Questions we are addressing

- How reproductively isolated are a group of populations?
- How many populations are in sample?

Allele Frequencies

- Allele frequencies (gene frequencies) = proportion of all alleles in an all individuals in the group in question which are a particular type
- Allele frequencies:
  - $p + q = 1$
- Expected genotype frequencies:
  - $p^2 + 2pq + q^2$

Hardy-Weinberg Equilibrium

- Null Model = population is in HW Equilibrium
  - Useful
  - Often predicts genotype frequencies well

Hardy-Weinberg Theorem

if only random mating occurs, then allele frequencies remain unchanged over time.
After one generation of random-mating, genotype frequencies are given by

AA	Aa	aa
p^2	2pq	q^2

p = freq (A)
q = freq (a)

Expected Genotype Frequencies

• The possible range for an allele frequency or genotype frequency therefore lies between (0 – 1)
• with 0 meaning complete absence of that allele or genotype from the population (no individual in the population carries that allele or genotype)
• 1 means complete **fixation** of the allele or genotype (fixation means that every individual in the population is homozygous for the allele -- i.e., has the same genotype at that locus).

ASSUMPTIONS

1) diploid organism
2) sexual reproduction
3) generations are non-overlapping
4) mating occurs at random
5) large population size
6) migration = 0
7) mutation = 0
8) no selection on genes

For a population to be in Hardy Weinberg Equilibrium, the observed genotype frequencies must match those predicted by the equation $p^2 + 2pq + q^2$

Finding "p": equals frequency of AA + ½(frequency of Aa)
Finding "q": equals frequency of aa + ½(frequency of Aa)

Then p^2 = predicted genotype frequency of AA
q^2 = predicted genotype frequency of aa
2pq = predicted genotype frequency of Aa

	f_{AA}	f_{Aa}	f_{aa}
Population 1	0.3	0.0	0.7
Population 2	0.2	0.2	0.6
Population 3	0.1	0.4	0.5

In each group, the allele frequency is actually the same

Population 1	A = 0.3 a = 0.7
Population 2	A = 0.2 + .5(0.2) = 0.3 a = 0.6 + .5(0.2) = 0.7
Population 3	A = 0.1 + .5(0.4) = 0.3 a = 0.5 + .5(0.4) = 0.7

Only **population 3** is in Hardy Weinberg equilibrium (need to do chi-square test)
$p^2 + 2pq + q^2 = 1$
$(0.3)^2 + 2\ (0.3)\ (0.7) + (0.7)^2 = 1$
$0.09 + 0.42 + 0.49 = 1$

IMPORTANCE OF HW THEOREM

If the only force acting on the population is random mating, allele frequencies remain unchanged and genotypic frequencies are constant.

Mendelian genetics implies that genetic variability can persist indefinitely, unless other evolutionary forces act to remove it

Population Substructure

- Many species naturally subdivide themselves into herds, flocks, colonies, schools etc.
- Patchy environments can also cause subdivision
- Subdivision decreases heterozygosity and generates genetic differentiation via:
  - » Natural selection
  - » Genetic drift

Wright's Fixation Index

- Equals the reduction in heterozygosity expected with random mating at one level of population hierarchy relative to another more inclusive level.

$$F_{ST} = (H_T - H_S)/ H_T$$

Step 1: Calculate mean heterozygosities at each population level

- Heterozygosity = mean percentage of heterozygous individuals per locus
- Assuming H-W, heterozygosity (H) = 2pq where p and q represent mean allele frequencies
- H_S = sum of all subpopulation heterozygosities divided by the total number of subpopulations

	Pop1	Pop2	Pop3
Sample size	20	20	20
AA	10	5	0
Aa	4	10	8
aa	6	5	12

	Pop1	Pop2	Pop3
Freq			
p	(20 + 1/2*8)/40 = 0.60	(10+1/2*20)/40 = .50	(0+1/2*16)/40 = 0.20
q	(12 + 1/2*8)/40 = 0.40	(10+1/2*20)/40 = .50	(24+1/2*16)/40 = 0.80

- Calculate H_S (2pq)
  - Pop1: 2*0.60*0.40 = 0.48
  - Pop2: 2*0.50*0.50 = 0.50
  - Pop3: 2*0.20*0.80 = 0.32

Pop	H_s		p	q	H_T		F_{ST}	
1	0.48		0.60	0.40				
2	0.50		0.50	0.50				
3	0.32		0.20	0.80				
Mean	0.43		0.43	0.57	0.49		0.12	

Interpreting F_{ST}

- Can range from 0 (no genetic differentiation) to 1 (fixation of alternative alleles).
- Wright's Guidelines:
  - 0 - 0.05, little differentiation
  - 0.05 – 0.15, moderate
  - 0.15 – 0.25, great
  - > 0.25, very great

F_{ST} for various organisms

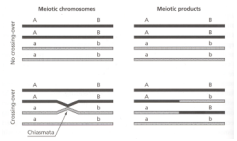
Organism	Number of Populations	Number Loci	Ht	Hs	Fst
Human (major races)	3	35	0.13	0.121	0.069
Yanomama Indian Villages	37	15	0.039	0.036	0.077
House mouse	4	40	0.097	0.086	0.113
Jumping rodent	9	18	0.037	0.012	0.676
Fruit fly	5	27	0.201	0.179	0.109
Horseshoe crab	4	25	0.066	0.061	0.076
Lycopod plant	4	13	0.071	0.051	0.282

Multiple loci

- Should be in HW equilibrium at all loci
- If recombination occurs freely, all combinations of alleles should be found equally

Linkage Disequilibrium

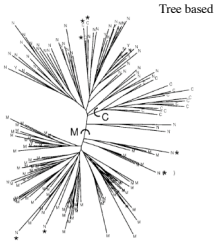
- $D = P_{AB}P_{ab} - P_{aB}P_{Ab}$
- D = 0 when population is in equilibrium



Inferring numbers of populations

- Phylogenetic/distance methods
  - Generate tree from population data, and look for obvious structure.

Tree based clustering



Inferring numbers of populations: clustering methods

- Minimize departures from HW
- Minimize linkage disequilibrium
- Assign individuals to K populations so achieve linkage equilibrium and HW equilibrium

Program Structure is a model based clustering method to infer population structure

- Main parameter is K, the number of populations
- Run models several times with different K values and compare likelihoods.

Inferring numbers of populations: clustering methods

- Minimize departures from HW
- Minimize linkage disequilibrium
- Assign individuals to K populations so achieve linkage equilibrium and HW equilibrium

Why care about population structure?

Why care about population structure?

- Epidemiological contexts
  - Risk factors, rates of disease and adverse drug response vary across populations
- Self-reporting or genotype?
  - What about mixed ancestry?

Why care about population structure?

- Association studies
  - If disease risk correlated with genetic ancestry, ignoring could increase false positives.

Why care about population structure?

- Conservation issues
  - If structure observed in a species, may want to conserve genetic diversity.

Human genetic structure

- 1056 individuals from 52 populations
- Genotype for 377 autosomal microsattelites
- Use Structure to infer population without using a priori knowledge of population origin.

What Are Microsatellites?

- Genetic markers based on variation of unique DNA sequences
- >4 nucleotide core element tandemly repeated, e.g.
  atatatatatatatatatat = $(at)_{10}$
- Allele size based on repeat number of core elements
- Often more variable than single nucleotide polymorphisms due to slippage during replication.

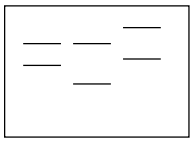
Primer Design

CGNATCCATATGCTGCAAGTGCTTGCCACTTGCAGTTTATGCTCTGAAACATGGAGCAGCAAGATGGCCCAACGAGCATA
TGTAAAGAAAGCATCAGGGACAGACAGACCAGAACGATCTGTTGACCACAACTTGACTCGTCTAATCCTCGCAACAAAAAC
TTGACTCATCTACAGACAACAACTAGCTATGTATGTTCCCATCATCAGCGCCACAACATGCCACTGATCAGGGACGCGAGCT
CATCAGATTTGGCAAAGACGTGCAGGCCTCGCACCTGAATTGGGAAGGCCAAATCTGCCTGGCCGGGAGAACCAGGGC
GCTGGGTCCGGACCTAAACCTGCCCGGCCGGCCAACTCCTCTGCTGCTGTTGGCCGGNCAAGCTCAGCTCTCGACAGTCTCA
CACCCACCCGACNAAAGCTAACGCCCGCCTTGGGGCCCAACCCAACGGAGAGCACAACCCACTANCCTTTCAAAGCCAAGC

OLIGO	3' exq
LEFT PRIMER	ACATCATCAGCGCACAACAT
RIGHT PRIMER	TGGGGTGTGAGACTGTGAGA

PRODUCT SIZE: 209

Size by electrophoresis

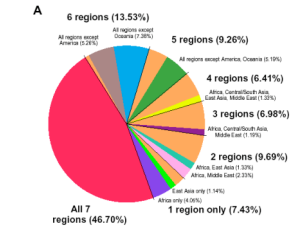


Variation within and between populations

- With-in population variation
  - 93 to 95%
- Between population variation
  - 3 to 5%

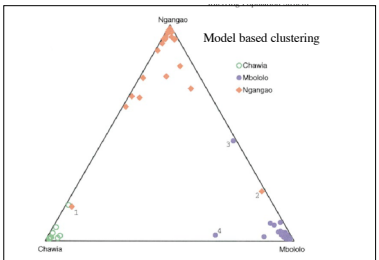
Sample	Number of regions	Number of populations	Variance components and 95% confidence intervals (%)		
			Within populations	Among populations within regions	Among regions
World	1	52	94.6 (94.3, 94.8)	5.4 (5.2, 5.7)	
World	5	52	93.2 (92.9, 93.5)	2.5 (2.4, 2.6)	4.3 (4.0, 4.7)
World	7	52	94.1 (93.8, 94.3)	2.4 (2.3, 2.5)	3.6 (3.3, 3.9)
World-B97	5	14	89.8 (89.3, 90.2)	5.0 (4.8, 5.3)	5.2 (4.7, 5.7)
Africa	1	6	96.9 (96.7, 97.1)	3.1 (2.9, 3.3)	
Eurasia	1	21	98.5 (98.4, 98.6)	1.5 (1.4, 1.6)	
Eurasia	3	21	98.3 (98.2, 98.4)	1.2 (1.1, 1.3)	0.5 (0.4, 0.6)
Europe	1	8	99.3 (99.1, 99.4)	0.7 (0.6, 0.9)	
Middle East	1	4	98.7 (98.6, 98.8)	1.3 (1.2, 1.4)	
Central/South Asia	1	9	98.6 (98.5, 98.8)	1.4 (1.2, 1.5)	
East Asia	1	18	98.7 (98.6, 98.9)	1.3 (1.1, 1.4)	
Oceania	1	2	93.6 (92.8, 94.3)	6.4 (5.7, 7.2)	
America	1	5	88.4 (87.7, 89.0)	11.6 (11.0, 12.3)	

Most alleles present in all regions (few "private" alleles).



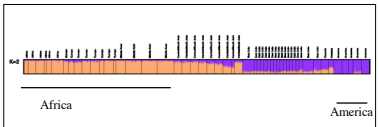
With so little variation between population, how to identify populations

- Look at allele frequencies.
- Cluster to identify groups with distinctive allele frequencies
  - Minimize departures from H-W equilibrium
  - Minimize LD

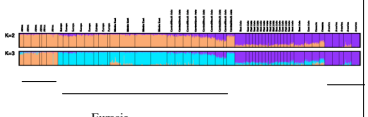


First cluster entire populations

- Run at a variety of K populations

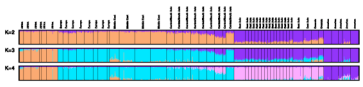


With K=2 find america and africa, with large genetic distance

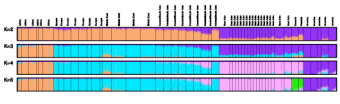


K = 3 parses out Europe

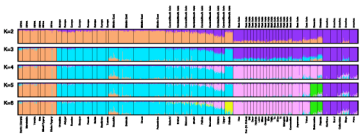
However, some inconsistency between runs of Structure
One run parsed out East Asia instead of Euroasia.
Which is "correct"?



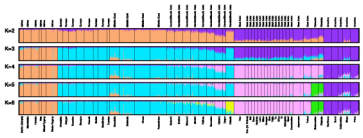
K = 4 identifies Africa, Eruasia, East asia, America



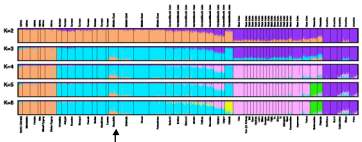
K = 5
Now have major geographic regions:
Africa, Euroasia, East Asia, Oceana, and Americas



K = 6
Identifies Kalash, an isolated tribe in northwest Pakistan
But, one run identified Karitiana instead of Kalash.



K = 6
Which should we choose?
What evidence did they present that this was statistically significant?
Makes clear biological sense.



What do the different mixtures of color mean?

Supplementary Table 1: Genetic distances between regional groups. The coancestry or F_{st} distance (*33*, p. 194) was estimated using GDA (*31*).

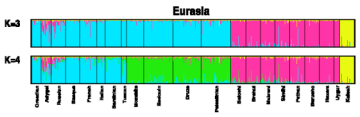
	Africa	Europe	Middle East	Central/ South Asia	East Asia	Oceania
Europe	0.040					
Middle East	0.033	0.005				
Central/South Asia	0.037	0.008	0.008			
East Asia	0.054	0.038	0.038	0.026		
Oceania	0.068	0.061	0.059	0.049	0.047	
America	0.101	0.079	0.081	0.068	0.060	0.102

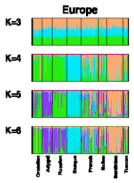
Example output from Structure

```
---------------------------------------------------------------------
Estimated Ln Prob of Data   = -4018.0
Mean value of ln likelihood = -3874.2
Variance of ln likelihood   = 287.4
Mean value of alpha         = 0.0432

Mean value of Fst_1         = 0.0703
Mean value of Fst_2         = 0.0643
Mean value of Fst_3         = 0.1004


Inferred ancestry of individuals:
     Label (Miss) :  Inferred clusters (and 90% probability intervals)
  1      1   (0)  :  0.414 0.568 0.018     (0.000,1.000) (0.000,1.000) (0.000,0.116)
  2      2   (0)  :  0.633 0.356 0.011     (0.000,1.000) (0.000,1.000) (0.000,0.061)
  3      3   (0)  :  0.378 0.610 0.012     (0.000,1.000) (0.000,1.000) (0.000,0.070)
  4      4   (0)  :  0.707 0.285 0.008     (0.000,1.000) (0.000,1.000) (0.000,0.039)
  5      5   (0)  :  0.314 0.650 0.036     (0.000,1.000) (0.000,1.000) (0.000,0.264)
  6      6   (0)  :  0.649 0.342 0.009     (0.000,1.000) (0.000,1.000) (0.000,0.053)
  7      7   (0)  :  0.659 0.330 0.010     (0.000,1.000) (0.000,1.000) (0.000,0.059)
  8      8   (0)  :  0.491 0.490 0.019     (0.000,1.000) (0.000,1.000) (0.000,0.131)
  9      9   (0)  :  0.587 0.375 0.038     (0.000,1.000) (0.000,1.000) (0.000,0.230)
```

Now, can we break down each population further

- Analyze individual populations for structures
- Why not just keep increasing K, like up to 52?





Highest likelihood K=3 found no structure.
Why are populations different in ability to identify structure?

Genetic drift occurs rapidly in small populations, particularly those isolated quickly accrue "distinctive" allele frequencies. This is the signal structure is good at identifying.

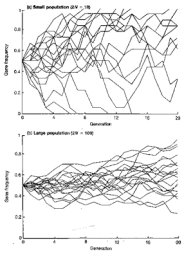
Principles of Neutral Theory

- If a population contains a neutral allele with frequency P_0, then the probability of fixation of the allele is P_0
  - Since initial frequency is 1/2N, the probability of fixation is 1/2N
- Can also calculate rates of fixation and loss
  - Both depend on effective population size

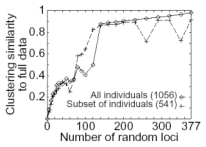
Table 1. Average time to fixation or loss of neutral alleles.

N^a	Average generations until loss[b]	Average generations until fixation[b]
50	7.4	160
100	8.5	320
250	9.9	800
500	11.1	1,600
1,000	12.2	3,200
5,000	14.7	16,000
10,000	15.8	32,000
50,000	18.4	160,000

[a] Tabulated values assume that $N_e = 0.8N$.
[b] Averages are for selectively neutral alleles destined to be lost or fixed, respectively.



Which is more important for analysis,
More individuals or more loci



Major conclusions

- Most variation is within-population differences
- Identified six main genetic clusters, five of which correspond to major geographic regions (?)
- General agreement between genetic and self-reported ancestry

Future studies

- Models of human migration
- Admixture in human origins
  - Did humans breed with other hominids during evolution?
- Application to other organisms