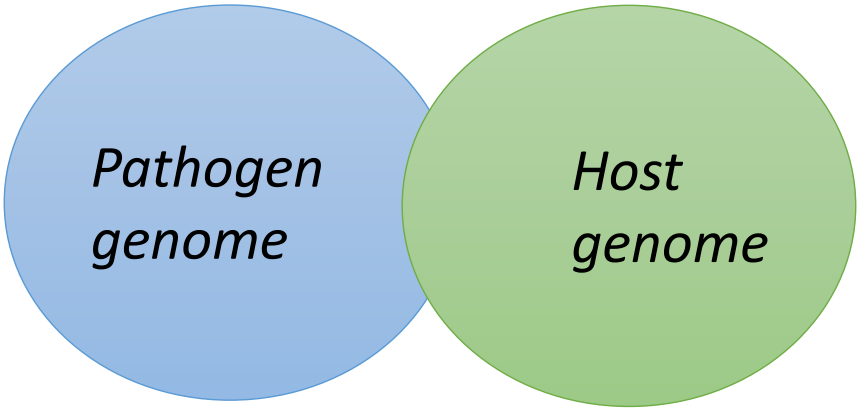
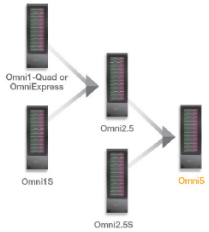
Infectious disease genomics

M.tuberculosis
(~4.4Mb)
P.falciparum
(~23Mb)



Human
(~2,900Mb)



Using whole genome sequencing

- Genetic determinants of transmission & virulence
- Markers for strain discrimination
- Drug resistance mechanisms
- Drug development
- Monitor emergence and spread

High throughput genotyping (>5 million markers)

- Steps underlying infection or disease are controlled by host genetic factors
- Heritability & complexity of phenotypes
- Host susceptibility - using genome association studies (GWAS) approach

Genomic variation

Small: SNPs, indels, microsatellites (human: N>10 million)

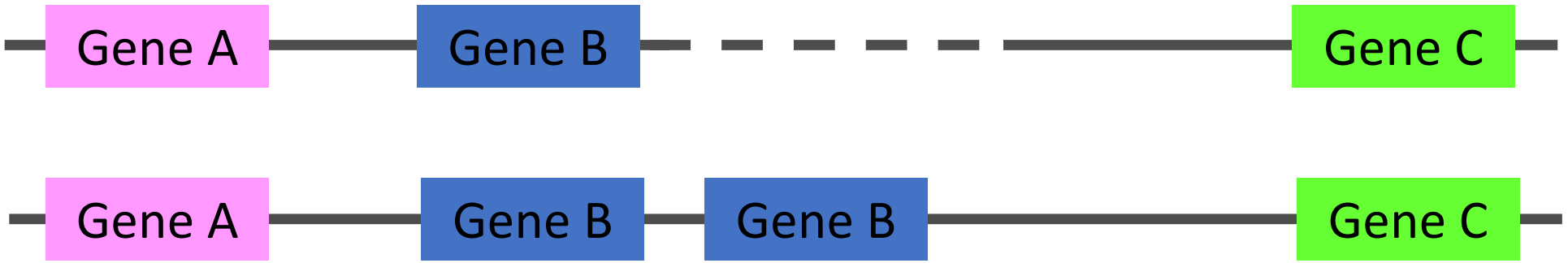
```
ACTCTACGATTTACGGTACTTAGGAGCATATGCTACT
ACTGTACGATTTACGGTACTTAG.AGCATATGCTACT
```

SNP: single nucleotide polymorphism

Indel: insertion / deletion

e.g. P. falciparum drug resistance *(PfCRT, PfDHFR, PfDHPS, PfKelch*)

Larger: Copy number variants, inversions, large indels (N>30k)



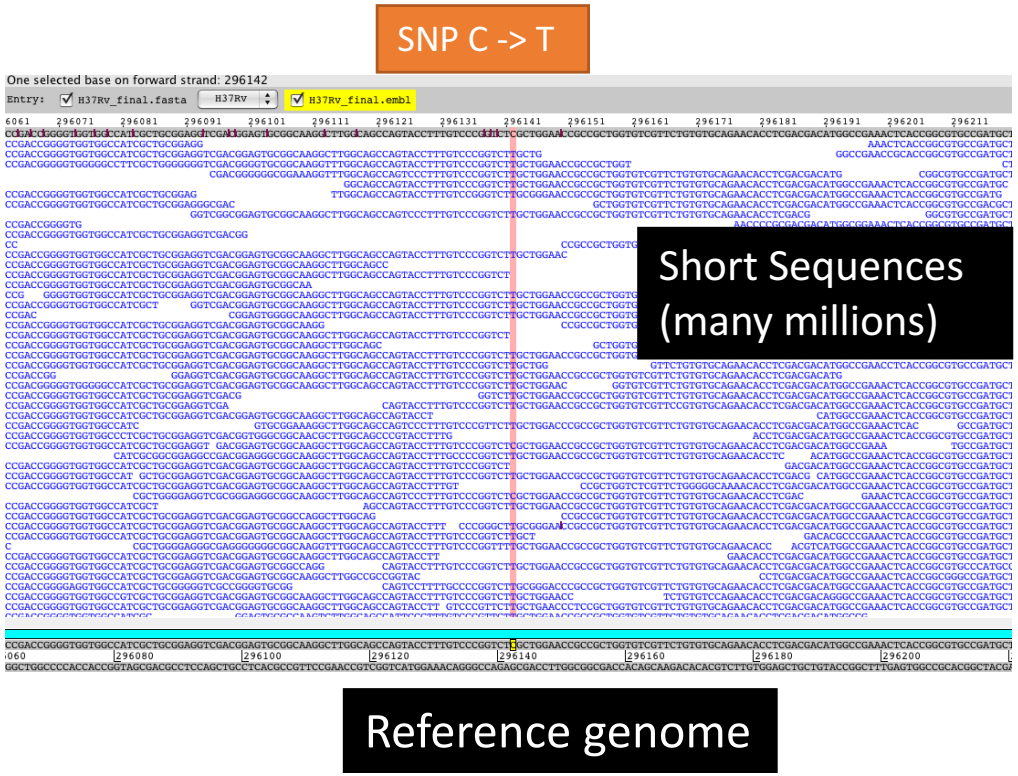
Copy number variation involving a large chunk of DNA that includes the whole of gene B

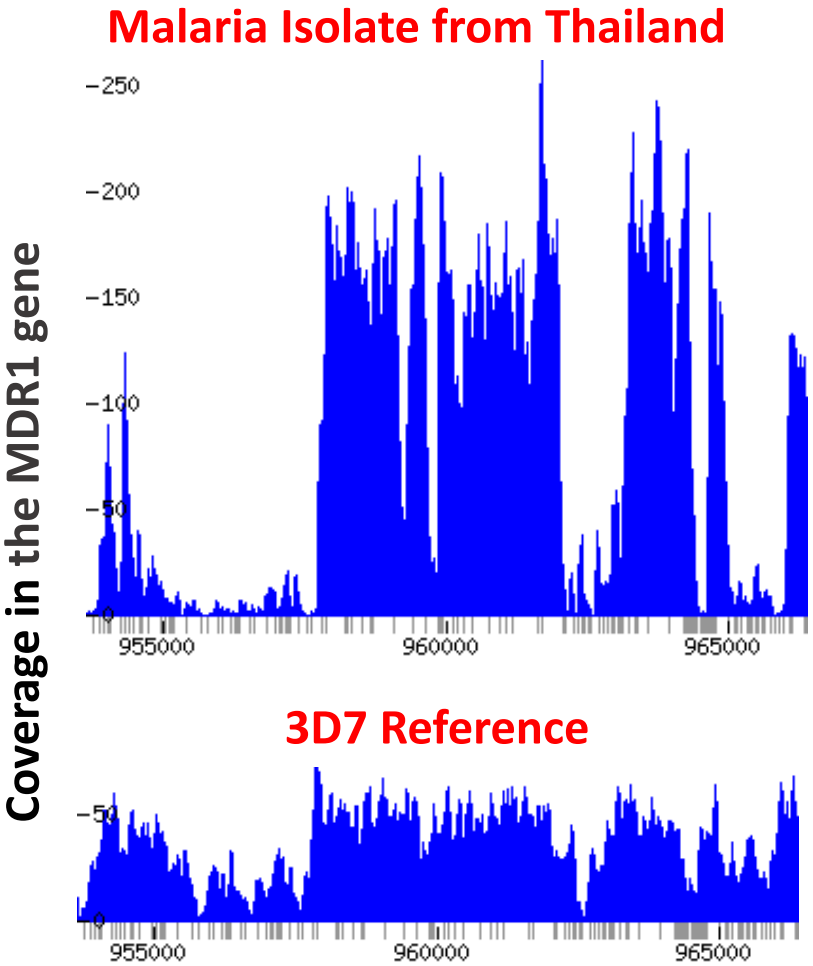
e.g. P. falciparum: PfMDR1 (Mefloquine) – CNV (3 – 10 copies)

Identifying variants

Alignment to a reference (or *de novo* assembly) on a per sample basis

~Zero coverage may imply deletions, excess coverage may imply CNVs





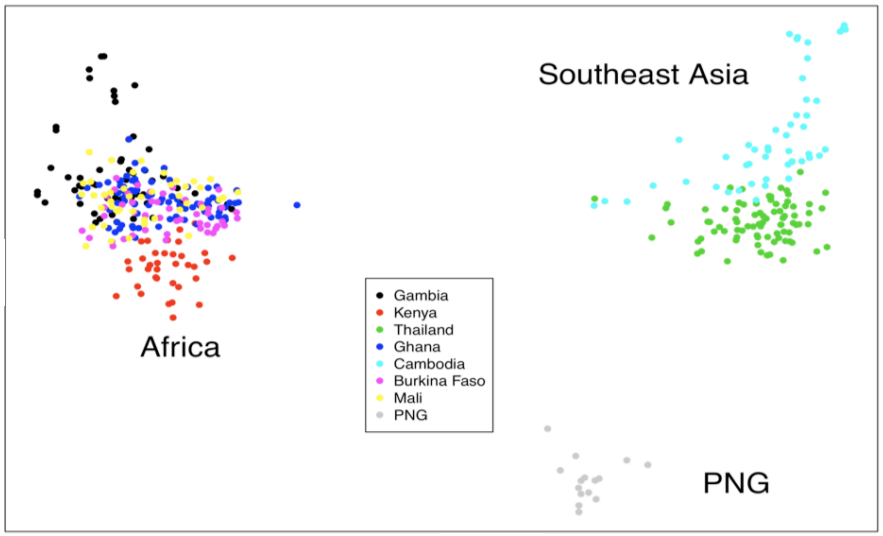
Useful for observing multiplicity of infection – “heterozygous” positions (Assefa et al, 2014)

Disentangling mixing proportions - Bayesian mixed models (Sobkowiak et al)

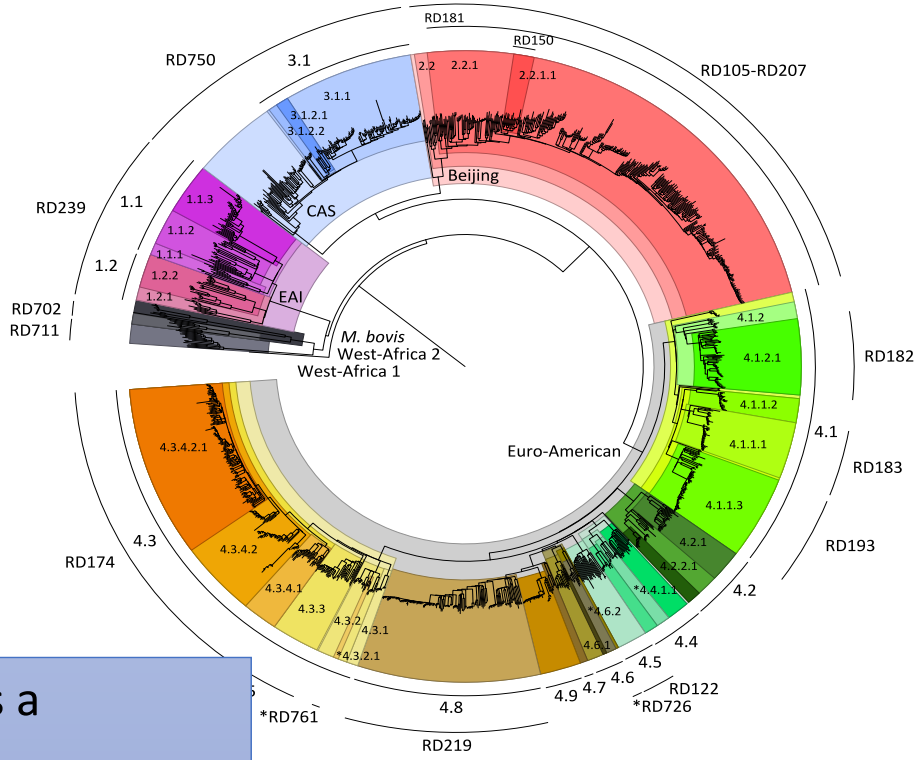
Poisson Hierarch. models (Sepulveda et al, 2013)

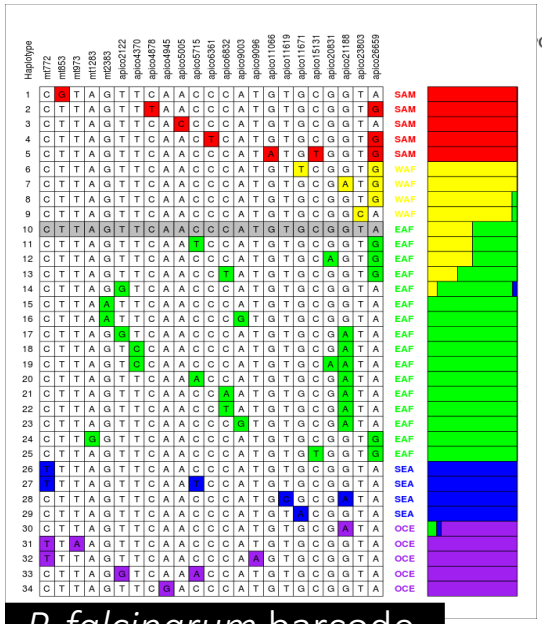
Population clustering using SNP variants

P. falciparum (n=3000; 700,000 SNPs)



M.tuberculosis (n=1,000; 70k SNPs; Coll et al. 2014)

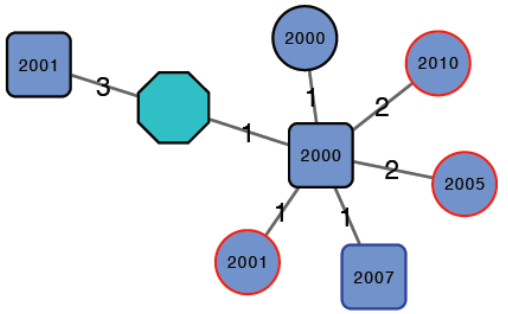




P. falciparum barcode (Preston et al, 2014)

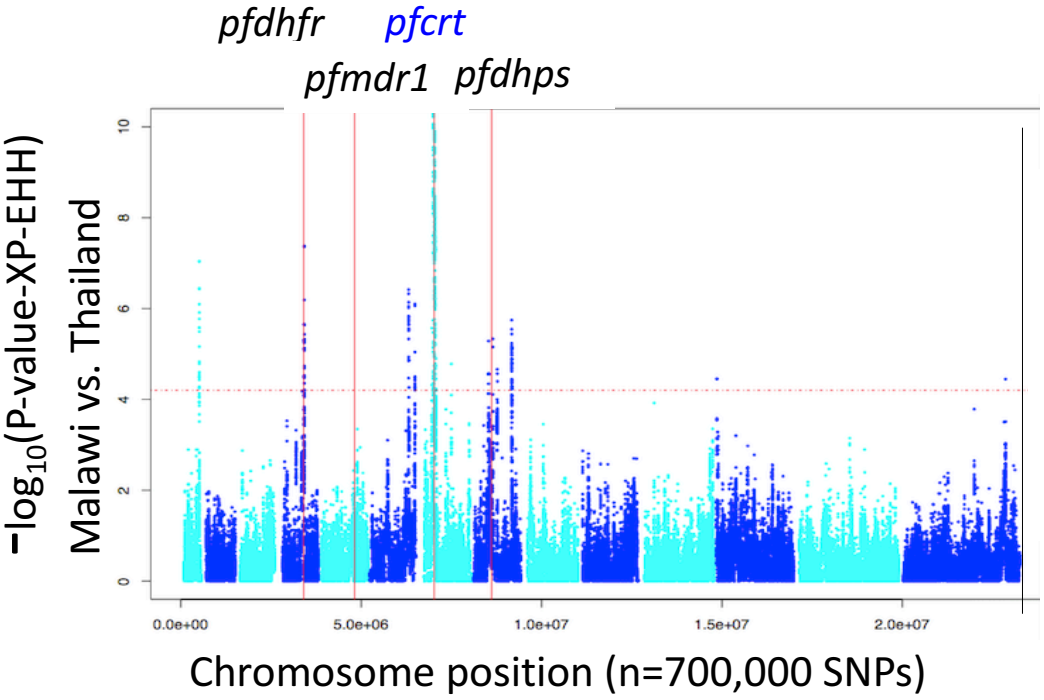
- Population structure is a confounder in GWAS
- Cluster strains agnostic to population label
- Identifying informative SNPs to barcode strains
- Near-genetically identical samples maybe part of transmission chains

TB transmission (n=2500) (Guerra et al, 2015i,ii)



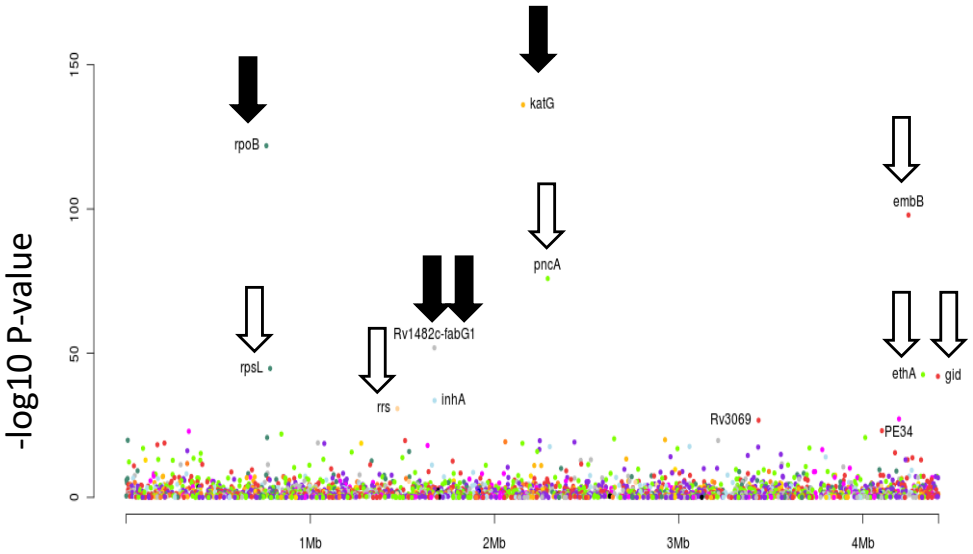
Genome-wide analyses

Malaria: Detecting recent positive selection (n=3,000)



Chromosome position (n=700,000 SNPs)

TB: Multi-drug resistance (GWAS, n=7500)



Genome position (100,000 SNPS)

Phenotype free

- Beneficial mutations sweep through populations (popgen methods: iHS, XP-EHH)
- Missing data (imputation: Samad et al, 2015)

Phenotypes with co-resistance

- Rare variants (aggregate mutations)
- Indels and compensatory mutations
- Phylogenetic structure (Mixed models)
- Resistance prediction (random forests)
- Convergent and micro-evolution