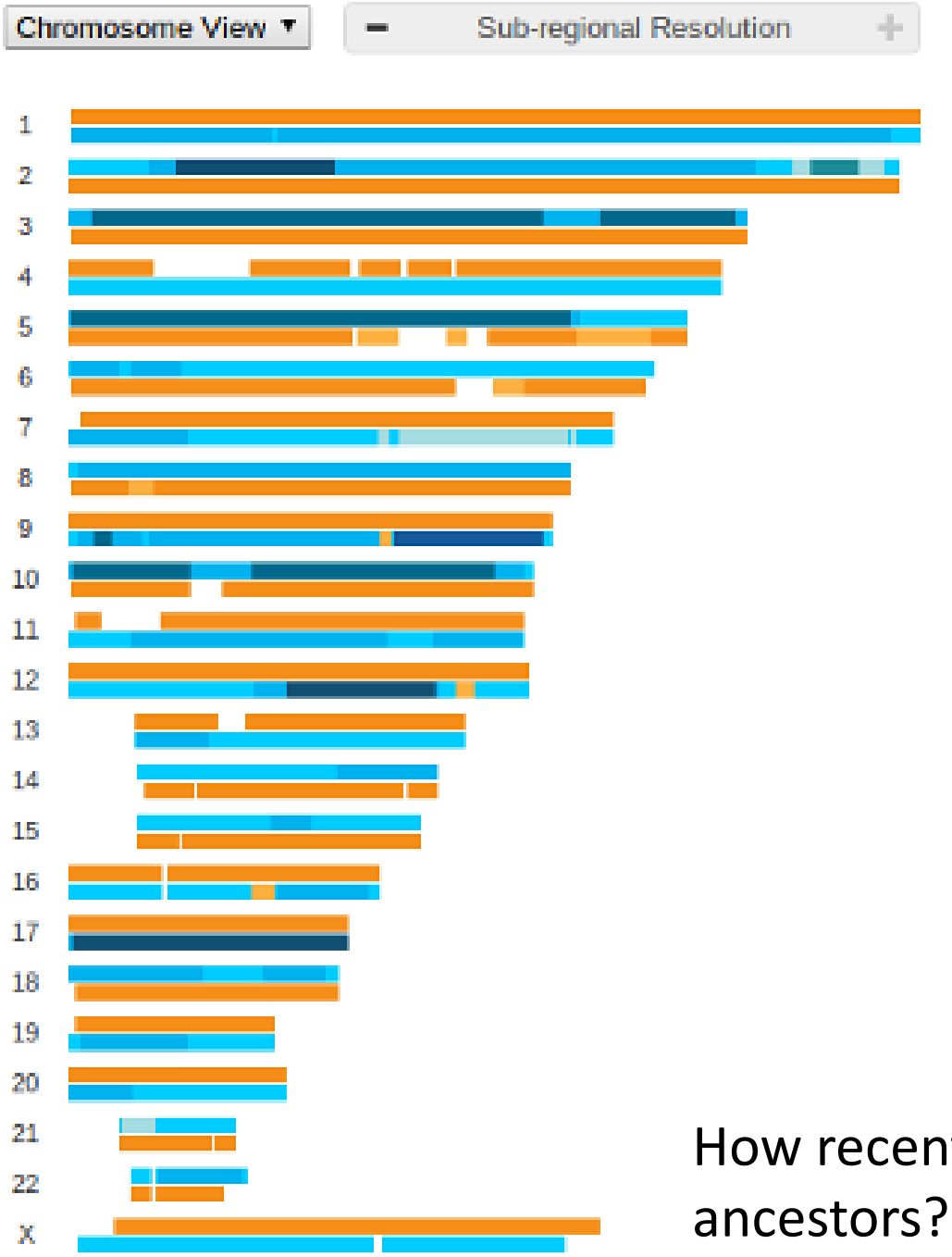
DNA segment and ancestry probability



Change confidence level



Jamie King	100%
European	**47.4%**
Iberian	19.7%
Ashkenazi Jewish	0.5%
Sardinian	0.2%
Broadly Southern European	21.1%
Broadly European	5.5%
Broadly Northwestern European	0.3%
East Asian & Native American	**41.8%**
Native American	34.4%
Manchurian & Mongolian	< 0.1%
Southeast Asian	< 0.1%
Broadly East Asian & Native American	6.8%
Broadly East Asian	0.5%
Sub-Saharan African	**5.2%**
West African	4.5%
East African	< 0.1%
African Hunter-Gatherer	< 0.1%
Broadly Sub-Saharan African	0.6%
Western Asian & North African	**1.3%**
North African & Arabian	1.0%
Broadly Western Asian & North African	0.3%
Unassigned	**4.4%**
No Data Available	**--**



Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.

49.2%	European
	Northern European
7.2%	French and German
2.8%	Scandinavian
17.5%	Nonspecific Northern Eur…
	Southern European
0.2%	Italian
1.2%	Nonspecific Southern Eur…
0.7%	Eastern European
19.5%	Nonspecific European
45.9%	South Asian
1.0%	East Asian & Native American
1.0%	East Asian
< 0.1%	Nonspecific East Asian & Nativ…
3.9%	Unassigned

How recent were the European and South Asian ancestors?

For matching ancestry - 23andMe

Reference data sets!!

Reflecting populations that existed before transcontinental travel and migration were common (at least 500 years ago). People who report four grandparents all born in the same country are included in the reference data.

Sep 22, 2018


DNA Story
Ethnicity Estimate
75% Great Britain
12% Iberian Peninsula
6 Other regions


TELUS
12:50 PM
90%
ancestry.ca
DNA Story
KH
Ethnicity Estimate
53% England, Wales, and Northwestern Europe
31% Ireland and Scotland
3 Other regions

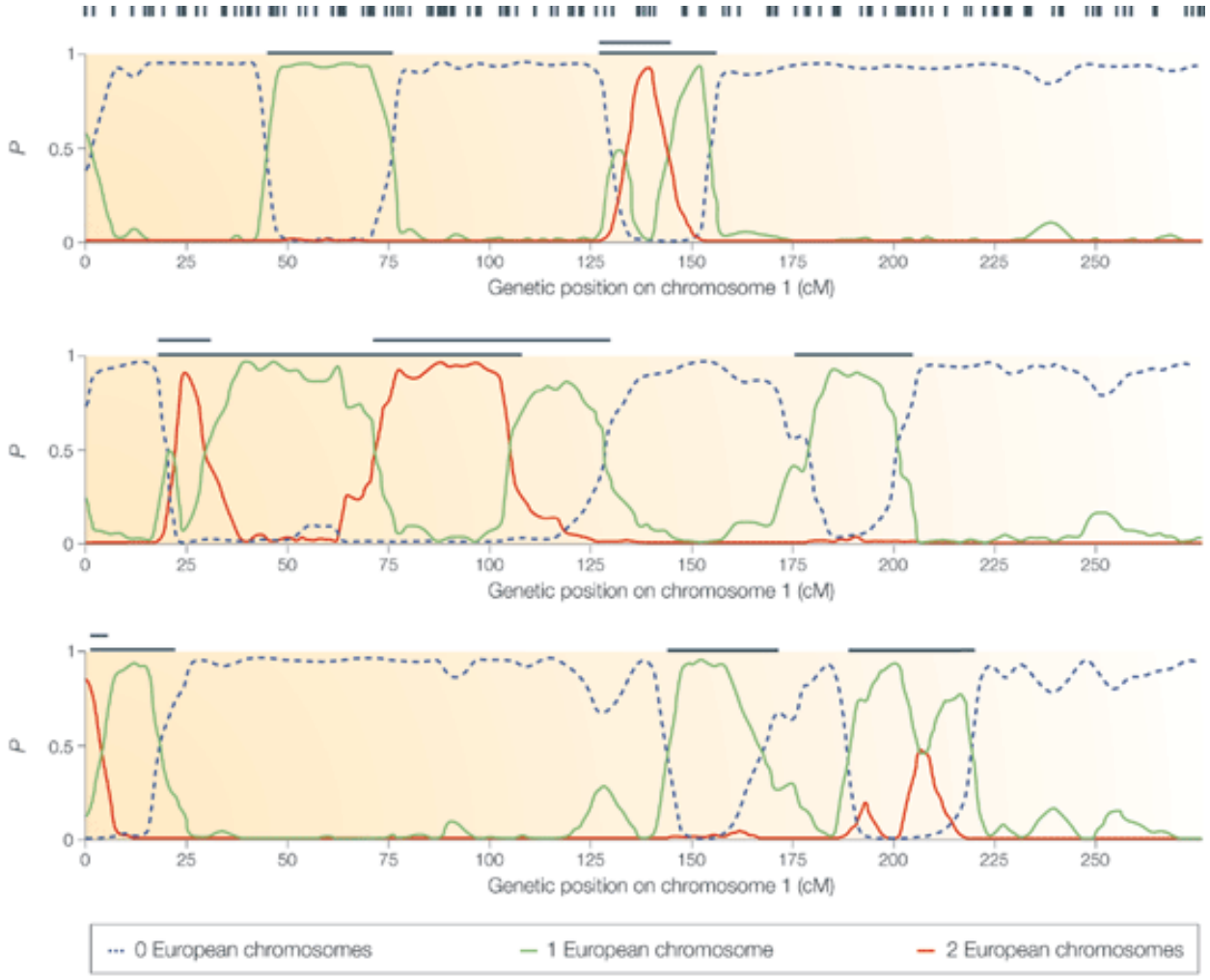
https://www.23andme.com/ancestry-composition-guide/

23andMe population precision and recall

POPULATION	PRECISION (%)	RECALL (%)
Sub-Saharan African	100	98
West African	98	93
Senegambian & Guinean	97	66
Coastal West African	93	65
Nigerian	90	66
Northern East African	100	93
Sudanese	98	79
Ethiopian & Eritrean	94	93
Somali	93	92
Congolese & Southern East African	98	92

Using LD to identify important regions

Genetic ancestry of African American

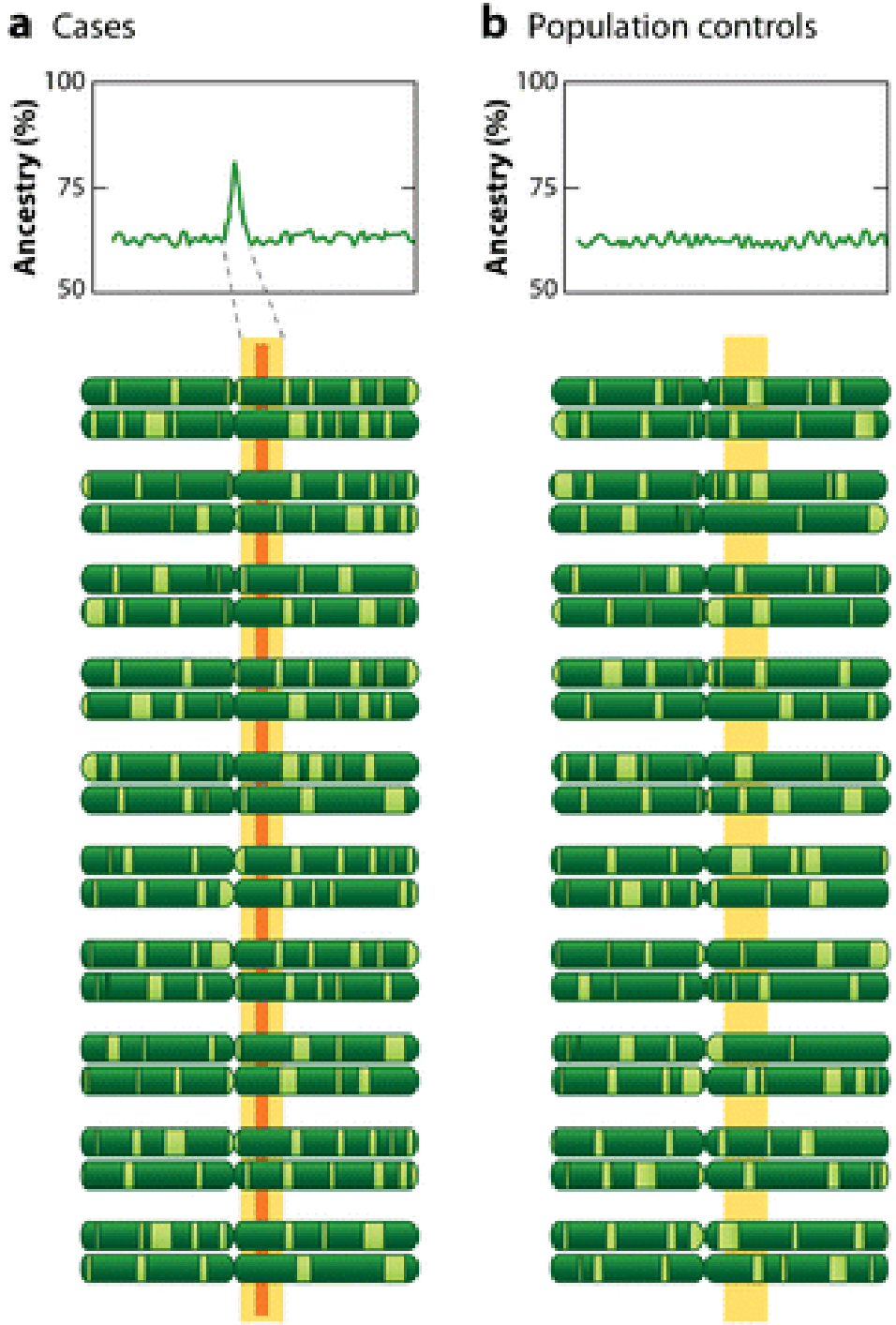


Smith, Nature Rev Genetics, 2005

Admixture mapping – a tool for gene discovery

The disease is inherited from the majority ancestry population (*dark green*), with the minority ancestry population shown in light green. The graphs show the percentage of ancestry derived from the dark green segment of chromosome.

In the region of the disease locus (*yellow bar*), there is an excess of majority ancestry blocks among cases, revealed as a spike in a graph of average ancestry for cases along the chromosome. The orange bar indicates the location of the disease gene.



Winkler CA, et al. 2010.
Annu. Rev. Genomics Hum. Genet. 11:65–89

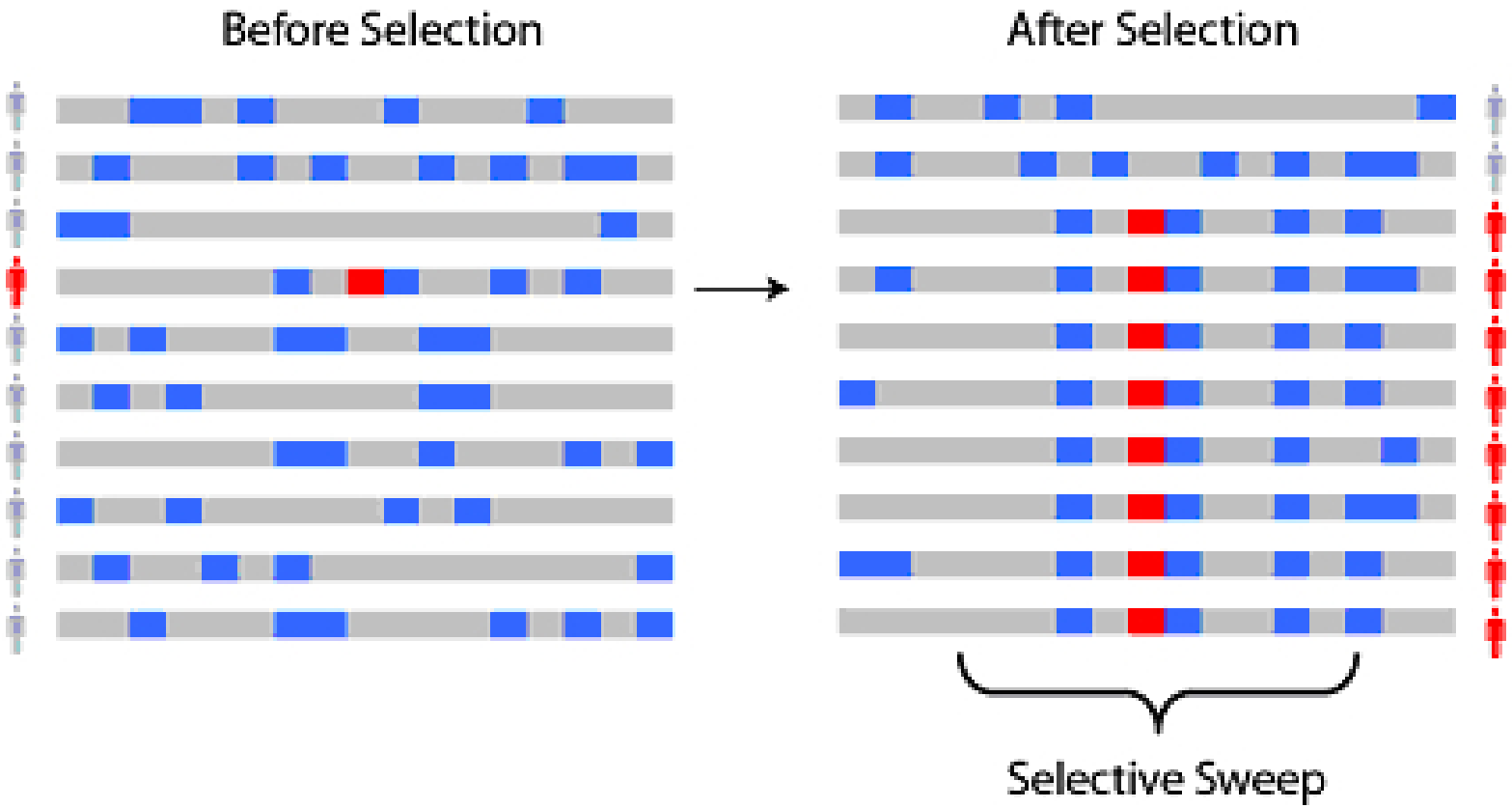
Admixture mapping – a tool for gene discovery

Table 1 | **Diseases with different risks in Africans and Europeans***

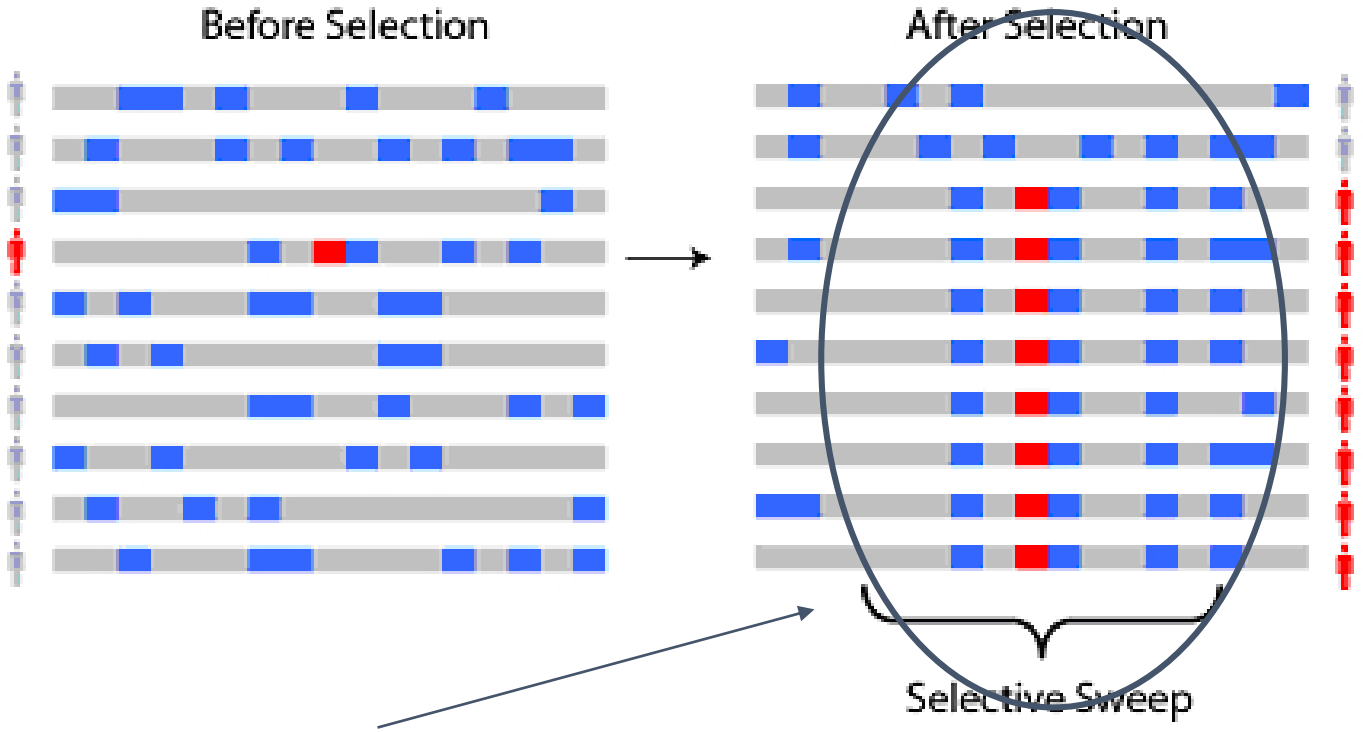
Disease or related trait	Population relative risk (African vs European)	95% Confidence interval	References
Lower relative risk in African-Americans			
Hepatitis C clearance	0.19	(0.10–0.38)	48
HIV vertical transmission	0.30	(0.10–0.90)	49
Multiple sclerosis	0.50	n.d.	50
Atrial fibrillation	0.51	(0.31–0.76)	51
Coronary artery disease	0.75	(0.60–0.95)	52
Carotid artery disease	0.62	(0.46–0.82)	52
Osteoporosis/BMD‡	Lower§	n.a.	53,54
Higher relative risk in African-Americans			
Lupus nephritis with systemic lupus erythematosus	3.13	(1.21–8.09)	55
Myeloma	3.14	(2.00–4.93)	56
Dementia	3.21	(2.18–4.73)	57
Prostate cancer	2.73	(2.13–3.52)	56
Hypertensive heart disease	2.80	(2.03–3.86)	56
Pregnancy-related death	2.65	(1.73–4.07)	58
Hypertension	2.61	(2.09–3.27)	52
Focal segmental glomerulosclerosis	2.49	(1.05–5.95)	59
Intracranial haemorrhage	2.10	(1.44–3.06)	56
Non-insulin dependent diabetes	1.99	(1.60–2.48)	52,60
End-stage renal disease	1.87	(1.47–2.39)	61
Stroke	1.57 1.30–5.00‖	(1.27–1.94) (1.00–1.61)	56 62
Hypertensive retinopathy	1.48	(1.08–2.03)	63
Lung cancer	1.48	(1.30–1.67)	56

Winkler CA, et al. 2010.
Annu. Rev. Genomics Hum. Genet. 11:65–89

Advantageous mutations -> selective sweep

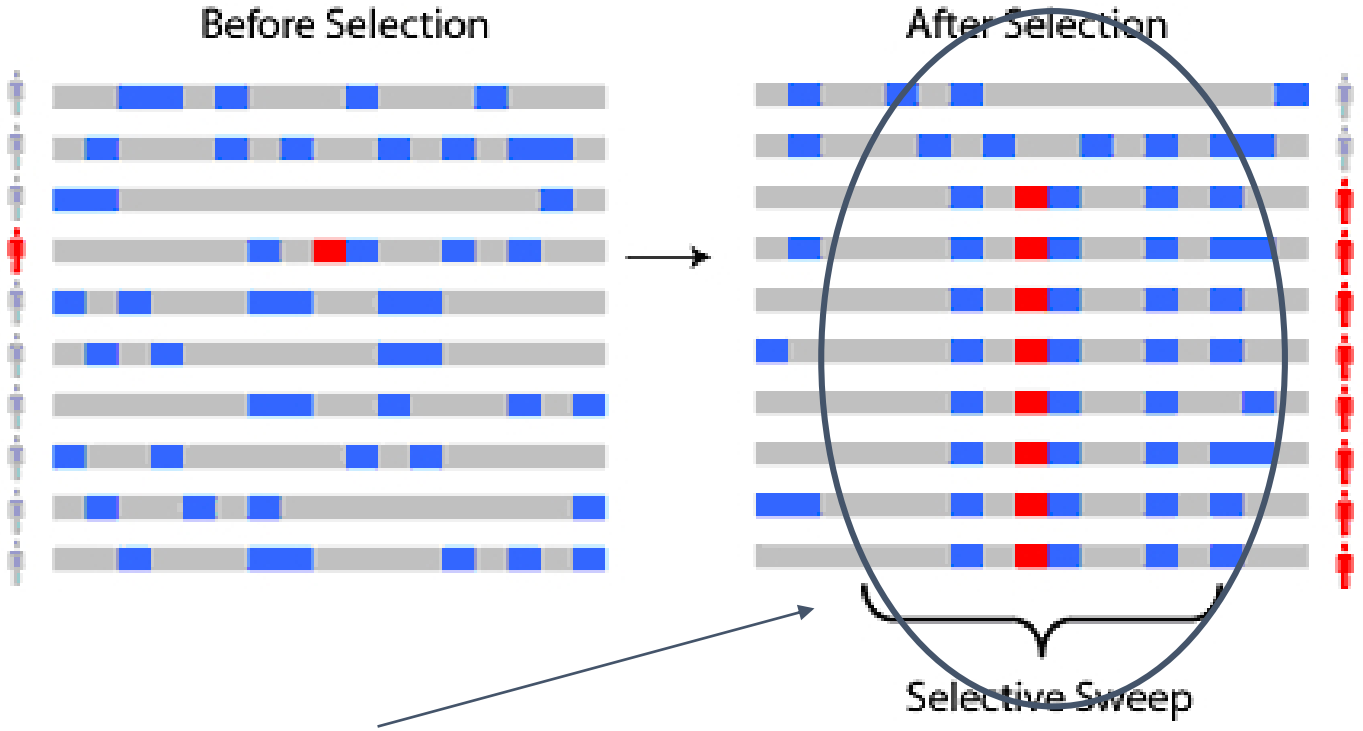


Advantageous mutations -> selective sweep



Why will these regions will also be the same?

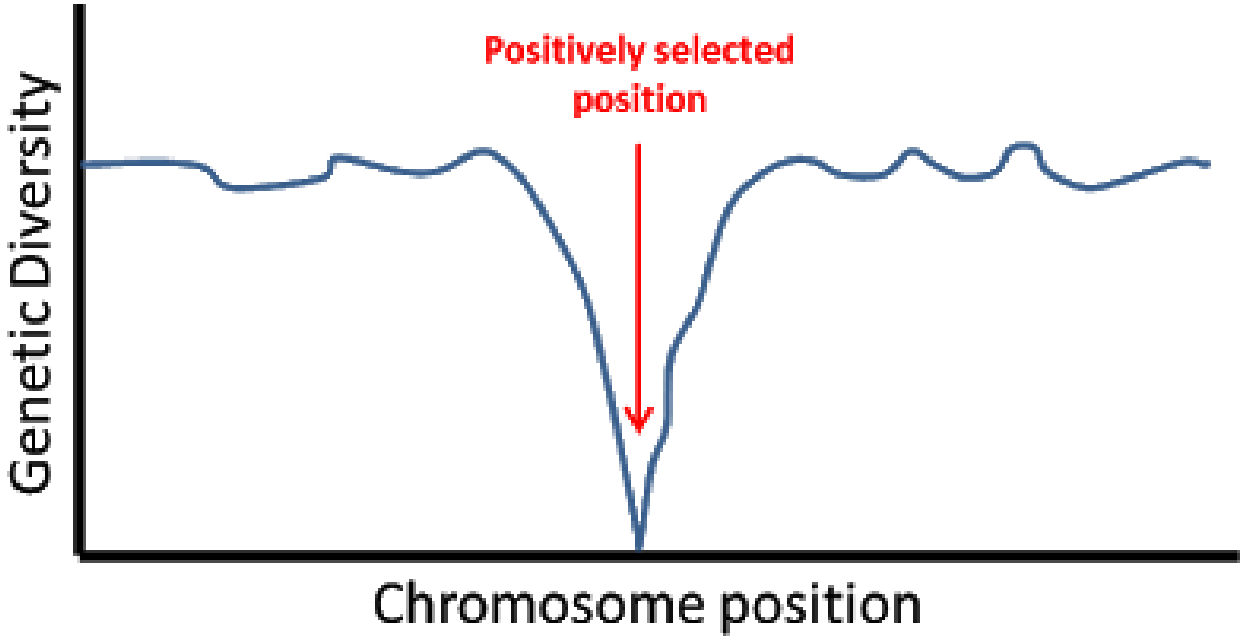
Advantageous mutations -> selective sweep

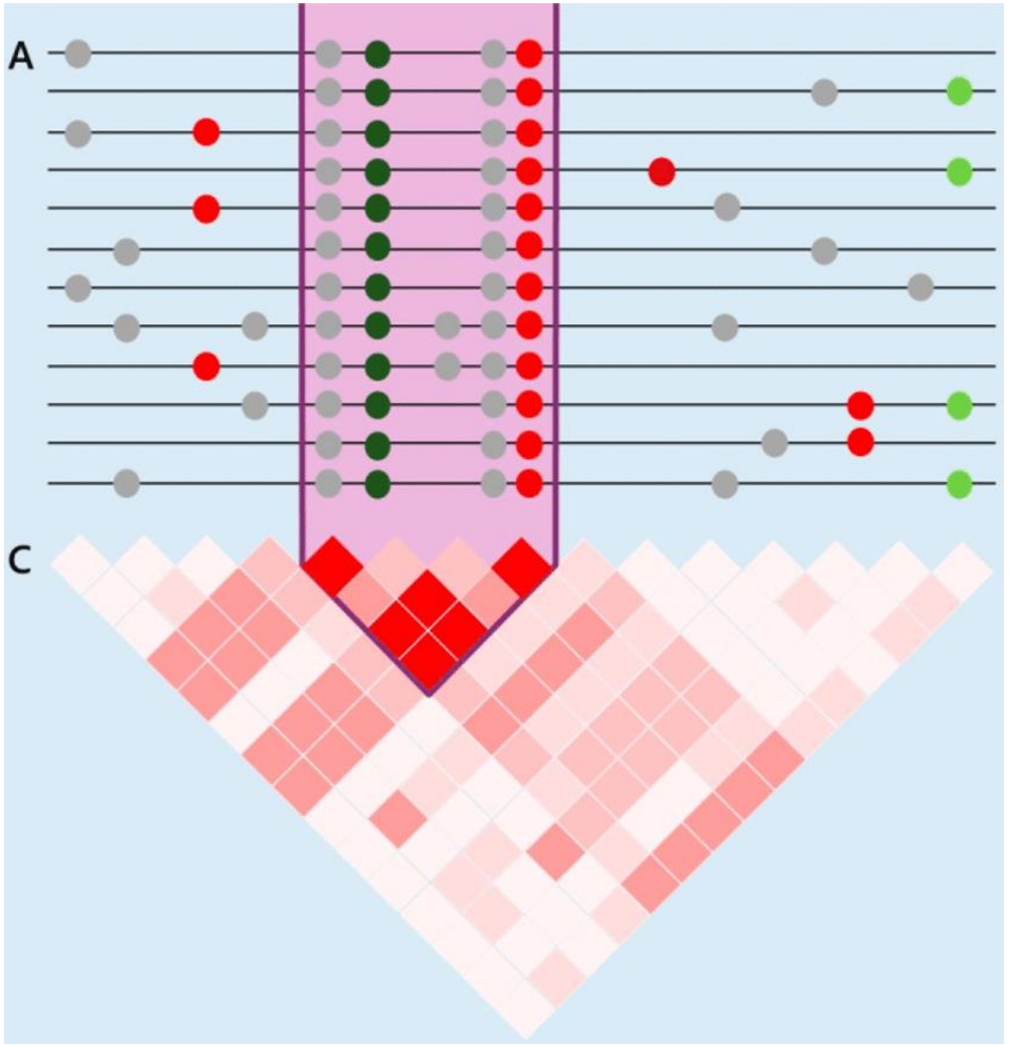


Why will these regions will also be the same?

Linkage disequilibrium!

Advantageous mutations -> selective sweep

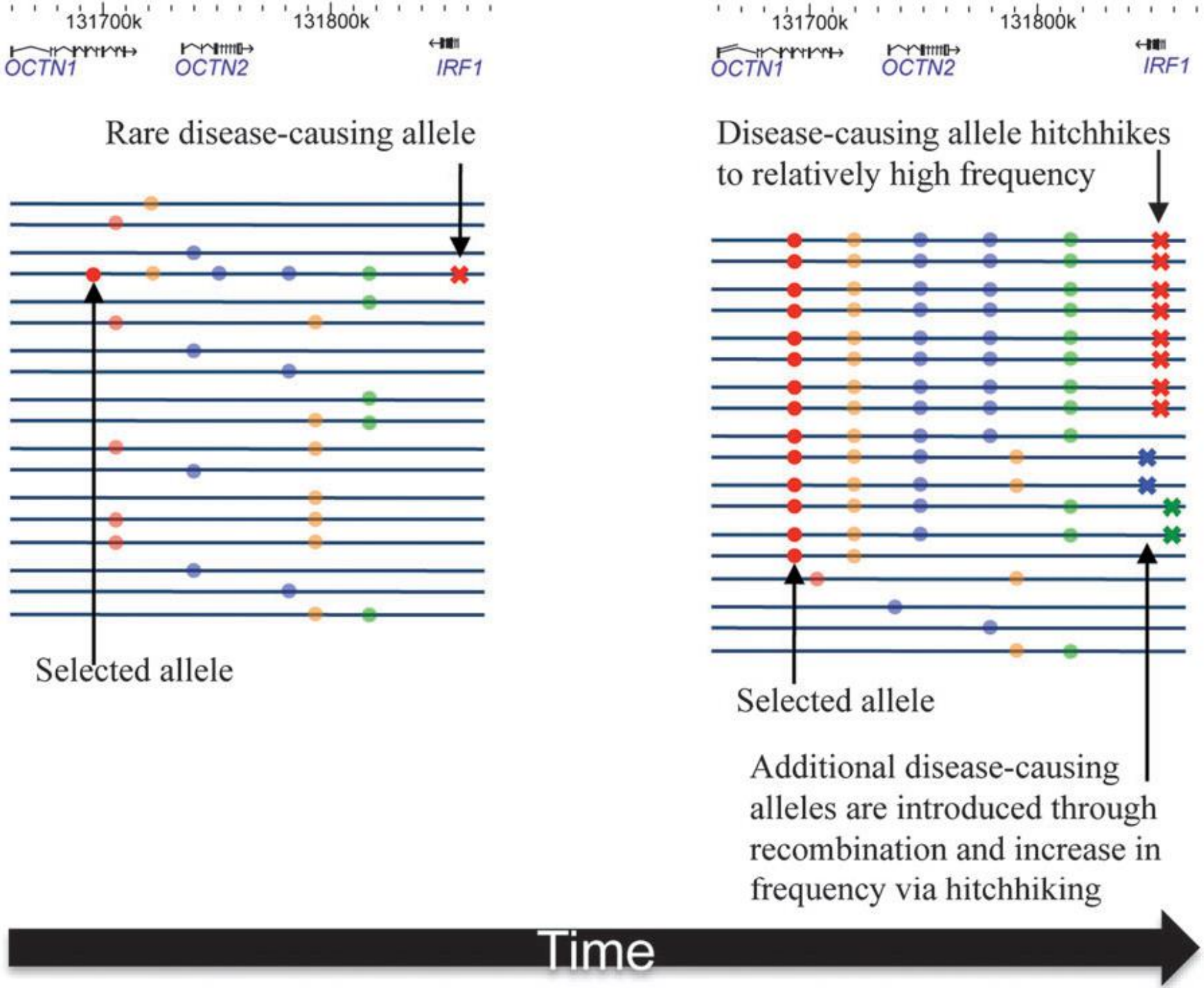




Summary

- Hardy Weinberg disequilibrium tests can indicate underlying population structure or selective pressure.
- Population structure can confound genetic association studies, but using principal component analysis can reveal and adjust.
- Leveraging population structure in admixture mapping can uncover loci associated with traits.

Genetic hitchhiking


131700k
131800k
OCTN1
OCTN2
IRF1
Rare disease-causing allele
Selected allele
131700k
131800k
OCTN1
OCTN2
IRF1
Disease-causing allele hitchhikes to relatively high frequency
Selected allele
Additional disease-causing alleles are introduced through recombination and increase in frequency via hitchhiking
Time