POPULATION STRUCTURE

1000 Genomes Data

	AFR	661	AFRICAN
1	ACB	96	African Caribbeans in Barbados
2	ASW	61	Americans of African Ancestry in SW USA
3	ESN	99	Esan in Nigeria
4	GWD	113	Gambian in Western Divisions in the Gambia
5	LWK	99	Luhya in Webuye, Kenya
6	MSL	85	Mende in Sierra Leone
7	YRI	108	Yoruba in Ibadan, Nigeria
	AMR	347	ADMIXED AMERICAN
8	CLM	94	Colombians from Medellin, Colombia
9	MXL	64	Mexican Ancestry from Los Angeles USA
10	PEL	85	Peruvians from Lima, Peru
11	PUR	104	Puerto Rican from Puerto Rico
	EAS	504	EAST ASIAN
12	CDX	93	Chinese Dai in Xishuangbanna, China
13	CHB	103	Han Chinese in Beijing, China
14	CHS	105	Southern Han Chinese
15	JPT	104	Japanese in Tokyo, Japan
16	KHV	99	Kinh in Ho Chi Minh City, Vietnam
	EUR	503	EUROPEAN
17	CEU	99	Utah Residents (CEPH) with Northern and Western European Ancestry
18	FIN	99	Finnish in Finland
19	GBR	91	British in England and Scotland
20	IBS	107	Iberian Population in Spain
21	TSI	107	Toscani in Italia
	SAS	489	SOUTH ASIAN
22	BEB	86	Bengali from Bangladesh
23	GIH	103	Gujarati Indian from Houston, Texas
24	ITU	102	Indian Telugu from the UK
25	PJL	96	Punjabi from Lahore, Pakistan
26	STU	102	Sri Lankan Tamil from the UK

Questions of Interest

- How much genetic variation is there? (animal conservation)

- How much migration (gene flow) is there between populations? (molecular ecology)

- How does the genetic structure of populations affect tests for linkage between genetic markers and human disease genes? (human genetics)

- How should the evidence of matching marker profiles be quantified? (forensic science)

- What is the evolutionary history of the populations sampled? (evolutionary genetics)

Statistical Analysis

It is possible to approach these data from purely statistical viewpoint.

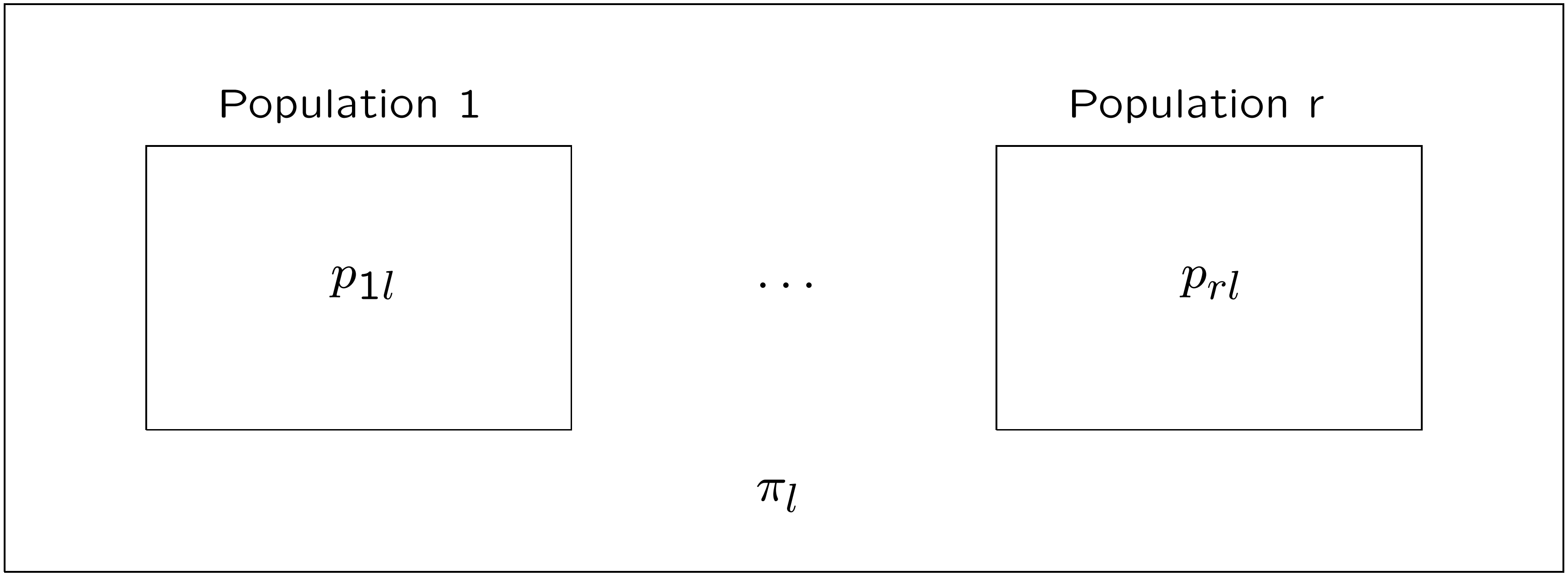
It is possible to test for differences in allele frequencies among populations.

It is also possible to use various multivariate techniques to cluster populations.

These statistical analyses may not answer the biological questions, and the alternative is to set up an evolutionary model that takes into account the history of the populations under study. This allows for a broader interpretation of the data.

Notation

Genetic Analysis: SNP l Allele Frequencies

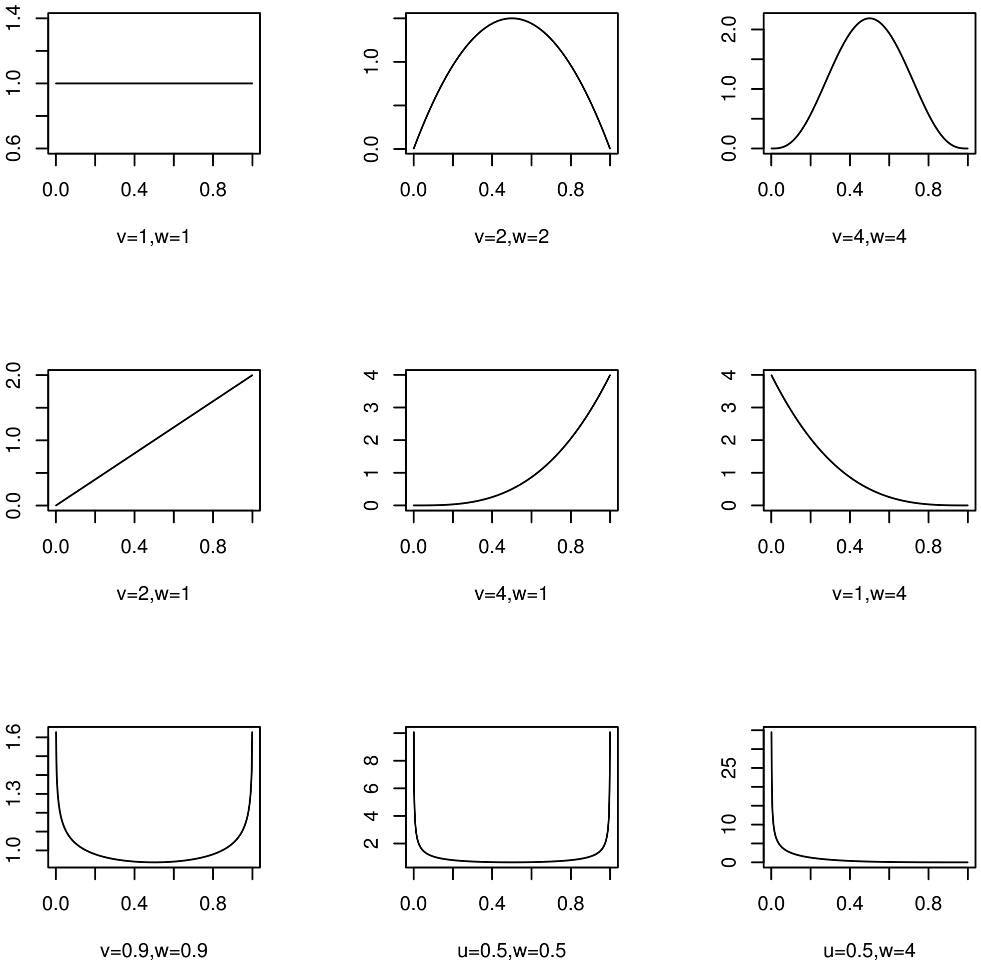


Among samples of n_i alleles from population i: counts for the SNP l reference allele follow a binomial distribution with mean p_{il} and variance $n_i p_{il}(1 - p_{il})$. Sample allele frequencies $\tilde{p}_{il}$ have expected values p_{il} and (under HWE) variances $p_{il}(1 - p_{il})/n_i$.

Among replicates of population i: p_{il} values follow a distribution with mean π_l and variance $\pi_l(1 - \pi_l)\theta^i$. Distribution sometimes assumed to be Beta.

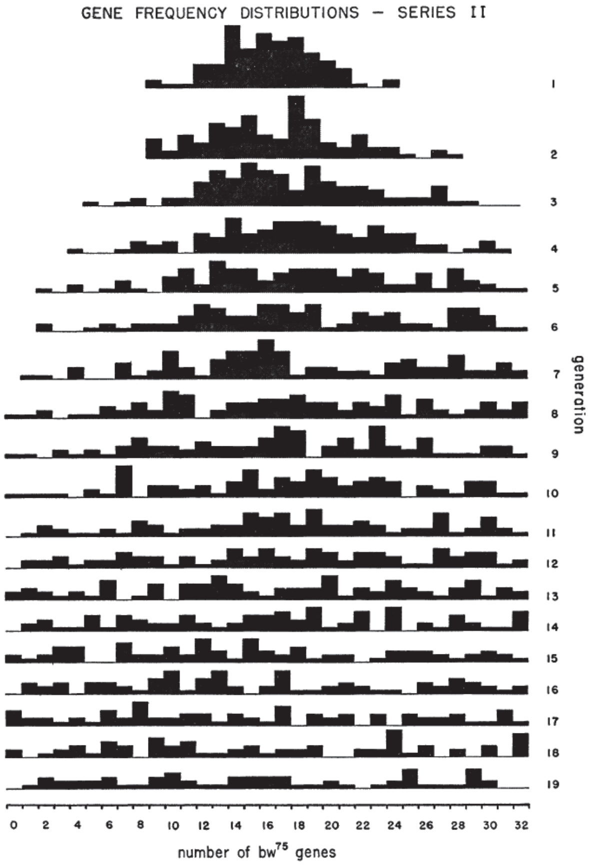
Beta distribution: Theoretical

The beta probability density is proportional to $p^{v-1}(1-p)^{w-1}$ and can take a variety of shapes.



Beta distribution: Experimental

The beta distribution is suggested by a *Drosophila* experiment with 107 replicate populations of size 16, starting with all heterozygotes:



Buri P. 1956. Evolution 10:367

What is θ?

Two ways of thinking about θ.

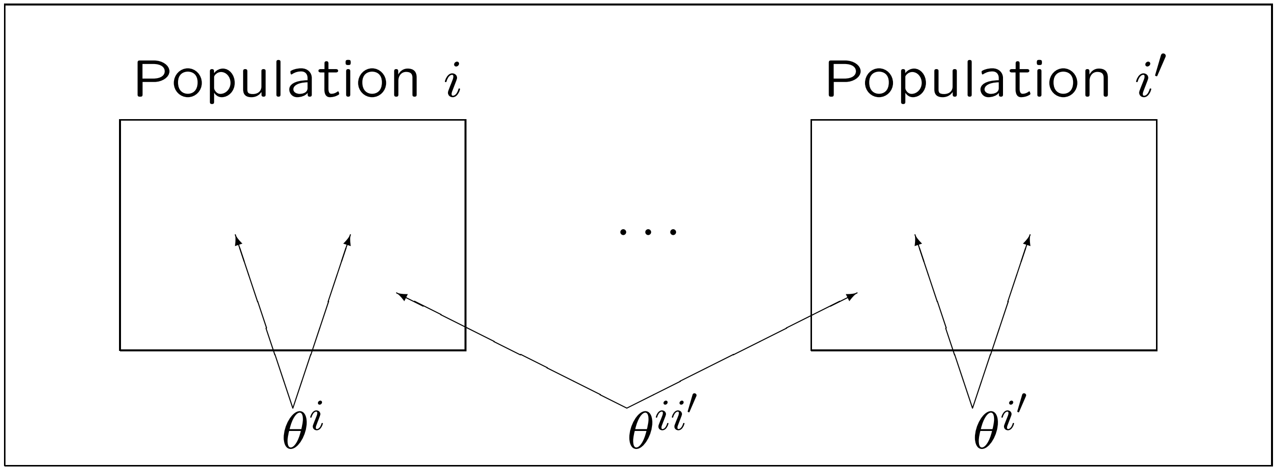
It measures the probability a pair of alleles are identical by descent: and this is with respect to some reference population.

The target alleles may be in specified populations, and this leads to characterization of population structure, of they may be in specified individuals and this leads to characterization of inbreeding and relatedness.

θ also describes the variance of allele frequencies among populations, or among evolutionary replicates of a single population.

Weir BS, Goudet J. 2017. Genetics 206:2085-2103.
Goudet J, Kay T, Weir BS. 2018. Molecular Ecology 27:4121-4135.

Allele-level θ's



θ's are ibd probabilities for pairs of alleles from specified populations.

θ_W^i is average of the within-population probabilities θ^i. Average over populations of θ_W^i is θ_W.

θ_B is average of the between-population-pair probabilities $\theta^{ii'}$.

Allelic Measure Predicted Values

Predicted Values of the θ's: Pure Drift

The estimation procedure for the θ's holds for all evolutionary scenarios, but the theoretical values of the θ's do depend on the history of the sampled populations.

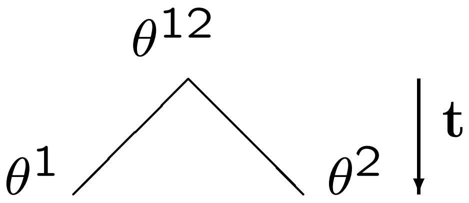
In the case of pure drift, where population i has constant size N_i and there is random mating, t generations after the population began drifting from an ancestral population in which $\theta^i = 0$

$$\theta^i(t) \;=\; 1 - \left(1 - \frac{1}{2N_i}\right)^t$$

If t is small relative to large N_i's, $\theta^i(t) \approx t/(2N_i)$, and $\theta_W(t) \approx t/(2N_h)$ where N_h is the harmonic mean of the N_i.

Drift Model: Two Populations

Now allow ancestral population itself to have ibd alleles with probability θ^{12} (the same value as for one allele from current populations 1 and 2):

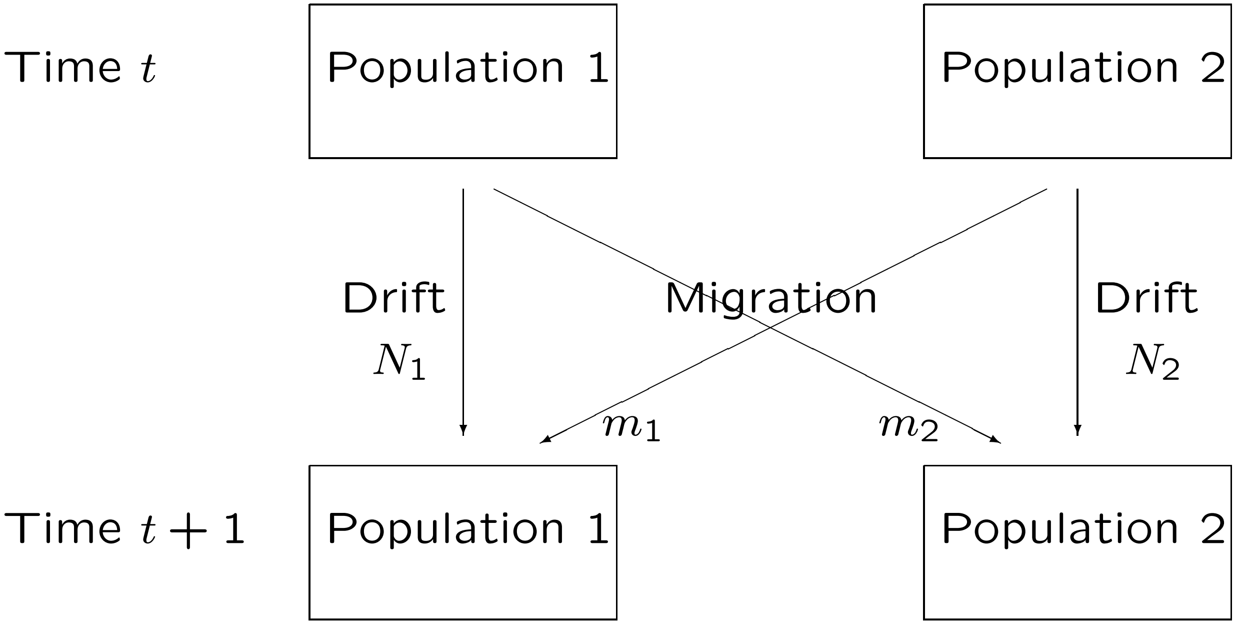


$$\theta^i \;=\; 1-(1-\theta^{12})\left(\frac{2N_i-1}{2N_i}\right)^t, \;\; i=1,2$$

It is possible to avoid needing to know the ancestral value θ^{12} by making θ^1, θ^2 *relative to* θ^{12}:

$$\beta^i = \frac{\theta^i-\theta^{12}}{1-\theta^{12}} \;=\; 1-\left(\frac{2N_i-1}{2N_i}\right)^t \approx \frac{t}{2N_i}, \;\; i=1,2$$

Two populations: drift, migration, mutation



Time t

Population 1

Population 2

Drift

N_1

Migration

Drift

N_2

m_1

m_2

Time $t+1$

Population 1

Population 2

There is also a probability μ that an allele mutates to a new type.

Aside: Drift, Mutation and Migration

It is possible to predict the values of $\theta^i, \theta^{ii'}$ and, therefore, the values of $\beta^i = (\theta^i - \theta^B)/(1-\theta^B)$.

For two populations, although $\theta^1, \theta^2, \theta^{12}$ are all non-negative probabilities, it is possible that both of $\beta^1 = (\theta^1 - \theta^{12})/(1-\theta^{12})$ and $\beta^2 = (\theta^2 - \theta^{12})/(1-\theta^{12})$ are positive, or that one of them is negative and the other one positive. The average $(\beta^1 + \beta^2)/2$ is non-negative.

Aside: Drift, Mutation and Migration

For populations 1 or 2 with sizes N_1 or N_2, if m_1 or m_2 are the proportions of alleles from population 2 or 1, the changes in the θ's from generation t to $t+1$ are

$$\begin{aligned}
\theta^1(t+1) &= (1-\mu)^2\Big[(1-m_1)^2\phi^1(t) + 2m_1(1-m_1)\theta^{12}(t) \\
&\quad +m_1^2\phi^2(t)\Big] \\
\theta^2(t+1) &= (1-\mu)^2\Big[m_2^2\phi^1(t) + 2m_2(1-m_2)\theta^{12}(t) \\
&\quad +(1-m_2)^2\phi^2(t)\Big] \\
\theta^{12}(t+1) &= (1-\mu)^2\Big[(1-m_1)m_2\phi^1(t) + [(1-m_1)(1-m_2) \\
&\quad + m_1m_2]\theta^{12}(t) + m_1(1-m_2)\phi^2(t)\Big]
\end{aligned}$$

where $\phi^i(t) = 1/(2N_i) + (2N_i - 1)\theta^i(t)/(2N_i)$ and μ is the infinite-allele mutation rate.

Drift and Mutation

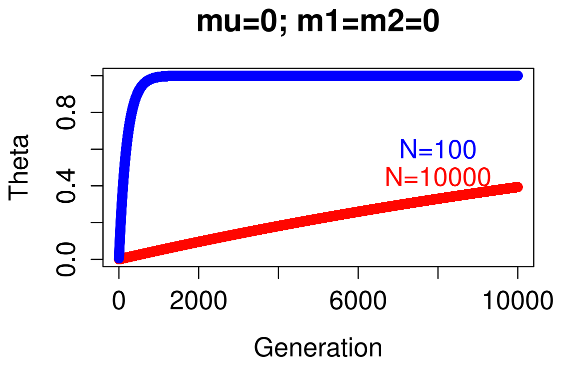
If there is no migration, the θ's tend to equilibrium values of

$$\begin{aligned}
\widehat{\theta}^1 &\approx \frac{1}{1+4N_1\mu} \\
\widehat{\theta}^2 &\approx \frac{1}{1+4N_2\mu} \\
\widehat{\theta}^{12} &= 0
\end{aligned}$$

so $\beta^i = \theta^i,\ i = 1, 2.$

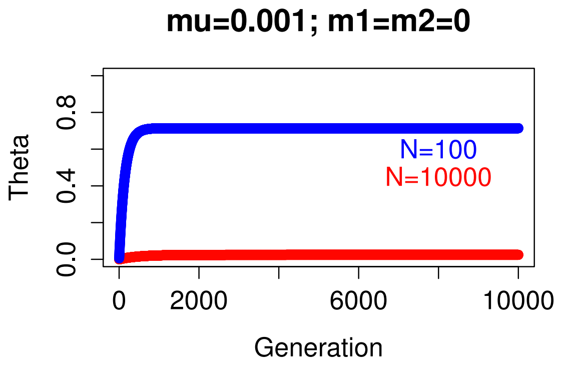
Drift, Mutation and Migration

The θ's are non-negative, but one of the β's may be negative.



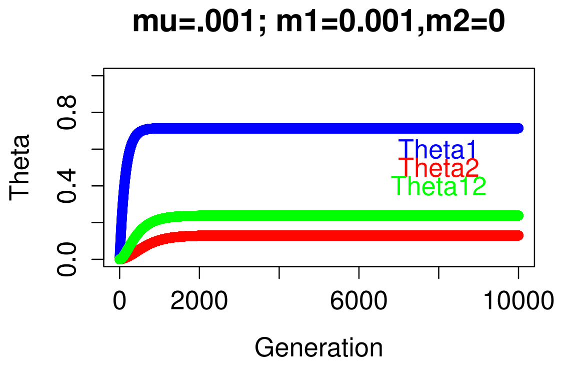
Drift Only

$\beta^1, \beta^2 > 0$



Drift and Mutation

$\beta^1, \beta^2 > 0$



Drift, Mutation and Migration

$\beta^1 > 0, \beta^2 < 0$

Multiple Populations

For random union of gametes, when pairing of alleles into individuals is not needed, the ibd probability θ_W^i for any distinct pair of alleles within population i *relative to* the ibd probability between populations is

$$\beta_{WT}^i = \frac{\theta_W^i - \theta_B}{1 - \theta_B}$$

This is the population-specific F_{WT}^i for alleles.

Averaging over populations:

$$\beta_{WT} = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

and this is the global F_{WT} for alleles. This is the quantity often referred to as "F_{ST}", but see later discussion in Kinship section.

Genotypes vs Alleles

So far, this treatment has ignored individual genotypic structure, leading to an analysis of population allele frequencies as opposed to genotypic frequencies.

θ^i is the probability two alleles drawn randomly from population i are ibd, and $\theta^{ii'}$ is the probability an allele drawn randomly from population i is ibd to an allele drawn from population i'.

Within population i, define θ^i_{jj} as the probability that two alleles drawn randomly from individual j are ibd, and $\theta^i_{jj'}$ as the probability that allele drawn randomly from individual j is ibd to an allele from individual j'.