Statistical analysis for GWAS: Population structure Alicia Martin Postdoctoral Research Fellow December 10, 2016

### Modules

- Serial founder effects
- Basic population structure
- Hardy-Weinberg equilibrium
- How genetic structure changes
- Linkage disequilibrium
- Effective population size
- Demographic models
- African origins and population structure

### Serial founder effects

## Historical human migration routes



Henn, Cavalli-Sforza, and Feldman (2012) PNAS



S Mallick et al. Nature 1–6 (2016) doi:10.1038/nature18964

Autosomal heterozygosity

0.0005

# Reduction in diversity due to serial founder effects



Henn, Cavalli-Sforza, and Feldman (2012) PNAS

# Serial founder effect model and assumptions Time

- Migration after the initial founder expansion has been limited There has been no substantial admixture from another highly diverged population
- Post-expansion demographic fluctuations have not decreased diversity substantially

# Decline in heterozygosi out-of-Africa

**Synonymous Heterozygotes**

**Nonsynonymous Heterozygotes**



# Basic population structure

# What is population structure?



Can be caused by multiple barriers to random mating: geography, language, ancestry

Random mating is an important assumption in pop gen and stat gen models, usually assess population structure first

Two commonly used methods of detecting structure are allele frequency-based clustering algorithms and principle component analysis

#### How does population stratification affect association analyses?  $H_{\alpha M}$ dee nomulation multiple populations (European Americans, African Americans and  $erct$  structure  $erct$ and the consequences of



tions and when might this pose problems for large-scale association

1Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. 2Wellcome Trust Center for Human Genetics, University of Oxford, Oxford

**substantial loss of power. The results of our analysis can guide the**

Recent advances in genotyping technologies and increases in genetic

marker availability have paved the way for association studies on

genomic scales1. A potential problem for every population-based asso-

ciation study is the presence of undetected population structure that

to missed real effects (**Fig. 1**). These concerns have influenced the

design, interpretation and funding of association studies during the

Disease more common in Population 2

- ‣ oversampling cases from this population relative to controls
- ‣ any allele that is more common in Pop 2 appears associated with the disease

Marchini et al., Nat Genet 2004

#### Population structure with clustering algorithms  $100 \Omega$

A

I'm 80% red and 20% blue!

**ACB** 

Each bar represents 1 individual. The number of colors is the number of potential ancestries. Proportion of different colors is the proportion of different ancestries for that individual

**ASW** 

أأفيدا وبالمعروف تدادر ماطعين كالوجيات وأ

#### Continental ancestry  $n \geq 1$ rately from the rest of the world. At K = 7, the new component occurs at highest proportions in tion. Hazara and Uygur share a similar profile of  $\cos 2\theta + \cos 2\theta$  $E[\mathbf{E}][\mathbf{A}][\mathbf{A}]$ share and America  $\gamma$  $\blacksquare$ can often be interpreted as are cented as arising from recent from rece  $\sim$ In the current setting, however, the estimated

ancestry. For example, Palestinians, Druze, and



Fig. 1. Individual ancestry and population dendrogram. (A) Regional ancestry inferred with the frappe program at  $K = 7$  (13) and plotted with the Distruct program (31). Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to his/ her ancestry coefficients in up to seven inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in  $\mathsf{plotting.}$  Maximum likelihood tree of  $51$  populations. Branches are colored according to  $\mathsf{plotting.}$ 

from South/Central Asia, separating this region

Daur 1104. Colombian JOT Science 319, 1100- $2<sub>0</sub>$  $O(X)$ Li, J.Z., et al. (2008).

Naxi

East Asian populations and correspond to a north-



#### Genes mirror geography  $\frac{1}{\sqrt{2}}$  same data, haplot decreases from south to north to nort  $\mathbf{A}$ Fig. 1 suggests, European DNA samples can be very informative reported origin may be reduced if finer-scale information on origin were available for each individual. Population structure poses a well-recognized challenge for diseaseassociation studies (for example, refs 11–13). The results obtained here reinforce that the geographic distribution of a sample is impor-

Supplementary Table 4). These numbers exclude individuals who

0.15%, first eigenvalue 5 4.09, second eigenvalue 5 2.04). However,



 $\overline{\phantom{a}}$  $I_{\text{one}}$  $\overline{a}$  $\overline{0}$ Novembre, J., et al. (2008). Nature 456, 98–101.

0.010

# Ancestry-specific PCA provides insights into admixture origins





Martin, A.R., et al. bioRxiv.<http://dx.doi.org/10.1101/070797>

## Fixation index (FsT)



Measures divergence across population pairs  $(S = subpopulations, T =$ total)

\* H = heterozygosity

$$
F_{ST} = 1 - \frac{H_S}{H_T}
$$

$$
= 1 - \frac{2p_S q_S}{2p_T q_T}
$$

Hardy-Weinberg Equilibrium

The Hardy–Weinberg equilibrium model states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.



## Parental allele frequencies



*p* = frequency of A allele *q* = frequency of a allele

*P* = frequency of AA genotype

*P* = frequency of AA genotype  $H = \text{frequency of Aa genotype}$ *Q* = frequency of aa genotype

# Hardy-Weinberg equilibrium



# Hardy-Weinberg: assumptions and violations

#### Assumptions

- ✓organisms are diploid
- ✓only sexual reproduction occurs
- ? generations are non overlapping
- ? mating is random
- ? population size is infinitely large
- ? allele frequencies are equal in the sexes
- ? there is no migration, mutation or selection

#### SLC24A5 - skin color



#### Implications:

Allele frequencies are constant, genetic diversity preserved HWE attained in just 1 generation of random mating

#### HWE in a realistic cohort

#### Tennessen et al (ESP): 2439 individuals \* 1351 Europeans, 1088 African Americans (80%, 20%)

 $P(derived|European) = 1$  $P(derived|African American) = 0.2$ 

 $p_{cohort} = \frac{1*1351+0.2*1088}{2439} = 0.643$  $q_{cohort} = 1 - p_{cohort} = 0.357$ 





# How genetic structure changes

How does population structure change? Changes in allele frequencies through time

mutation

migration

natural selection

genetic drift

non-random mating

How does population structure change? Changes in allele frequencies through time mutation migration natural selection genetic drift non-random mating spontaneous change in DNA Human mutation rate:  $-1.2 \times 10^{-8}$  / bp ‣ ~80-100 total *de novo* variants ‣ <1 *de novo* coding variant How does population structure change? Changes in allele frequencies through time

mutation

migration

individuals moves into population, introduce new alleles ("gene flow")



How does population structure change? Changes in allele frequencies through time

> certain genotypes produce more/less offspring

differences in survival and  $reproduction \rightarrow differences$ in "fitness"

Many kinds: balancing (e.g. sickle-cell), positive (e.g. height), negative (most common), etc



mutation

migration

natural selection

genetic drift

non-random mating

How does population structure change? Changes in allele frequencies through time genetic change by chance mutation alone migration occurs in small populationsnatural selection genetic drift non-random mating 

രൈരെ



AFR (autosome)

EUR (autosome)

\* genetic drift

AFR (autosome)

A

I

non-random mating  $\mathbf \P$ ng<br>—  $\bf C$ ●  $n$  $\overline{\phantom{0}}$ ●● ●  $\overline{H}$ 0.50 Oct 200 Oct 200 Oct 200 Oct 200  $\mathbf{n}$ 

 $\overline{C}$ 

● ● ● ● ● ●  $\bullet$ ● ● ●  $\sqrt{\phantom{.0066}}$ ● ● ● ● ● ● ●  $\sqrt{2}$ ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●  $\triangleleft$ ● ● ●  $\bullet$ ●  $\blacksquare$ ● ● ● ● ● ● ●  $\bullet$ ● ● ● ● ● ●  $\bullet$ ●● ● ● ● ●  $\circ$ ● ●  $\bullet$ ● ● ●  $\bullet$ ● ● ● ● ● ●  $\bullet$ ●● ● ● **D - アープー** ●  $\bullet$ ●  $\sqrt{2}$ ● ● ● ● ● ●  $\bullet$ ●  $\bullet$ ● ● ● ● ● ● ● 0.00 0.00 0.25 0.50 0.75 1.00 1.00 0.00 0.25 0.50 0.75 1.00 0.00 0.25 0.50 0.75 1.00 EUR (autosome) EUR (chrX)  $P$ Hispanic/Latinos  $\triangle$  PUR ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●  $\overline{\mathcal{A}}$ ● ● ● ● ● ● **ア** ● ● ● ● ● ● ● ●● ● ● ● ● ● ● ● ● ● 。<br>《 ● ● 0.25 0.50 0.75  $EUR$  (chr $X$ )



●

●

0.00 0.25 0.50 0.75 1.00 NAT (autosome) in de la propriété de la propri

●

●

●

# Linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at different loci. Loci are said to be in LD when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.

Recombination is the process or act of exchanges of DNA between chromosomes, resulting in a different genetic combination and ultimately to the formation of unique gametes with chromosomes that are different from those in parents.

# Calculation of linkage disequilibrium

Suppose we have the following sequences:





# Calculation of linkage disequilibrium

Covariance between A and B alleles at two loci:  $D_{AB} = p_{AB} - p_{A}p_{B}$ 

Common statistic for summarizing LD: Decay of LD over time (t in generations):  $r^2 =$  $D^2$  $p_A(1-p_A)p_B(1-p_B)$ 

$$
D_t = (1 - r)^t D_0
$$
  
P(recombination in one generation)







# Effective population size

The effective population size  $(N_e)$  is the population size that would result in the same rate of drift in an idealized constant population size, obeying our modeling assumptions, as that observed in our true population. (n 5 96) (Supplementary Table 4). population size that would distortion on the terminal branches of the tree compared to point mutations based our modeling assumption Information, the human branch in particular showed no excess  $\mathbf{I}$ specific pseudogenes24 (Supplementary Information).  $\lim_{n \to \infty} \mathfrak{a}(\mathbf{N})$  is the  $\frac{1}{2}$  understand  $\frac{1}{2}$  use the range of  $\frac{1}{2}$  $result in the same rate of$ ted in the same rate of  $\frac{1}{2}$ cas that observed in our range. These efforts will greatly enhance conservation planning and maintain genetic diversity in wild populations for future generations.

panzee. The Nigeria–Cameroon chimpanzee population size appears

drastic declines in suitable habitat in recent years26, along with declines

in local population sizes of up to 75% (ref. 27). These observations

Despite their rich evolutionary history, great apes have experienced

loss-of-function mutations $\mathcal{A}$ . We thus characterized thus characterized the distribution of distribution of

terminal N<sup>e</sup> correspond to the effective population size after the last split event.

the values used in this figure can be found in Supplementary Table 5. The

fixed loss-of-function mutation mutations among different species of great apesitive  $\mathcal{L}_{\mathcal{A}}$ 

identifying nonsense and frameshift mutations resulting frameshift  $\mathcal{N}$ 



 $\overline{a}$ (scaled in units of 4 Prado-Martinez, J., et al. (2013). Nature 1–5.

### Assumptions of pop gen models affecting N<sub>e</sub> when violated

- There are equal numbers of males and females, all of whom are able to reproduce
- All individuals are equally likely to produce offspring, and number of offspring the each produces varies no more than expected by chance
- Mating is random
- The number of breeding individuals is constant from one generation to the next.

Essentially all violations to pop gen models decrease Ne

#### Ne with unequal numbers of breeding males and females  $4N_{m}N_{f}$







# Identity-by-descent



#### Fine-scale birth record data enables refined view of population history view di population dlo e



 $\blacksquare$ Lapland maintained very little growth  $T$  =  $\frac{1}{1}$   $\frac{1}{1}$   $\cdots$ Time (generations) for extended period 2: Southwest coastal region started growing longer ago 12: Lapland maintained very little growth



# Demographic models



# Wright-Fisher model



Non-overlapping generations

Binomial sampling of alleles

Finite, constant N

Basis of the coalescent



Graham Coop's pop gen notes:<http://bit.ly/2fEXzUe>

Figure 8: Loss of heterozygosity over time, in the absence of new mutations. A diploid

# Wright-Fisher model

- Diploid population of size N has 2N alleles
- Probability that two alleles have same parent: 1/2N
- Probability different parent: 1-1/2N
- Probability two alleles coalesce before mutation:

$$
\frac{1}{2N} \int_0^\infty e^{-t(2\mu + 1/(2N))} dt = \frac{1/(2N)}{1/(2N) + 2\mu} = \frac{1}{1 + 4N\mu}
$$

\* Population-scaled mutation rate:  $\theta = 4N_e\mu$ From the binomial:  $E[K_1] = Np$  $Var[K_1] = Np(1-p)$ 

# Loss of heterozygosity in a bottlenecking population



Graham Coop's pop gen notes:<http://bit.ly/2fEXzUe>

generation I color I color I color i color so we can track their descendants, the international color track the<br>Seconda

#### Demographic model from 1000 Genomes data interval  $\mathcal{A}$  , and inference based on the exon pilot  $\Gamma$   $\Gamma$   $\sim$  98  $\sim$ In general, the gain in precision was strongest for the parameters in more and more and in low-coverage data yielded an unrealistic TB = 14 kya split.  $\overline{10000}$  $\mathbf{h}$  and  $\mathbf{h}$  are parameters and  $\mathbf{h}$  and  $\mathbf{h}$  are parameters and  $\mathbf{h}$ predictions; given a demographic model, we can predict, for example, the number of synonymous variants to be discovered in only two parameters per population, and in  $\Gamma$  $10^{6}$  model trom sites are accurately determined, and errors occur independently debatable: we expect at least some correlations in the coverage at n americato account such contractions would

We found that the bulk of the discrete that the discrete that the discrete that the discrete that the discrete

a maximum likelihood estimate of TB  $=$  140 kya (95% confidence)



# African origins and population structure

What do we know about African population history?

# Anatomically modern humans originated in Africa





Klein 1999 White 2003

### Hominid evolution



#### Timing of population divergence within Africa  $\blacksquare$  distribution of  $\blacksquare$ ulation reduction in the set  $T_{\rm eff}$  search for generally with unusually  $\sim$ we convergence the single greatest FST value ∼2.3 million SNPs (14). Although genome-wide  $\sim$   $\sim$   $\sim$   $\sim$  $it$  nonulation extraordinary fraction of ancestry from Bantuspeakers (South Africa) in the Nama (Fig. 4D and  $\mathbf{S}$   $\mathbf{$ introgression and, potentially, ensuing selection. Because of their early divergence, signals of

skin lesions, and an elevated risk of skin cancer

- \* Oldest divergence is between KhoeSan populations and everyone else (120-90 kya) San groups was moderate (0.012 to 0.034), the everyone else (120-90 k were all  $\alpha$  -defined in the same region on the same region on the same region on  $\mathbf{p}$  $\mathfrak{t}$
- \* Divergence between Central and Eastern Africans: 70-45 kya bivergence between en  $tral$  and  $Factor$  A fricance  $\pi$  $f$ requency designations shared and  $\frac{1}{2}$ 
	- Eurasians derive from Eastern African populations

their early and extensive contact with European



## Linguistic structure

- 5 major language families in Africa
- Expansion of Niger-Congo language 4,000 years ago
- Most isolated and most controversial language family is Khoisan



## Population structure



# Population samples

- Samples assayed on multiple genotyping platforms: Illumina 550K.v2 & 600K, Affymetrix 6.0, HapMap3
- $*$  50,000 500,000 SNPs across the genome
- Datasets are publicly available ([http://www](http://www-evo.stanford.edu/repository/paper0002/)[evo.stanford.edu/](http://www-evo.stanford.edu/repository/paper0002/) [repository/paper0002/](http://www-evo.stanford.edu/repository/paper0002/))



Henn, B.M., et al. (2011). PNAS. 108, 5154–5162.

#### Structure within Africa



Henn, B.M., et al. (2011). PNAS. 108, 5154-5162.

#### Structure and FsT







- African populations are highly structured (pre-Bantu expansion)
- Time depth of structure is unresolved (~120-40 kya)
- Despite recent gene flow, underlying structure and diversity is detectable Annual<br>Annual<br>Ex Access provided by Harvard University on 12/04/16. For personal use only.



Solid horizontal lines indicate gene flow between populations and the dashed horizontal line indicates recent

 $A_{\text{max}}$   $D_{\text{avg}}$   $C_{\text{maxima}}$   $T_{\text{sum}}$   $C_{\text{max}}$   $A_{\text{max}}$ Campbell, M.C., & Tishkoff, S.A. (2008). Annu Rev Genomics Hum Genet 9, 403–433.

## Takeaways

- Complexities to population structure (LD, allele frequency differences, etc). Need to consider for ALL genetic methods:
	- GWAS has potential to confound associations
	- RVAS difficulty accounting for rare structure
	- Genetic risk prediction
	- ...many more