Logistics

Reminder: Reading for Week 4 is Hahn pp. 79-93

Next Quiz: is Wednesday (September 28) covering LD/recombination (last week’s lecture material)

Assignment 1 due: Thursday October 6 at midnight.

Definitions

Population: a group of freely interbreeding individuals

Subpopulation: typically used interchangeably with population

Panmixia: random mating within a population

Population structure: The outcome of population differentiation (due primarily to low levels of migration and genetic drift)

Gene flow: The exchange of alleles between populations

How do we define populations in practice and measure differences between them?

Hardy-Weinberg Equilibrium (HWE)

If we assume random mating no selection, no drift, no mutation, and no migration then we can calculate expected genotype frequencies from allele frequencies

Expectations are derived from the “random union of gametes”

Example: biallelic locus with A and a alleles

If we denote the frequency of A as p and a as q, then we can derive expected genotype frequencies at HWE

$E(p_{AA}) = p^2$

$E(p_{Aa}) = 2pq$

$E(p_{aa}) = q^2$

where $E(p_{AA})$, $E(p_{Aa})$, $E(p_{aa})$ are the expected genotype frequencies at HWE

Hardy-Weinberg Equilibrium (HWE)

If we assume random mating no selection, no drift, no mutation, and no migration then we can calculate expected genotype frequencies from allele frequencies

Expectations are derived from the “random union of gametes”

A locus is at HWE when genotype frequencies match expectations under the random union of gametes

Example: biallelic locus with A and a alleles

If we denote the frequency of A as p and a as q, then we can derive expected genotype frequencies at HWE

$E(p_{AA}) = p^2$

$E(p_{Aa}) = 2pq$

$E(p_{aa}) = q^2$

where $E(p_{AA})$, $E(p_{Aa})$, $E(p_{aa})$ are the expected genotype frequencies at HWE

The Wahlund effect

Population structure is a form of nonrandom mating that causes deviations from HWE

The Wahlund effect is the deviation from HWE expectations when:

(1) multiple differentiated populations are sampled

(2) expectations of genotype frequency under HWE are derived without knowledge of existing population structure between them

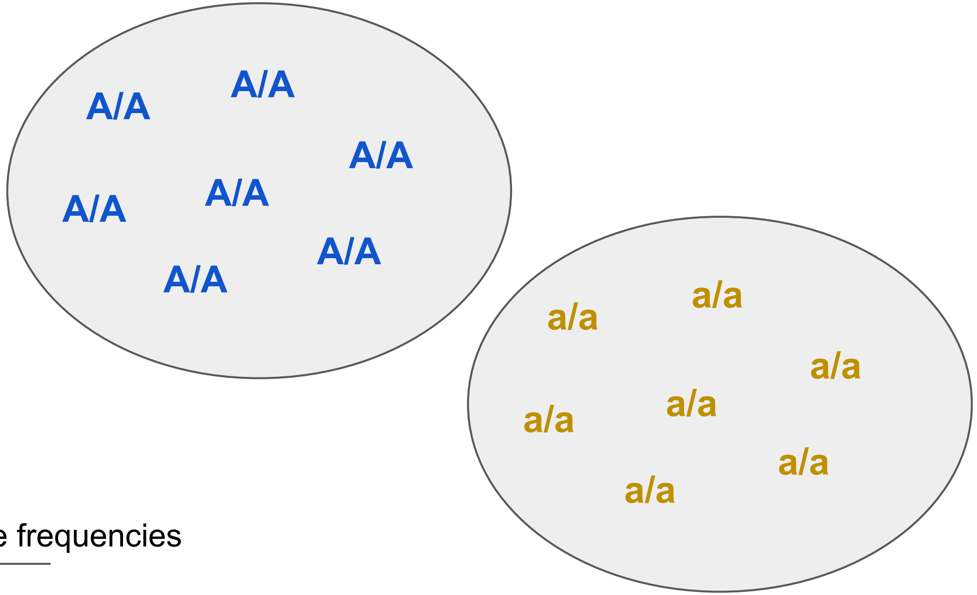
The Wahlund effect

Non random mating

Example: Consider two populations that are fixed for alternate alleles A with frequency p and a with frequency q

All individuals are homozygous for A in population 1 (p = 1, q = 0 in population 1) and a in population 2 (p = 0, q = 1 in population 2)

If we **combine** both populations, global allele frequencies are p = 0.5 and q = 0.5

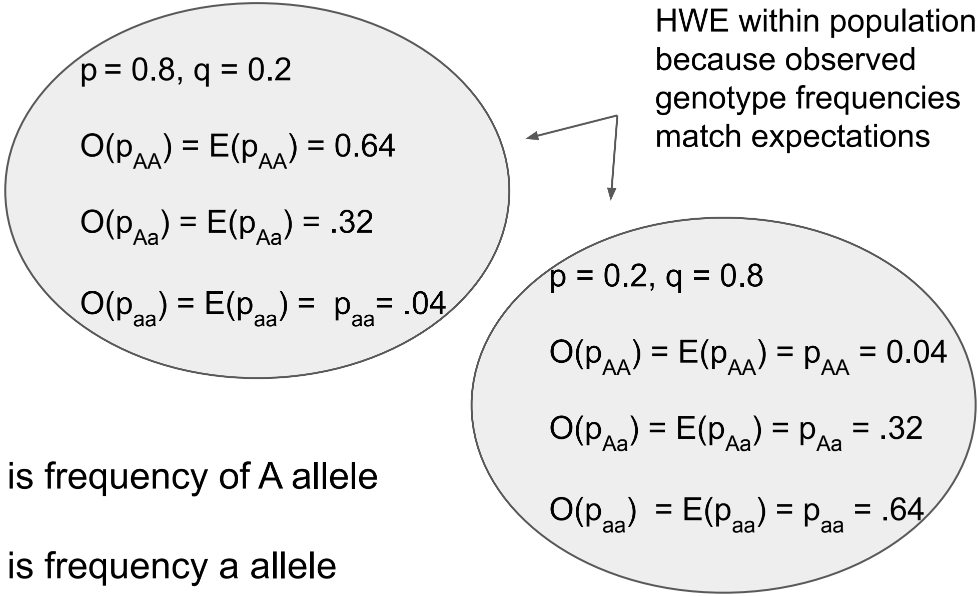


Expected genotype frequencies (under HWE)	Observed genotype frequencies
$E(p_{AA}) = p^2 = 0.25$	$Obs(p_{AA}) = 0.5$
$E(p_{Aa}) = 2pq = 0.5$	$Obs(p_{Aa}) = 0$
$E(p_{aa}) = 0.25$	$Obs(p_{aa}) = 0.5$

Wahlund Effect is observed when combining differentiated populations

The Wahlund effect

imgigration up
Fst down



Two differentiated populations. Both are at HWE within populations because observed genotype frequencies match expectations

$p = 0.8, q = 0.2$

$O(p_{AA}) = E(p_{AA}) = 0.64$

$O(p_{Aa}) = E(p_{Aa}) = .32$

$O(p_{aa}) = E(p_{aa}) = p_{aa} = .04$

$p = 0.2, q = 0.8$

$O(p_{AA}) = E(p_{AA}) = p_{AA} = 0.04$

$O(p_{Aa}) = E(p_{Aa}) = p_{Aa} = .32$

$O(p_{aa}) = E(p_{aa}) = p_{aa} = .64$

p is frequency of A allele

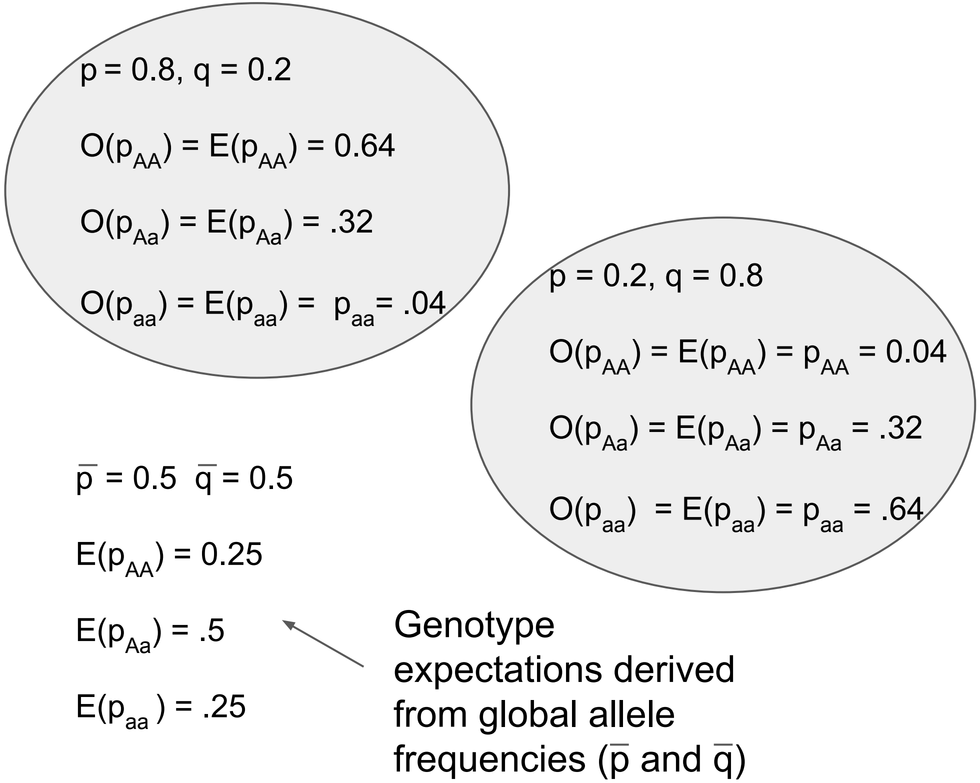
q is frequency a allele

$O(p_{AA}, p_{Aa}, p_{aa})$ are observed genotype frequencies

$E(p_{AA}, p_{Aa}, p_{aa})$ are expected genotype frequencies

The Wahlund effect

Expected 0.5
Observed 0.32



$p = 0.8$, $q = 0.2$

$O(p_{AA}) = E(p_{AA}) = 0.64$

$O(p_{Aa}) = E(p_{Aa}) = .32$

$O(p_{aa}) = E(p_{aa}) = \ p_{aa} = .04$

$p = 0.2$, $q = 0.8$

$O(p_{AA}) = E(p_{AA}) = p_{AA} = 0.04$

$O(p_{Aa}) = E(p_{Aa}) = p_{Aa} = .32$

$O(p_{aa}) \ = E(p_{aa}) = p_{aa} = .64$

$\bar{p} = 0.5 \ \ \bar{q} = 0.5$

$E(p_{AA}) = 0.25$

$E(p_{Aa}) = .5$

$E(p_{aa}) = .25$

Genotype expectations derived from global allele frequencies ($\bar{p}$ and $\bar{q}$)

Note: for this example, assume equal sample sizes for the two populations

Variance in allele frequencies

The variance in allele frequencies among populations (σ^2) is a natural way to quantify the Wahlund effect and population differentiation

The variance in allele frequencies (σ^2) is the deviation in allele frequencies of sample populations from the global mean (e.g., $\overline{p}$)

Key point: Higher variance in allele frequencies between populations (i.e., the greater the frequencies differ from the mean, p), the greater the deficit of heterozygotes (i.e., the stronger the Wahlund effect)

The mean allele frequency across populations is defined as:

$$\overline{p} = \frac{\sum_{i=i}^{n} p_i}{n}$$

where p_i is the frequency of the A allele in population *i*
n is the number of populations

The variance in allele frequency is defined as:

$$\sigma^2 = \sigma_p{}^2 = \sigma_q{}^2 = \frac{\sum (p_i - \overline{p})^2}{n} = \frac{\sum p_i{}^2}{n} - \overline{p}^2$$

Variance in allele frequencies as a measure of the Wahlund effect

Here we show how expectations for genotype frequencies at HWE are related to the variance in allele frequencies (σ^2)

$$E(p_{AA}) = \frac{\sum p_i^2}{n} = \bar{p}^2 + \sigma^2$$

$$E(p_{Aa}) = \frac{\sum 2 p_i q_i}{n} = 2\left(\frac{\sum p_i}{n} - \frac{\sum p_i^2}{n}\right) = 2\overline{pq} - 2\sigma^2$$

$$E(p_{aa}) = \bar{q}^2 + \sigma^2$$

$$E(p_{AA}) = \bar{p}^2 + \sigma^2$$

$$E(p_{Aa}) = 2\bar{p}\bar{q} - 2\sigma^2$$

$$E(p_{aa}) = \bar{q}^2 + \sigma^2$$

If σ^2 is 0, then genotype frequencies are at HWE.

Variance in allele frequencies is a measure of the Wahlund effect

Key point: the higher the variance in allele frequencies, the greater the deficit in heterozygotes (i.e., the greater the Wahlund effect)

p_1	q_1	p_2	q_2	p	q	σ^2	$E(p_{Aa})$	$O(p_{Aa})$
1	0	0	1	0.5	0.5	.25	0.5	0
0.8	0.2	0.2	0.8	0.5	0.5	.09	0.5	0.32
0.6	0.4	0.4	0.6	0.5	0.5			

Measuring population differentiation: F_{ST}

σ^2 might be a good measure of population differentiation, except that the variance depends on the allele frequency

Alleles at high frequency (e.g. p = 0.5) will have greater variance (σ^2) in frequency across populations than low frequency alleles (e.g. p = 0.01)

Therefore, we define F_{ST}, which is the σ^2 normalized by the average allele frequencies

capture wahlund effect

$$F_{ST} = \frac{\sigma^2}{\bar{p}\,\bar{q}}$$

difference of two groups up
Fst up

$$E(p_{AA}) = \bar{p}^2 + \overline{pq}F_{ST}$$

$$E(p_{Aa}) = 2\bar{p}\bar{q} - 2\overline{pq}F_{ST}$$

$$E(p_{aa}) = \bar{q}^2 + \overline{pq}F_{ST}$$

If $F_{ST} = 0$ then genotypes are at HWE, if $F_{ST} = 1$ then zero heterozygotes

Alternate approaches to calculating F_{ST}

Formulae to calculate F_{ST} can also be derived in terms of expected heterozygosity

$$G_{ST} = \frac{H_T - \overline{H}_S}{H_T} = 1 - \frac{\overline{H}_S}{H_T}$$

H_T is expected heterozygosity combining populations

H_S is the average expected heterozygosity within subpopulations

Interpreting F_{ST}

F_{ST} varies from 0 (=no differentiation) to 1 (=complete differentiation)

Wright (1978) suggested how to interpret levels of differentiation (see table)

F_{ST} is a **relative measure of differentiation** because it is strongly influenced by within population diversity (i.e., it is inflated when within population diversity is low)

As a result, F_{ST} will also vary between regions of the genome (e.g., low diversity regions will have higher F_{ST})

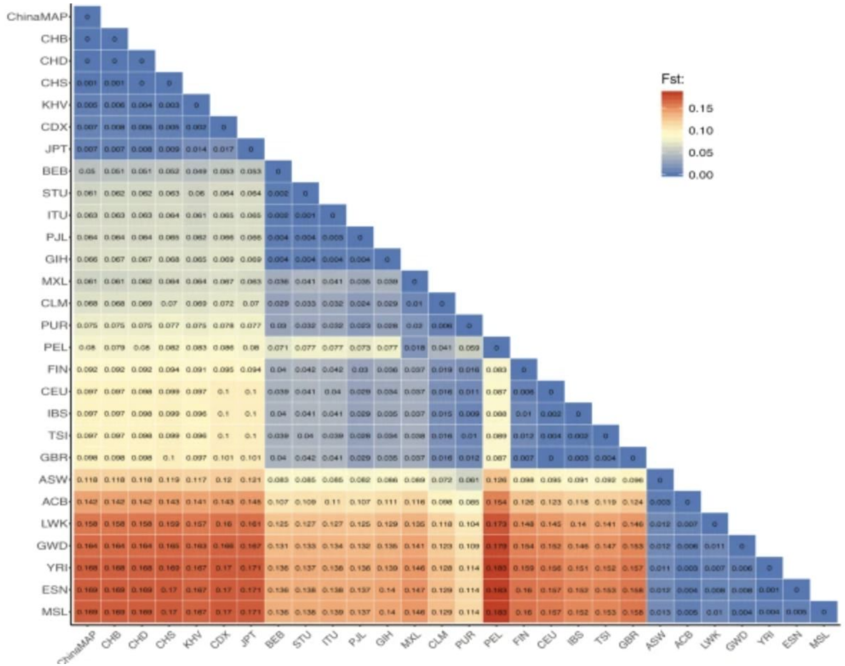
F_{ST} range	Interpretation (Wright 1978)
0.05 to 0.15	“Moderate differentiation”
0.15 to 0.25	“Great genetic differentiation”
>0.25	“Verg great genetic differentiation”

Example: F_{ST} between worldwide populations

Pairwise F_{ST} between populations and averaged across entire genome

ChinaMAP samples are most differentiated from YRI and other African populations and least from CHB (Han) and other Asian populaitons

Highest F_{ST} = 0.18 is between PEL (Peruvians in Lima) vs. YRI (Yorubans)



Cao et al. (2020) Cell Research

Example: F_{ST} of a risk allele for diffuse-type gastric cancer

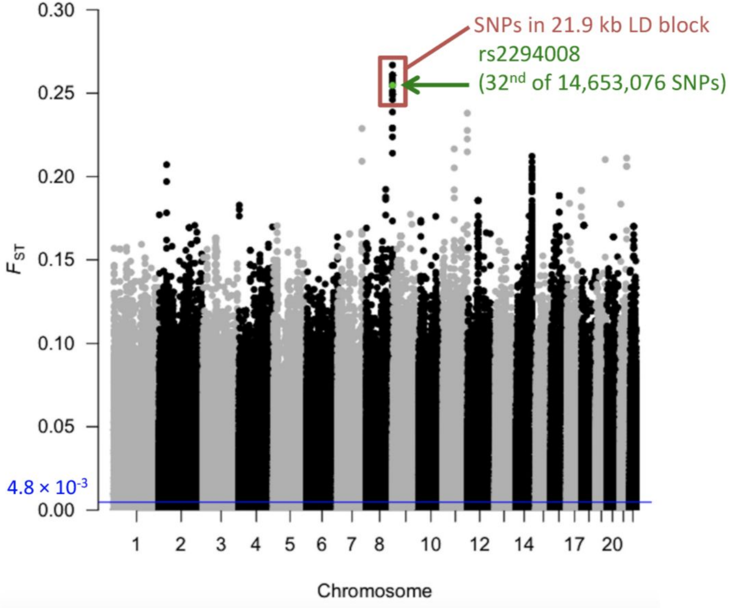
F_{ST} is frequently used to identify unusually differentiated regions of the genome

rs2294008 is SNP in the *PSCA* gene and is a risk allele for diffuse-type gastric cancer

SNP is an ATG->ACG missense change at first codon position (effect is “start lost”)

C allele reaches its highest frequency in Japanese (>0.6) in 1000 Genomes Project

F_{ST} ~ 0.26 at rs2294008 between Japanese and Han Chinese



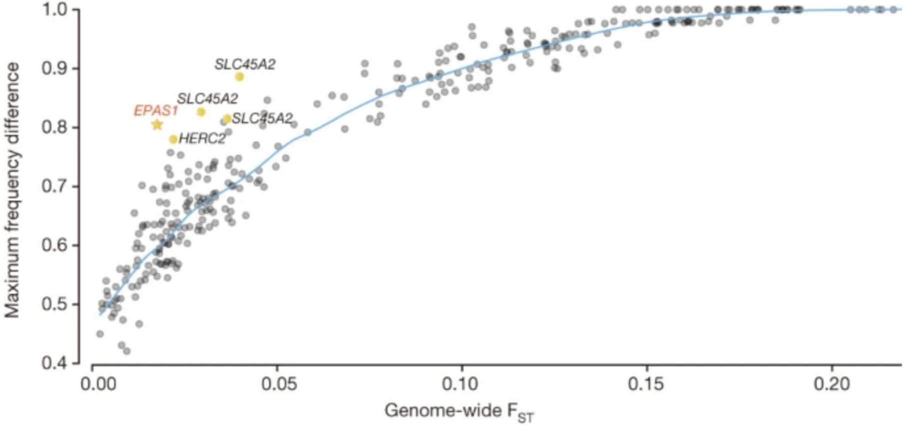
Iwasaki et al. (2020) Genes

Example: Use of F_{ST} to identify locally adapted loci

X-axis is genome-wide F_{ST} for pairs of human populations

Y-axis is the highest frequency difference observed in a genomewide sample of each population pair

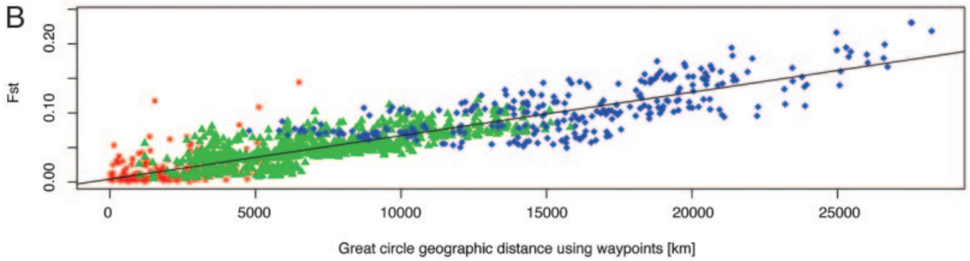
Loci with large allele frequency differences between otherwise undifferentiated populations is a signature of local adaptation



Huerta-Sanchez et al. (2014) Nature

Example: F_{ST} increases with geographic distance between populations

At migration-drift equilibrium, theory predicts a correlation between population differentiation and geographic distance.



Red points represent F_{ST} between populations within the same region

Green points represent F_{ST} between African vs. Eurasian populations

Blue points are comparisons between Native American and Oceania

Ramachandran et al. (2005) PNAS

Confusion surrounding F_{ST} measures and how to estimate it

There is considerable confusion concerning F_{ST}

Please read:

Bhatia et al. (2013) Genome Research

Jost (2008) Molecular Ecology

Alternate measures of population differentiation

A common alternative to quantifying population differentiation is d_{XY}

d_{XY} is the average pairwise differences between each chromosome in population X and each chromosome in population Y

$$d_{XY} = \sum_{ij} x_i y_j d_{ij}$$

Where x_i is the frequency of the ith haplotype in population X

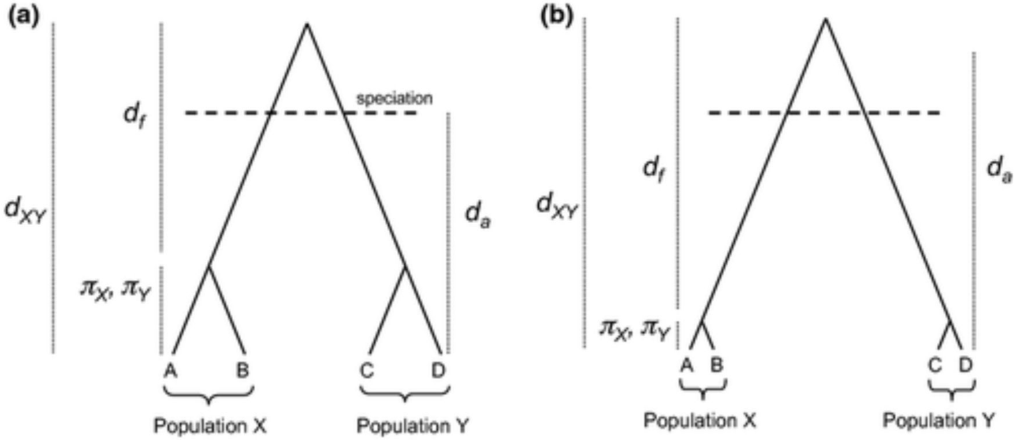
y_j is the frequency of the jth haplotype in Y

d_{ij} is the pairwise distance between haplotype i and j

Alternate measures of population differentiation

d_{XY} is an absolute measure of differentiation because it does not depend on within population diversity

This can be seen visually by comparing panel (a) which shows genealogy with high diversity ("pi") within populations with (b) which has low diversity ("pi") within populations



Hahn and Cruickshank (2014) Molecular Ecology

Is there statistical support for population differentiation?

Chi-square contingency test of independence of allele counts between two populations

Permutation-type tests

Logistics

Reading for Week 4 is Hahn pp. 79-93

Next Quiz: Wednesday 10/12/2022 12:30 - 1:45 pm (covers Week 4 and Week 5)

Assignment 1 due: Thursday October 6 at midnight.

Alternate approaches to calculating F_{ST}

Formulae to calculate F_{ST} can also be derived in terms of expected heterozygosity

$$G_{ST} = \frac{H_T - \overline{H_S}}{H_T} = \frac{1 - \overline{H_S}}{H_T}$$

H_T is expected heterozygosity combining populations

$\overline{H_S}$ is the average expected heterozygosity within subpopulations

Alternate approaches to calculating F_{ST}

Formulae to calculate F_{ST} can also be derived in terms of expected heterozygosity

FST originally derived for bi-allelic loci

G_{ST} is a reformulation of FST that accomodates multi-allelic (such as microsatellite markers i.e., short tandem repeats)

This formulation calculates FST in terms of the proportion of heterozygosity within populations

If mean $H_S = H_T$ then $F_{ST} = 0$, but if mean $H_S < H_T$ than there is a deficit of heterozygotes attributable to Wahlund Effect

$$F_{ST} = G_{ST} = \frac{H_T - \overline{H}_S}{H_T} = \frac{1 - \overline{H}_S}{H_T}$$

H_T is expected heterozygosity combining populations

$\overline{H}_S$ is the average expected heterozygosity within subpopulations

Example: F_{ST} of a risk allele for diffuse-type gastric cancer

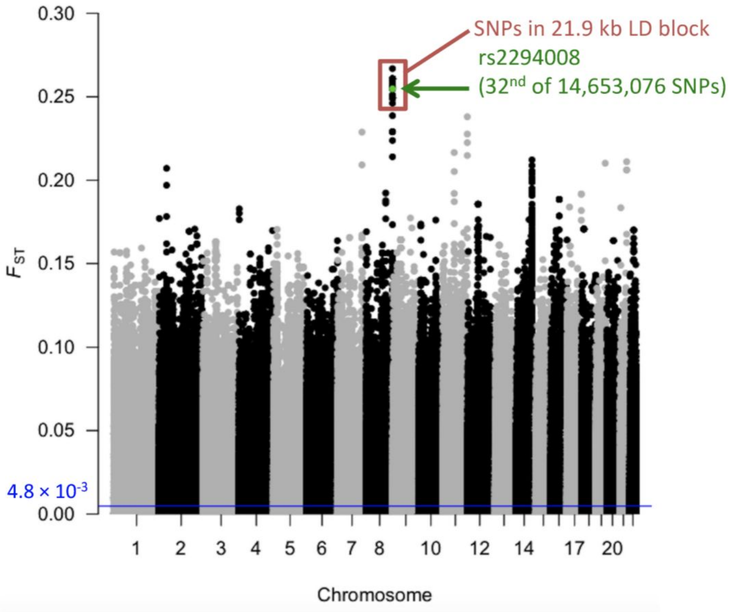
F_{ST} is frequently used to identify unusually differentiated regions of the genome

rs2294008 is SNP in the *PSCA* gene and is a risk allele for diffuse-type gastric cancer

SNP is an ATG->ACG missense change at first codon position (effect is “start lost”)

C allele reaches its highest frequency in Japanese (>0.6) in 1000 Genomes Project

F_{ST} ~ 0.26 at rs2294008 between Japanese and Han Chinese



Iwasaki et al. (2020) Genes

Interpreting F_{ST}

F_{ST} varies from 0 (=no differentiation) to 1 (=complete differentiation)

Wright (1978) suggested how to interpret levels of differentiation (see table)

F_{ST} is a **relative measure of differentiation** because it is strongly influenced by within population diversity (i.e., it is inflated when within population diversity is low)

As a result, F_{ST} will also vary between regions of the genome (e.g., low diversity regions will have higher F_{ST})

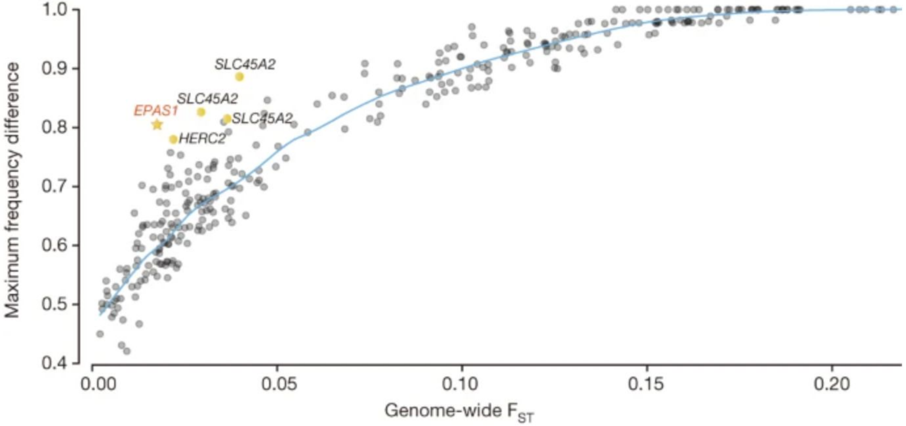
F_{ST} range	Interpretation (Wright 1978)
0.05 to 0.15	“Moderate differentiation”
0.15 to 0.25	“Great genetic differentiation”
>0.25	“Verg great genetic differentiation”

Example: Use of F_{ST} to identify locally adapted loci

X-axis is genome-wide F_{ST} for pairs of human populations

Y-axis is the highest frequency difference observed in a genomewide sample of each population pair

Loci with large allele frequency differences between otherwise undifferentiated populations is a signature of local adaptation



Huerta-Sanchez et al. (2014) Nature

Confusion surrounding F_{ST} measures and how to estimate it

There is considerable confusion concerning F_{ST}

Please read:

Bhatia et al. (2013) Genome Research

Jost (2008) Molecular Ecology

Alternate measures of population differentiation

A common alternative to quantifying population differentiation is d_{XY}

d_{XY} is the average pairwise differences between each chromosome in population X and each chromosome in population Y

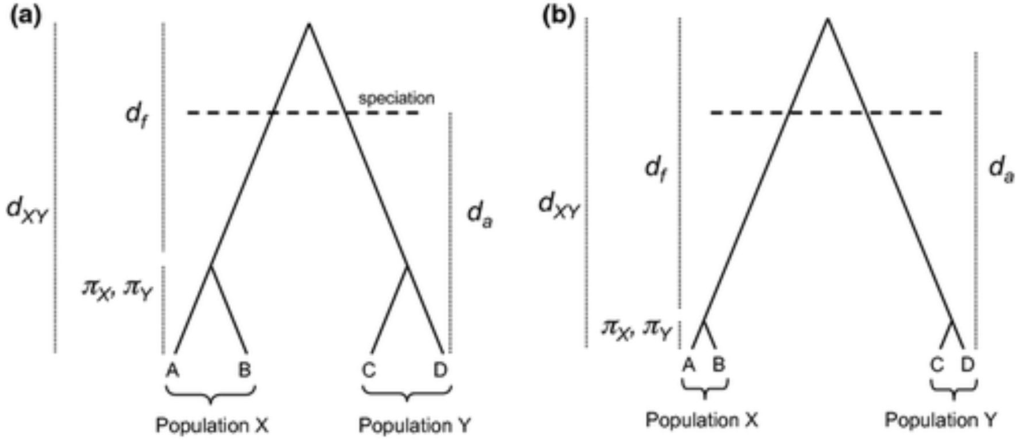
$$d_{XY} = \sum_{ij} x_i y_j d_{ij}$$

Where x_i is the frequency of the ith haplotype in population X
y_j is the frequency of the jth haplotype in Y
d_{ij} is the pairwise distance between haplotype i and j

Alternate measures of population differentiation

d_{XY} is an absolute measure of differentiation because it does not depend on within population diversity

This can be seen visually by comparing panel (a) which shows genealogy with high diversity ("pi") within populations with (b) which has low diversity ("pi") within populations



Hahn and Cruickshank (2014) Molecular Ecology

Is there statistical support for population differentiation?

Chi-square contingency test of independence of allele counts between two populations

Permutation-type tests