Introduction to genetic association studies in Africa

Dr Kirk Rockett
Introductions

Public databases and resources for genetics

Basic genotype data summaries and analyses

GWAS QC

GWAS association analyses

GWAS results and interpretation

Epidemiology

Basic principles of measuring disease in populations

population genetics

Principal components analyses

Bioinformatics

Genetics

Public databases and resources for genetics

meta-analysis and power of genetic studies

whole genome sequencing and fine-mapping

Basic genotype data summaries and analyses
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

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
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)

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

- Gene A → Gene B → Gene C

Duplications:

- Gene A → Gene B → Gene B → Gene C

Inversions:

- Gene A → Gene B ← Gene B → Gene C

Complex rearrangements:

- Gene A → Gene C ← Gene B ← Gene B
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

- 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

Effect size (OR)

Allele frequency
Variation in resistance & susceptibility to disease

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

GWAS studies

Effect size (OR)

Big effects

5.0

Intermediate effects

4.0

3.0

2.0

1.0

0.0

0.1

0.2

0.3

0.4

0.5

Allele frequency

Rare variants

Common variants with large effects (presumably for specific traits).

Common variants with small effect

Rare, small effects (really hard to find)

Highly penetrant Mendelian mutations

GWAS studies
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?*

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

- Number of samples
- Effect size
- Allele frequency
- LD
Variation in resistance & susceptibility to disease

\[ \text{Power} \approx N \quad \beta^2 \quad f(1-f) \quad r^2 \]
Finding loci that influence disease

• Consider a position in the genome that shows variation between individuals, for example ...

\[
\begin{align*}
A & T G A C T C G T A \\
A & T G A C A C G T A
\end{align*}
\]

• 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)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>HbAA (normal)</th>
<th>HbAS (sickle trait)</th>
<th>HbSS (sickle cell disease)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe malaria cases</td>
<td>2700</td>
<td>35</td>
<td>13</td>
</tr>
<tr>
<td>Population</td>
<td>3689</td>
<td>588</td>
<td>22</td>
</tr>
</tbody>
</table>

\[ N = 7047 \]
\[ f = 0.07 \ (7\%) \]
Example: sickle and severe Malaria
Gambian data (MalariaGEN consortium)

<table>
<thead>
<tr>
<th></th>
<th>TT</th>
<th>AT</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe malaria cases</td>
<td>2700</td>
<td>35</td>
<td>13</td>
</tr>
<tr>
<td>Population</td>
<td>3689</td>
<td>588</td>
<td>22</td>
</tr>
</tbody>
</table>

Odds ratio = \( \frac{3689 \times 35}{2700 \times 588} = 0.08 \)

\[ P < 2 \times 10^{-16} \]

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.
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)
• 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!

...and 6 non-endemic countries: France, Germany, Italy, Sweden, UK, USA

Building a resource of DNA and clinical data from ~100,000 subjects
Recruitment of 13,000 cases of severe malaria

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

**Cases and controls from:**
- Burkina Faso
- Cameroon
- Gambia
- Ghana (Navrongo)
- Ghana (Kumasi)
- Kenya
- Malawi
- Mali
- Nigeria
- Papua New Guinea
- Tanzania
- Vietnam
**Consistent effects despite phenotypic heterogeneity**

### HbAS effect in severe malaria

<table>
<thead>
<tr>
<th>Country</th>
<th>Cases (n/N)</th>
<th>Cntls (n/N)</th>
<th>Rockett et al. (2014) Nature Genetics 46: 1197</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gambia</td>
<td>32/2542</td>
<td>460/3332</td>
<td></td>
</tr>
<tr>
<td>Mali</td>
<td>4/453</td>
<td>28/344</td>
<td></td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>21/865</td>
<td>73/729</td>
<td></td>
</tr>
<tr>
<td>Ghana (Navrongo)</td>
<td>19/6820</td>
<td>50/484</td>
<td></td>
</tr>
<tr>
<td>Ghana (Kumasi)</td>
<td>32/1495</td>
<td>271/2042</td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>9/77</td>
<td>9/40</td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>32/621</td>
<td>99/576</td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td>57/2261</td>
<td>594/3941</td>
<td></td>
</tr>
<tr>
<td>Tanzania</td>
<td>5/428</td>
<td>75/452</td>
<td></td>
</tr>
<tr>
<td>Malawi</td>
<td>2/1388</td>
<td>132/2696</td>
<td></td>
</tr>
<tr>
<td>All severe malaria</td>
<td>213/10685</td>
<td>1791/14641</td>
<td></td>
</tr>
</tbody>
</table>


**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

<table>
<thead>
<tr>
<th>Country</th>
<th>Cases (O/total)</th>
<th>Cntls (O/total)</th>
<th>Rockett et al. (2014) Nature Genetics 46: 1197</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gambia</td>
<td>1000/2345</td>
<td>1664/3624</td>
<td></td>
</tr>
<tr>
<td>Mali</td>
<td>130/445</td>
<td>143/336</td>
<td></td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>321/854</td>
<td>326/729</td>
<td></td>
</tr>
<tr>
<td>Ghana (Navrongo)</td>
<td>263/674</td>
<td>227/556</td>
<td></td>
</tr>
<tr>
<td>Ghana (Kumasi)</td>
<td>548/1480</td>
<td>992/1988</td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>27/78</td>
<td>24/40</td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>267/608</td>
<td>312/572</td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td>1061/2254</td>
<td>2131/3899</td>
<td></td>
</tr>
<tr>
<td>Tanzania</td>
<td>189/423</td>
<td>221/455</td>
<td></td>
</tr>
<tr>
<td>Malawi</td>
<td>615/1414</td>
<td>1298/2607</td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td>272/788</td>
<td>1000/2517</td>
<td></td>
</tr>
<tr>
<td>Papua New Guinea</td>
<td>139/385</td>
<td>76/239</td>
<td></td>
</tr>
<tr>
<td>All severe malaria</td>
<td>4832/11948</td>
<td>8414/17652</td>
<td></td>
</tr>
</tbody>
</table>

O blood group
Protective effect of rs8176719 against severe malaria
P=10^{-32}
Attempt #1: GWAS of Severe Malaria in Gambia (2009)

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

Within a 40 sq mile area of The Gambia we find complex population structure.

Population structure can give rise to false positive genetic associations.
Importance of population structure

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td><img src="image1" alt="Aa" /></td>
<td><img src="image2" alt="Aa" /></td>
<td><img src="image3" alt="aa" /></td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td><img src="image4" alt="Aa" /></td>
<td><img src="image5" alt="Aa" /></td>
<td><img src="image6" alt="aa" /></td>
</tr>
</tbody>
</table>
Importance of population structure

<table>
<thead>
<tr>
<th>Subpopulation A</th>
<th>Subpopulation B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td><strong>Cases</strong></td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td><strong>Controls</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>aa</th>
<th>Aa</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td><img src="yellow" alt="aa" />, <img src="green" alt="Aa" />, <img src="red" alt="AA" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="green" alt="χ² = 2.1" /> (p = 0.34)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td><img src="yellow" alt="aa" />, <img src="green" alt="Aa" />, <img src="red" alt="AA" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="green" alt="χ² = 16.3" /> (p &lt; 0.001)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>aa</th>
<th>Aa</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td><img src="yellow" alt="aa" />, <img src="green" alt="Aa" />, <img src="red" alt="AA" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="green" alt="χ² = 1.57" /> (p = 0.46)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td><img src="yellow" alt="aa" />, <img src="green" alt="Aa" />, <img src="red" alt="AA" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="green" alt="χ² = 2.1" /> (p = 0.34)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Corrected by principal components analysis

Inflation factor = 1.25

Inflation factor = 1.03

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

Low LD acts to attenuate GWA signals of association

- HbS signal is $P = 4 \times 10^{-7}$ (causal variant $P = 10^{-28}$)
- No signal at $ABO$
Targetted resequencing

---

Data and references for the targetted resequencing study are shown in the diagram. Key features include:

- **Gene Regions**: Annotation of genes such as $HBB$, $HBG1$, $HBE1$, $HBD$, and $HBG2$
- **SNP Marks**: Indication of specific SNPs like rs334, rs11036711, and rs11036238
- **Recombination Rate**: Graphs depicting recombination rate with $-\log_{10}(P)$ on the y-axis

---

*Note: This diagram is an example of how targetted resequencing can be applied in genomics research, focusing on identifying and analyzing genetic variations.*

- 5,000 cases and 7,000 controls from Gambia, Kenya and Malawi.

- Imputed to ~1.3M variants from the publicly 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.

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.

<table>
<thead>
<tr>
<th></th>
<th>PC 1</th>
<th>PC 2</th>
<th>PC 3</th>
<th>PC 4</th>
<th>PC 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gambia</td>
<td>1.35e-08</td>
<td>7.80e-05</td>
<td>0.00742</td>
<td>0.03446</td>
<td>6.44e-08</td>
</tr>
<tr>
<td>Malawi</td>
<td>1.37e-05</td>
<td>0.037366</td>
<td>0.047264</td>
<td>0.00541</td>
<td>0.846552</td>
</tr>
<tr>
<td>Kenya</td>
<td>&lt; 2e-16</td>
<td>0.16672</td>
<td>3.72e-08</td>
<td>0.31626</td>
<td>0.00596</td>
</tr>
</tbody>
</table>


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.


Where we see the most signal

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

<table>
<thead>
<tr>
<th>Region</th>
<th>Chromosome</th>
<th>Regional test Bayes factor</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR51F1 (HBB region)</td>
<td>11</td>
<td>&gt; 10^{11}</td>
<td>Sickle Signal</td>
</tr>
<tr>
<td>ABO</td>
<td>9</td>
<td>4920</td>
<td>O blood group signal</td>
</tr>
<tr>
<td>BET1L</td>
<td>11</td>
<td>319</td>
<td></td>
</tr>
<tr>
<td>C10orf57</td>
<td>10</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>MYOT</td>
<td>5</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>SMARCA5</td>
<td>4</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>ATP2B4</td>
<td>1</td>
<td>103</td>
<td>Red cell calcium channel</td>
</tr>
</tbody>
</table>

---

**LETTER**

**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^{1}, Daniel Ansong^{3}, Sampson Antwi^{3}, Emanuel Asafo-Adjei^{2}, Samuel Blay Nguah^{3}, Kingsley Osei Kwakye^{5}, Alex Osei Yaw Akoto^{5}, Justice Sylverken^{5}, Michael Brendel^{1,2}, Kathrin Schuldt^{1}, Christina Loley^{5}, Andre Franke^{5}, Christian G. Meyer^{3}, Tsiri Agbenyega^{5}, Andreas Ziegler^{2} & Rolf D. Horstmann^{1}

*doi:10.1038/nature11334*
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.
<table>
<thead>
<tr>
<th></th>
<th>Europe</th>
<th>Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of LD</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Variability of LD</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Finding signals of association by genome-wide SNP typing</td>
<td>easy</td>
<td>difficult</td>
</tr>
<tr>
<td>Localising causal variants by genome sequencing</td>
<td>difficult</td>
<td>?easy</td>
</tr>
</tbody>
</table>
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.)