Diversity in Genetics

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

jan.graffelman@upc.edu

March 15, 2018

Contents

1. Genetic markers
2. Genetic equilibria
3. Genetic diversity

Genetic markers

- The study of diversity in Genetics is possible due to the existence of among others, "positions" in the genome that show variability over individuals.
- A "position" that shows variation from one individual to another is called a **marker**.
- A variety of genetic markers exists
    - Classical markers observable as phenotypes (e.g. blood group polymorphisms: MN locus, ABO locus, Rhesus factor)
    - Allozymes. Charge differences between variants of proteins that can be separated by electrophoresis.
    - Restriction Fragment Length Polymorphism (RFLP). Specific short DNA sequences that are recognized and cut by restriction enzymes.
    - Short Tandem Repeats (STRs, also called microsatellites). Short DNA motifs of a few base pairs that consecutively repeat a number of times.
    - Single Nucleotide Polymorphism (SNP). Variation in one specific position in the DNA sequence.
    - Copy Number Variation (CNV). Variation in the number of copies of a genetic due to duplication.
    - ....

Information on (multiple) markers is registered for a set of individuals in a database.

A bit of terminology

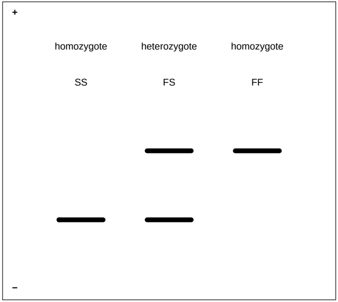
- A chromosome is a structure in the nucleus of a cell, made of protein and DNA. Chromosomes come in homologous pairs. A human being has 23 pairs of chromosomes.
- An allele is one of the alternative forms of a gene or marker.
- A locus refers to the position of a gene or marker on a specific chromosome.
- Diploid individuals have two copies of the genome, one from their father and one from the mother.
- An individual with two identical alleles for a gene or marker is called homozygote.
- An individual with two different alleles for a gene or marker is called heterozygote.
- Each individual has a particular genotype. The genotype is the allelic makeup of an individual, consisting of the two alleles (one from the father and one from the mother) the indiviual has at a particular locus.
- Diploid individuals produces reproductive cells by meiosis. The reproductive celss (egg cells, sperm cells) contain only one copy of the genome and are therefore haploid.
- At a locus with k alleles, there exit $\frac{1}{2}k(k+1)$ genotypes.

Allozymes

Allozymes are variants of an enzyme, that arise as a consequence of the existence of different alleles at a locus.

```
       Allele 1

DNA    ...AGC CTA...
       ...TCG GAT...

Enzyme Val-Phe-Tyr-Leu

       Allele 2

DNA    ...AGA CTA...
       ...TCT GAT...

Enzyme Val-Phe-Glu-Leu
```

Banding pattern in a gel after electrophoresis:



Restriction Fragment Length Polymorphism

- Restriction enzymes are enzymes capable of cutting DNA strands at a specific recognition sequence. E.g.

BamHI	GGATCC	EcoRI	GAATTC
	CCTAGG		CTTAAG

- DNA sequences of interest are digested with a restriction enzymes. The resulting segments will vary in size and can be made visible by electrophoresis.
- Gives rise to absence/presence data. Individuals have or do not have the recognition sequence.

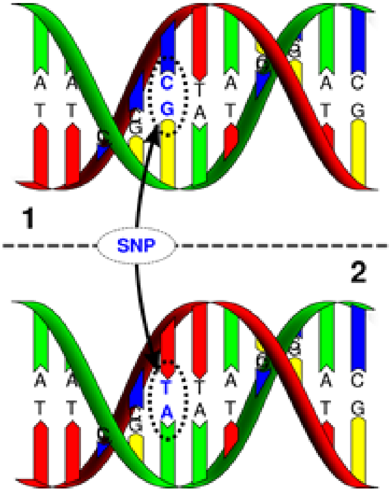
Microsatellites or STRs (Short Tandem Repeat)

- Microsatellites consist of short sequences (e.g. ATT) that repeat a certain number of times (e.g. ATTATTATTATT).
- Individuals vary in the number of repeats they have.
- Individuals can have the same number of repeats on each homologous chromosome (homozygote), or a different number (heterozygote).
- Produces count data, with a limited number of outcomes.
- Data for an individual are pairs of integers.

A glance at a STR database

	Id	STR1	STR2	STR3	STR4	STR5	STR6	STR7	STR8	STR9	· · ·
1	794	129	264	142	156	157	171	205	183	196	· · ·
2	794	155	292	146	156	166	179	205	187	196	· · ·
3	795	145	288	138	168	157	171	205	195	196	· · ·
4	795	150	292	142	172	166	175	210	203	196	· · ·
5	796	155	292	138	156	157	167	205	183	184	· · ·
6	796	155	300	142	156	169	171	205	199	196	· · ·
7	797	150	264	142	156	157	171	205	187	196	· · ·
8	797	155	292	146	176	163	175	205	187	196	· · ·
9	798	150	292	138	156	157	171	205	183	187	· · ·
10	798	155	300	146	160	166	171	205	207	190	· · ·
11	799	155	296	146	152	157	167	205	179	196	· · ·
12	799	155	296	146	176	157	171	210	183	196	· · ·
13	800	145	264	138	156	157	163	205	187	190	· · ·
14	800	160	296	146	156	157	171	210	199	196	· · ·
15	801	155	264	142	156	157	175	205	183	196	· · ·
16	801	155	292	146	184	166	179	209	199	199	· · ·
17	802	145	292	138	176	157	159	193	183	187	· · ·
18	802	155	296	142	180	166	171	201	187	187	· · ·
19	803	155	280	142	172	166	163	205	183	196	· · ·
20	803	155	300	142	176	169	175	213	187	196	· · ·
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Single Nucleotide Polymorphisms (SNPs, "snips")



- One DNA strand is the reference strand
- The depicted SNP is a C/T polymorphism
- The depicted individual is heterozygote
- The possible genotypes are C/C, C/T and T/T.

Single nucleotide polymorphisms (SNPs)

- Variation in one position in the DNA strand.
- Typically bi-allelic (e.g. (A/A, A/T, T/T).
- Number of SNPs in the human genome:
  - 3.1 million SNPs (2007, www.hapmap.org)
  - 15 million SNPs (2010, www.1000genomes.org)
  - ...
- Has become the most popular genetic marker.

Single nucleotide polymorphisms (SNPs)

A "typical" data set of SNPs:

Id	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	···
NA18524	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18526	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18529	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18532	CC	CC	TT	TT	TT	AC	CG	AC	CT	GG	···
NA18537	CC	CC	TT	TT	AT	CC	CC	AC	CT	GG	···
NA18540	CC	CC	CT	TT	AT	CC	CG	AC	CT	GG	···
NA18542	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18545	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18547	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18550	CC	CC	CT	TT	AT	CC	CC	AC	CT	GG	···
NA18552	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18555	CC	CC	TT	TT	TT	CC	CG	AC	CT	GG	···
NA18558	CC	NN	CC	TT	TT	CC	CG	CC	CT	GG	···
NA18561	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18562	CC	CC	TT	TT	AT	AC	CG	AC	CT	GG	···
NA18563	CC	CC	CT	TT	AA	CC	CC	AA	CT	GG	···
NA18564	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18566	CC	CC	TT	TT	TT	AC	CC	AC	CT	GG	···
NA18570	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
NA18571	CC	CC	TT	TT	AT	AC	CC	AC	CT	GG	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

SNP data is multivariate categorical data.

Haplotypes

- A haplotype is combination of alleles on consecutive loci that are inherited as a block.
- A haplotype may span a short or a large stretch of DNA, depending on how much recombination occurs.
- If data are collected on genotypes, then haplotypes are unknown, but can be estimated.
- The use of haplotypes allows a reduction in the number of variables.

Descriptive statistics for genetic markers

- A genetic marker that shows no variation in a particular sample is called fixed or monomorphic.
- Genetic markers are basically categorical variables.
- Genetic markers are described by the genotype frequencies and the allele frequencies.
- The minor allele frequency (MAF) is the minimum of all allele frequencies for a given locus.

Descriptive statistics of a SNP

- Categorical variable, usually 3 categories.
- Allele counts or frequencies reported (n_A, n_B).
- Minor Allele Frequency (MAF).
- Genotype counts or frequencies reported (n_{AA}, n_{AB}, n_{BB}).
- Observed and Expected Heterozygosity (H_o and H_e).

Example:

rs5939319		
AA	477	0.733
AB	164	0.252
BB	10	0.015
NA	1	-

rs5939319		
A	1118	0.859
B	184	0.141

$\text{MAF} = \min(0.859, 0.141) = 0.141$

$$H_o = 0.252 \quad H_e = 2 \cdot 0.859 \cdot 0.141 = 0.242$$

Example with R

Of 1000 individuals their MN blood group has been determined as MM,MN or NN, obtaining the respective counts (298,489,213). We calculate genotype and allele frequencies.

```
> x <- c(MM=298,MN=489,NN=213)
> N <- sum(x)
> x/N
   MM    MN    NN
0.298 0.489 0.213
> fM <- (2*x[1]+x[2])/(2*N)
> fM
    MM
0.5425
> fN <- (2*x[3]+x[2])/(2*N)
> fN
    NN
0.4575
> 1-fM
    MM
0.4575
> min(fM,fN)
0.4575
```

Marker description with the R-package genetics

```
> install.packages("genetics")
> library(genetics)
> marker <- c(rep("M/M",298),rep("M/N",489),rep("N/N",213))
> table(marker)
marker
M/M M/N N/N
298 489 213
> marker.geno <- genotype(marker)
> summary(marker.geno)

Number of samples typed: 1000 (100%)

Allele Frequency: (2 alleles)
  Count Proportion
M  1085       0.54
N   915       0.46


Genotype Frequency:
    Count Proportion
M/M   298       0.30
M/N   489       0.49
N/N   213       0.21

Heterozygosity (Hu)  = 0.4966358
Poly. Inf. Content   = 0.3731872
```

Genetic equilibria

- Within one marker: Hardy-Weinberg equilibrium
- Between markers: Linkage equilibrium

Hardy-Weinberg equilibrium

- A biological population of n individuals.
- A bi-allelic genetic marker.
- One locus with alleles A and B, frequencies p and q.
- Three genotypes AA, AB, BB frequencies f_{AA}, f_{AB} and f_{BB}.

			♀	
			p	q
			A	B
♂	p	A	p^2	pq
	q	B	pq	q^2

f_{AA}	f_{AB}	f_{BB}
p^2	$2pq$	q^2

- Equilibrium achieved in one generation.
- Note that the allele frequency of A in the new generation is $p^2 + pq = p(p + q) = p$.

Hardy-Weinberg assumptions

- The organism under study is diploid.
- There is sexual reproduction.
- Non-overlapping generations.
- Random mating (w.r.t the trait under study).
- Population size is very large.
- Migration is negligible.
- Mutation can be ignored.
- Natural selection does not affect the trait under study.
- There is no genotyping error.

Basic law

- Genetic markers are, in general, expected to follow the HW law.
- If they do not follow the law, one (or more) of the HWE assumptions is/are violated.
- The most likely cause for disequilibrium is genotyping error.
- Markers need to be checked for HWE as part of a quality control procedure.

Hardy-Weinberg Equilibrium

$$\begin{array}{ccc} f_{AA} & f_{AB} & f_{BB} \\ p^2 & 2pq & q^2 \end{array}$$

Alternatively:

$$f_{AB}^2 = 4 \, f_{AA} \, f_{BB}$$

Hardy-Weinberg equilibrium for multiple alleles

If a marker has three alleles (e.g. the bloodgroup system A, B and O), with frequencies p_1, p_2 and p_3 with $p_1 + p_2 + p_3 = 1$, then under random mating we would obtain the genotype frequencies

				♀	
			p_1	p_2	p_3
			A	B	O
	p_1	A	p_1^2	$p_1 p_2$	$p_1 p_3$
♂	p_2	B	$p_2 p_1$	p_2^2	$p_2 p_3$
	p_3	O	$p_3 p_1$	$p_3 p_2$	p_3^2

In general, for a k-alleles system, homozygotes $A_i A_i$ will have frequency p_i^2, and heterozygotes $A_i A_j$ will have frequency $2 p_i p_j$.

Statistical Tests for Hardy-Weinberg Equilibrium

- Classical χ^2 test.
- Exact test (based on $P(N_{AB} \mid N_A)$).
- Likelihood ratio test.
- Non-parametric test.
- Bayesians tests.
- ...

Classical χ^2 test for Hardy-Weinberg equilibrium

- The counts n_{AA}, n_{AB} and n_{BB} are regarded as a sample from a multinomial distribution.
- Expected counts under HWE are np^2, $n2p(1-p)$ and $n(1-p)^2$.
- A chi-square statistic for goodness-of-fit can be used

$$X^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

- The reference distribution is a χ_1^2 distribution.
- If we define the deviation from independence $D = \frac{1}{2}(n_{AB} - e_{AB})$, then

$$X^2 = \frac{D^2}{p^2(1-p)^2 n}$$

Example

- For an A/T polymorphism with counts AA=46, AT=39 and TT=15 we have

$$\hat{p}_A = \frac{2 \cdot 46 + 39}{200} = 0.655$$

- Expected counts under HWE

$$\begin{aligned} e_{AA} &= n\hat{p}_A^2 = 100 \cdot (0.655)^2 = 42.9025 \\ e_{AT} &= 2n\hat{p}_A(1 - \hat{p}_A) = 2 \cdot 100 \cdot 0.655 \cdot 0.345 = 45.195 \\ e_{TT} &= n(1 - \hat{p})^2 = 100 \cdot (0.345)^2 = 11.9025 \end{aligned}$$

- $$X^2 = \frac{(46 - 42.9025)^2}{42.9025} + \frac{(39 - 45.195)^2}{45.195} + \frac{(15 - 11.9025)^2}{11.9025} = 1.8789$$

- $$p - \text{value} = P\left(\chi_1^2 \geq 1.8789\right) = 0.1704601$$

Example in R

```
> library(HardyWeinberg)
> x <- c(46,39,15)
> names(x) <- c("AA","AT","TT")
> results <- HWChisq(x,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  1.878892 p-value =  0.1704601 D =  -3.0975
>
```

Linkage Disequilibrium (LD)

- LD: an association between the alleles at different sites in the genome.
- Maybe a consequence of the physical closeness of the sites, but not necessarily so.
- LD is an important concept in disease-marker association studies.

Linkage Disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE)

- Both concepts refer to association between alleles
- HWE refers to association between alleles at the same locus (within one marker)
- LD refers to association between alleles at different loci (between markers)

Measures of LD

- D
- Lewontin's $D' = \frac{D}{D_{max}}$
- R^2
- χ^2 statistic of a contingency table
- $p-$value in a chi-square test or in an exact test
- ...

Haplotype

- A haplotype is a combination of alleles at adjacent loci on a chromosome that are transmitted together to the next generation.
- In practice, a haplotype often refers to a set of SNPs on a single chromosome that are statistically associated.
- A haplotype map of the human genome has been constructed (www.hapmap.org).

LD

- Consider a population of n individuals
- Consider two sites (two bi-allelic markers)
- One marker with alleles A and a, and one marker with alleles B and b.
- Allele frequencies p_A, p_a, p_B and p_b.

		SNP2		
		B	b	
SNP1	A	$p_A p_B$	$p_A p_b$	p_A
	a	$p_a p_B$	$p_a p_b$	p_a
		p_B	p_b	1

Expected probabilities of each haplotype under independence

LD

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

Observed probabilities of each haplotype in presence of LD

$$D = p_{AB} - p_A p_B$$

D'

- D' is an attempt to standardize D.

$$D' = \frac{D}{D_{max}}$$

$$D_{max} = \begin{cases} \min\left(p_A p_b, p_a p_B\right) & D > 0 \quad \text{(coupling)} \\ \min\left(p_A p_B, p_a p_b\right) & D < 0 \quad \text{(repulsion)} \end{cases}$$

- $-1 \leq D' \leq 1$.
- $D' \approx 0$: low LD
- $|D'|$ close to 1 : high LD.

R^2 and χ^2 statistic

- The genotype data can be recoded as indicator data, creating indicators for the carriers of the A and B allele.
- R^2 is the squared correlation between these indicators.
- R^2 is related to the χ^2 statistic of a 2×2 contingency table: $R^2 = \chi^2/n$.
- The χ^2 statistic is related to D

$$R^2 = \chi^2/n = \frac{D^2}{p_A p_B p_a p_b}$$

Computing LD in R

```
> library(genetics)
> X[84,]
NA18524 NA18526 NA18529 NA18532 NA18537 NA18540 NA18542 NA18545 NA18547 NA18550
   "AT"    "AT"    "AT"    "AT"      NA      NA      NA    "AT"    "AT"    "AT"
NA18552 NA18555 NA18558 NA18561 NA18562 NA18563 NA18564 NA18566 NA18570 NA18571
     NA    "AT"      NA    "AT"      NA    "AT"      NA      NA      NA    "TT"
NA18572 NA18573 NA18576 NA18577 NA18579 NA18582 NA18592 NA18593 NA18594 NA18603
     NA    "AT"    "AT"    "AT"      NA    "AT"    "AT"    "AT"    "AT"    "AT"
NA18605 NA18608 NA18609 NA18611 NA18612 NA18620 NA18621 NA18622 NA18623 NA18624
     NA    "AT"    "AT"      NA    "TT"    "AT"    "TT"    "AT"    "TT"    "AT"
NA18632 NA18633 NA18635 NA18636 NA18637
   "AT"      NA    "TT"    "TT"    "AT"
> snp53 <- genotype(X[53,],sep="")
> snp84 <- genotype(X[84,],sep="")
> LD(snp53,snp84)

Pairwise LD
-----------
                    D        D'      Corr
Estimates: 0.04183712 0.2075121 0.1705739

              X^2   P-value  N
LD Test: 1.803919 0.1792395 31
>
```

Computing LD in R

```
summary(snp53)

Number of samples typed: 45 (100%)

Allele Frequency: (2 alleles)
  Count Proportion
T    45        0.5
C    45        0.5


Genotype Frequency:
    Count Proportion
T/T    12       0.27
T/C    21       0.47
C/C    12       0.27

Heterozygosity (Hu)  = 0.505618
Poly. Inf. Content   = 0.375

> p53T <- summary(snp53)$allele.freq[1,2]
> p53C <- summary(snp53)$allele.freq[2,2]
> p53T
[1] 0.5
> p53C
[1] 0.5
>
```

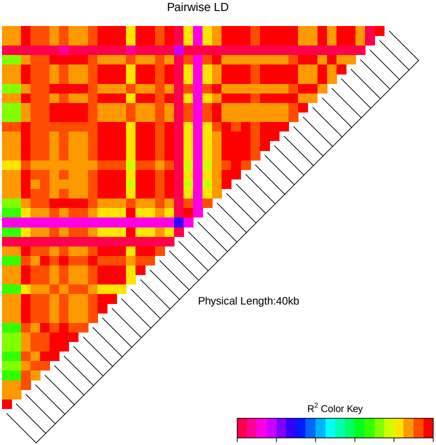
Computing LD in R

```
> summary(snp84)

Number of samples typed: 31 (68.9%)

Allele Frequency: (2 alleles)
   Count Proportion
T     37        0.6
A     25        0.4
NA    28         NA


Genotype Frequency:
    Count Proportion
T/T     6       0.19
T/A    25       0.81
NA     14         NA

Heterozygosity (Hu)  = 0.4891592
Poly. Inf. Content   = 0.3654593
```

Computing LD in R

```
> p84T <- summary(snp84)$allele.freq[1,2]
> p84A <- summary(snp84)$allele.freq[2,2]
> p84T
[1] 0.5967742
> p84A
[1] 0.4032258
>
> out <- LD(snp53,snp84)
> attributes(out)
$names
[1] "call"    "D"       "D’"      "r"       "R^2"     "n"       "X^2"
[8] "P-value"

$class
[1] "LD"

> out$"X^2"
[1] 1.803919
> out$"R^2"
[1] 0.02909546
> out$n
[1] 31
> out$"X^2"/(2*out$n)
[1] 0.02909546
> D <- out$D
> D^2/(p53T*p53C*p84T*p84A)
[1] 0.02909546
> D/min(p53T*p84A,p53C*p84T)
[1] 0.2075121
```

Graphics for LD: the LD heatmap



Genetic diversity

Genetic diversity can be measured within and between biological populations.

- intrapopulation genetic diversity
- interpopulation genetic diversity

Measures of intrapopulation genetic diversity

- Proportion of polymorphic loci.
- Average expected heterozygosity (H_e).
- Allelic richness (A).
- Effective number of alleles (A_e).
- Shannon's index of diversity.
- Simpson's index of diversity.

Proportion of polymorphic loci

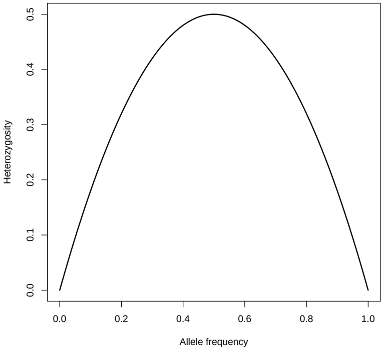
- A locus is considered polymorphic if the frequency of the most common allele is below some threshold, e.g. 0.95 or 0.99.
- $P = \frac{m_p}{M}$ with M the total of loci studied, and m_p the number satisfying the criterion.

Expected Heterozygosity (H_E)or Gene Diversity

For one locus:

$$H_E = 1 - \sum_{i=1}^{I} p_i^2$$

Heterozygosity for a bi−allelic locus



- I is the number of alleles, and p_i the frequency of allele i.
- For a bi-allelic marker, maximum heterozygosity is reached if $p = 0.5$.
- Note this is 1 minus the sum of the frequencies of the homozygotes expected under HWE.
- H_E often estimated by:

$$\hat{H}_E = 1 - \sum_{i=1}^{I} \hat{p}_i^2$$

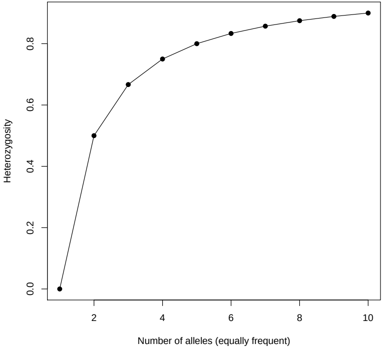
  with $\hat{p}_i^2$ the sample allele frequency of allele i.
- An unbiased estimator for H_E is given by

$$\hat{H}_E = \frac{2 \cdot n \left(1 - \sum_{i=1}^{I} \hat{p}_i^2\right)}{2n - 1}$$

- For multiple loci:

$$H_E = 1 - \frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{I} p_{li}^2$$

Heterozygosity and the number of alleles



- The maximal attainable heterozygosity depends on the number of alleles.
- The maximum is attained when all alleles have equal frequencies:

$$H_E = 1 - \sum_{i=1}^{I} \left(\frac{1}{I} \right)^2 = 1 - I \left(\frac{1}{I} \right)^2 = \frac{I-1}{I}$$

- This should be kept in mind when comparing groups or populations.
- Microsatellites have higher heterozygosities, simply because these markers tend to have more alleles.

Observed Heterozygosity (H_o)

- The observed heterozygosity (H_o) is the fraction of heterozygotes in a sample.
- In general, observed heterozygosity is different from expected heterozygosity.

Allelic richness (A)

- The allelic richness (A) is the mean number of alleles per locus.
- Monomorphic loci (with only one allele) are included in the computation.

$$A = \frac{1}{K} \sum_{i=1}^{K} n_i$$

with n_i the number of alleles at locus i.

Results from an empirical study

Observed and expected heterozygosities in two African elephant populations

Population	n	A	H_o	H_E
Addo park ($N = 325$)	105	1.89	0.192	0.180
Kruger park ($N > 8000$)	108	3.89	0.422	0.444

The Addo park population originated from 16 individuals

Effective number of alleles (A_e)

$$A_e = \frac{1}{1 - H_E} = \frac{1}{\sum_{i=1}^{I} p_i^2}$$

A numerical example

Ind.	SNP1	SNP2	SNP3	SNP4
1	AA	GG	AA	AA
2	AA	GG	AT	AA
3	TT	GG	TT	TT
4	TT	GG	AG	AT
5	AT	GG	GG	AT
# Alleles	2	1	3	2
p_1	0.50	1.00	0.40	0.60
p_2	0.50	-	0.30	0.40
p_3	-	-	0.30	-
H_E	0.56	0.00	0.73	0.53
A_E[a]	2.00	1.00	2.94	1.92

[a] using biased estimator for H_E

Calculating genetic diversity indexes in R

```
> library(gstudio)
> AA <- locus( c("A","A") )
> AT <- locus( c("A","T") )
> TT <- locus( c("T","T") )
> SNP1 <- c(AA,AA,TT,TT,AT)
> GG <- locus( c("G","G") )
> SNP2 <- c(GG,GG,GG,GG,GG)
> AG <- locus( c("A","G") )
> SNP3 <- c(AA,AT,TT,AG,GG)
> SNP4 <- c(AA,AA,TT,AT,AT)
> Population <- c(rep("Pop-A",5))
> df <- data.frame( Population, SNP1=SNP1,SNP2=SNP2,SNP3=SNP3,SNP4=SNP4)
> A  <- genetic_diversity( df, mode="A")
> Ae <- genetic_diversity( df, mode="Ae")
> Ho <- genetic_diversity( df, mode="Ho")
> He <- genetic_diversity( df, mode="He")
> X <- cbind(A[,2],Ae[,2],Ho[,2],He[,2])
> rownames(X) <- A[,1]
> colnames(X) <- c("A","Ae","Ho","He")
> X
     A       Ae  Ho   He
SNP1 2 2.000000 0.2 0.50
SNP2 1 1.000000 0.0 0.00
SNP3 3 2.941176 0.4 0.66
SNP4 2 1.923077 0.4 0.48
```

Shannon's index of diversity

$$H' = -\sum_i p_i \ln\left(p_i\right)$$

- p_i now refers to the frequency of the ith allele.
- H' can be averaged over loci.

Simpson's index of diversity

- Simpson's index of diversity is used at the genotype level, and called Genotypic diversity (D_G).
- $$D_G = 1 - \frac{\sum n_i(n_i - 1)}{N(N-1)}$$

  with n_i the number of individuals of genotype I, and N the total sample size.

Interpopulation diversity: two methods

When samples from several populations are taken, diversity measures can be decomposed into a within population component, and a between population component.

Two sets of statistics are currently being used for this purpose:

- Wright's F statistics
- Nei's gene diversity indices (not covered)

Data layout for interpopulation diversity study

Marker	Population	Genotype frequencies		
		AA	AB	BB
SNP 1	1	p_{111}	p_{112}	p_{113}
	2	p_{121}	p_{122}	p_{123}
	3	p_{131}	p_{132}	p_{133}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J	p_{1J1}	p_{1J2}	p_{1J3}
SNP 2	1	p_{211}	p_{212}	p_{213}
	2	p_{221}	p_{222}	p_{223}
	3	p_{231}	p_{232}	p_{233}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J	p_{2J1}	p_{2J2}	p_{2J3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
SNP m	1	p_{m11}	p_{m12}	p_{m13}
	2	p_{m21}	p_{m22}	p_{m23}
	3	p_{m31}	p_{m32}	p_{m33}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J	p_{mJ1}	p_{mJ2}	p_{mJ3}

Wright's F statistics

Wright's F_{ST} statistic is measure of genetic differentation of a set of populations.

$$0 \leq F_{ST} \leq 1$$

0.00	no genetic divergence between populations
0.00 - 0.05	small genetic differentiation
0.05 - 0.15	moderate genetic differentiation
0.15 - 0.25	large genetic differentiation
0.25 - 1.00	very large genetic differentiation
1.00	populations are fixed for different alleles

Computing F_{ST}

A set of F-statistics is computed on the basis of different Heterozygosites. In particular

$$F_{IT} = 1 - \frac{H_I}{H_T} \quad \text{dearth or excess of average heterozygotes in a group of populations}$$

$$F_{IS} = 1 - \frac{H_I}{H_S} \quad \text{dearth or excess of average heterozygotes in each population}$$

$$F_{ST} = 1 - \frac{H_S}{H_T} \quad \text{degree of differentation of allele frequencies among populations}$$

These F statistics satisfy

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$$

Heterozygosities

The heterozygosities are defined as follows:

- H_T expected heterozygosity using averaged allele frequency (total gene diversity)
- H_I mean of observed heterozygosities over a group of populations (intrapopulation gene diversity)
- H_S mean expected heterozygosity over subpopulations.

Numerical example (1/3)

	AA	AB	BB	p	q	2pq
Pop 1	0.25	0.50	0.25	0.50	0.50	0.50
Pop 2	0.35	0.30	0.35	0.50	0.50	0.50
				0.50	0.50	

Ht	Hi	Hs
0.50	0.40	0.50

Fit	Fis	Fst
0.20	0.20	0.00

- No genetic differentiation
- Deficiency of heterozygotes in population 2

Numerical example (2/3)

	AA	AB	BB	p	q	2pq
Pop 1	0.25	0.50	0.25	0.50	0.50	0.50
Pop 2	0.49	0.42	0.09	0.70	0.30	0.42
				0.60	0.40	

Ht	Hi	Hs
0.48	0.46	0.46

Fit	Fis	Fst
0.04	0.00	0.04

- Small genetic differentiation
- Both populations in HWE ($F_{IS} = 0$)

Numerical example (3/3)

	AA	AB	BB	p	q	2pq
Pop 1	1.00	0.00	0.00	1.00	0.00	0.00
Pop 2	0.00	0.00	1.00	0.00	1.00	0.00
				0.50	0.50	

Ht	Hi	Hs
0.50	0.00	0.00

Fit	Fis	Fst
1.00		1.00

- Maximal genetic differentiation
- Both populations in HWE

Notes

- Wright's F-statistics are not like F-statistics in ANOVA
- When there are multiple markers, these statistics are often averaged

An empirical example: MN Blood groups (1/2)

	MM	MN	NN
Eskimos Greenland	1160	780	141
Europe France	290	518	192
Oceania Micronesia	228	436	298

Blood group counts for the MN locus for 3 human populations

An empirical example: MN Blood groups (2/2)

	AA	AB	BB	p	q	2pq
Pop 1	0.56	0.37	0.07	0.74	0.26	0.38
Pop 2	0.29	0.52	0.19	0.55	0.45	0.50
Pop 3	0.24	0.45	0.31	0.46	0.54	0.50
				0.59	0.41	

Ht	Hi	Hs
0.49	0.45	0.46

Fit	Fis	Fst
0.08	0.02	0.06

Moderate genetic differentiation

Computing F_{ST} in R

```
library(gstudio)
PopInd <- c(rep("ESK",2081),rep("EUR",1000),rep("OCE",962))
MM <- locus( c("M","M") )
MN <- locus( c("M","N") )
NN <- locus( c("N","N") )

ESK <- c(MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,.....
EUR <- c(MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,.....
OCE <- c(MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,MM,.....

 out <- Gst(c(ESK,EUR,OCE), PopInd, nperm=99 )
> out
         Gst        Hs        Ht P
1 0.01883205 0.4577466 0.4665324 0
> FST <- 1- out$Hs/out$Ht
> FST
[1] 0.01883205
```

Software and Bibliography

- R-packages `DEMEtics`, `gstudio`, `genetics`, `HardyWeinberg`.

- Lowe, A., Harris, S. & Ahston, P. (2004) *Ecological genetics: design, analysis and application*, Blackwell Publishing.
- Weir, B. S. (1996) *Genetic Data Analysis II*, Sinauer Associates, Massachusetts.