Population Genetics

Section 4

(1.5 hours)

Learning Objectives

- Understand the importance of Hardy Weinberg equilibrium and how to calculate deviance.
- Describe population substructure and how it can confound results. Also understand methods for adjusting for it in analysis.

Population genetics principles

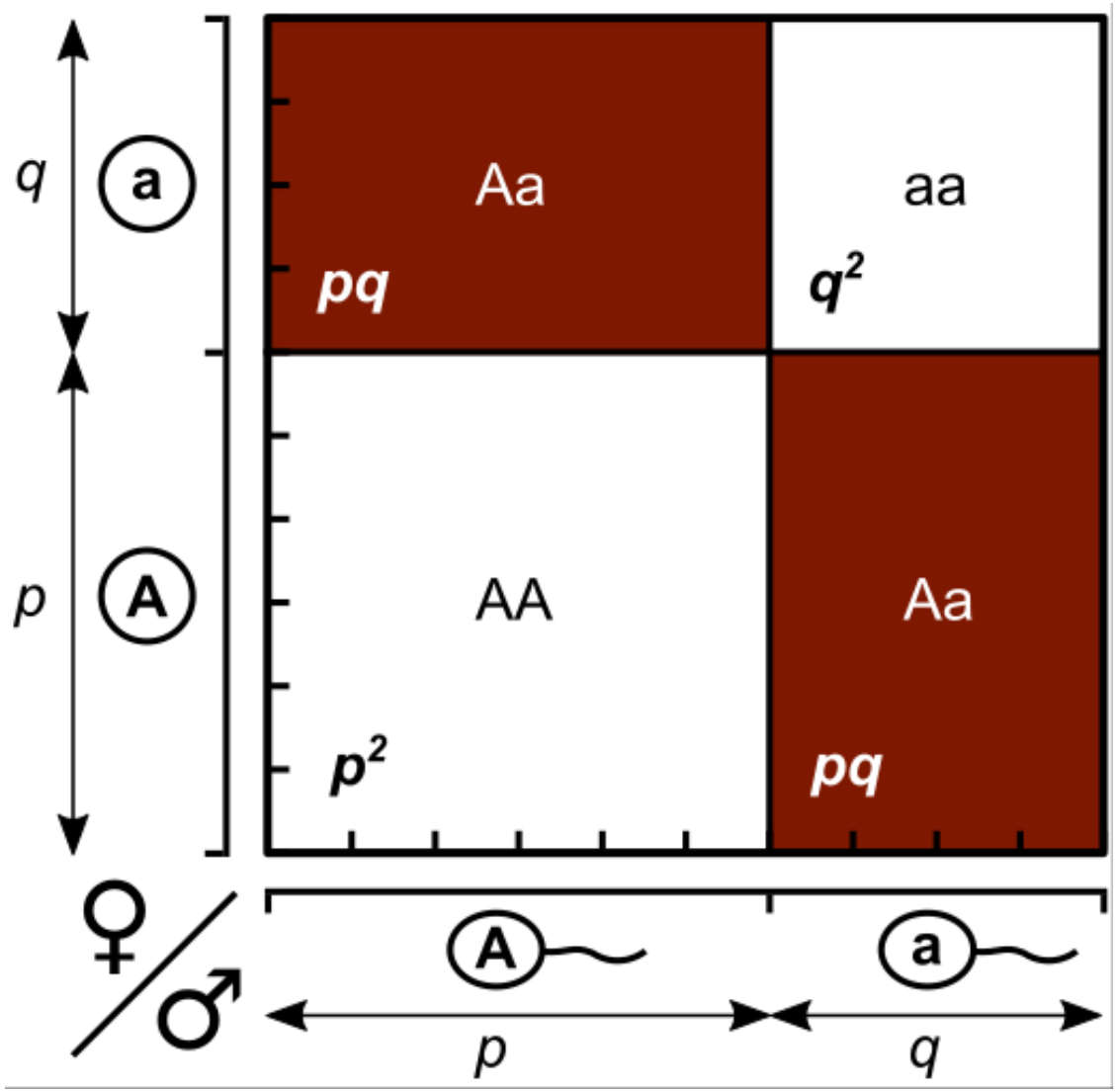
- Overall patterns of genetic variants within and between populations.
- Discipline originally developed to study evolution.
- Reflects interplay between genetic variation, phenotypes, and environmental pressures.
- Subject to mutation, mating and migration.

Single mating pair and offspring

	A	a
A	AA	Aa
a	Aa	aa

¼ (AA) + $^{2}/_{4}$ (Aa) + ¼ (aa)

Expected genotype combinations



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

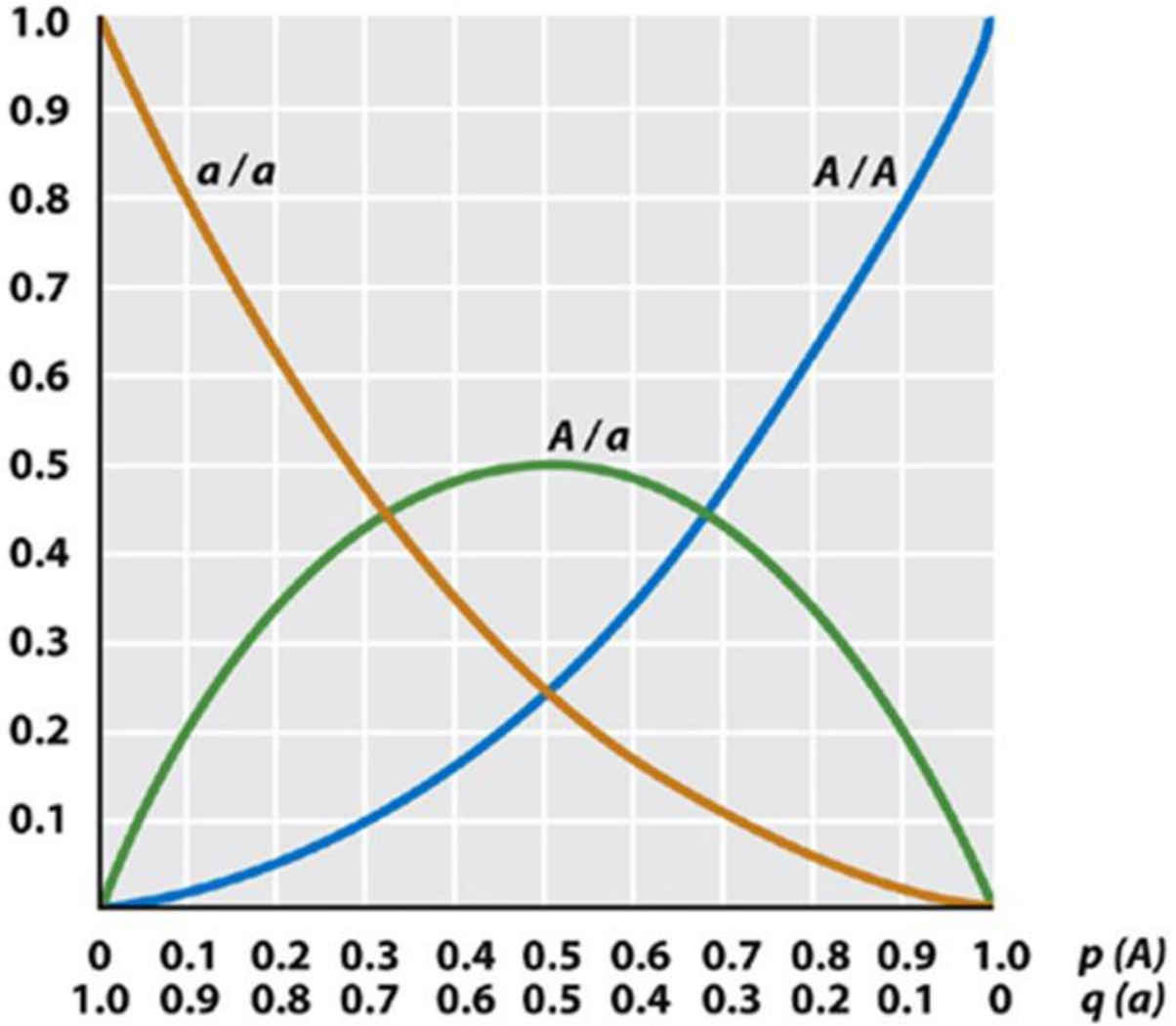


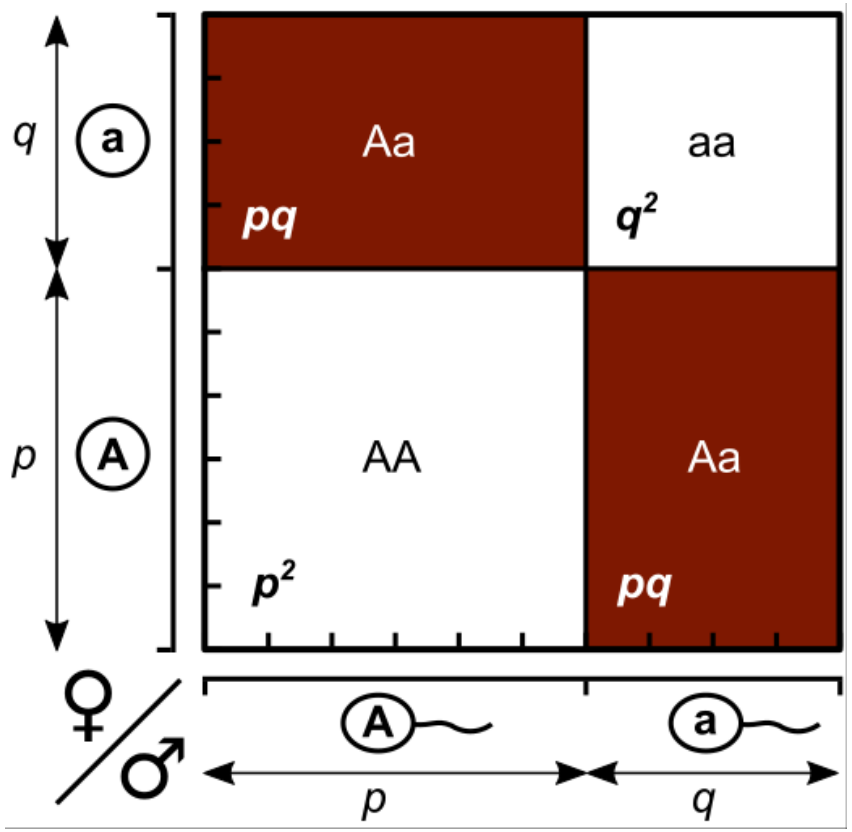
Figure 17-5
Introduction to Genetic Analysis, Ninth Edition

The Hardy-Weinberg law under the assumption of non-evolving allele frequencies

- The Hardy-Weinberg Law provides two equations allowing us to relate the expected allele and genotype frequencies to each other
- Assume a SNP with

  alleles A (frequency p)

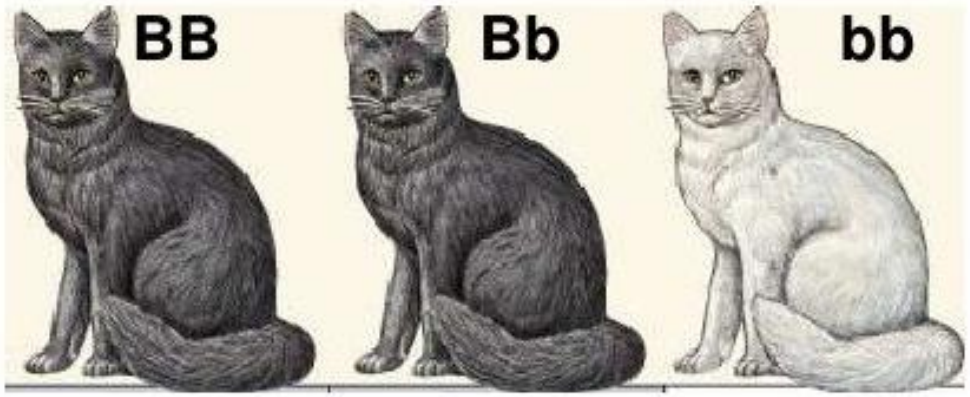
  alleles a (frequency q)
- p+q=1 (allele frequencies)
- $p^2+2qp+q^2=1$ (genotype frequencies)



HWE example

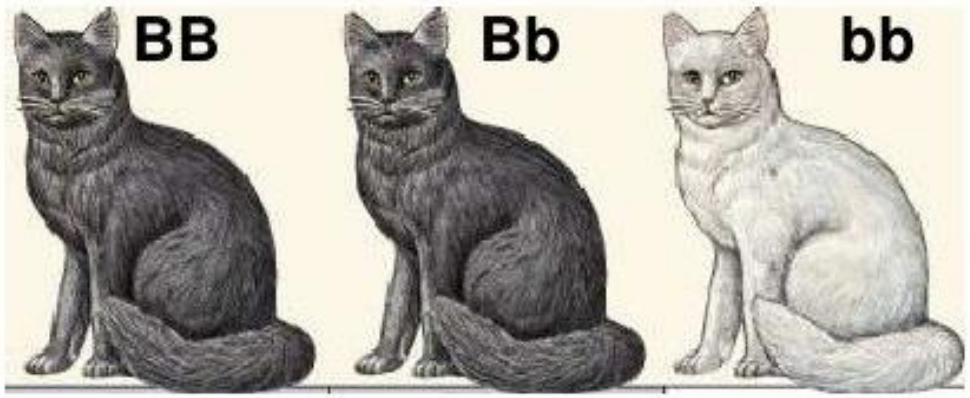
- Assume 100 cats (200 alleles) with alleles B and b. B allele is dominant and results in black coloring. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?

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



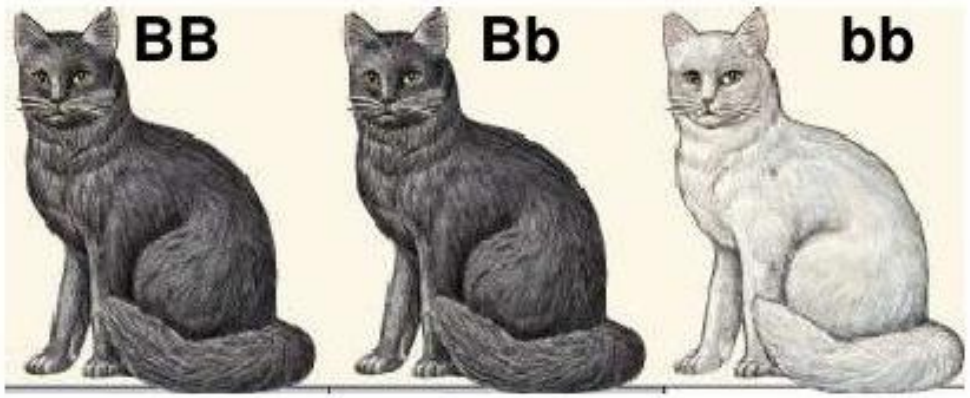
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$



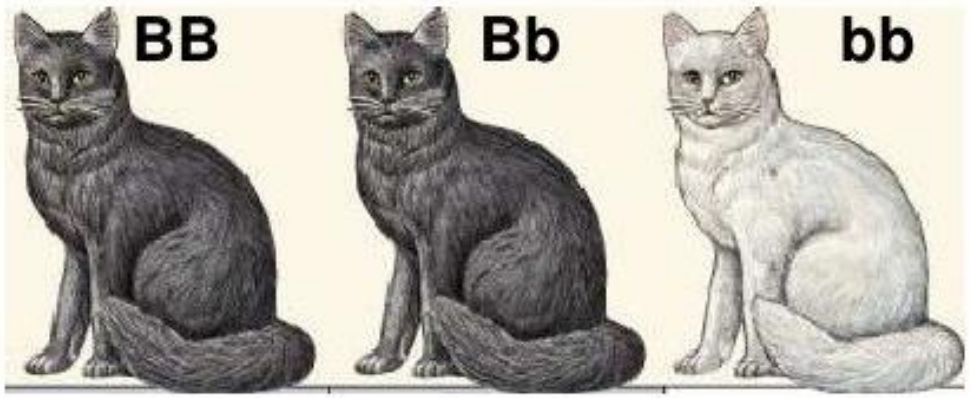
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- $q=0.4$, $p=0.6$



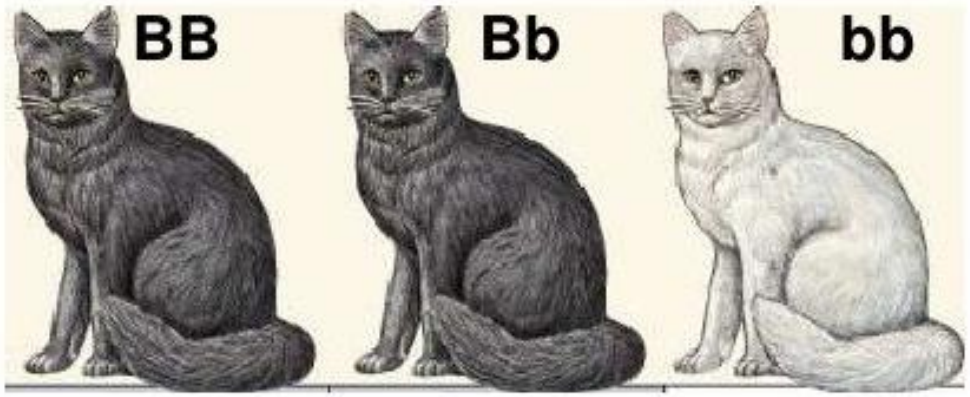
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- $q=0.4$, $p=0.6$
- $p^2=0.36$



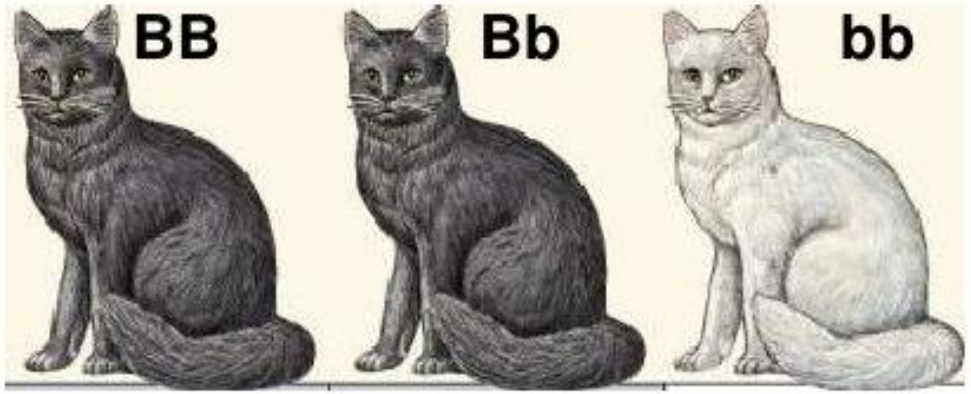
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- $q=0.4$, $p=0.6$
- $p^2=0.36$
- $2pq=2x0.6x0.4=0.48$



HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- $q=0.4$, $p=0.6$
- $p^2=0.36$
- $2pq=2x0.6x0.4=0.48$
- 16 white cats and 84 black cats (28 Bb, 36 BB)



Calculate expected genotype frequencies

The ability to taste bitterness of phenylthiourea (PTC - a chemical in mustards and broccoli) is due to genotype at a single SNP. The taste is due to a dominant inheritance of allele T.

You sampled 215 individuals

150 could detect the bitter taste of PTC

65 could not.

Calculate the frequencies of T and C (recessive allele, no bitterness taste), and expected genotypes.

Calculate expected genotype frequencies

- Lay out what we know:
- $q^2 = 65/215$
- $p^2 + 2pq = 150/215$
- $p + q = 1$
- $p^2 + 2pq + q^2 = 1$

Calculate expected genotype frequencies

- Lay out what we know:
- $q^2 = 65/215$
- $p^2 + 2pq = 150/215$
- $p + q = 1$

- q = sqrt(65/215) = 0.55

Calculate expected genotype frequencies

- Lay out what we know:
- $q^2 = 65/215$
- $p^2 + 2pq = 150/215$
- $p + q = 1$

- q = sqrt(65/215) = 0.55
- $p + 0.55 = 1$
- $p = 1- 0.55 = 0.45$

How many are expected to be TT vs TC?

- Lay out what we know:
- $q^2 = 65/215 = 0.30$
- $p^2 + 2pq = 150/215$
- $p + q = 1$

- $q = \text{sqrt}(65/215) = 0.55$
- $p + 0.55 = 1$
- $P = 1- 0.55 = 0.45$

- $p^2 = (0.45)^2 = 0.20$
- $2pq = 2*(0.45*0.55) = 0.50$

- $TT = 0.20*215 = 43$
- $TC = 0.50 *215 = 107$

How different are the frequencies?

Compare expected and observed genotype frequencies:

$$x^2 = \sum \frac{(0bserved - Expected)^2}{Expected}$$

Compare to x^2 for degree of freedom, p < 0.05 = 3.841

If < 3.841 then population is not out of HWE

If > 3.841 then population IS out of HWE

How different are the frequencies?

Compare expected and observed genotype frequencies:

$$x^2 = \sum \frac{(0bserved - Expected)^2}{Expected}$$

Compare to x^2 for degree of freedom, p < 0.05 = 3.841

Genotype	Count	Allele	Frequency
AA	30	A	0.575
Aa	55	a	0.425
aa	15		
Total	100		

Calculate the χ2 value.

How different are the frequencies?

Genotype	Count
AA	30
Aa	55
aa	15
Total	100

Allele	Frequency
A	0.575
a	0.425

Calculate the χ2 value.

Genotype	Observed	Expected
AA	30	33 (0.575*0.575*100)
Aa	55	49
aa	25	18
Total	100	100

How different are the frequencies?

Genotype	Observed	Expected
AA	30	33 (0.575*0.575*100)
Aa	55	49
aa	25	18
Total	100	100

$$x^2 = \sum \frac{(0bserved - Expected)^2}{Expected}$$

$$x^2 = \sum \frac{(AAO-AAE)^2}{AAE} + \frac{(AaO-AaE)^2}{AaE} + \frac{(aaO-aaE)^2}{aaE}$$

$$x^2 = \sum \frac{(30-33)^2}{33} + \frac{(55-49)^2}{49} + \frac{(25-18)^2}{18}$$

How different are the frequencies?

Genotype	Observed	Expected	$(O-E)^2/E$
AA	30	33	0.27
Aa	55	49	0.73
aa	25	18	0.50
Sum	100	100	**1.50**

$1.5 < 3.84$

therefore, frequencies are not different from expected

What if HWE is violated?

- We could have genotyping error.
- Population substructure.
  - Mating is not random.
  - Immigration, emigration, population mixing.
- Natural Selection.
- New mutations.
- Small population size.

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

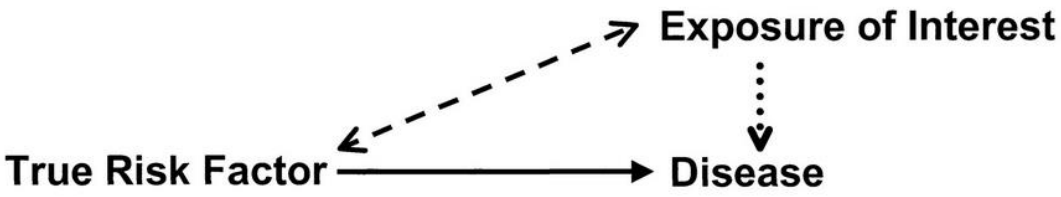
Assume we conduct a case-control GWAS…

- 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?

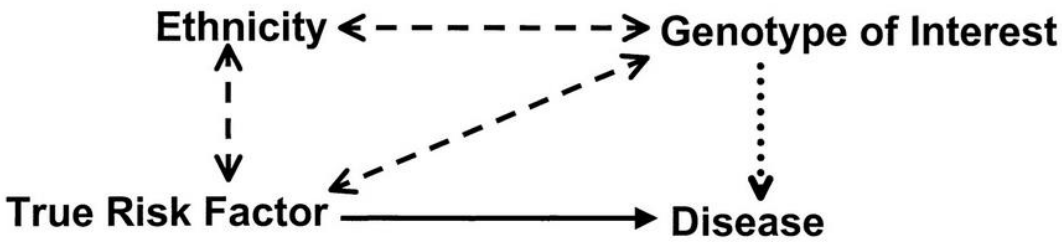
Population Substructure

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

Confounding

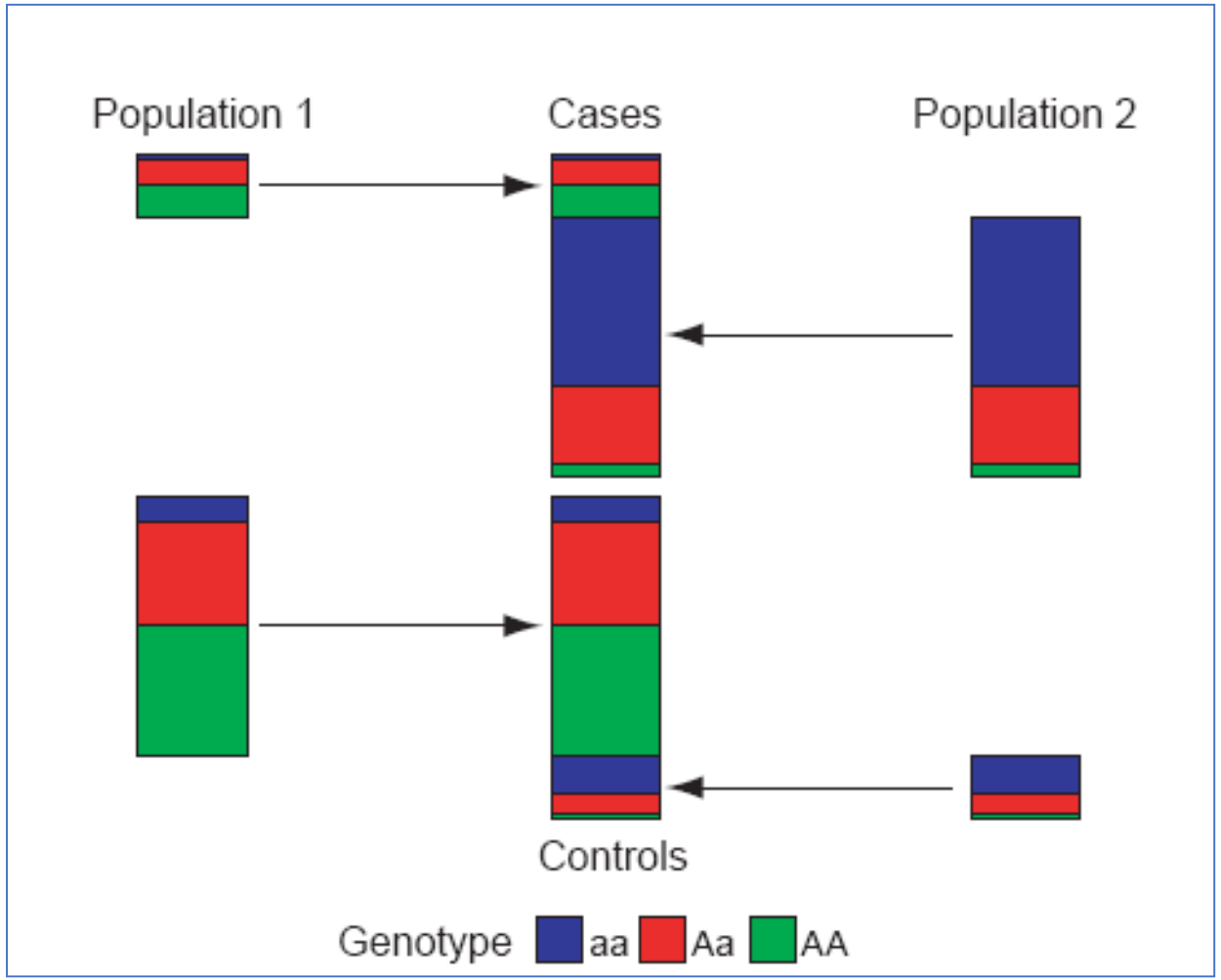


Population Stratification



Population Stratification - Confounding by ancestry

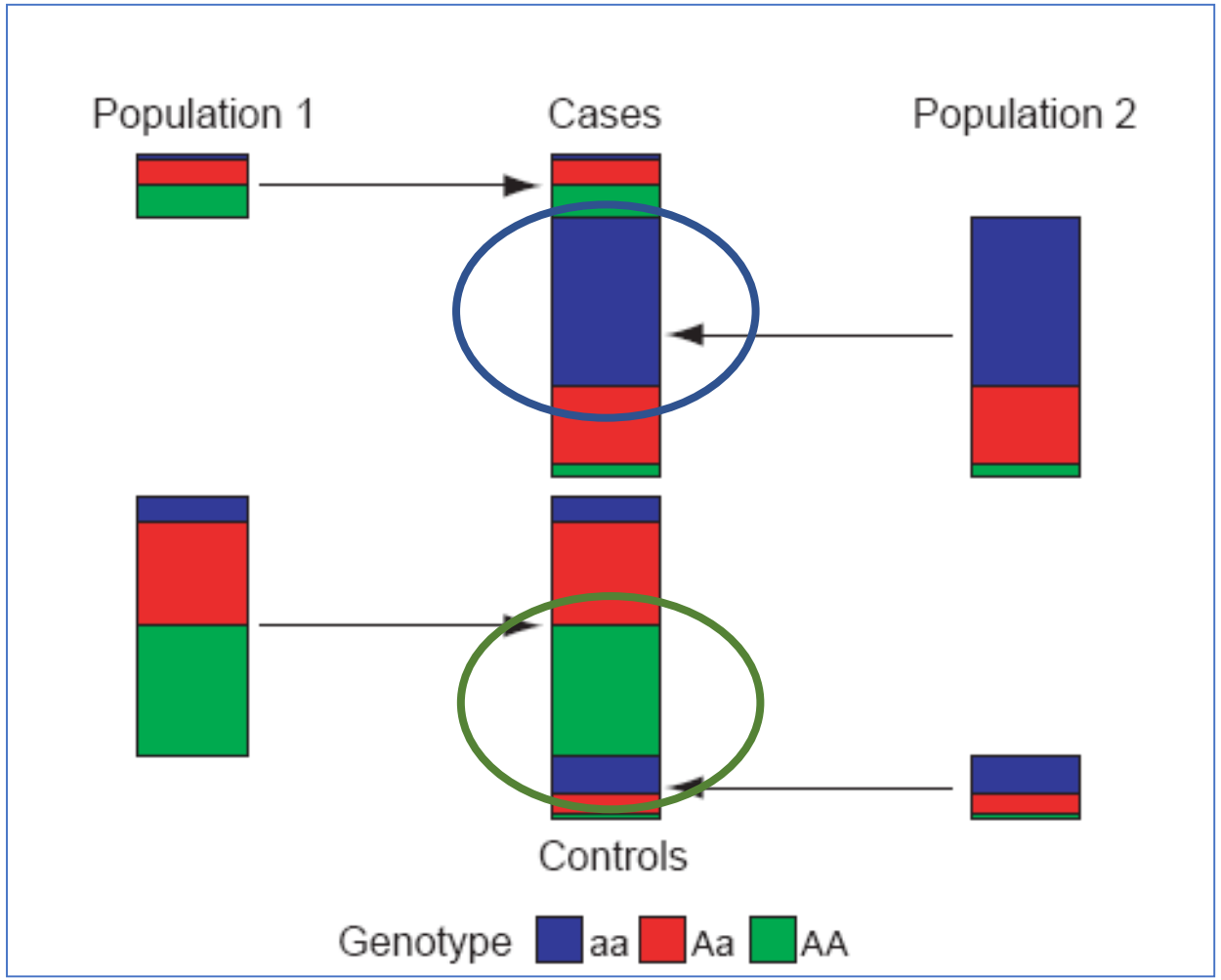
Group differences in ancestry
AND outcome



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

Assume we conduct a case-control GWAS…

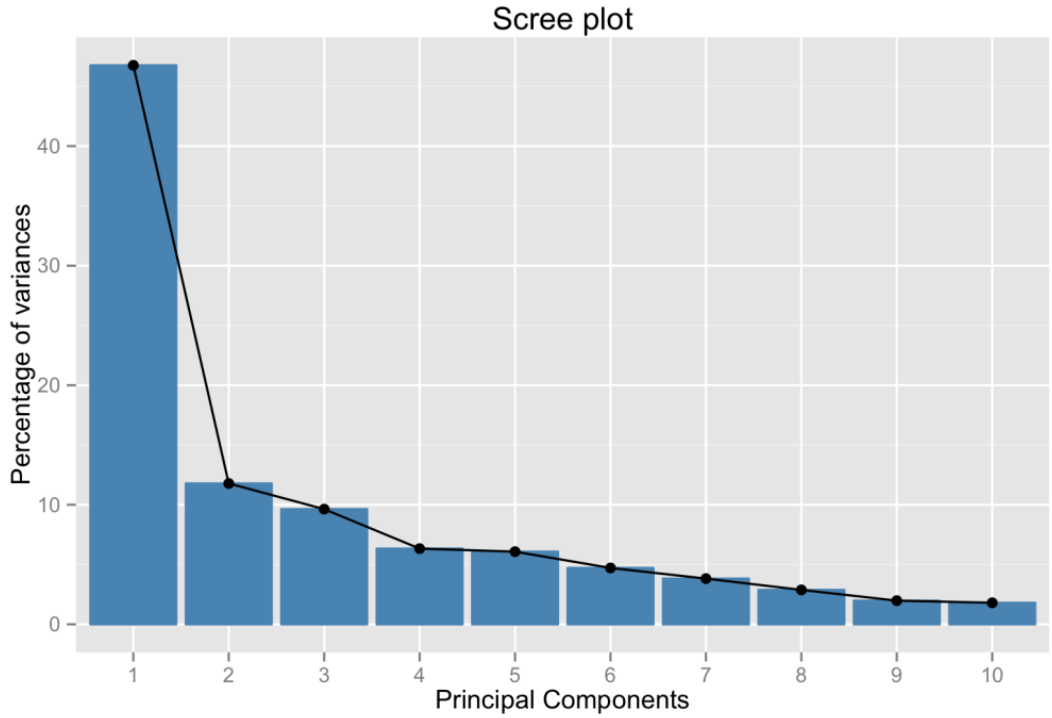
- Our cases were collected in Africa
- Our controls were collected in Asia
- If we find multiple SNPs that are significantly more/less common in cases than controls, do we believe that these results are due to association with disease or population differences?

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

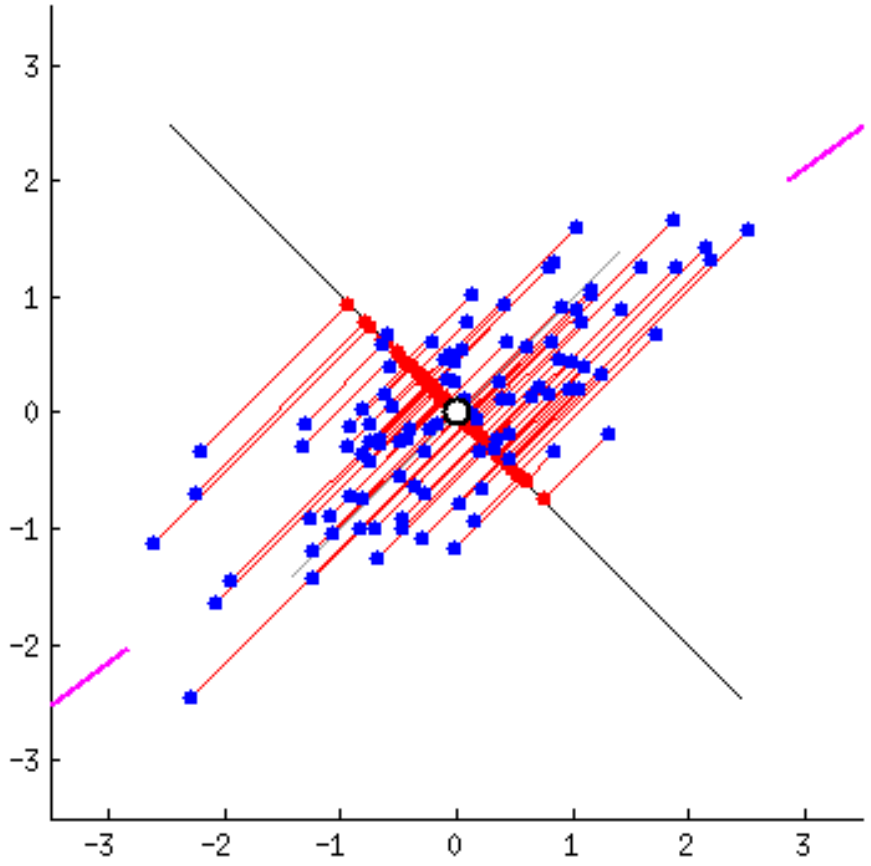
We can use genetic data to determine ancestry and to adjust for ancestry in association studies.

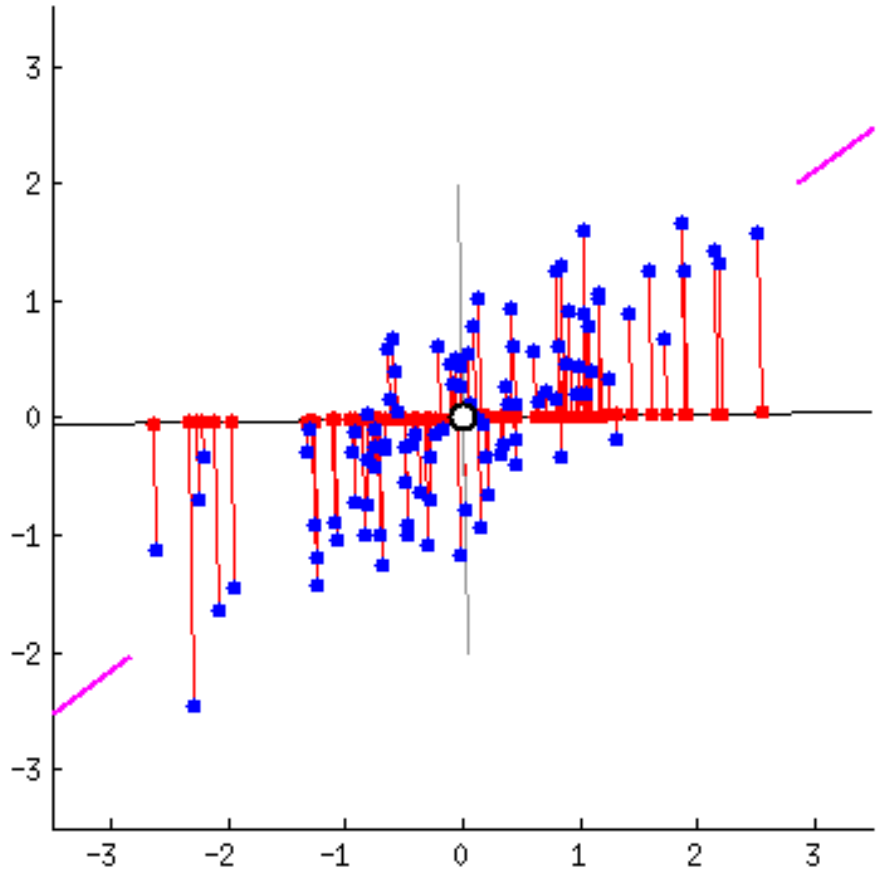
Principal Component Analysis (PCA)

- 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.



PCA minimizes error and maximizes variance





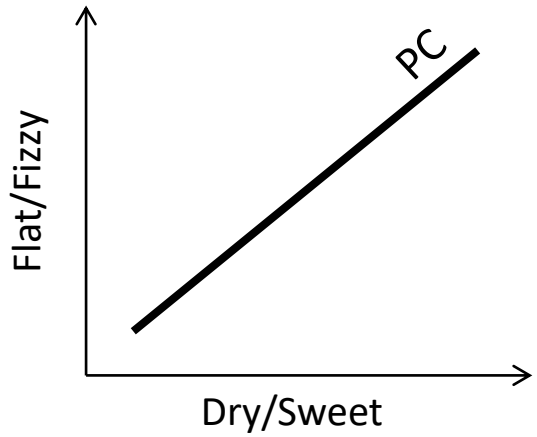
PCA features: Apple Ciders

- Sweet, fizzy
- Sweet, fizzy, pear
- Cloudy, dry, flat, alcoholic
- Cloudy, sweet, fizzy
- Sweet, fizzy, alcoholic
- Sweet, strawberry

PCA features: Apple Ciders

- Sweet, fizzy
- Sweet, fizzy, pear
- Cloudy, dry, flat, alcoholic
- Cloudy, sweet, fizzy
- Sweet, fizzy, alcoholic
- Sweet, strawberry

- PC sweet/dry/fizzy/flat, which often show up together (colinear) and can be rolled into one feature.



PCA features: Apple Ciders

- PC0.3
- Pear +PC2.8
- Cloudy, alcoholic + PC0.9
- Cloudy + PC2.1
- Alcoholic +PC2.7
- Strawberry +PC1.4

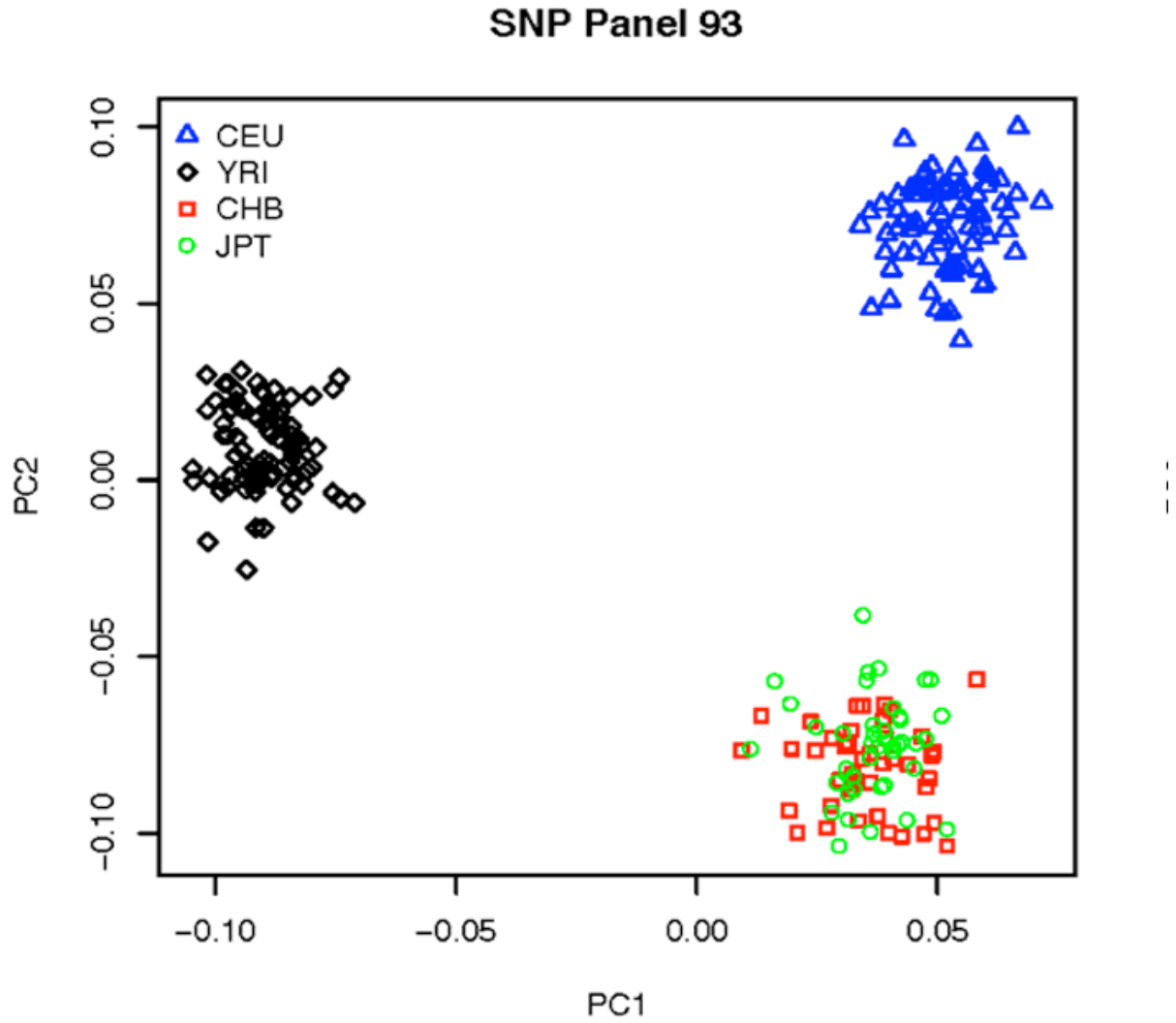
- PC is a feature that collapses the variability, so now we have more manageable view of the differences between our cider choices.

Principal Component Analysis (PCA)

- PCA is a standard tool to detect large patterns in the data.

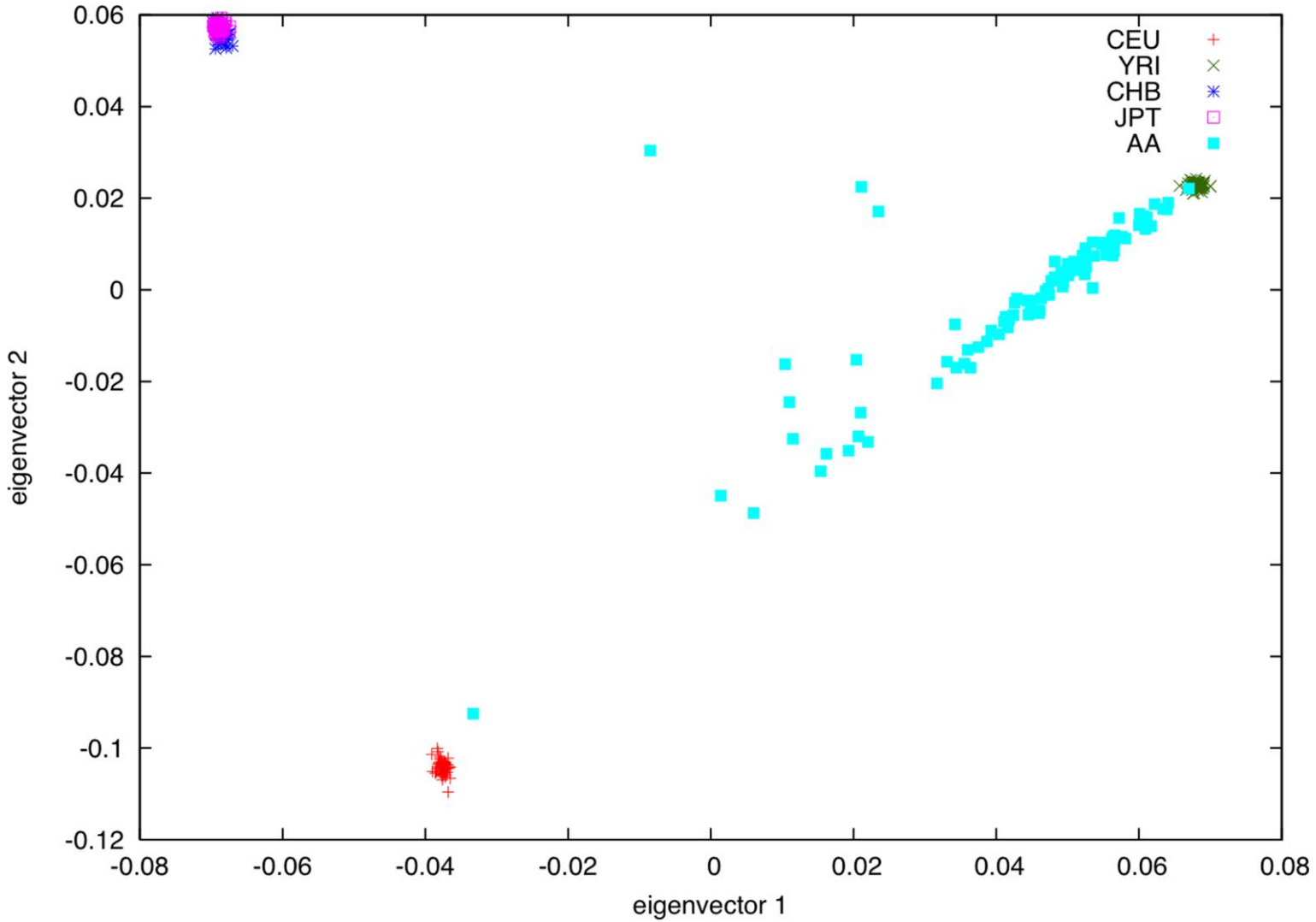
- We can make use of PCA in our studies to identify ancestry for samples in our data
  - Population genetics
  - Identify “population outliers”
  - Identify any other structure that is not obvious
  - Adjust analyses to avoid confounding

The first two PCs can help distinguish ancestral populations



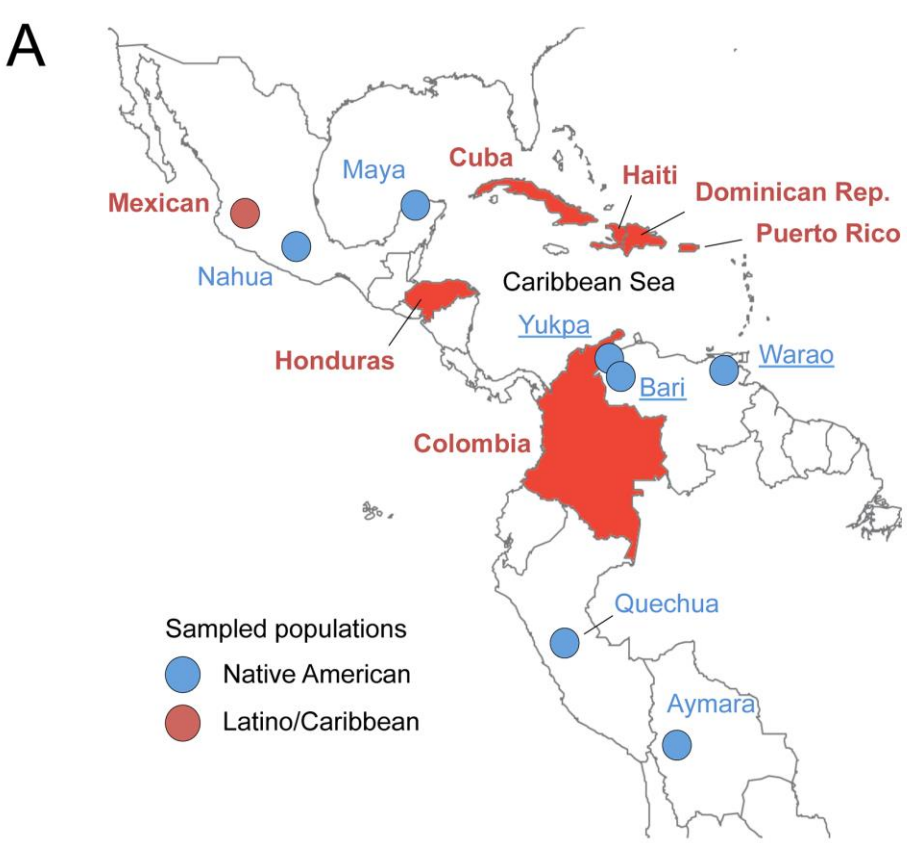
Hou, PLoS One 2011

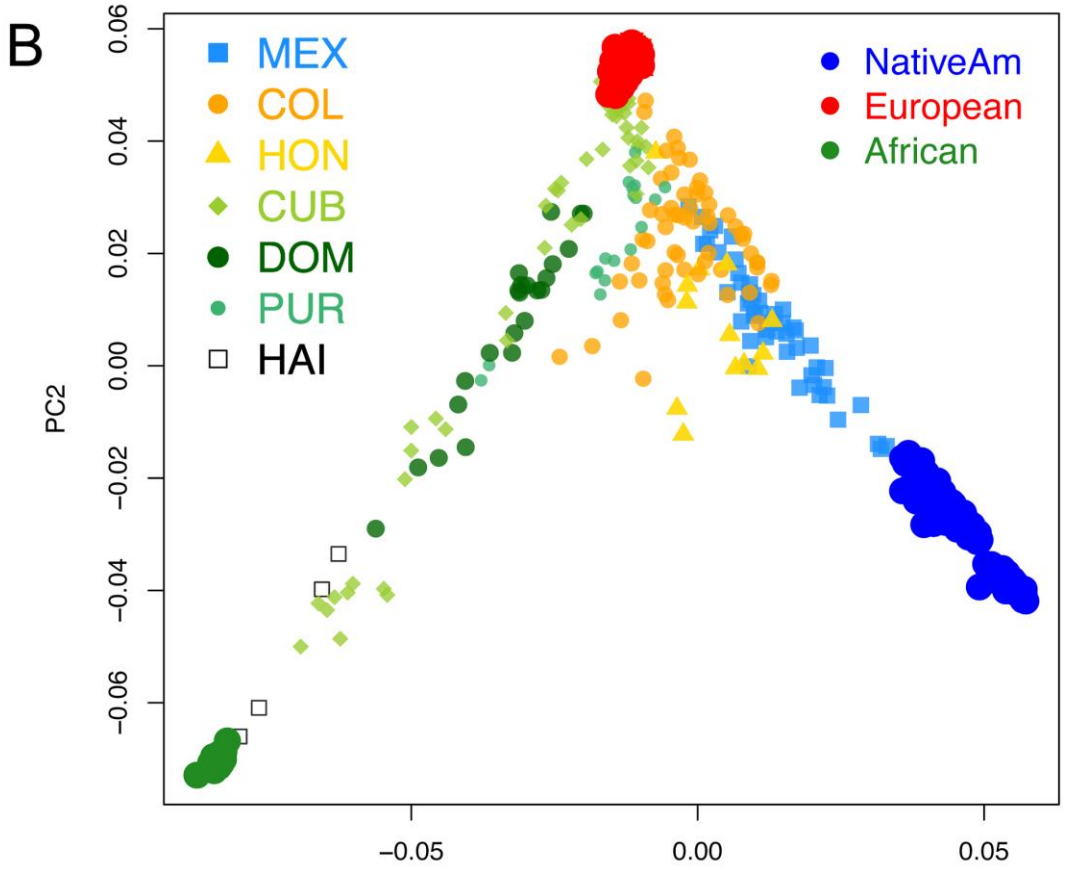
The first two PCs can help distinguish ancestral populations



Price, PLoS Genetics 2008

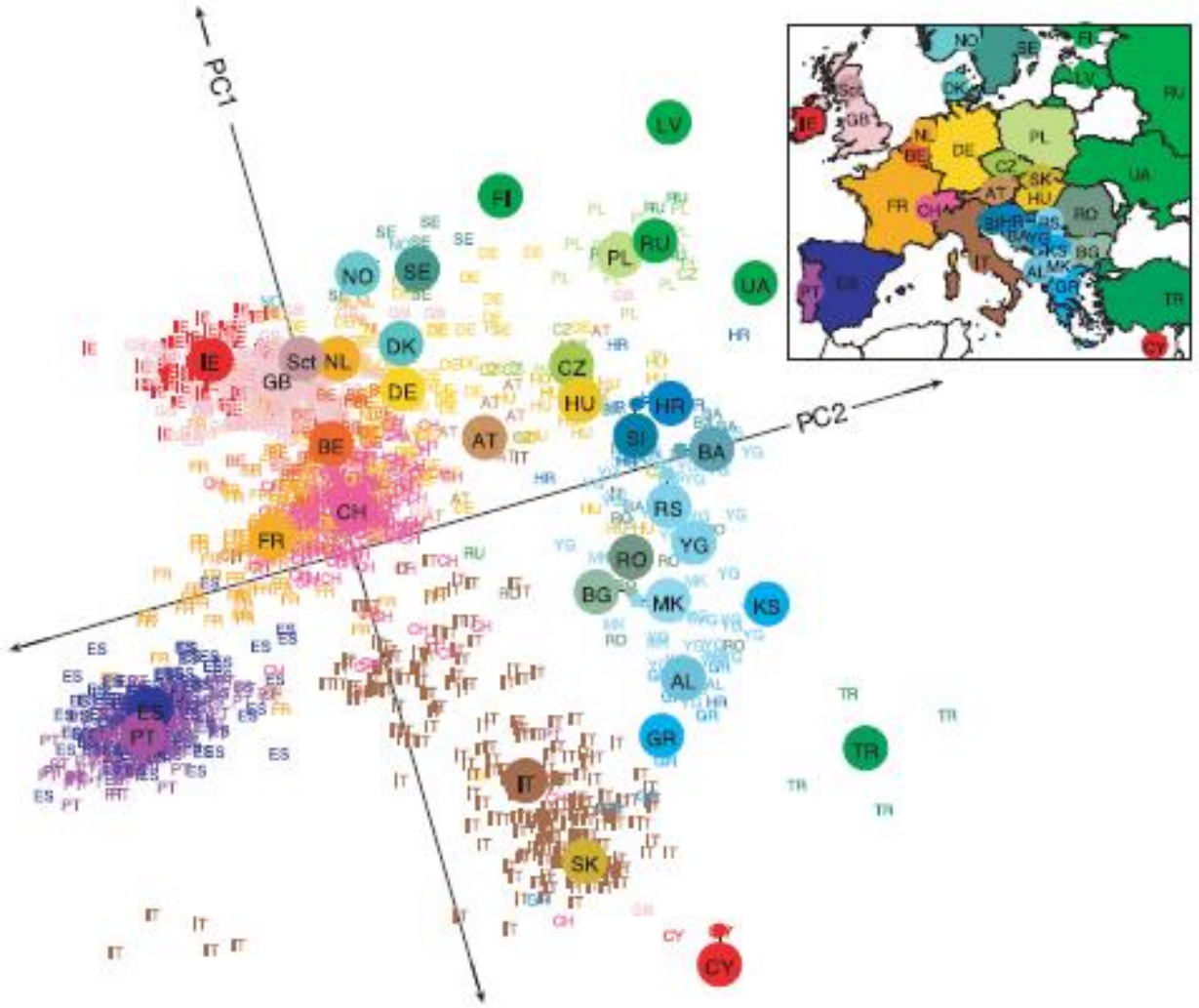
Population Structure of the Caribbean





Moreno-Estrada, PLoS Genetics 2013

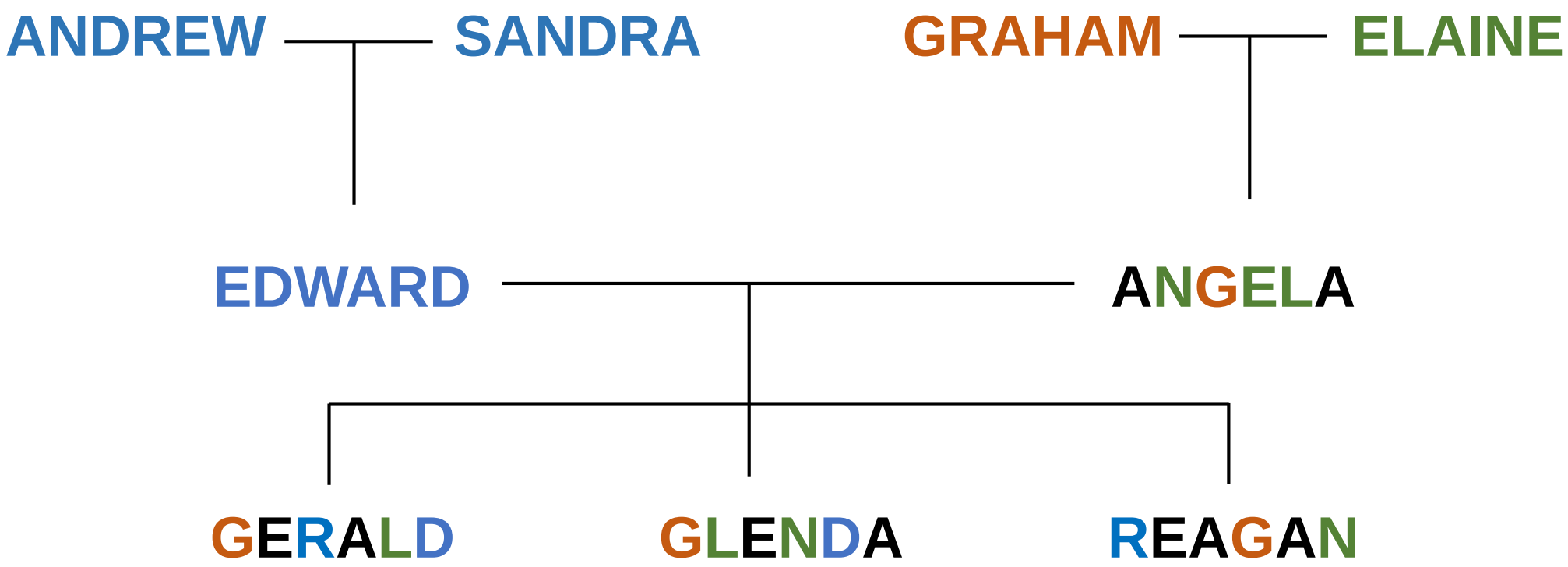
Population Structure in Europe



1,387 samples
~200K SNPs

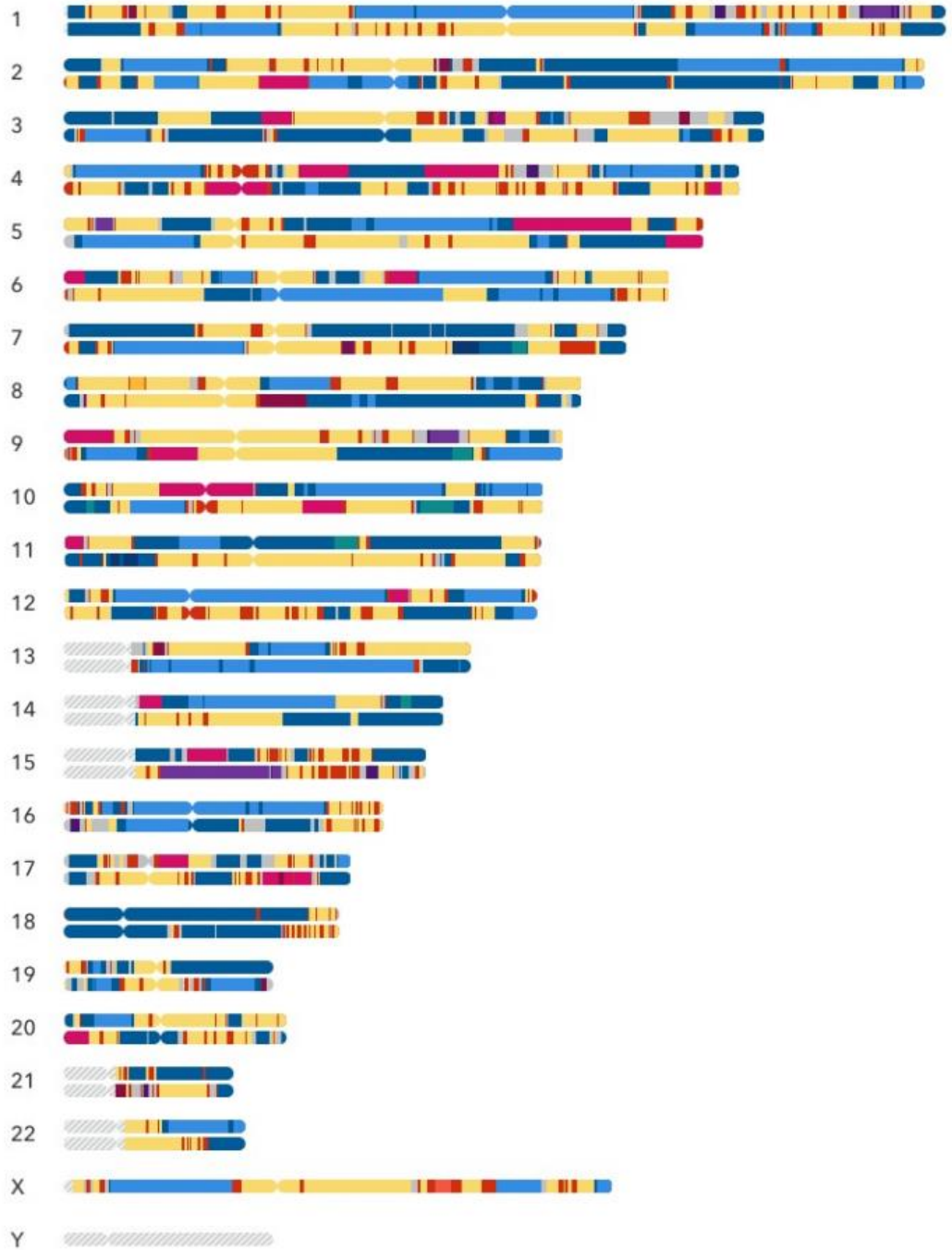
Novembre, Nature 2008

Ancestry determination in admixture



Adapted from ancestry.com

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 of admixed populations

European Americans		Proportion female contribution
$f_{African,male} = 0.04\%$	$f_{African,female} = 0.1\%$	71%
$f_{European,male} = 49.9\%$	$f_{European,female} = 48.7\%$	49%
$f_{N.American,male} = 0.02\%$	$f_{N.American,female} = 0.2\%$	91%

Table S4: **Best fit estimates of African American, Latino, and European American ancestry contributions for males and females, by ancestral population.** For African Americans, we estimate a male:female European ratio of 3.6, meaning that of European ancestors to African Americans, over three times as many were male as were female. Proportion female contribution is calculated for each cohort, for each ancestry, as $\frac{f_{female}}{f_{female}+f_{male}}$.

Bryc, AJHG 2015

Ancestry of admixed populations

African Americans		Proportion female contribution
$f_{African,male} = 31\%$	$f_{African,female} = 42.2\%$	58%
$f_{European,male} = 18.8\%$	$f_{European,female} = 5.2\%$	22%
$f_{N.American,male} = 0.2\%$	$f_{N.American,female} = 0.6\%$	75%

Latinos		Proportion female contribution
$f_{African,male} = 2.3\%$	$f_{African,female} = 3.9\%$	63%
$f_{European,male} = 40.7\%$	$f_{European,female} = 24.4\%$	37%
$f_{N.American,male} = 7.0\%$	$f_{N.American,female} = 11.0\%$	61%

European Americans		Proportion female contribution
$f_{African,male} = 0.04\%$	$f_{African,female} = 0.1\%$	71%
$f_{European,male} = 49.9\%$	$f_{European,female} = 48.7\%$	49%
$f_{N.American,male} = 0.02\%$	$f_{N.American,female} = 0.2\%$	91%

Table S4: **Best fit estimates of African American, Latino, and European American ancestry contributions for males and females, by ancestral population.** For African Americans, we estimate a male:female European ratio of 3.6, meaning that of European ancestors to African Americans, over three times as many were male as were female. Proportion female contribution is calculated for each cohort, for each ancestry, as $\frac{f_{female}}{f_{female}+f_{male}}$.

Bryc, AJHG 2015

Ancestry of admixed populations

African Americans		Proportion female contribution
$f_{African,male} = 31\%$	$f_{African,female} = 42.2\%$	58%
$f_{European,male} = 18.8\%$	$f_{European,female} = 5.2\%$	22%
$f_{N.American,male} = 0.2\%$	$f_{N.American,female} = 0.6\%$	75%

Latinos		Proportion female contribution
$f_{African,male} = 2.3\%$	$f_{African,female} = 3.9\%$	63%
$f_{European,male} = 40.7\%$	$f_{European,female} = 24.4\%$	37%
$f_{N.American,male} = 7.0\%$	$f_{N.American,female} = 11.0\%$	61%

European Americans		Proportion female contribution
$f_{African,male} = 0.04\%$	$f_{African,female} = 0.1\%$	71%
$f_{European,male} = 49.9\%$	$f_{European,female} = 48.7\%$	49%
$f_{N.American,male} = 0.02\%$	$f_{N.American,female} = 0.2\%$	91%

Table S4: **Best fit estimates of African American, Latino, and European American ancestry contributions for males and females, by ancestral population.** For African Americans, we estimate a male:female European ratio of 3.6, meaning that of European ancestors to African Americans, over three times as many were male as were female. Proportion female contribution is calculated for each cohort, for each ancestry, as $\frac{f_{female}}{f_{female}+f_{male}}$.

Bryc, AJHG 2015

Ancestry of admixed populations

African Americans		Proportion female contribution
$f_{African,male} = 31\%$	$f_{African,female} = 42.2\%$	58%
$f_{European,male} = 18.8\%$	$f_{European,female} = 5.2\%$	22%
$f_{N.American,male} = 0.2\%$	$f_{N.American,female} = 0.6\%$	75%

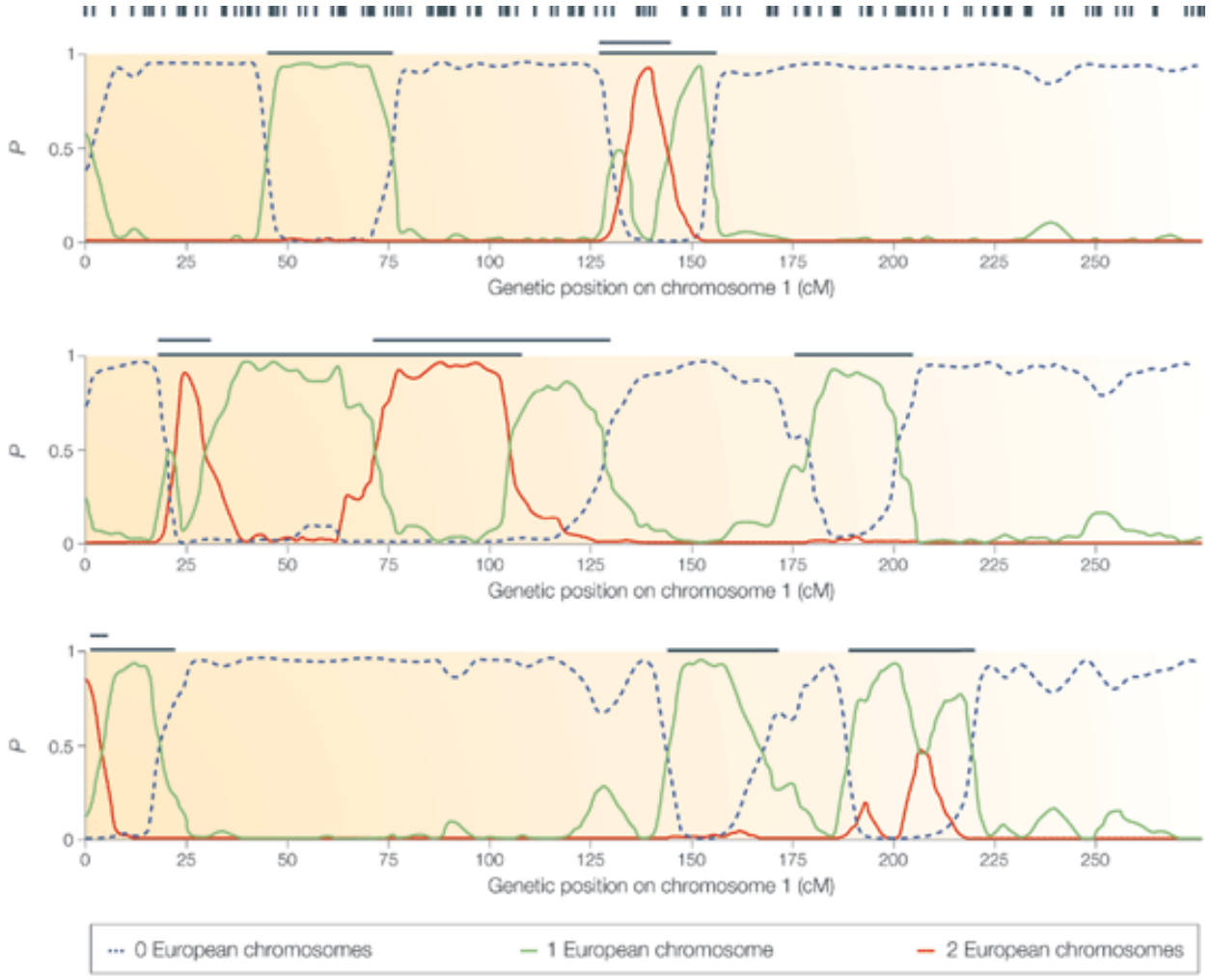
Latinos		Proportion female contribution
$f_{African,male} = 2.3\%$	$f_{African,female} = 3.9\%$	63%
$f_{European,male} = 40.7\%$	$f_{European,female} = 24.4\%$	37%
$f_{N.American,male} = 7.0\%$	$f_{N.American,female} = 11.0\%$	61%

European Americans		Proportion female contribution
$f_{African,male} = 0.04\%$	$f_{African,female} = 0.1\%$	71%
$f_{European,male} = 49.9\%$	$f_{European,female} = 48.7\%$	49%
$f_{N.American,male} = 0.02\%$	$f_{N.American,female} = 0.2\%$	91%

"European Americans might have ten times as many female Native American ancestors as male, and African Americans might have four times as many female Native American ancestors as male."

Bryc, AJHG 2015

Genetic ancestry of African American

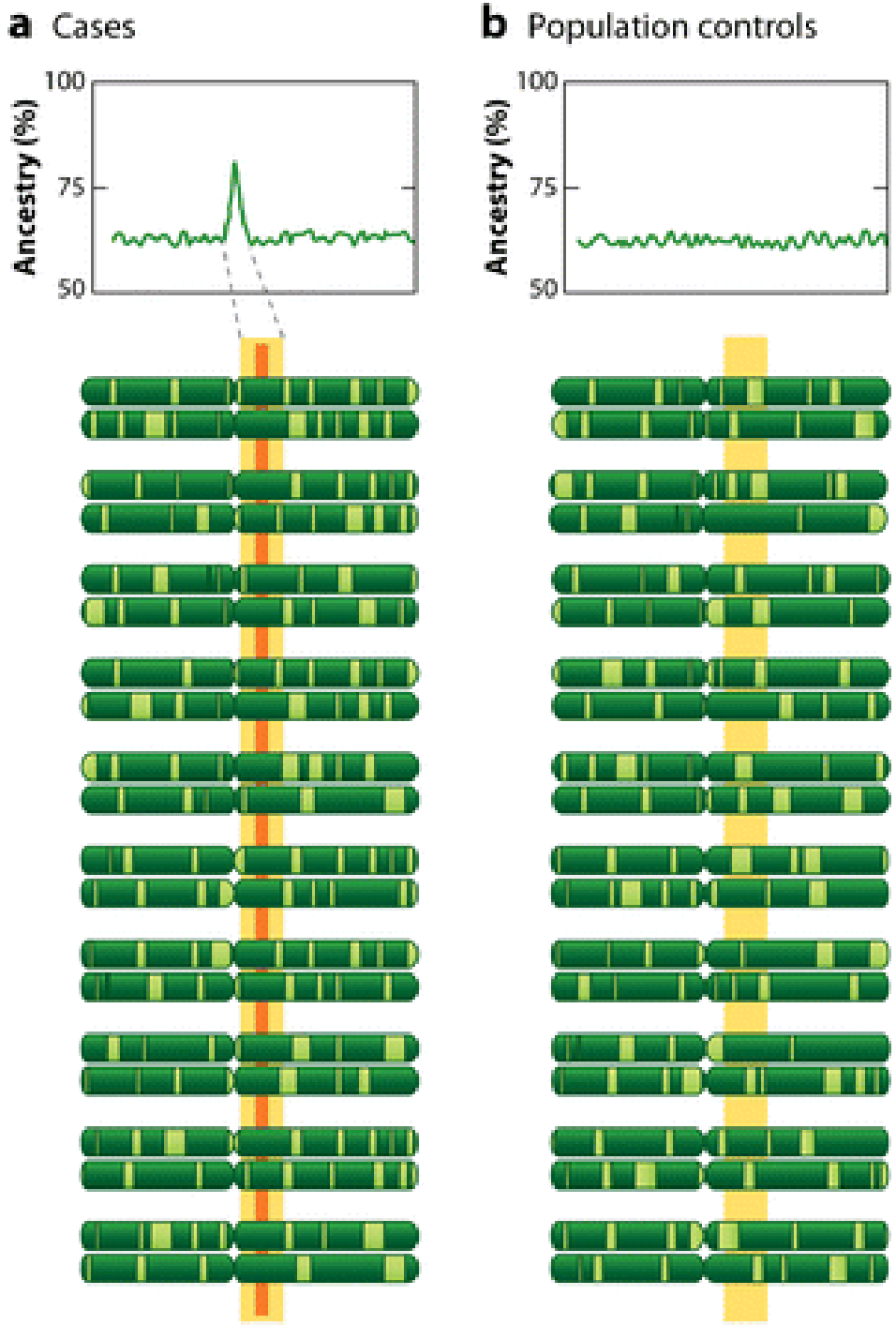


Smith, Nature Rev Genetics, 2005

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

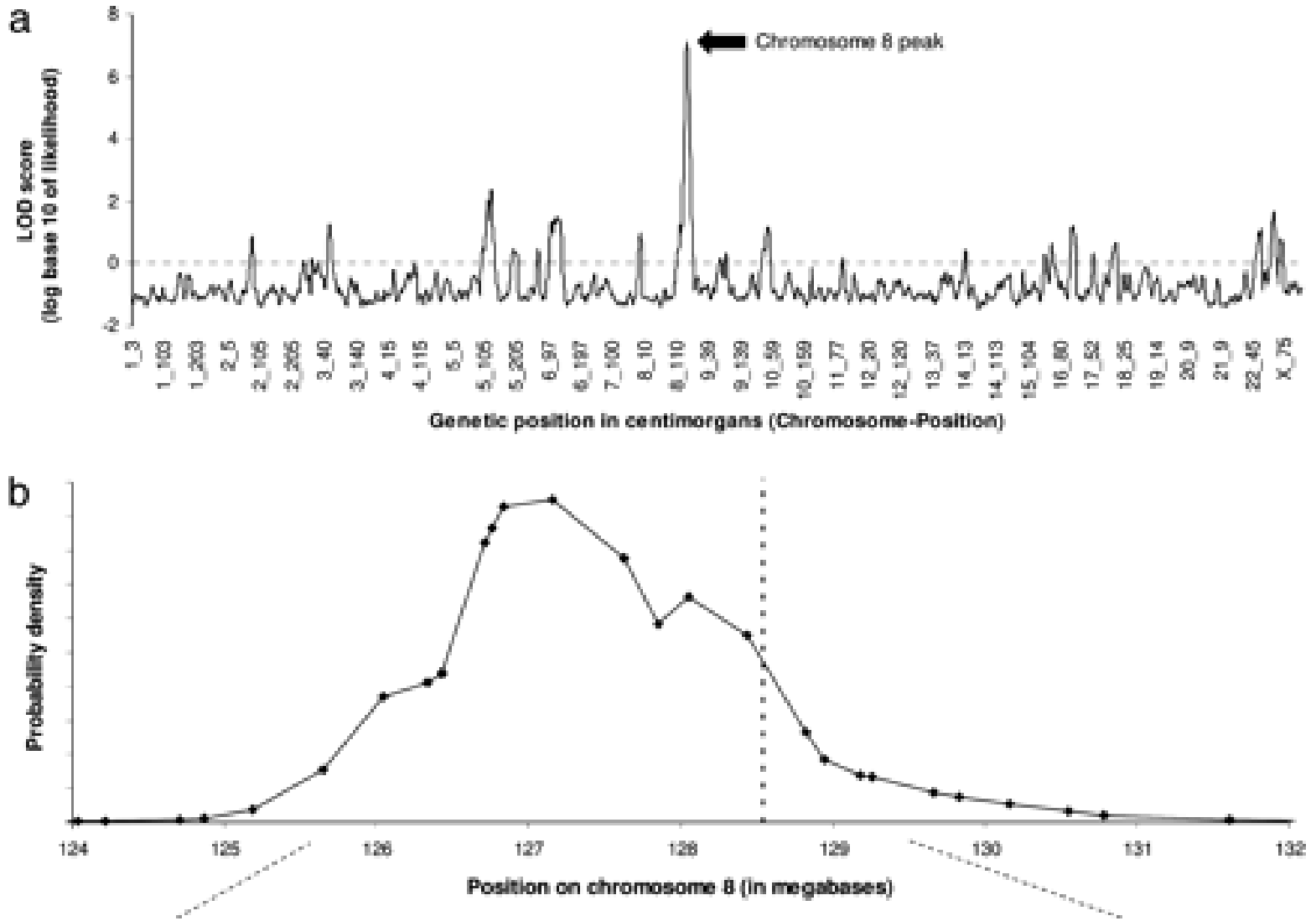
Admixture mapping – a tool for gene discovery

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

Whole-genome admixture scan identified the 8q24 locus in prostate cancer - 1,597 prostate cancer cases and 873 controls



Freedman, PNAS 2006

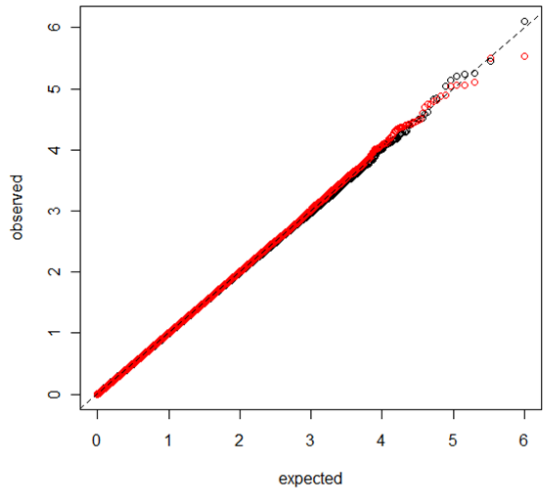
Assess potential population stratification

- Most of the genetic markers on the genome (e.g. a GWAS) are likely not associated with the disease

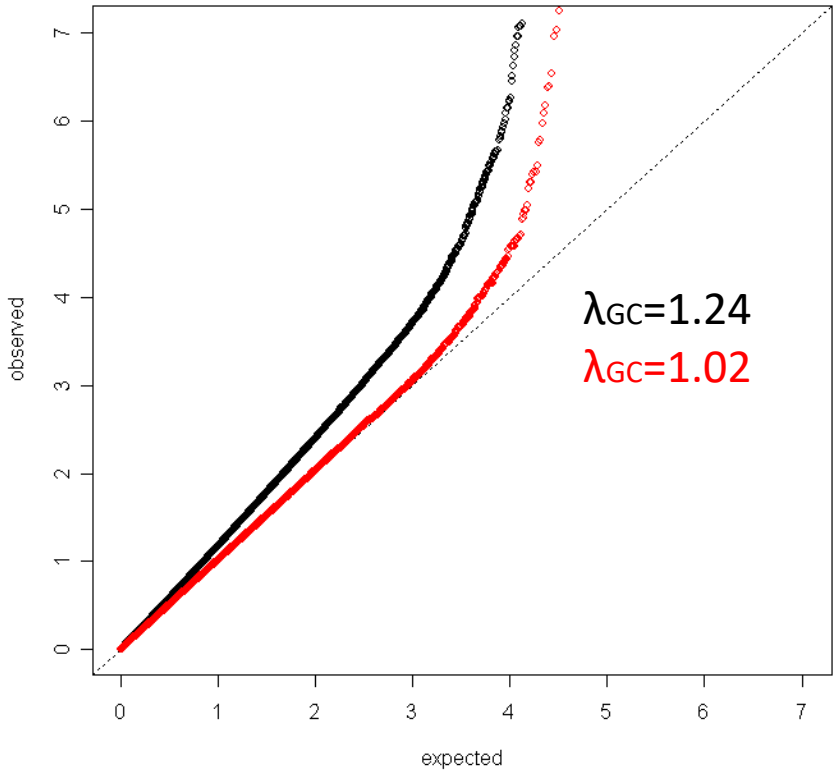
- The genomic control parameter (λ_{GC}) summarizes systematic inflation from a large number of association test results

$$\lambda_{GC} = \frac{The\ median\ of\ the\ observed \chi^2\ statistics}{The\ median\ of\ the\ \chi^2\ statistics\ under\ the\ NULL}$$

For a 1 d.f. χ^2test, the denominator is **0.455**



Hair Color in Nurses Health Study (n=2,287)



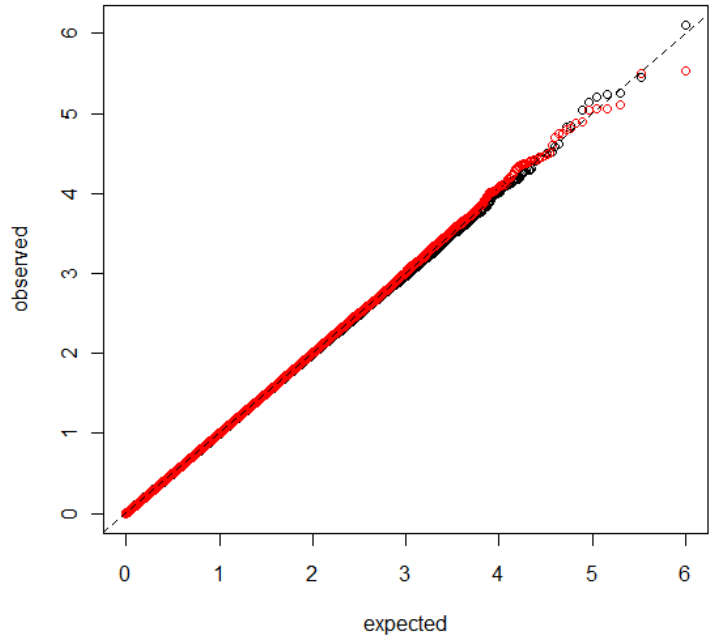
Han, PloS Genetics 2008

QQ plot for a GWAS of dark-light hair color in US European-ancestry subjects from the NHS.

The black points are the p-values from the unadjusted tests.

The red points are from **principal-component adjusted tests**.

QQ plot for a GWAS of breast cancer in the same NHS samples (breast cancer risk does not correlate with European ancestry)



Hunter Nat Genet 2007

A few notes about λ_{GC}

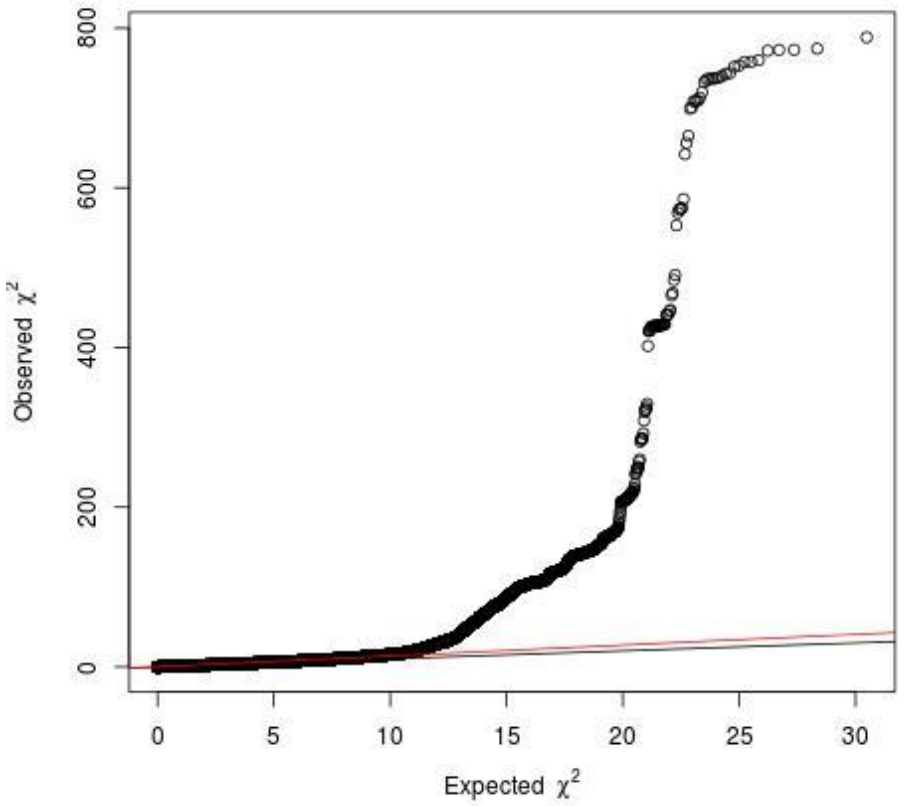
- λ_{GC} should be close to 1 if no bias exists.
  - Rule of thumb: In limited sample sizes, λ_{GC} <1.05 is often ok, above 1.1 deserves attention

- λ_{GC} scales with sample size
  - Under a polygenic model, many SNPs with small effect sizes will be detected with very large sample size -> expect λ_{GC} to increase
  - λ_{GC} of 1.06 is a much bigger concern in studies with hundreds of samples compared to studies with thousands of samples

- A standard approach is to correct for inflation by dividing all test statistics by λ_{GC}
  - Drawback: Affects all SNPs, so SNPs that are not affected by bias are overpenalized and SNPs that are very affected by bias are underpenalized

N=122,977 cases and 105,974 controls

Inflation factor 1.37



Michailidou, Nature 2017

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.