Wellcome Trust Advanced Courses; Genomic Epidemiology in Africa, 21st – 26th June 2015
Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Durban, South Africa

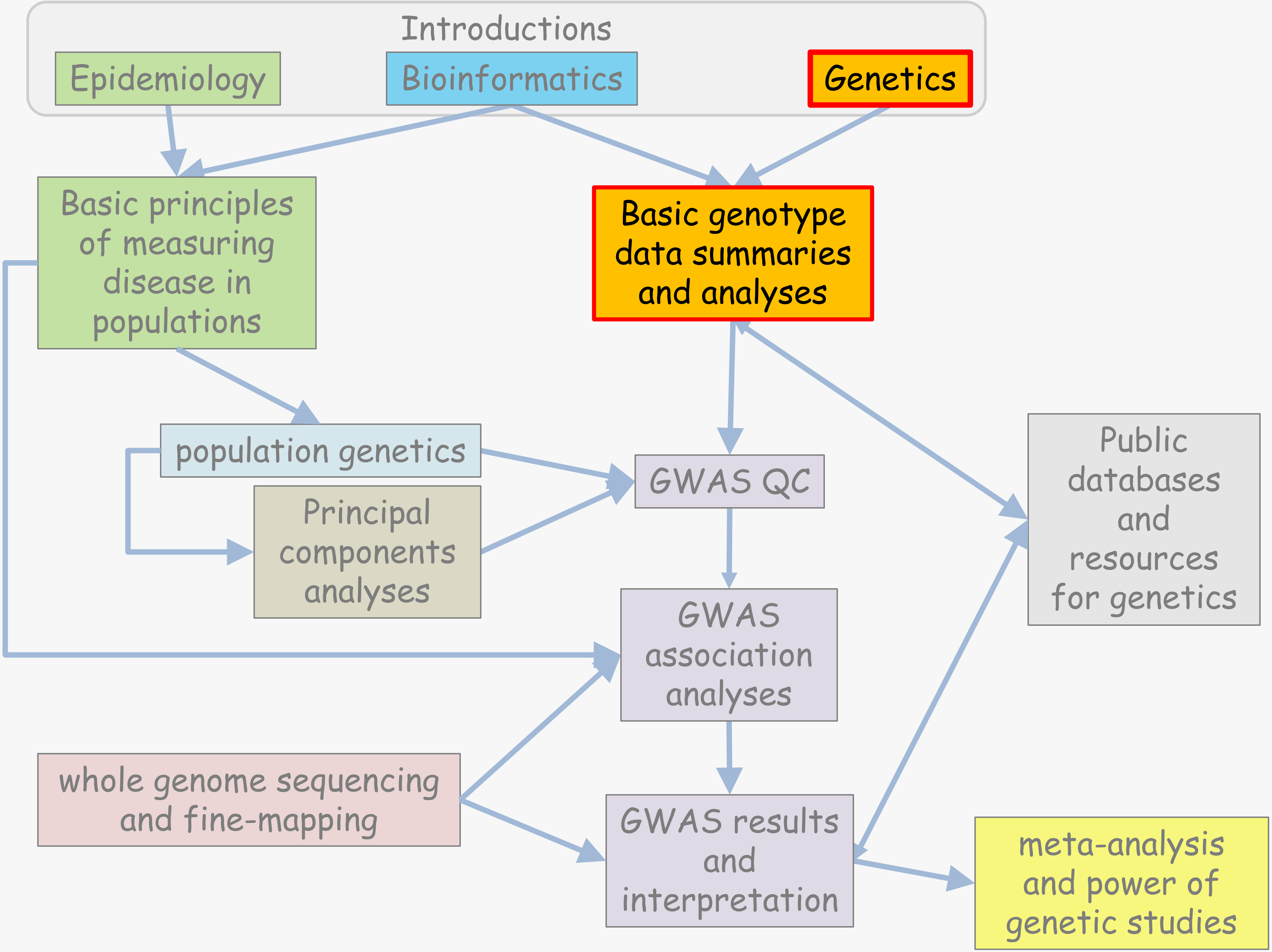
Introduction to genetic association studies in Africa

Dr Kirk Rockett

Nuffield Department of Medicine

The Wellcome Trust Centre for Human Genetics

University of Oxford



Introductions

Epidemiology

Bioinformatics

Genetics

Basic principles of measuring disease in populations

Basic genotype data summaries and analyses

population genetics

Principal components analyses

GWAS QC

Public databases and resources for genetics

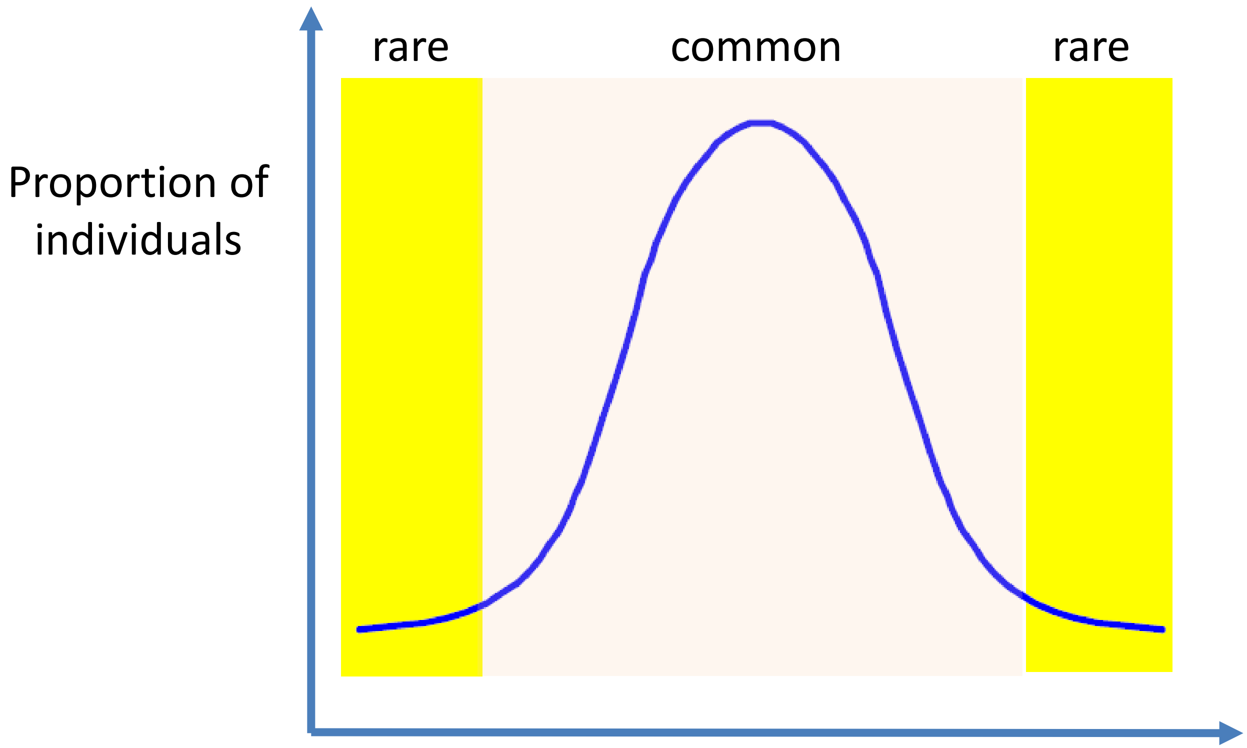
GWAS association analyses

whole genome sequencing and fine-mapping

GWAS results and interpretation

meta-analysis and power of genetic studies

A complex trait



Variation due to age, sex, environmental factors (e.g. diet), and genetic variation.

- A small proportion of variation is caused by **rare gene defects causing major disruption of normal physiological processes. These tend to be found at the extremes of the** distribution.

- Most variation is probably due to **multiple common variants that slightly alter normal physiological processes. It is** challenging to pin down the variants responsible because, at an individual level, they do not have strong effects.

Variation in resistance & susceptibility to disease

Why should we look for common variants with small effects?

- These variants may not contribute much to overall risk.
- ***But*** they may lead to new insights into etiology of disease – e.g. mechanisms of immunity, disease, drug action, erythrocyte invasion and other critical host – parasite interactions.
- …and new drug targets.
- We now have the scientific tools to do it.

Genetic variation

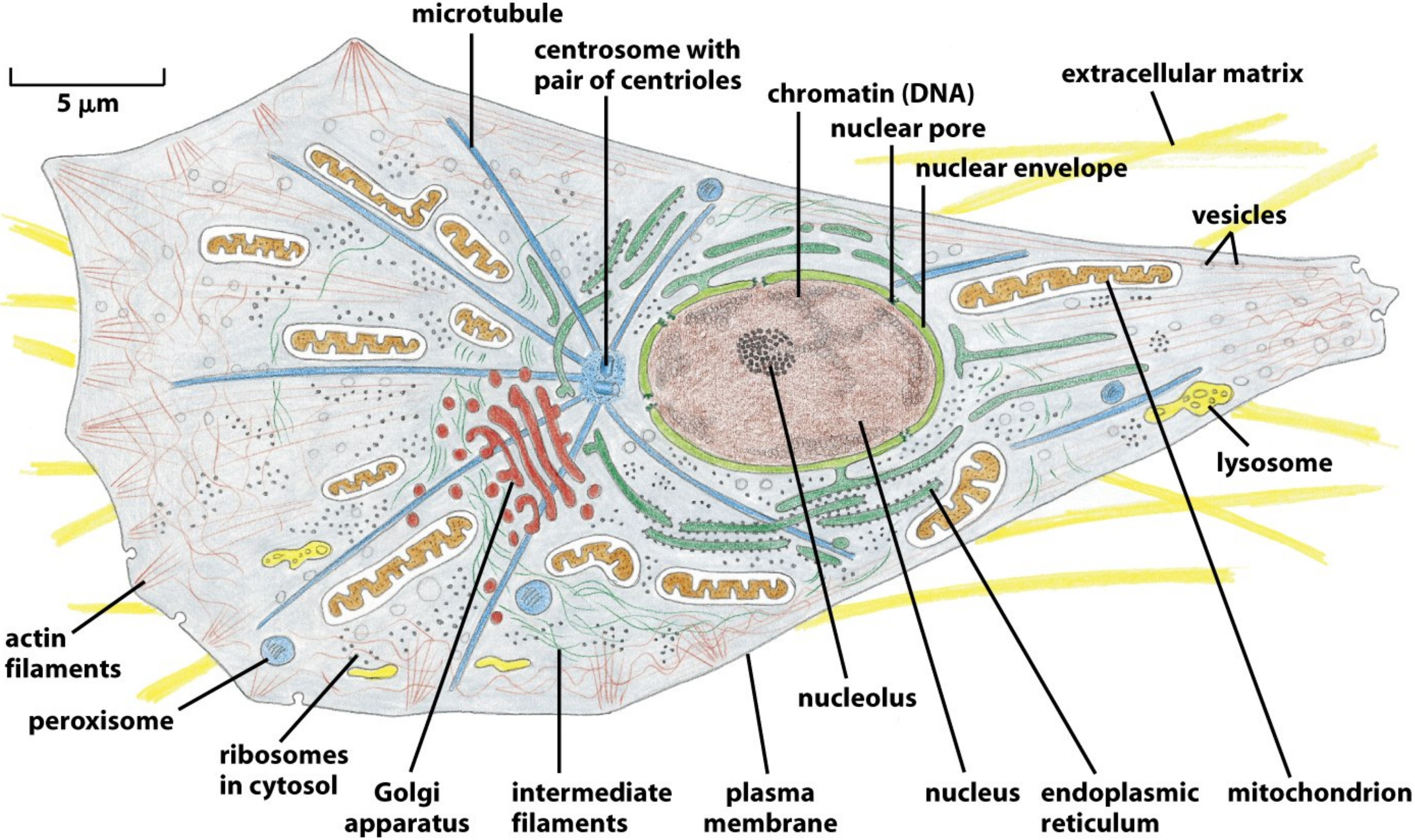
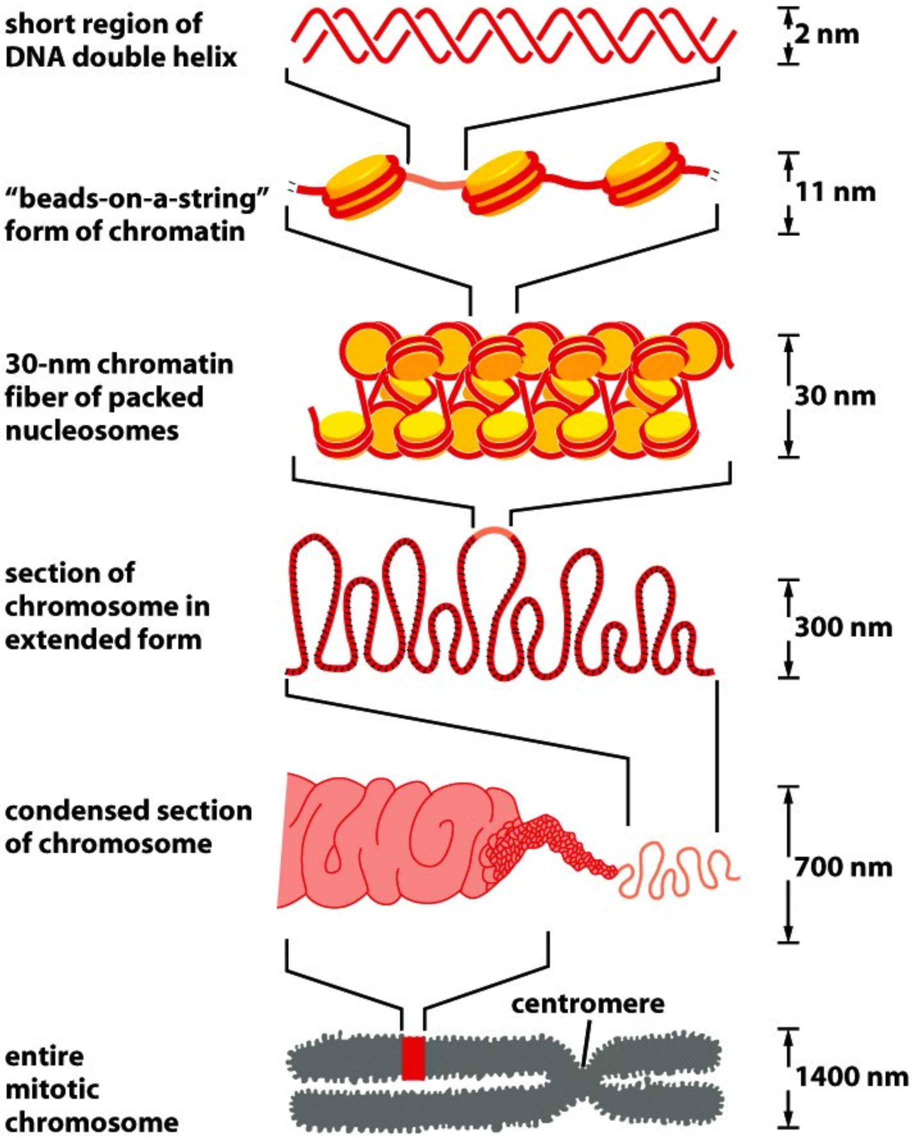


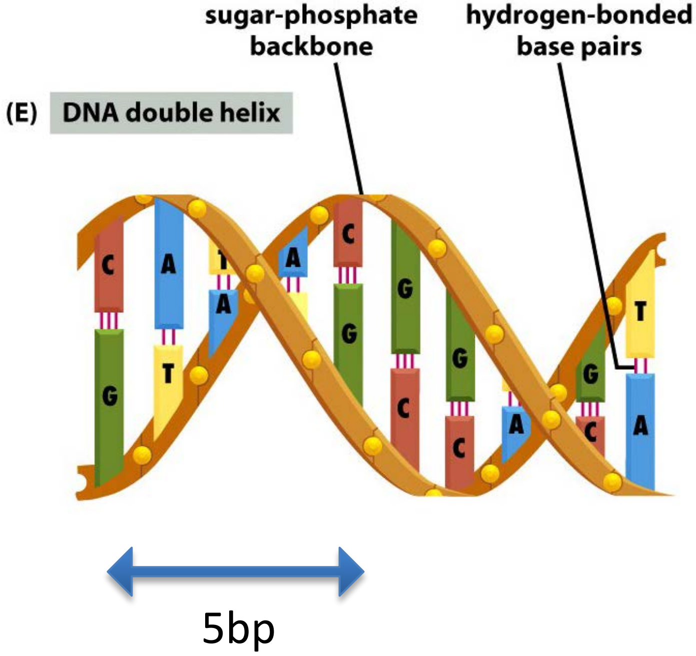
Figure 1-30 Molecular Biology of the Cell 5/e (© Garland Science 2008)

DNA structure overview


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm

NET RESULT: EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 10,000-FOLD SHORTER THAN ITS EXTENDED LENGTH

Figure 4-72 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Genetic variation in the human genome

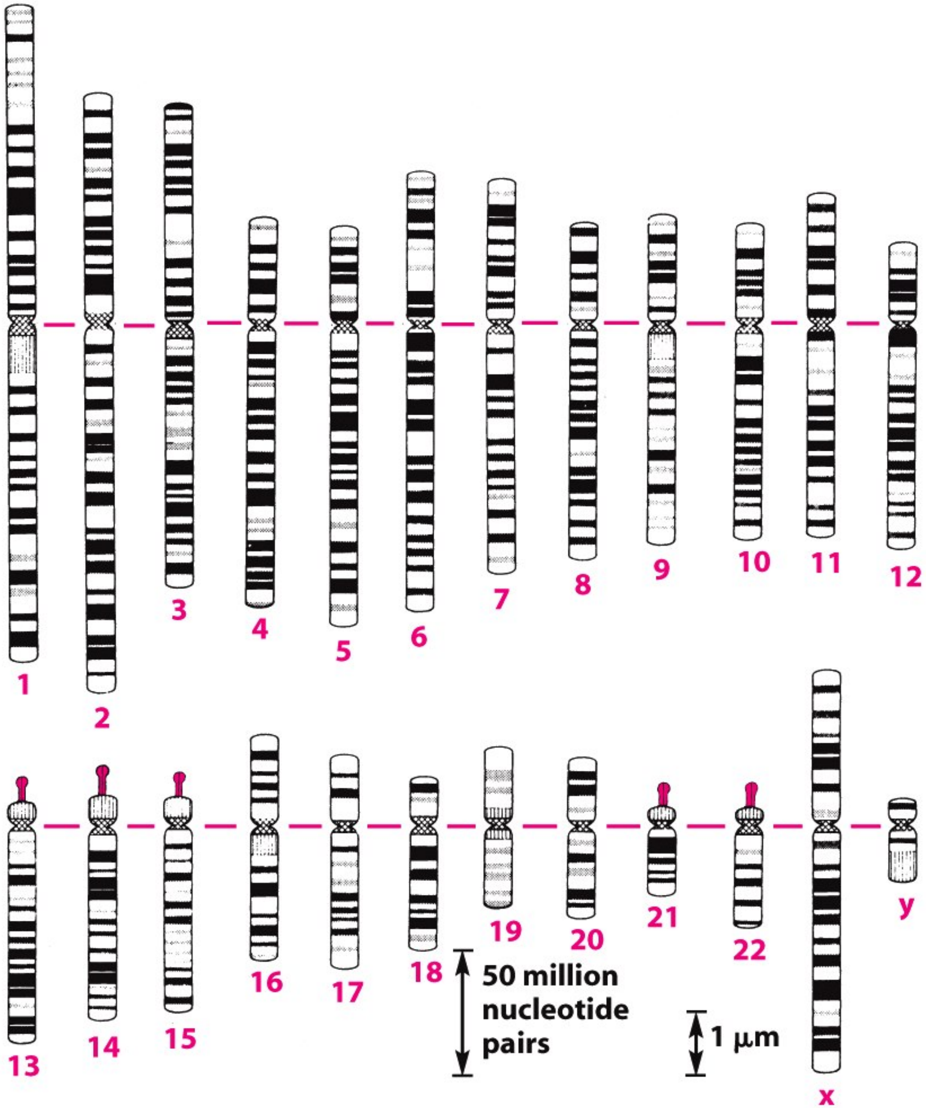


Figure 4-11 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Common forms of variation in the human genome

There are many different variants including

small variations in the DNA sequence, e.g.

- a small ‘spelling mistake’
- deletion or insertion of a few characters

large structural variations, e.g.

- deletion of a large part of DNA sequence
- multiple copies of a section of DNA sequence, with variable copy number

Common forms of variation in the human genome

Most variants are single nucleotide polymorphisms (SNPs)

```
ACTCTACGATTTACGGTACTTAGGAGCATATGCTACT
ACTGTACGATTTACGGTACTTAG.AGCATATGCTACT
```

SNP
single nucleotide
polymorphism

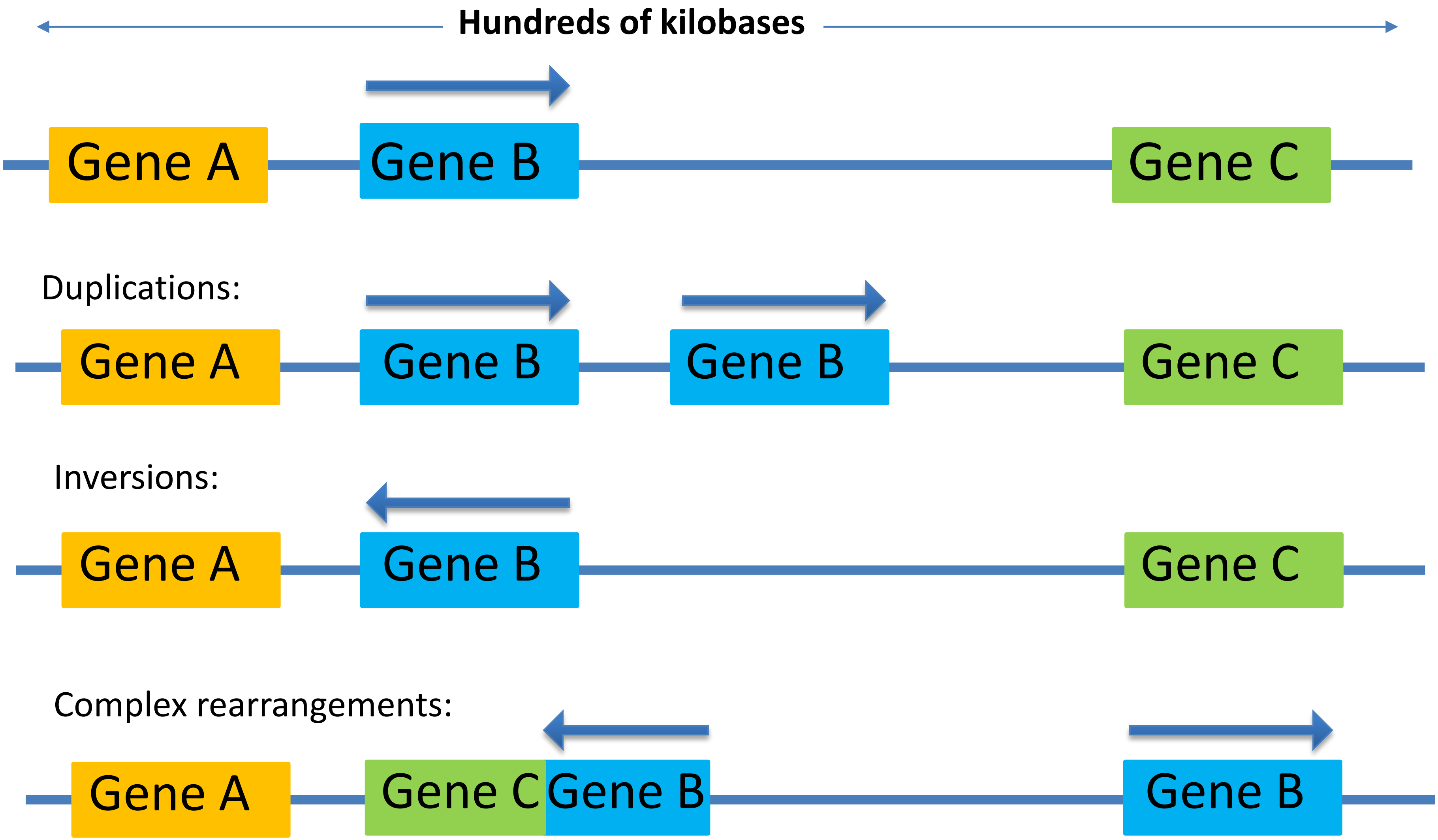
indel
insertion /
deletion

About 38 million SNPs found across the human genome worldwide – one every 84bp.

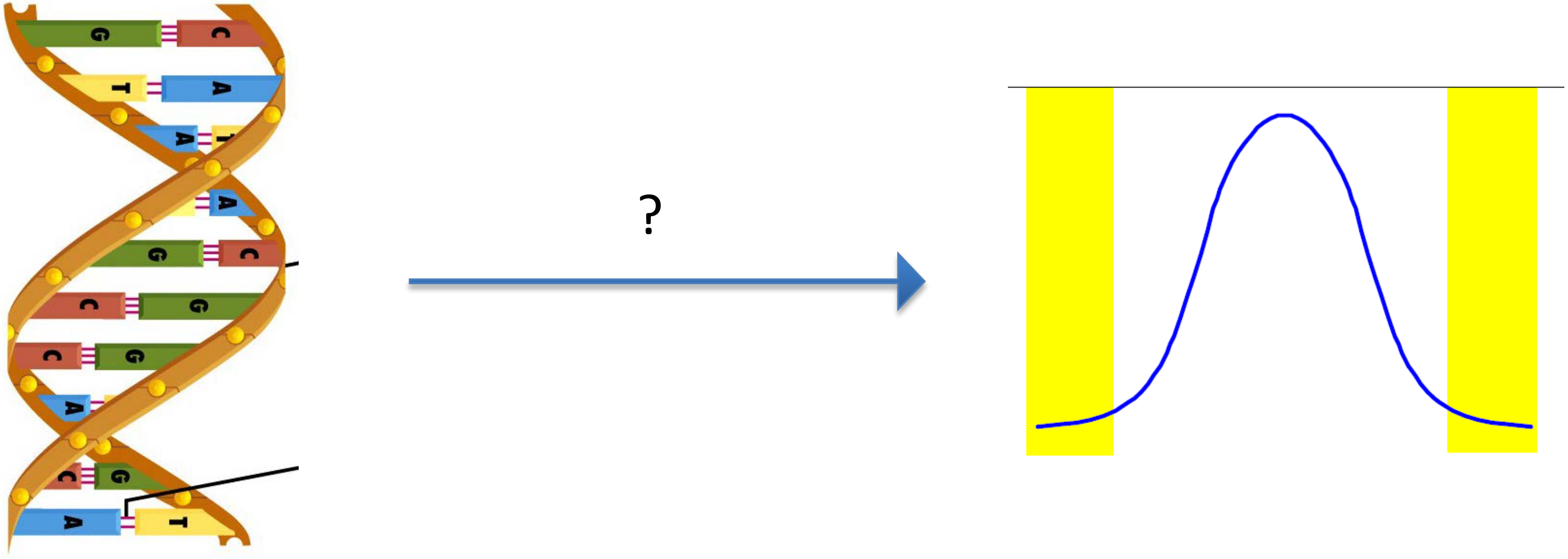
Maybe ~2 million small indels worldwide – about one every 1,600bp.

Common forms of variation in the human genome

Structural variants



Finding loci that influence disease



?

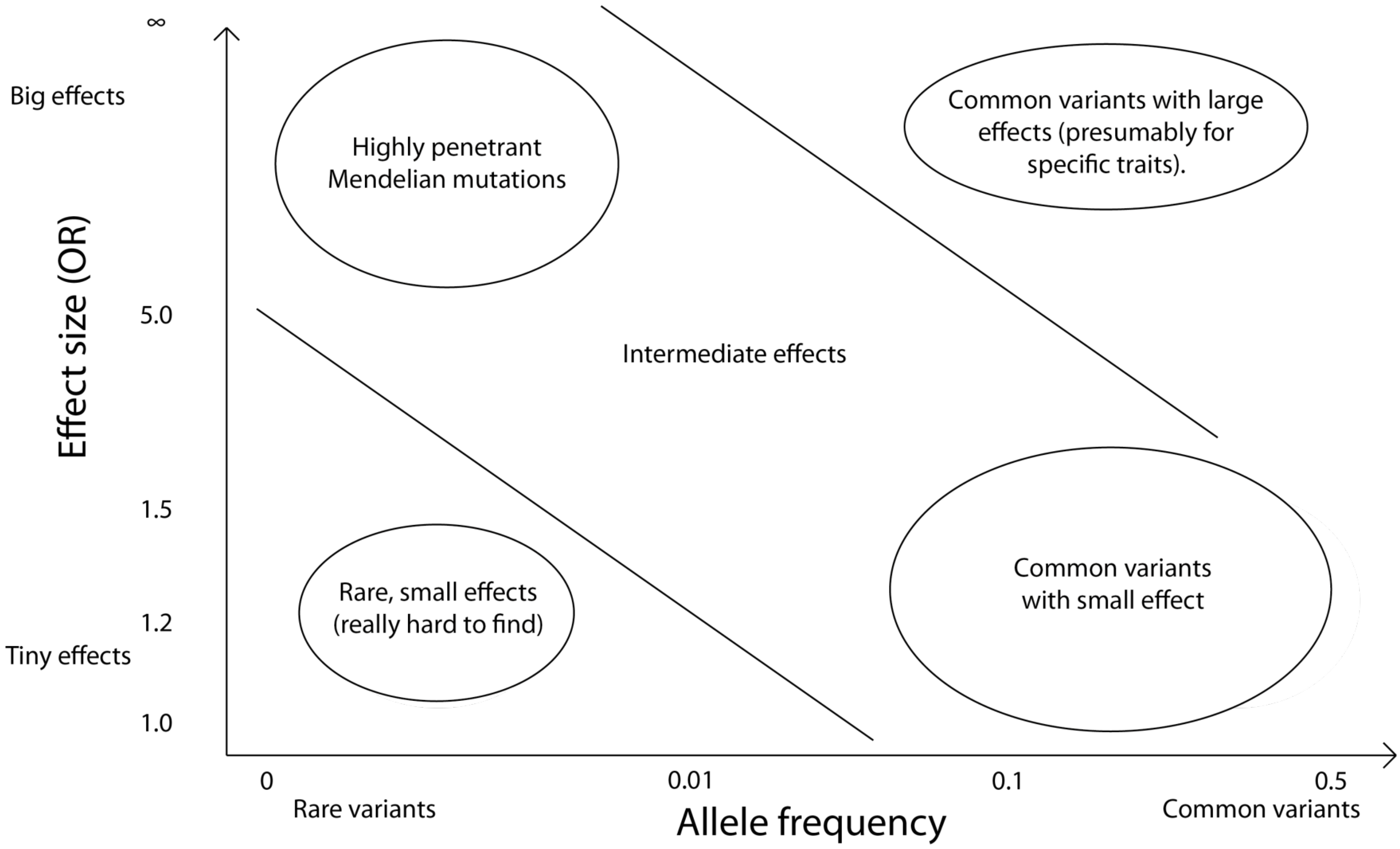
Finding loci that influence disease

Association studies broadly fall into two categories:

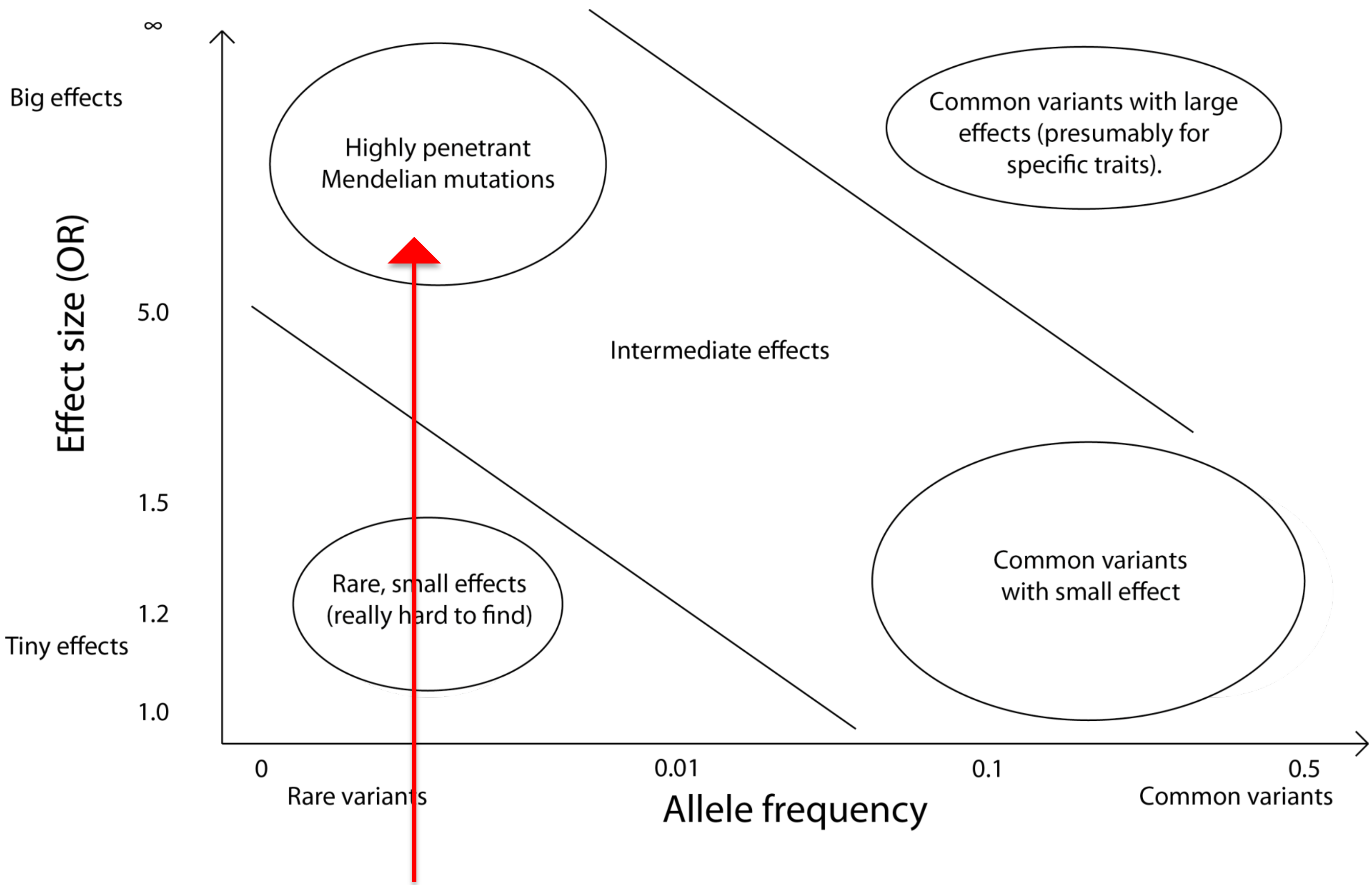
- Family-based studies
- Case/control studies

Mixed designs are also possible.

Variation in resistance & susceptibility to disease

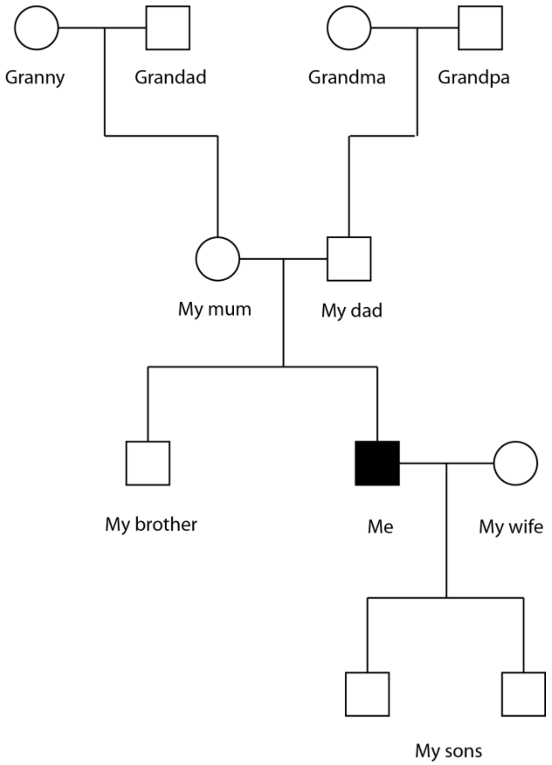


Variation in resistance & susceptibility to disease


∞
Big effects
Effect size (OR)
5.0
1.5
1.2
Tiny effects
1.0
Highly penetrant Mendelian mutations
Common variants with large effects (presumably for specific traits).
Intermediate effects
Rare, small effects (really hard to find)
Common variants with small effect
0
0.01
0.1
0.5
Rare variants
Allele frequency
Common variants

Family (linkage and/or sequencing) studies

Family-based association analysis



Compare *probands* (e.g. cases) with other family members, such as parents.

Pros:

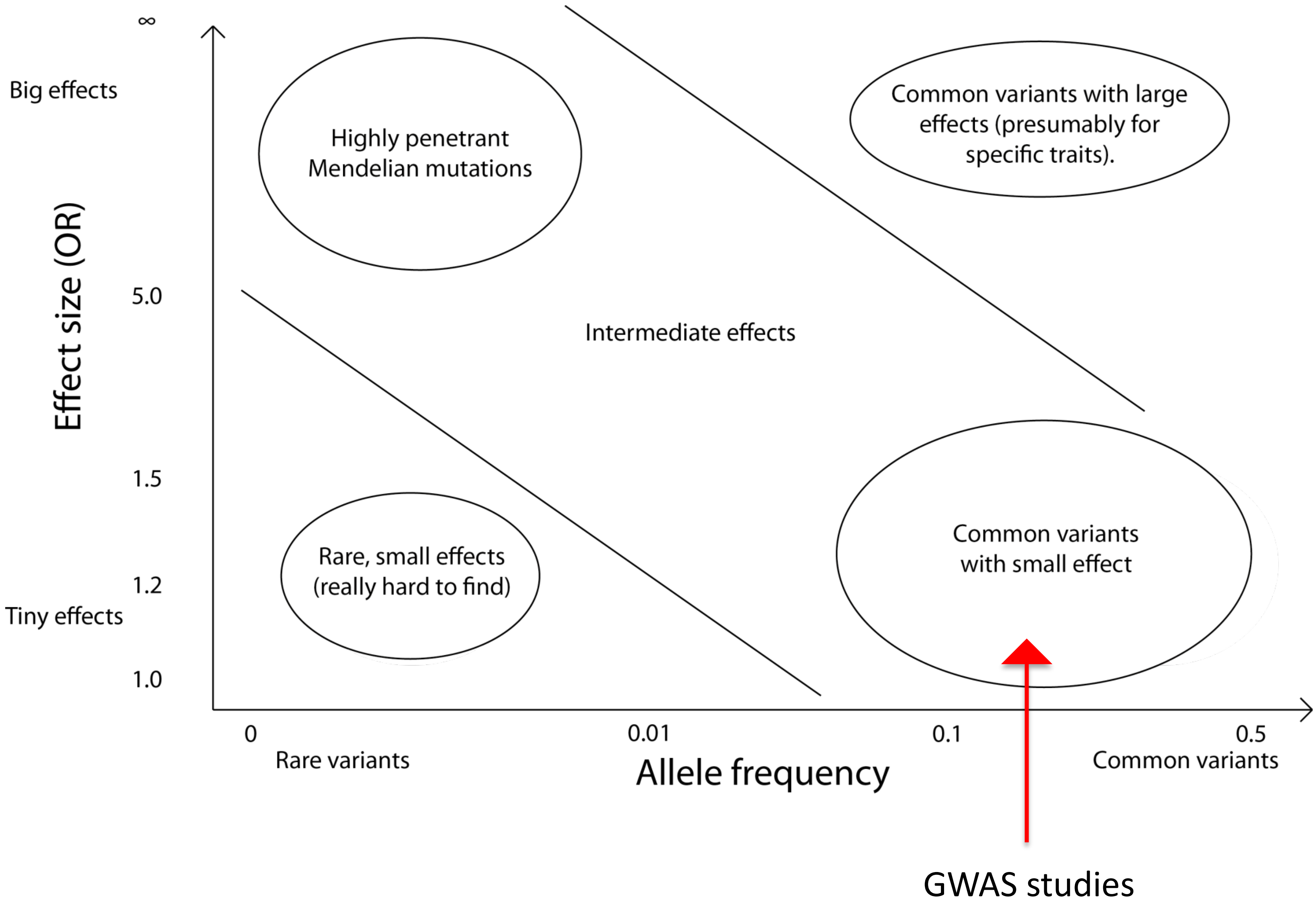
- Robust against potential confounding factors, such as population structure or environmental effects.
- Great when looking for variants with *big* effects.
- Extended family designs can go where other designs can't(*).

Cons:

- Can be harder difficult to collect large samples.
- For common variants / complex trait association there is potentially reduced power (for equal sample size)

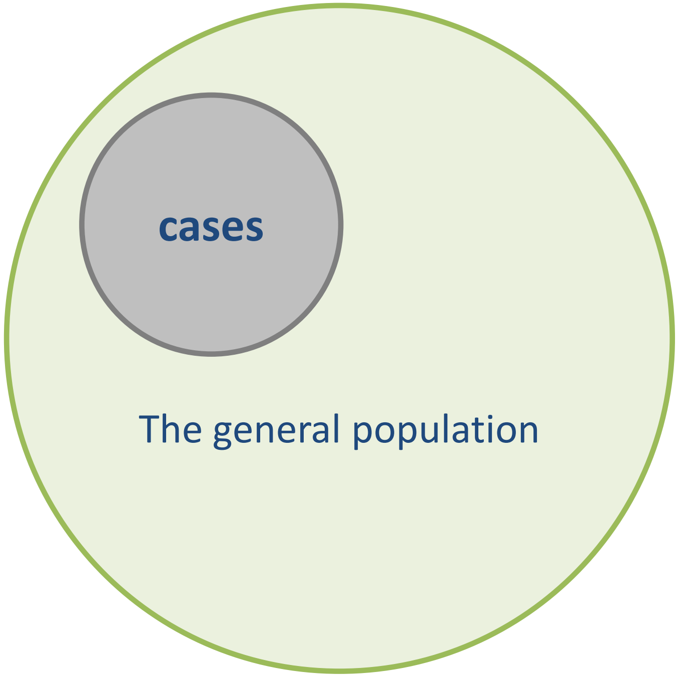
(*) e.g. Kong et al, "*Parental origin of sequence variants associated with complex diseases*", Nature 462 (2009)

Variation in resistance & susceptibility to disease


∞
Big effects
Effect size (OR)
5.0
1.5
1.2
Tiny effects
1.0
Highly penetrant Mendelian mutations
Common variants with large effects (presumably for specific traits).
Intermediate effects
Rare, small effects (really hard to find)
Common variants with small effect
0
Rare variants
0.01
Allele frequency
0.1
0.5
Common variants
GWAS studies

Case/control association analysis

Compare disease-affected individuals (*cases*) with unaffected individuals (*controls*).



Pros:
Large sample sizes can be realised => powered to detect small effects.

Cons:
Potential confounding effects from differential selection of cases and controls – (e.g. cases and controls should be ethnically matched where possible).

Most of this course will focus on case/control designs.

What do we need to know to detect our effect?

Or what POWER do we have to detect an effect

A heuristic for statistical power

Power = *how likely are we to find a real effect?*

$$Power \approx N\ \beta^2\ f(1-f)\ r^2$$

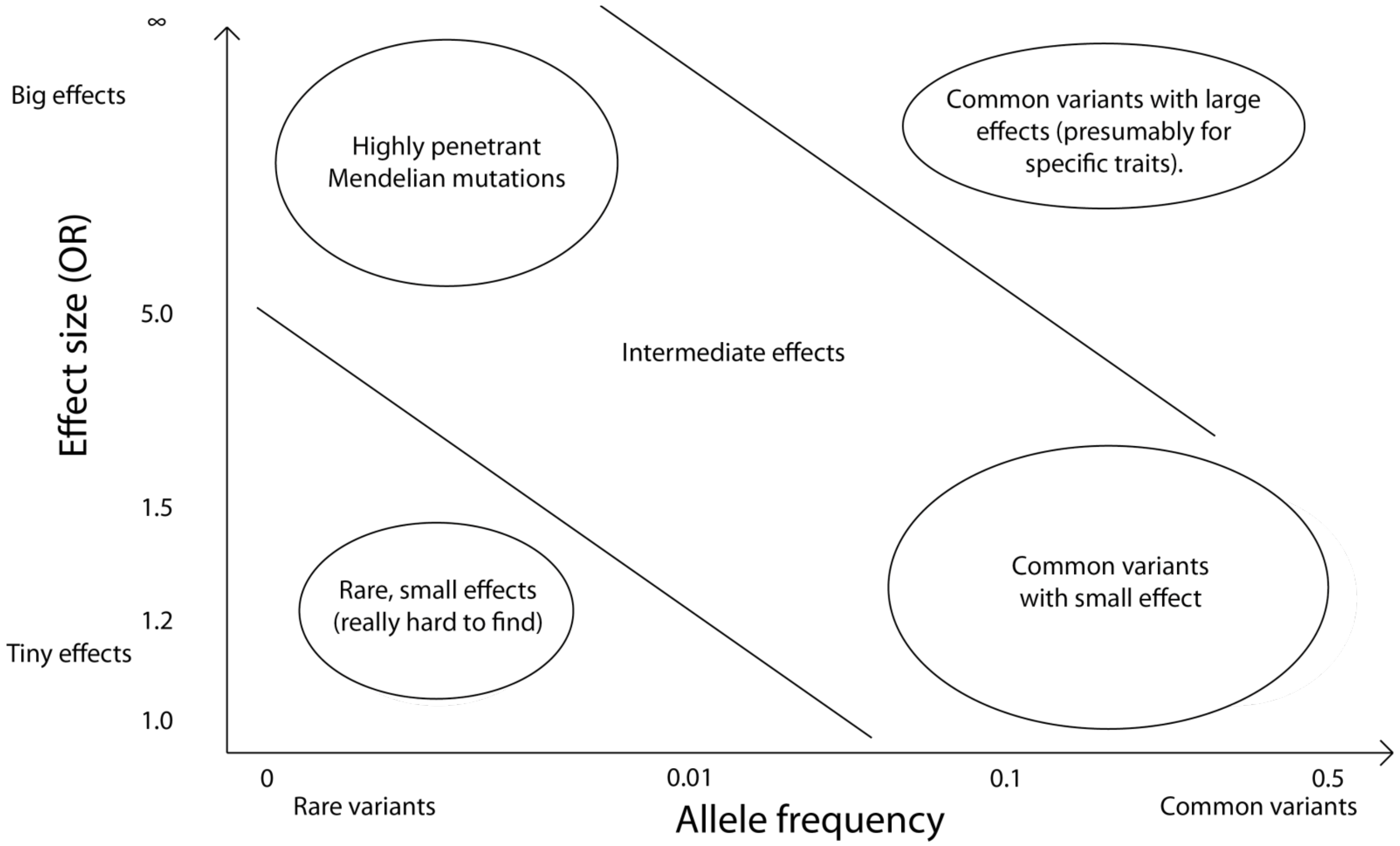
Effect size

Number of samples

LD

Allele frequency

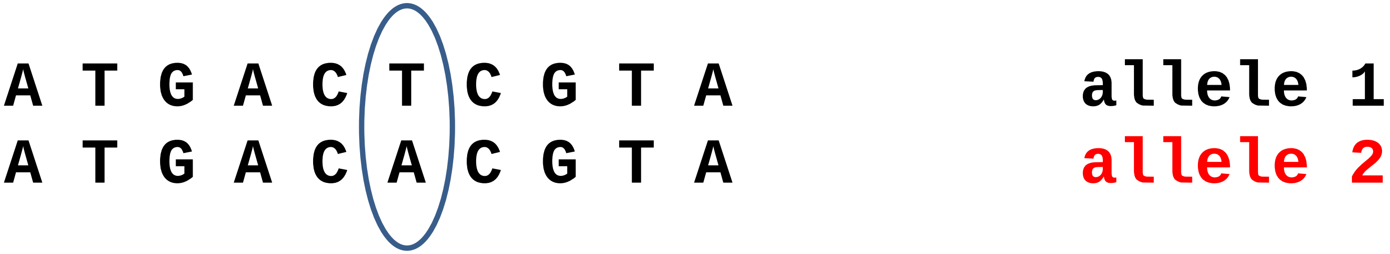
Variation in resistance & susceptibility to disease


∞
Big effects
Effect size (OR)
5.0
1.5
1.2
Tiny effects
1.0
Highly penetrant Mendelian mutations
Common variants with large effects (presumably for specific traits).
Intermediate effects
Rare, small effects (really hard to find)
Common variants with small effect
0
Rare variants
0.01
Allele frequency
0.1
0.5
Common variants

$$Power \approx N \ \ \beta^2 \ f(1\text{-}f) \ \ r^2$$

Finding loci that influence disease

- Consider a position in the genome that shows variation between individuals, for example …



```
A T G A C T C G T A          allele 1
A T G A C A C G T A          allele 2
```

- Each of the different variant forms is called an **allele**

- We are looking for alleles that are associated with **high or low risk of disease**

Example: sickle and severe Malaria

Gambian data (MalariaGEN consortium)

	Genotype		
	HbAA (normal)	HbAS sickle trait	HbSS sickle cell disease
	TT	**AT**	**AA**
Severe malaria cases	2700	35	13
Population	3689	588	22

$N = 7047$

$f = 0.07$ (7%)

Example: sickle and severe Malaria

Gambian data (MalariaGEN consortium)

	TT	AT	AA
Severe malaria cases	2700	35	13
Population	3689	588	22

Odds ratio = 3689*35 / 2700 * 588 = 0.08

$P < 2x10^{-16}$

e.g. chisq.test in R

Individuals with AT (sickle) genotype have 10-fold lower risk of malaria than those with TT (wild-type) genotype.

Genome-wide association analysis (GWAS) in a nutshell

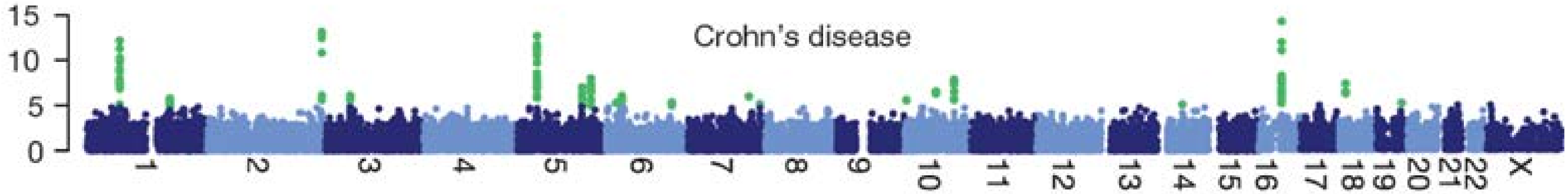
Aim:

- Find common variants influencing disease by performing this test at millions of variants across the human genome.
- Typical modern experiment: type 2.5M variants in thousands of cases and thousands of population controls. Use estimated genome-wide relationships to control for population structure.
- This design exploits linkage disequilibrium to assess variants that are not directly typed.

Key concept: linkage disequilibrium

Genome-wide association (GWA) analysis in a nutshell

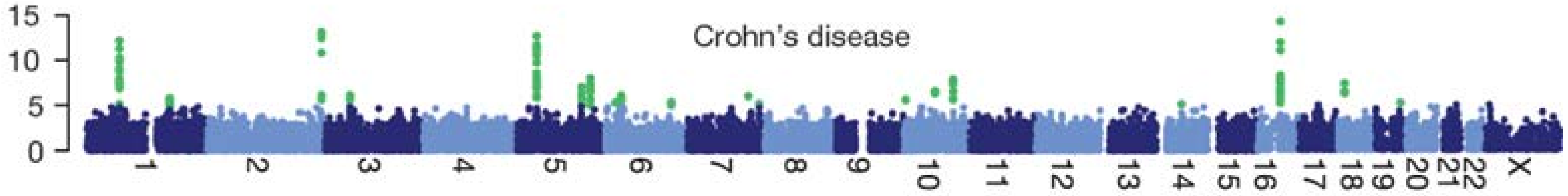
Amazingly, it works! E.g: 2,000 cases and 3,000 controls typed at 500k variants:



"Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls"
The Wellcome Trust Case Control Consortium Nature 447 (2007)

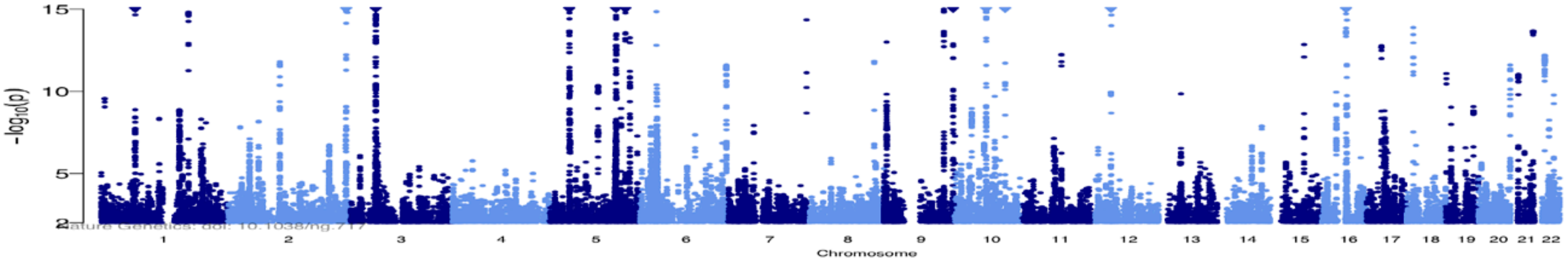
Genome-wide association (GWA) analysis in a nutshell

Amazingly, it works! E.g: 2,000 cases and 3,000 controls typed at 500k variants:



"Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls" The Wellcome Trust Case Control Consortium Nature 447 (2007)

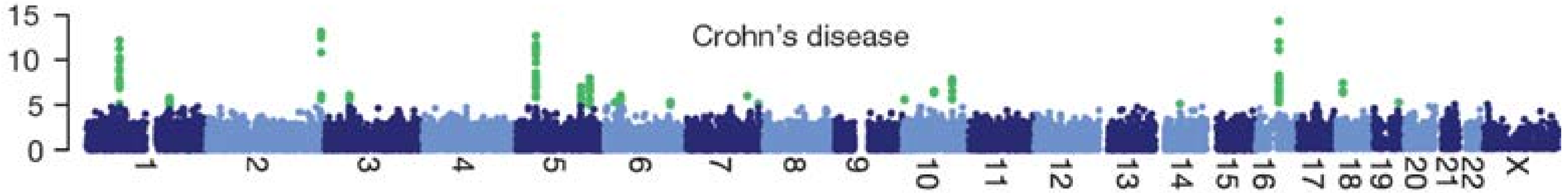
With 6,000 cases and 15,000 controls imputed to 1 million variants:



"Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci", Franke et al Nature Genetics 42 (2010)

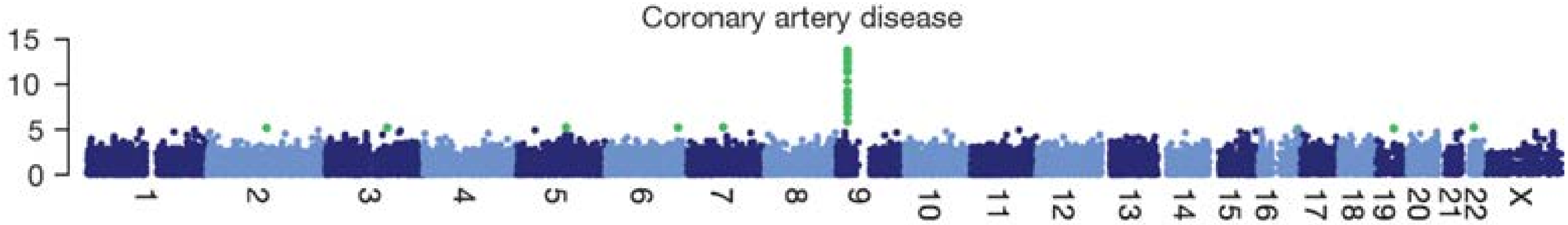
Genome-wide association (GWA) analysis in a nutshell

Amazingly, it works! E.g: 2,000 cases and 3,000 controls typed at 500k variants:



"Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls"
The Wellcome Trust Case Control Consortium Nature 447 (2007)

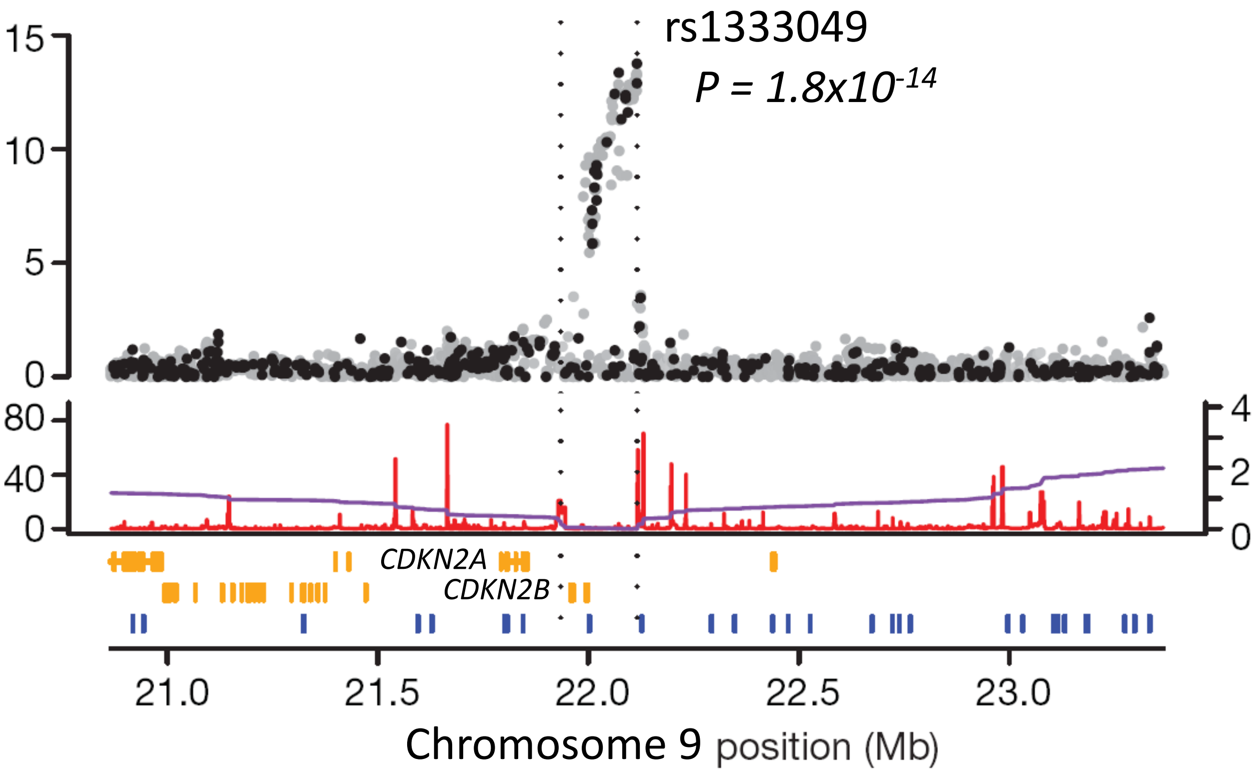
Different diseases have different architectures:



"Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls"
The Wellcome Trust Case Control Consortium Nature 447 (2007)

Wellcome Trust Case Control Consortium

Discovery of a common genetic variant that affects risk of coronary artery disease

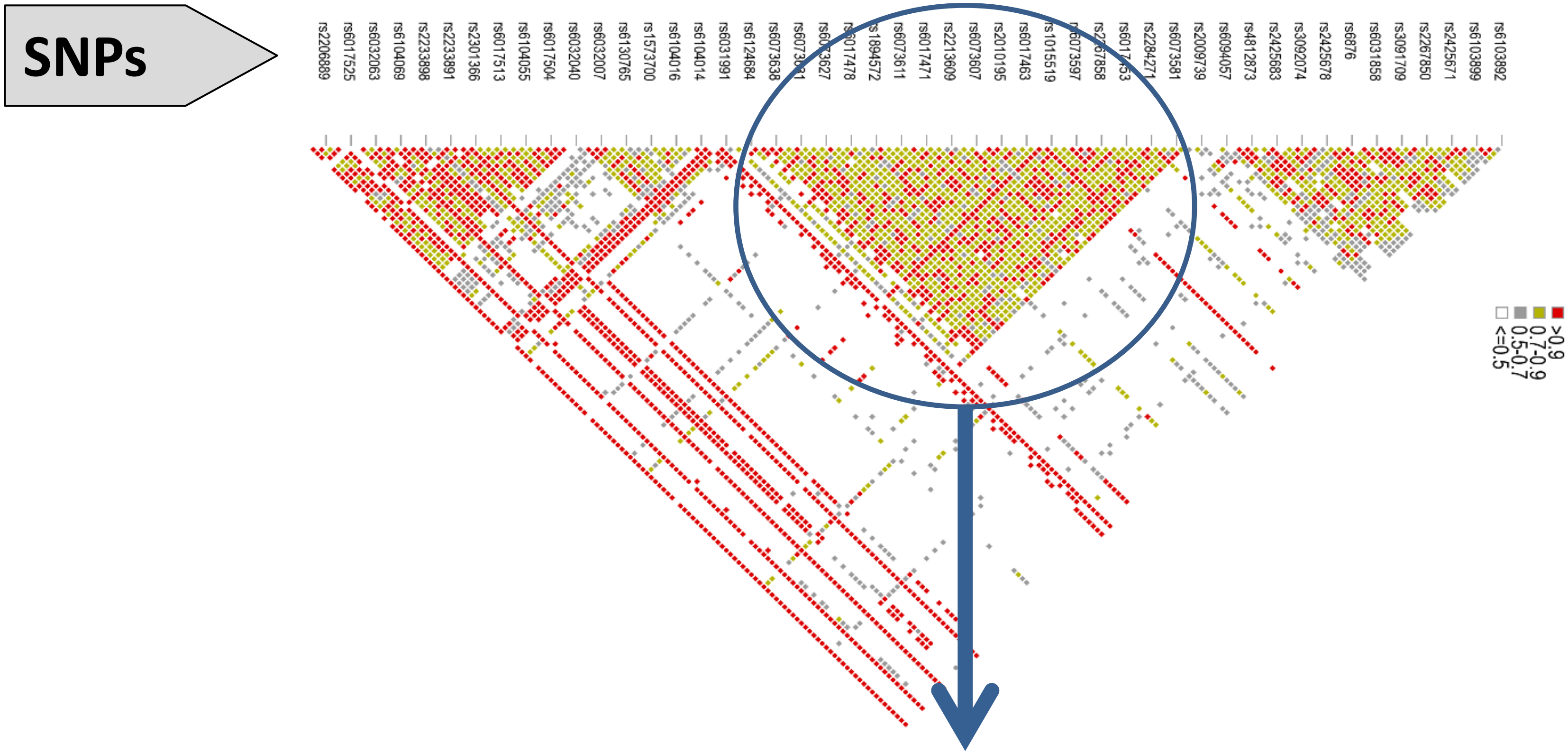


Best SNP marker was rs1333049

- OR ~ 1.47: one copy of the risk allele (present in half the population) increases "risk" of coronary artery disease by ~50%

- two copies of risk allele (present in quarter of population) almost doubles "risk" of coronary artery disease (OR 1.47 * 1.47)

Each population has a distinct pattern of genome variation



- Most SNPs are correlated with surrounding SNPs. This is known as **linkage disequilibrium**

- Linkage disequilibrium reflects the common combinations of variants (haplotypes) that exist in the population

GWAS in Africa

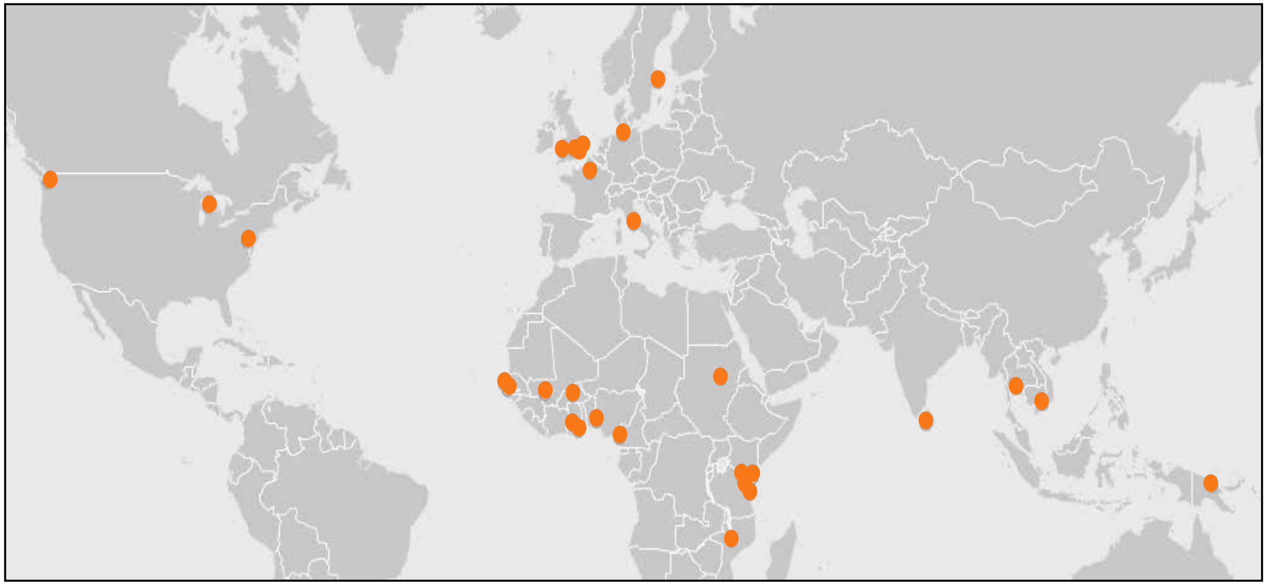
A number of factors make GWAS particularly challenging in Africa.

- Genome diversity much higher in African than other populations – more SNPs, more structure, more haplotypes.
- Low levels of LD…
- …and differences in LD between populations means power to detect untyped causal loci is reduced.
- A unique burden of infectious disease - the full story might involve two or more genomes at once!

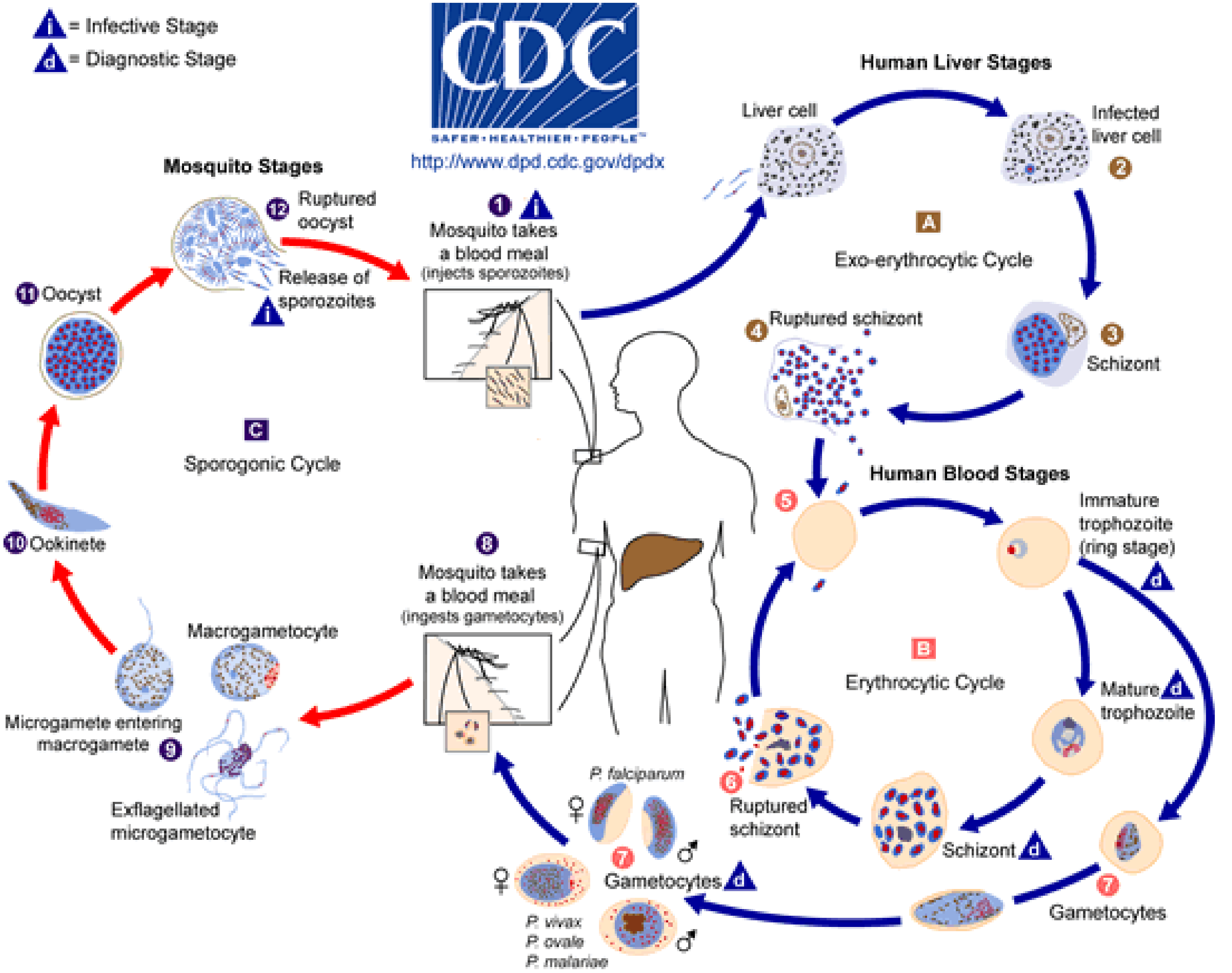
Malaria Genomic Epidemiology Network

MalariaGEN

www.malariagen.net



- **Investigators in 16 malaria endemic countries:** Burkina Faso, Cambodia, Cameroon, Gambia, Ghana, Ghana, Kenya, Malawi, Mali, Nigeria, Papua New Guinea, Senegal, Sudan, Tanzania, Thailand, Vietnam.
- **…and 6 non-endemic countries:** France, Germany, Italy, Sweden, UK, USA
- Building a resource of DNA and clinical data from ~100,000 subjects


i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Human Liver Stages
Liver cell
Infected liver cell
2
A
Exo-erythrocytic Cycle
3 Schizont
4 Ruptured schizont
Mosquito Stages
12 Ruptured oocyst
Release of sporozoites
i
11 Oocyst
1 i
Mosquito takes a blood meal
(injects sporozoites)
C
Sporogonic Cycle
10 Ookinete
8
Mosquito takes a blood meal
(ingests gametocytes)
Macrogametocyte
Microgamete entering macrogamete
9
Exflagellated microgametocyte
Human Blood Stages
5
Immature trophozoite (ring stage)
d
B
Erythrocytic Cycle
Mature trophozoite
d
6
Ruptured schizont
Schizont
d
7
Gametocytes
P. falciparum
♀
♂
7
Gametocytes
d
♀
♂
P. vivax
P. ovale
P. malariae

Recruitment of 13,000 cases of severe malaria

Cases and controls from:

- Burkina Faso
- Cameroon
- Gambia
- Ghana (Navrongo)
- Ghana (Kumasi)
- Kenya
- Malawi
- Mali
- Nigeria
- Papua New Guinea
- Tanzania
- Vietnam

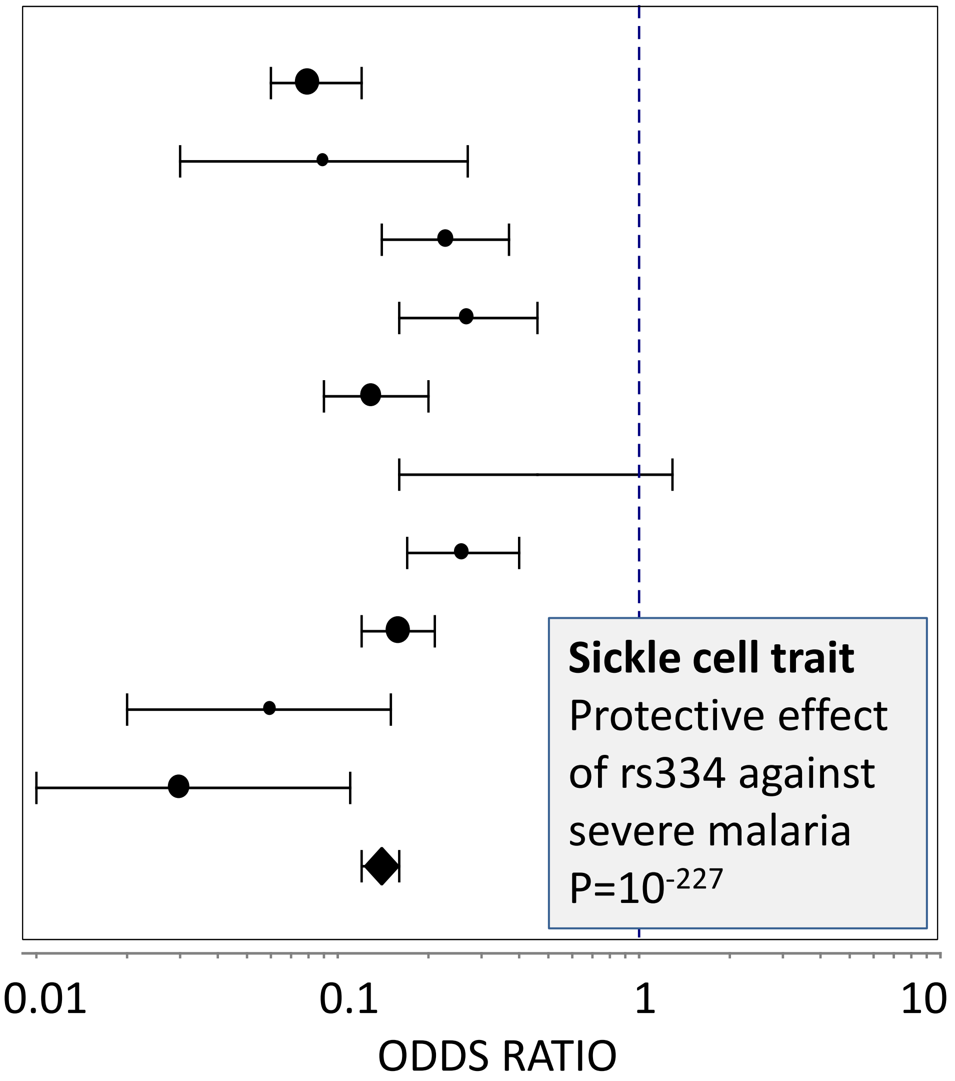
Question: In communities where every child is repeatedly infected with malaria, why do some children die and not others?

Consistent effects despite phenotypic heterogeneity

HbAS effect in severe malaria

Country	Cases (n/N)	Cntls (n/N)
Gambia	32/2542	460/3332
Mali	4/453	28/344
Burkina Faso	21/865	73/729
Ghana (Navrongo)	19/6820	50/484
Ghana (Kumasi)	32/1495	271/2042
Nigeria	9/77	9/40
Cameroon	32/621	99/576
Kenya	57/2261	594/3941
Tanzania	5/428	75/452
Malawi	2/1388	132/2696
All severe malaria	213/10685	1791/14641

Rockett *et al.* (2014) Nature Genetics 46: 1197



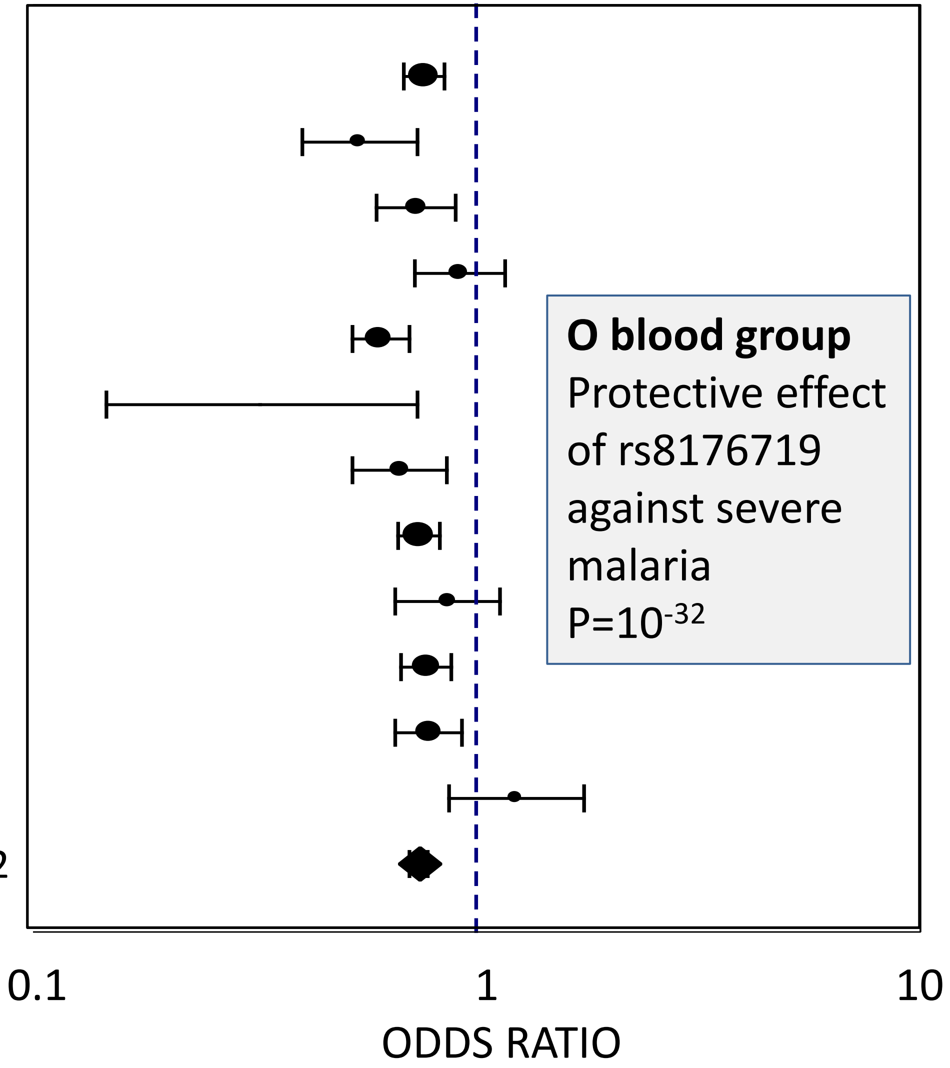
Sickle cell trait
Protective effect of rs334 against severe malaria
$P=10^{-227}$

Consistent effects despite phenotypic heterogeneity

O blood group effect in severe malaria

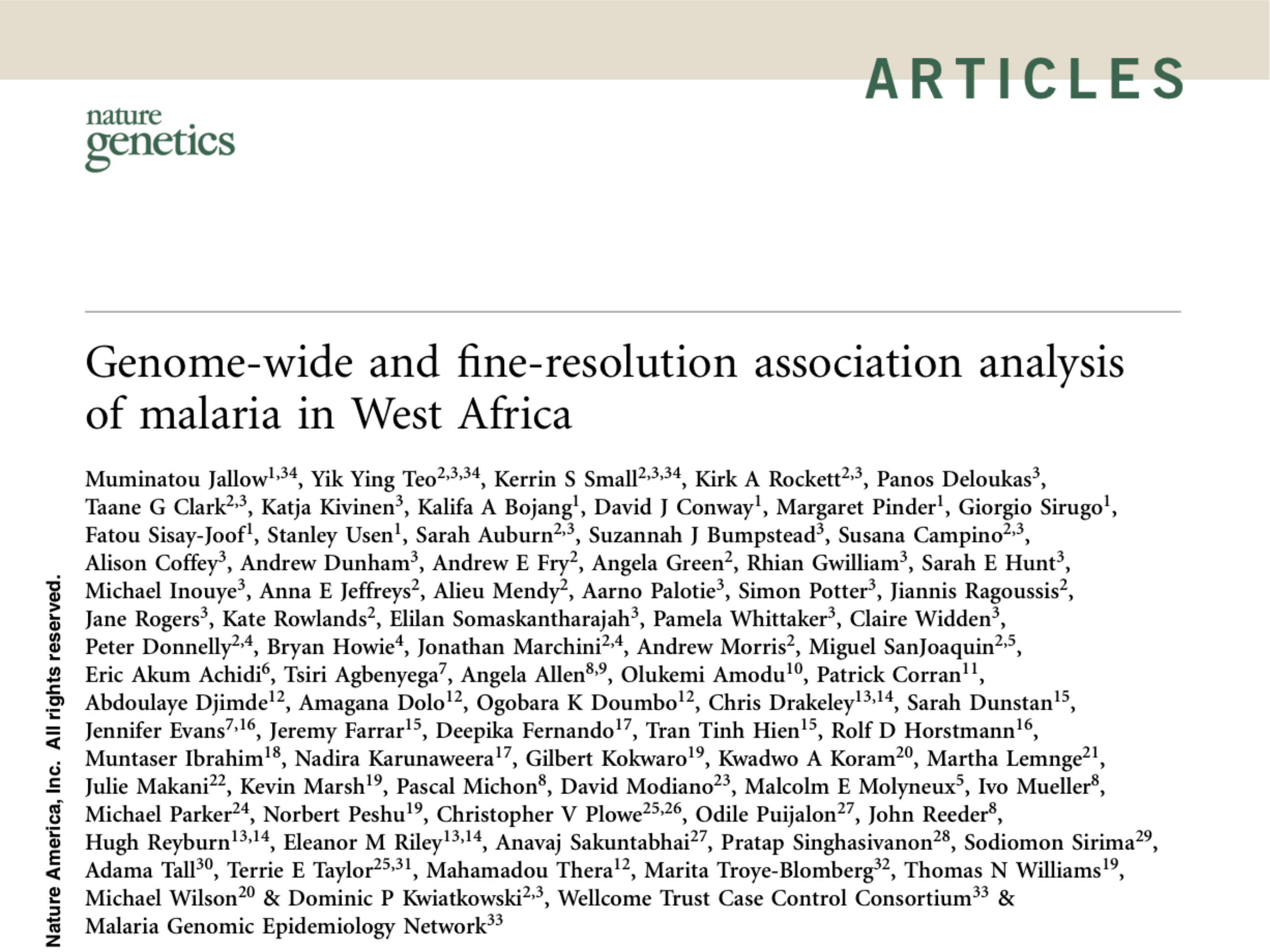
Rockett *et al.* (2014) Nature Genetics 46: 1197

Country	Cases (O/total)	Cntls (O/total)
Gambia	1000/2345	1664/3624
Mali	130/445	143/336
Burkina Faso	321/854	326/729
Ghana (Navrongo)	263/674	227/556
Ghana (Kumasi)	548/1480	992/1988
Nigeria	27/78	24/40
Cameroon	267/608	312/572
Kenya	1061/2254	2131/3899
Tanzania	189/423	221/455
Malawi	615/1414	1298/2607
Vietnam	272/788	1000/2517
Papua New Guinea	139/385	76/239
All severe malaria	4832/11948	8414/17652



O blood group
Protective effect of rs8176719 against severe malaria
$P=10^{-32}$

Attempt #1: GWAS of Severe Malaria in Gambia (2009)



ARTICLES

nature genetics

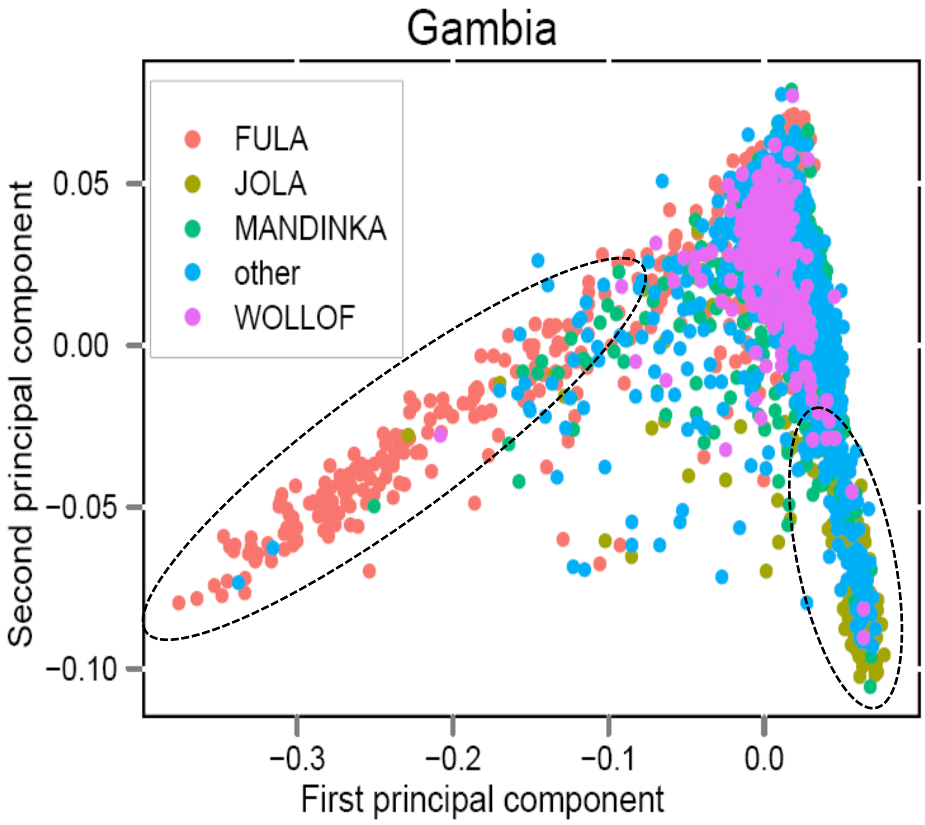
Genome-wide and fine-resolution association analysis of malaria in West Africa

Muminatou Jallow[1,34], Yik Ying Teo[2,3,34], Kerrin S Small[2,3,34], Kirk A Rockett[2,3], Panos Deloukas[3], Taane G Clark[2,3], Katja Kivinen[3], Kalifa A Bojang[1], David J Conway[1], Margaret Pinder[1], Giorgio Sirugo[1], Fatou Sisay-Joof[1], Stanley Usen[1], Sarah Auburn[2,3], Suzannah J Bumpstead[3], Susana Campino[2,3], Alison Coffey[3], Andrew Dunham[3], Andrew E Fry[2], Angela Green[2], Rhian Gwilliam[3], Sarah E Hunt[3], Michael Inouye[3], Anna E Jeffreys[2], Alieu Mendy[2], Aarno Palotie[3], Simon Potter[3], Jiannis Ragoussis[2], Jane Rogers[3], Kate Rowlands[2], Elilan Somaskantharajah[3], Pamela Whittaker[3], Claire Widden[3], Peter Donnelly[2,4], Bryan Howie[4], Jonathan Marchini[2,4], Andrew Morris[2], Miguel SanJoaquin[2,5], Eric Akum Achidi[6], Tsiri Agbenyega[7], Angela Allen[8,9], Olukemi Amodu[10], Patrick Corran[11], Abdoulaye Djimde[12], Amagana Dolo[12], Ogobara K Doumbo[12], Chris Drakeley[13,14], Sarah Dunstan[15], Jennifer Evans[7,16], Jeremy Farrar[15], Deepika Fernando[17], Tran Tinh Hien[15], Rolf D Horstmann[16], Muntaser Ibrahim[18], Nadira Karunaweera[17], Gilbert Kokwaro[19], Kwadwo A Koram[20], Martha Lemnge[21], Julie Makani[22], Kevin Marsh[19], Pascal Michon[8], David Modiano[23], Malcolm E Molyneux[5], Ivo Mueller[8], Michael Parker[24], Norbert Peshu[19], Christopher V Plowe[25,26], Odile Puijalon[27], John Reeder[8], Hugh Reyburn[13,14], Eleanor M Riley[13,14], Anavaj Sakuntabhai[27], Pratap Singhasivanon[28], Sodiomon Sirima[29], Adama Tall[30], Terrie E Taylor[25,31], Mahamadou Thera[12], Marita Troye-Blomberg[32], Thomas N Williams[19], Michael Wilson[20] & Dominic P Kwiatkowski[2,3], Wellcome Trust Case Control Consortium[33] & Malaria Genomic Epidemiology Network[33]

MalariaGEN wellcometrust

Importance of population structure

Principal components analysis

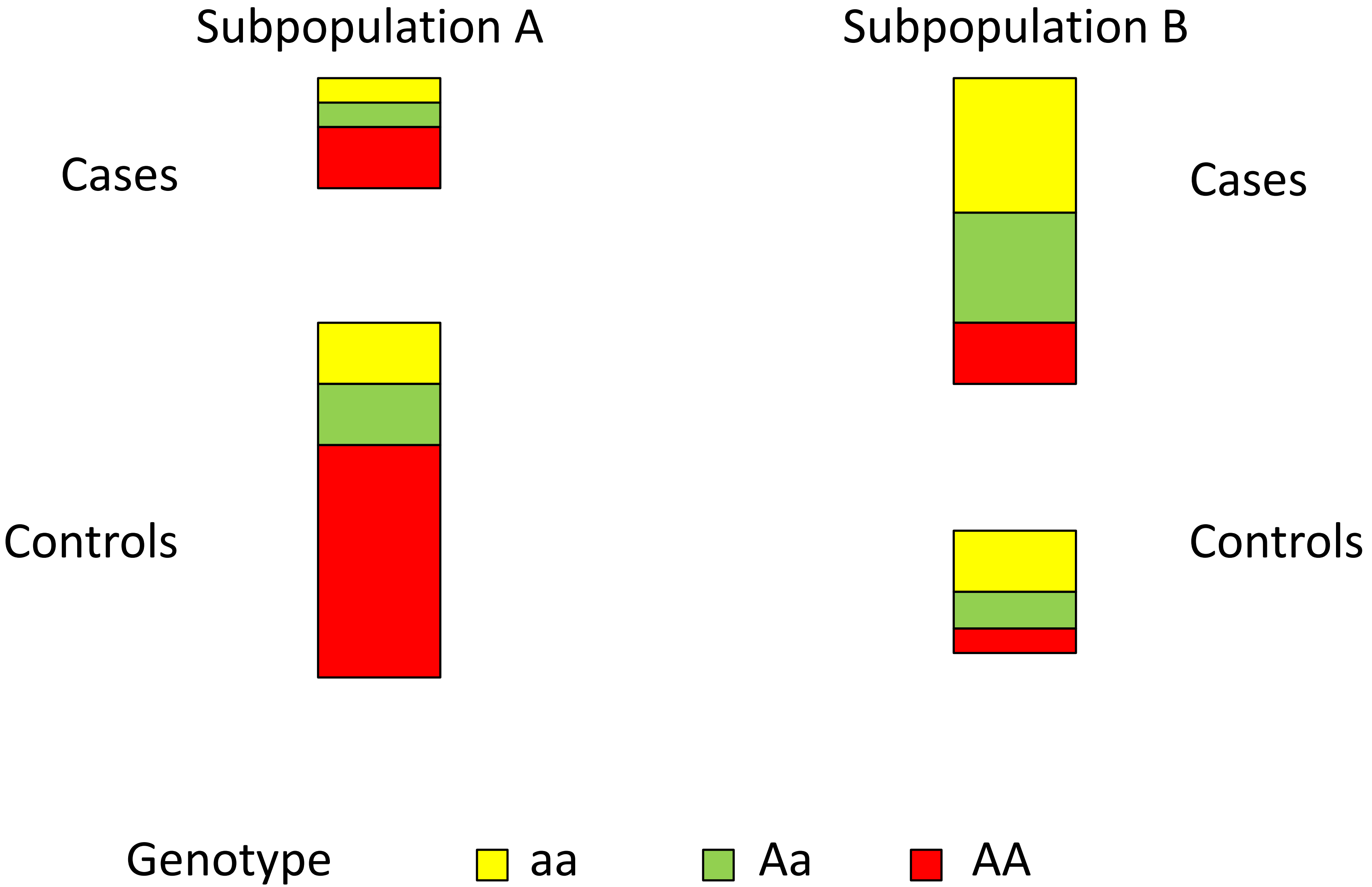


- Within a 40 sq mile area of The Gambia we find complex population structure
- Population structure can give rise to false positive genetic associations

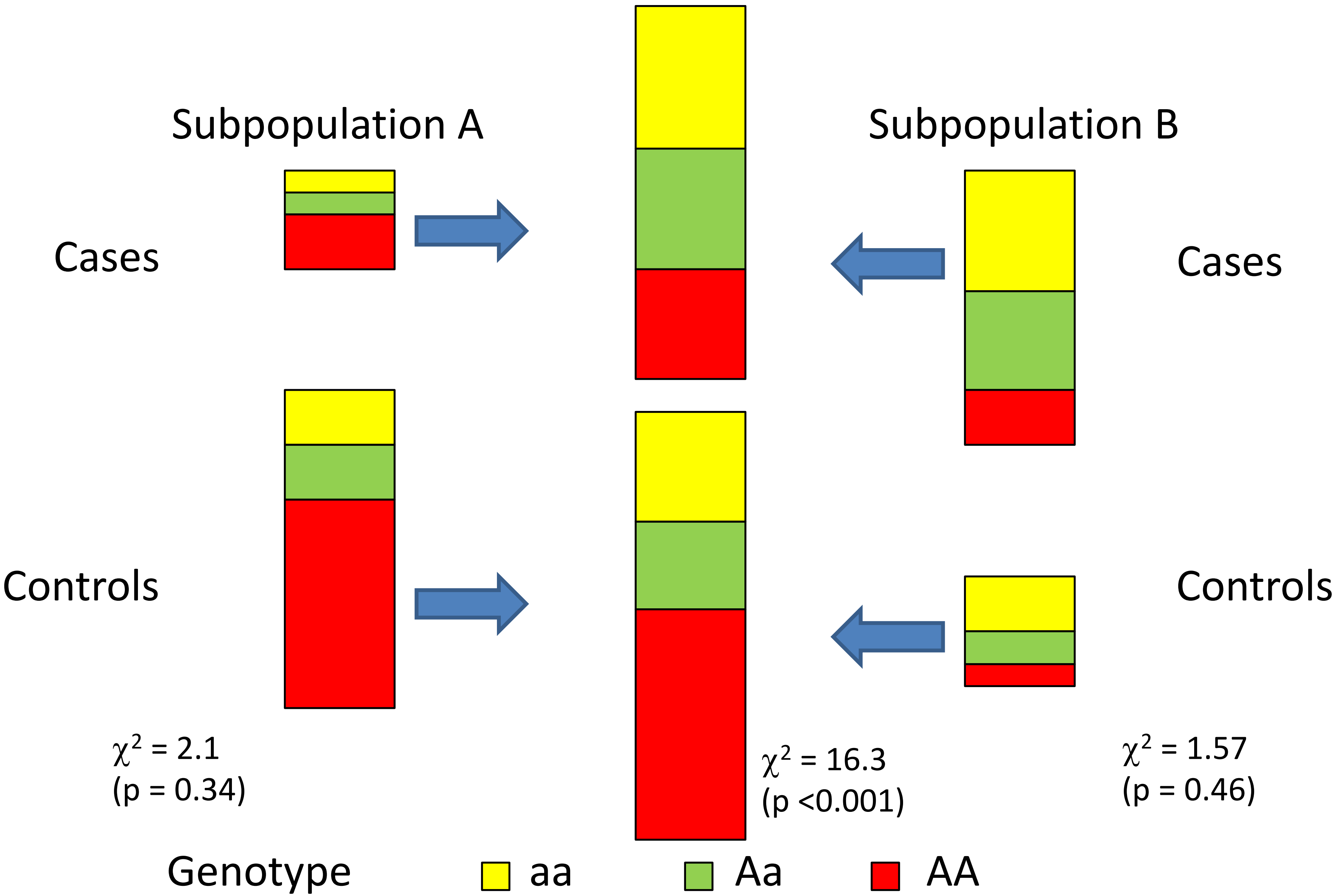
MalariaGEN wellcometrust

Jallow *et al.* (2009) Nature Genetics 41: 657

Importance of population structure



Importance of population structure



Subpopulation A

Subpopulation B

Cases

Cases

Controls

Controls

$\chi^2 = 2.1$
(p = 0.34)

$\chi^2 = 16.3$
(p <0.001)

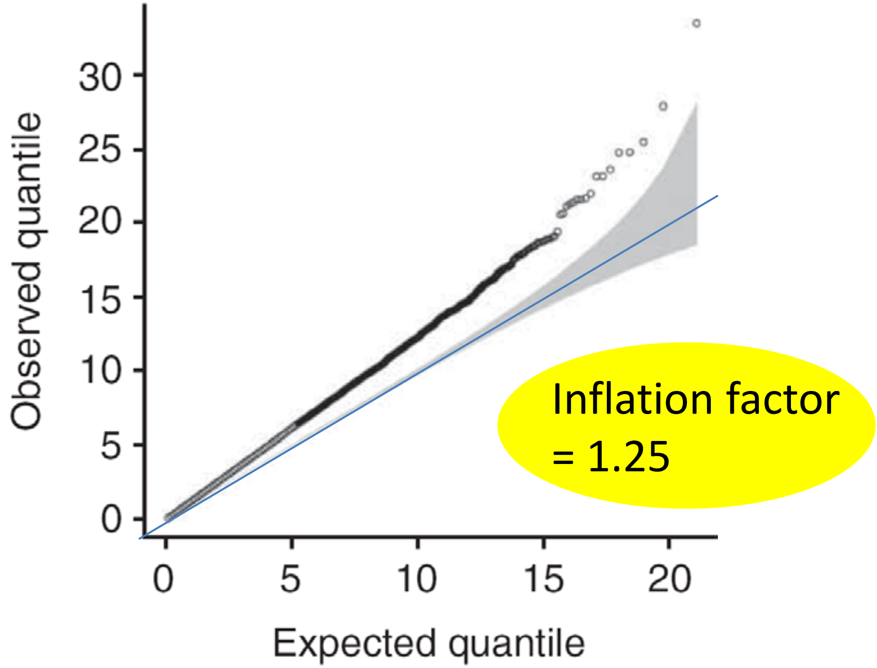
$\chi^2 = 1.57$
(p = 0.46)

Genotype 🟨 aa 🟩 Aa 🟥 AA

Importance of population structure

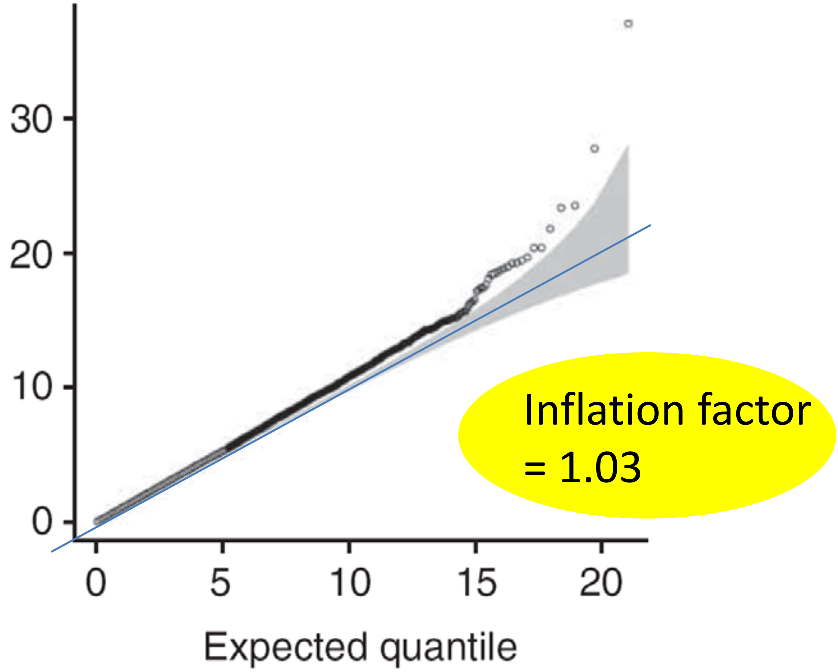
Quantile-quantile plot of chi-squared statistic comparing what we observed *versus what we'd expect if no disease association*

Uncorrected



Inflation factor = 1.25

Corrected by principal components analysis



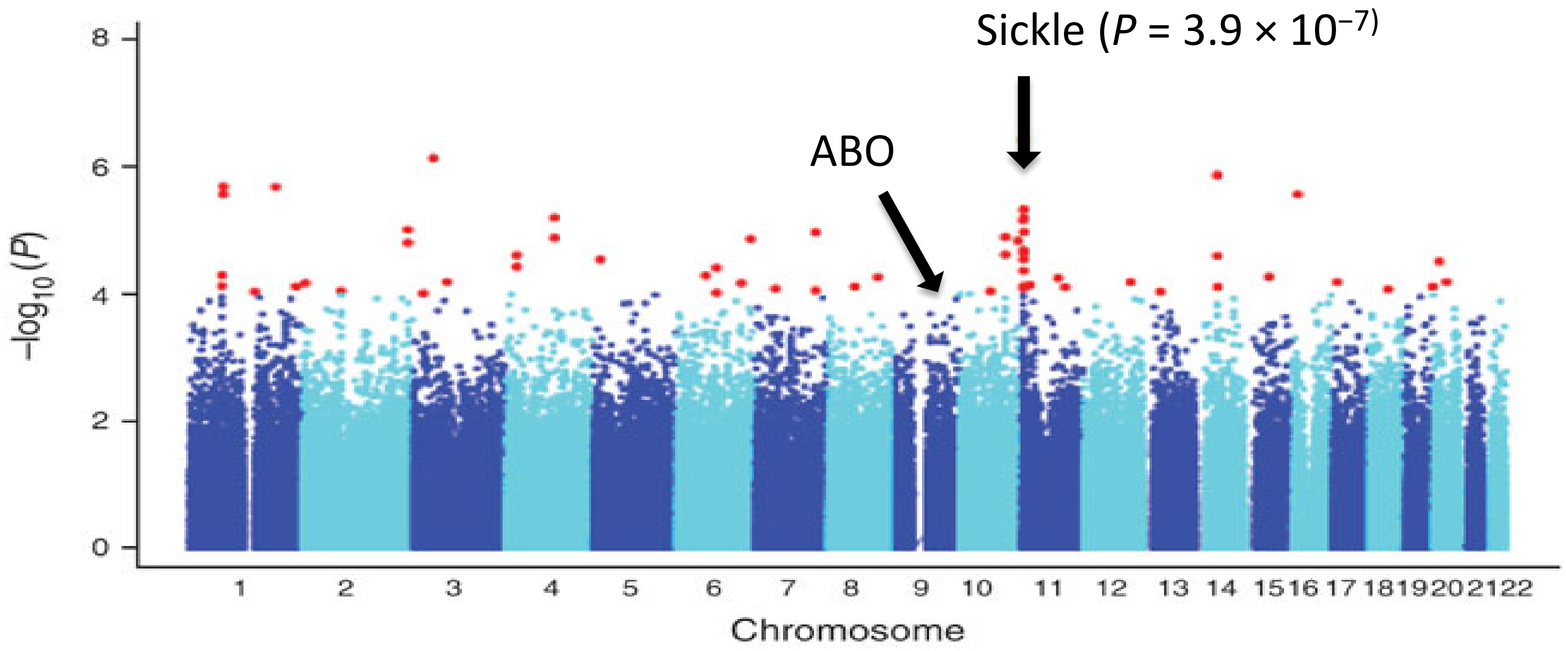
Inflation factor = 1.03

MalariaGEN wellcome trust

Jallow *et al.* (2009) Nature Genetics 41: 657

GWA studies of severe malaria
Study of 500,000 SNPs in 2,500 Gambian children

Jallow *et al.* (2009) Nature Genetics 41: 657



Sickle ($P = 3.9 \times 10^{-7}$)

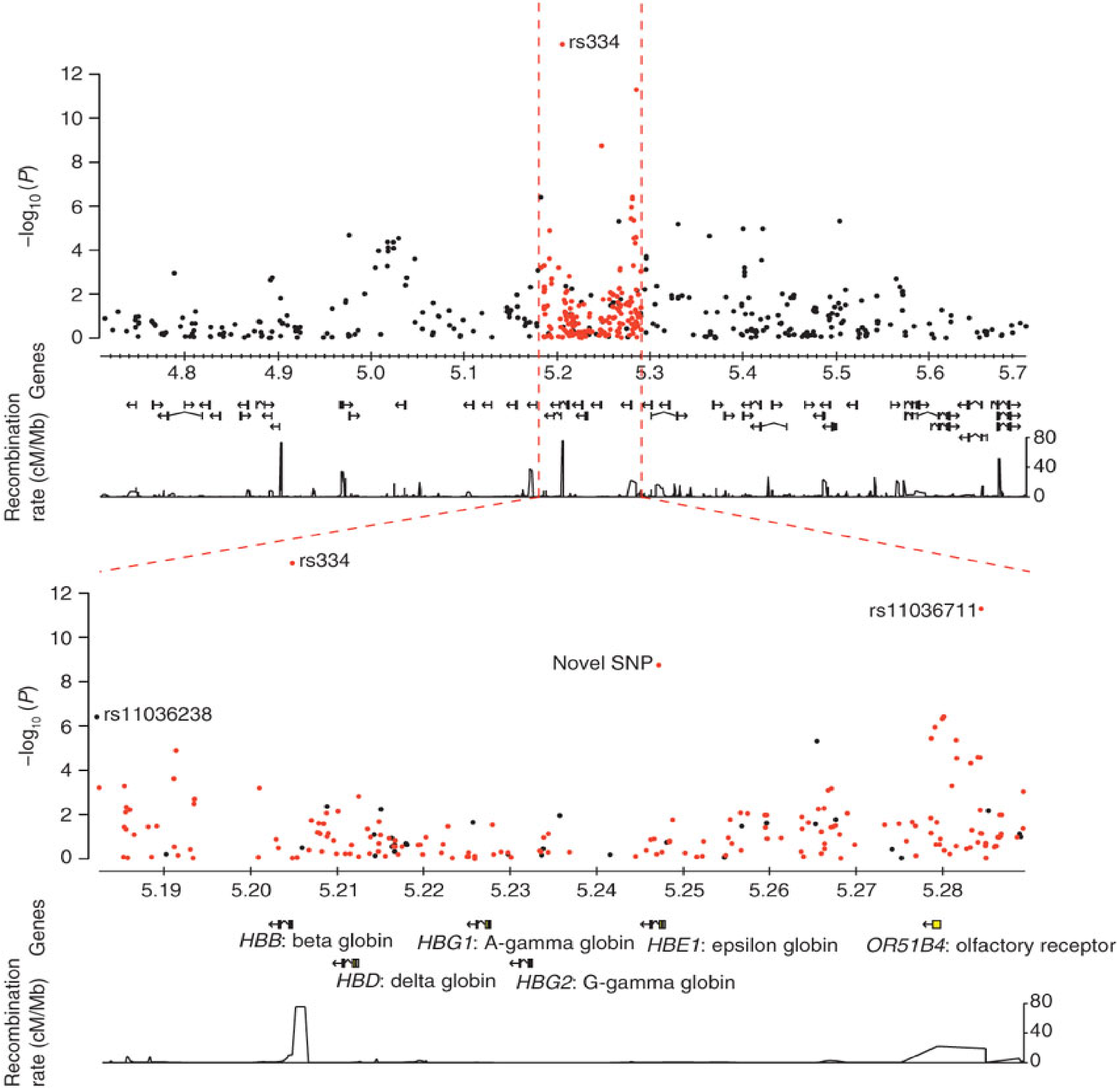
ABO

Low LD acts to attenuate GWA signals of association

- HbS signal is *$P=4x10^{-7}$ (causal variant $P=10^{-28}$)*
- No signal at *ABO*

MalariaGEN wellcome trust

Targetted resequencing

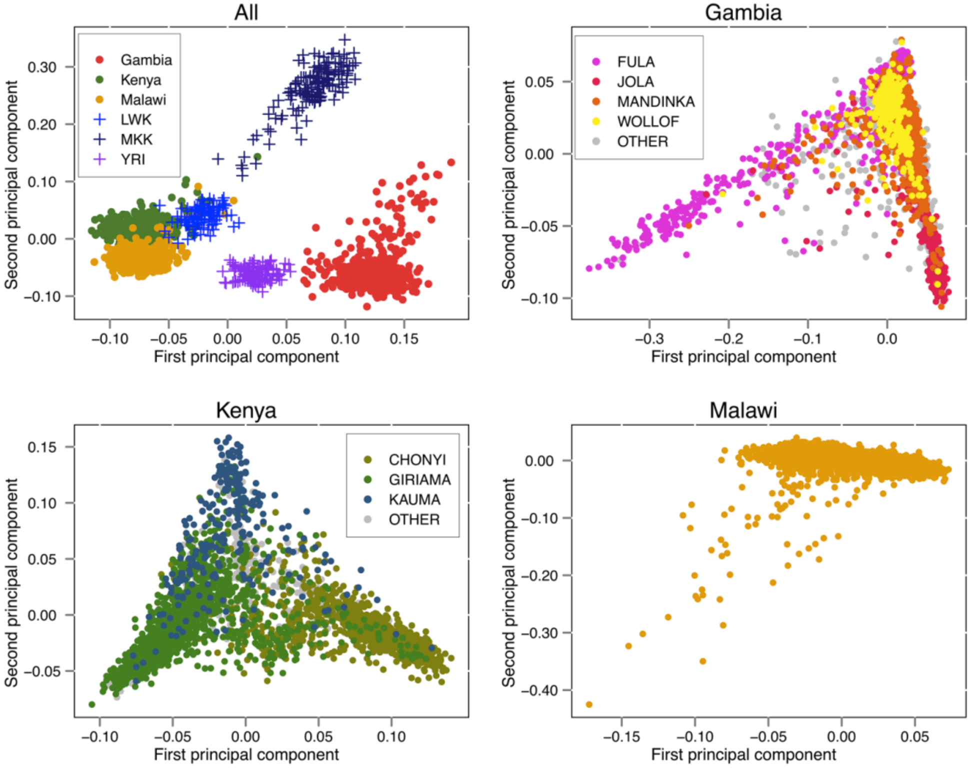


MalariaGEN

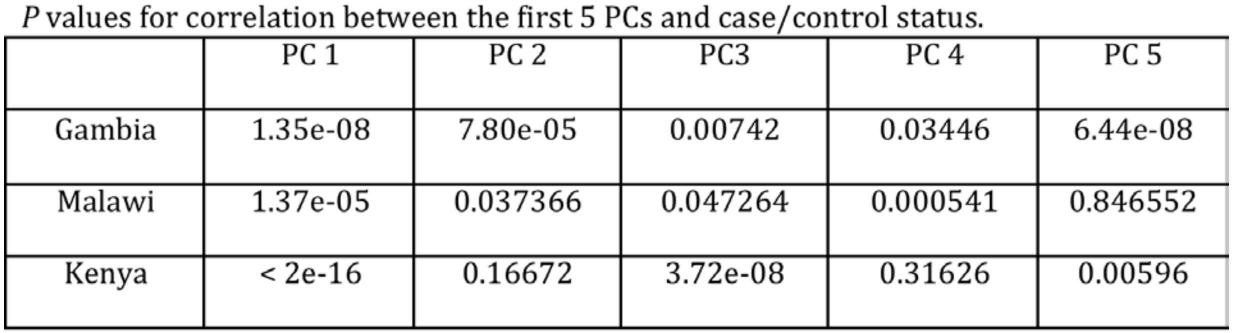
Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).

- 5,000 cases and 7,000 controls from Gambia, Kenya and Malawi.
- Imputed to ~1.3M variants from the publically available HapMap reference panel.
- Novel methods to allow for heterogeneity and differences in haplotype background: heterogeneity Bayes factors, and region-based tests that take into account all variants in each region.

Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).



Control for the extensive structure using a mixed model that takes into account relatedness at all levels. (PCs also used for comparison with similar results.)



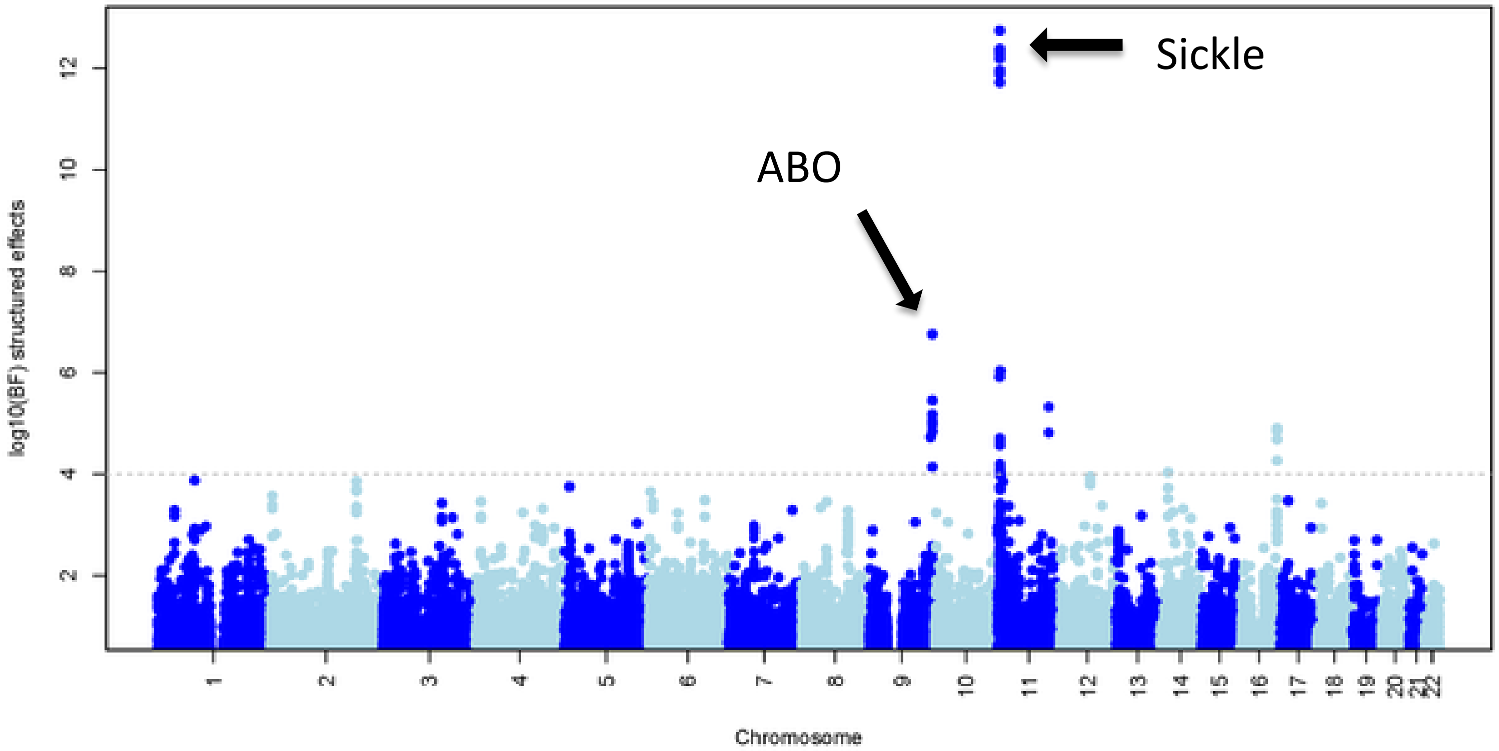
P values for correlation between the first 5 PCs and case/control status.

	PC 1	PC 2	PC3	PC 4	PC 5
Gambia	1.35e-08	7.80e-05	0.00742	0.03446	6.44e-08
Malawi	1.37e-05	0.037366	0.047264	0.000541	0.846552
Kenya	< 2e-16	0.16672	3.72e-08	0.31626	0.00596

"*Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations*", Band G, et al. PLoS Genetics (2013)

MalariaGEN

Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).

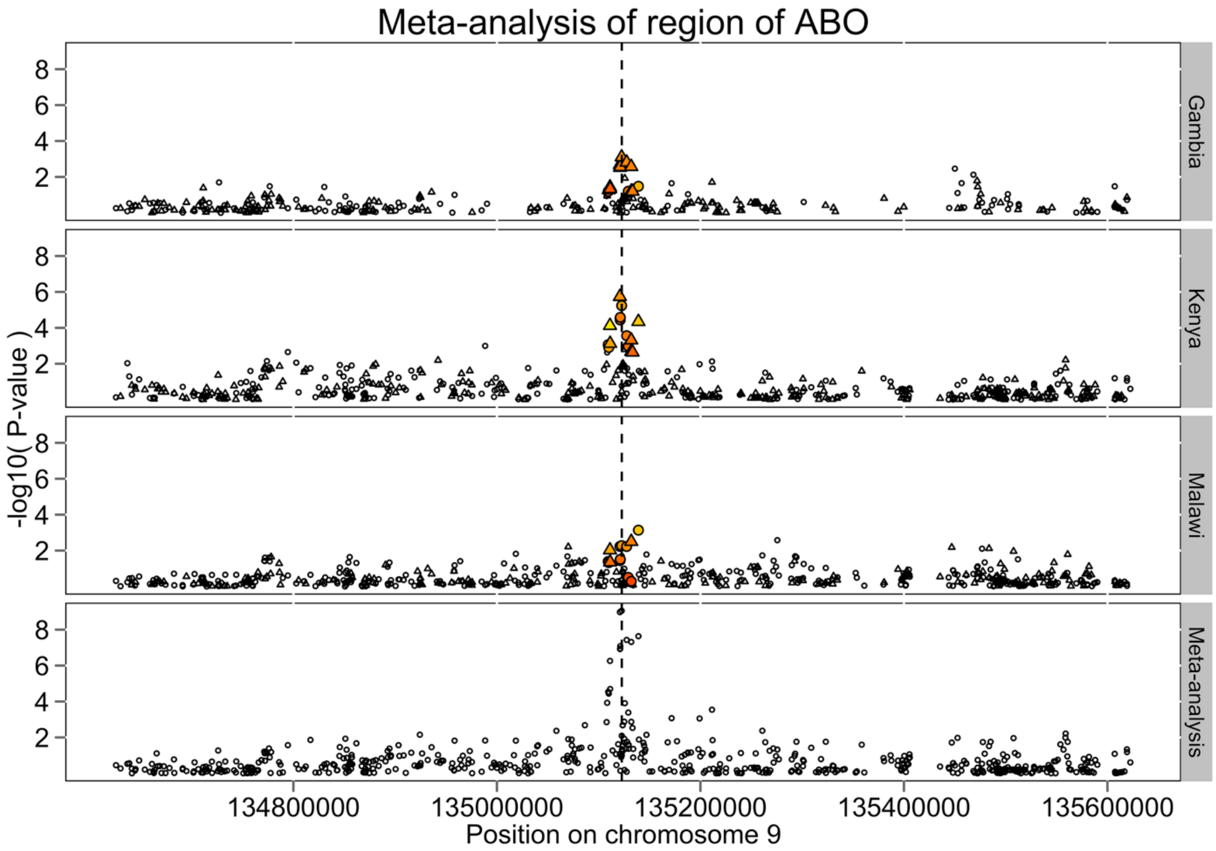


5000 cases and 7000 controls from Gambia, Kenya and Malawi.
Use of imputation into publically available reference set (HapMap) to assess association at 1.3M variants.

"*Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations*", Band G, et al. PLoS Genetics (2013)

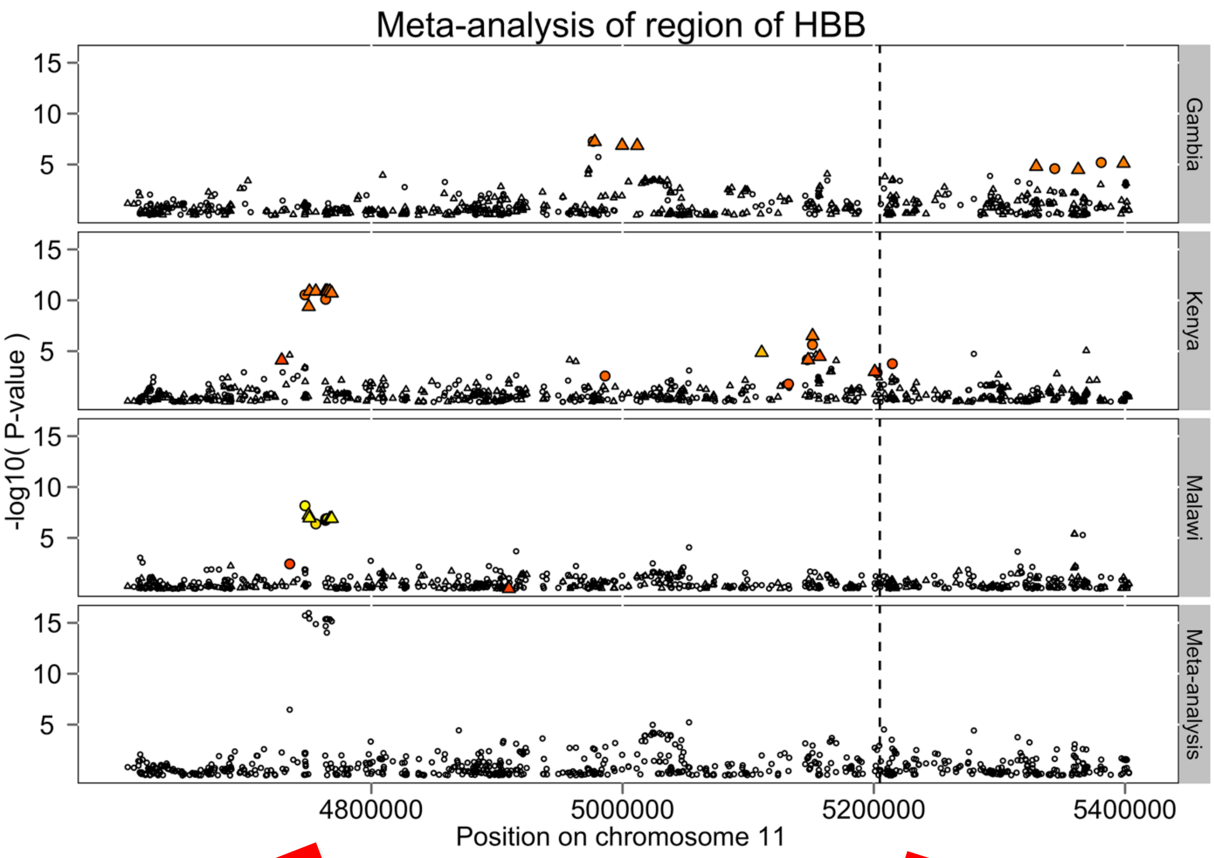
MalariaGEN

Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).



MalariaGEN

Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).



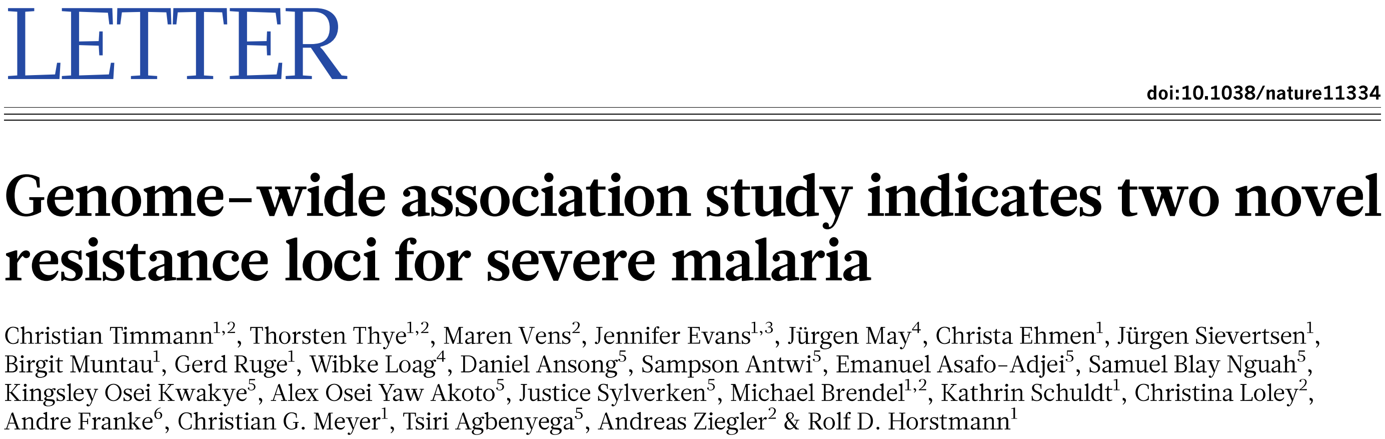
Where we see the most signal

Where sickle is

MalariaGEN

Attempt #2: GWAS of severe malaria in three African populations (Gambia, Kenya and Malawi) (2013).

Region	Chromosome	Regional test Bayes factor	
OR51F1 (HBB region)	11	> 10^{11}	Sickle Signal
ABO	9	4920	O blood group signal
BET1L	11	319	
C10orf57	10	243	
MYOT	5	112	
SMARCA5	4	110	
ATP2B4	1	103	Red cell calcium channel



LETTER

doi:10.1038/nature11334

Genome-wide association study indicates two novel resistance loci for severe malaria

Christian Timmann[1,2], Thorsten Thye[1,2], Maren Vens[2], Jennifer Evans[1,3], Jürgen May[4], Christa Ehmen[1], Jürgen Sievertsen[1], Birgit Muntau[1], Gerd Ruge[1], Wibke Loag[4], Daniel Ansong[5], Sampson Antwi[5], Emanuel Asafo-Adjei[5], Samuel Blay Nguah[5], Kingsley Osei Kwakye[5], Alex Osei Yaw Akoto[5], Justice Sylverken[5], Michael Brendel[1,2], Kathrin Schuldt[1], Christina Loley[2], Andre Franke[6], Christian G. Meyer[1], Tsiri Agbenyega[5], Andreas Ziegler[2] & Rolf D. Horstmann[1]

Attempt #3 (2015?): GWAS of severe malaria in eight populations in sub-Saharan Africa

- Approx. 10,000 cases and 10,000 controls (across 11 countries).

- Typed at 2.5M variants and imputed up to 40M variants from the phase 3 1000 Genomes reference panel.

- Starting to find new loci. Some evidence that there are rarer, bigger effects around, differing between populations.

- Data is being made publically available – we have an ongoing effort to develop web-based tools for data sharing.

GWAS Summary

- Power to detect association depends on sample size, effect size, frequency, and density of markers. Bigger is better!

- Careful QC and control for confounding factors is essential.

- High diversity and patterns of LD make GWAS in Africa particularly challenging.

GWAS : the hare and the tortoise?

	Europe	Africa
Level of LD	high	low
Variability of LD	low	high
Finding signals of association by genome-wide SNP typing	easy	difficult
Localising causal variants by genome sequencing	**difficult**	**?easy**

Next-generation sequencing will transform genome-wide association analysis

In the near term

- The 1000 Genomes Project is including 2 MalariaGEN study sites (Gambia, Vietnam) in addition to at least 6 other African populations.
- Other groups working to create Africa-specific reference panels (e.g. AGVP, H3Africa).
- By combining GWAS data with population-specific sequence data, we can **boost** signals of association and **localise** causal variants.

In the longer term

- GWAS-by-sequencing will replace GWAS-by-SNP-typing.
- This will particularly benefit studies in Africa and multiethnic studies.

What’s next?

As a warm-up for a full GWAS analysis later in the week, the next practical shows you how to perform association analyses on individual SNPs using R. (Based on MalariaGEN data.)