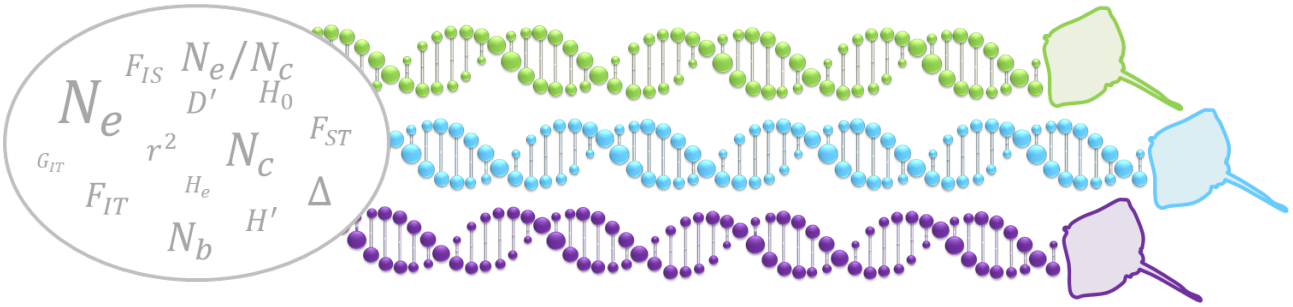


Introduction to population genetics for fisheries scientists

Literature review 2016

Florianne Marandel

Ifremer, Unité Ecologie et Modèles pour l'Halieutique,

Nantes, France

GENO POPTAILLE

ANR

Ifremer

Table of contents

1. Introduction

The raw material of **evolution*** and adaptation to local environments is the genetic variability of individuals in local **population** (Smith and Smith, 2001) which ultimately leads to speciation. Genetic variability can be created by **mutation** and **recombination** but its distribution is defined by evolutive forces such as **migration**, **selection** and **genetic drift**. The pattern of genetic variation can be investigated to understand evolution and adaptation of a species or a population (Lowe et al., 2004; Hamilton, 2009). Such investigations are not recent. Ecological genetics were already a central element of research as soon as 1922 but over the last 30 years, the application of genetics in ecology has increased for a wide variety of biological problems (Dudgeon et al., 2012; Lowe et al., 2004; Portnoy and Heist, 2012). The development of molecular markers, such as SNP (Single Nucleotide Polymorphism) or microsatellites was also fundamental for the spread of genetic analyses. Moreover the increasing affordability of new analyses and the pressing need to address critical conservation and management issues led to the development of genetic analyses in numerous fields including investigations of stock structure and population demography in fisheries. A great virtue of genetic approaches is that a small tissue sample collected from the living or deceased animal at any age contains its complete nuclear and mitochondrial genomic information (Dudgeon et al., 2012). One famous example is the extraction of DNA from frozen woolly mammoths (Smith and Smith, 2001). DNA analysis in general allows the investigation of genetic relationships at several organisational levels (species, population, individual) and can be applied to numerous ecological problematic like natural selection, **mating system** or migration (Smith and Smith, 2001; Lowe et al., 2004; Dudgeon et al., 2012).

Contemporary ecological genetics investigate the origin and maintenance of the genetic variation within and between populations. Population size, **population structure**, the interactions between local selection and genetic drift are some of the main issues studied nowadays (Ryman and Utter, 1987; Lowe et al., 2004; Nikolic, 2009; Beaumont et al., 2010; Dudgeon et al., 2012; Portnoy and Heist, 2012). Among them, genetic variation is a great way to study the effect of loss in a population. Populations of many plants and animal species are being reduced to small, sometimes isolated, populations leading genetic deterioration. Genetic analysis allows us to quantify the loss of genetic diversity and the loss of adaptative potential and thus provide advice for the conservation and the management of such species (Smith and Smith, 2001). When the genetic population structure of a species is known, the distribution of sub-populations can be estimated and used for harvest regulation (Ryman and Utter, 1987).

Fisheries management requires understanding biological principles underlying resource dynamics. Management has focused for many years on ecology and population dynamics to the

*Words in bold are in glossary

detriment of understanding **population genetics**. Ecological and population dynamics can be seen as short-term focus and genetics as both long-term and short-term focus. Genetics has not been forgotten in all studies, and its importance has maybe been more admitted for fishes than for other vertebrates. As early as the beginning of the XXe century, genetics was studied for fish sub-populations. In 1983, 15% of the concerns relating to the genetics of animals referred to fish (Ryman and Utter, 1987). In parallel, estimation of the **effective population size** (N_e) in fisheries management and marine conservation is quite recent and is emerging since the beginning of the XXI century (Dudgeon et al., 2012) while it has featured in terrestrial conservation efforts for decades (Schwartz et al., 2007). Genetic monitoring can estimate a population's effective size to evaluate abundance and genetic health, complementing conventional stock assessment methods (Hamilton, 2009). N_e is indeed a fundamental parameter in evolutionary biology (Hare et al., 2011) and conservation biology (Waples and Do, 2010). It is also a potential indicator for conservation and fisheries management and reference points exist to be compared with estimated N_e (Smith and Smith, 2001; Portnoy and Heist, 2012). Effective population size indicates a population's current and future viability (Hare et al., 2011). The aim is to preserve high genetic variability and a sufficient effective population size to maximise the adaptive potential in the face of new environmental conditions. However, using genetics can be a tough job, especially because of the specialised vocabulary and the many assumptions required for practical applications.

In this literature review, some fundamental aspects of population genetics are presented to introduce the notion of effective population size. Estimation methods, ecological processes affecting it, and its use in fishery research are also discussed. The definition of population used is the genetic one (but see glossary).

2. Basics of genetics

Inherited characteristics of a species and variations in individuals are transmitted from parents to offspring. The sum of the hereditary information carried by an individual is the **genotype** which directs the development of the individual and underlies its morphological, physiological and behavioral characteristics. The universal support of the hereditary information is the desoxyribonucleic acid, the DNA, present in every cell of the organism (Schleif, 1993; Hartl and Jones, 1998). It is a complex molecule with two strands in the shape of a double helix. Strands are linked by nitrogenous bases, which are paired: adenine (A) and thymine (T) (the purines bases), cytosine (C) and guanine (G) (the pyrimidine bases). The hereditary information is coded by the sequential pattern in which the base pairs occur. Each species is unique with its own base pairs arrangement and its own number of base pairs (Smith and Smith, 2001; Beaumont et al., 2010).

In eukaryotic cells, DNA is present in larger units called chromosomes going by pair. Humans possess 23 pairs of chromosomes for example. Each chromosome carries units of heredity, the **genes** which are also paired in the body cells. One version of a gene is inherited from the mother and the other one from the father, where each one forms a **haplotype**. Processes leading to a new eggs cell are summarized in Fig.1 which shows how maternal and paternal haplotypes are transmitted to the next generation during the fertilisation. Processes such as replication and meiosis are not explicated in details. The position of a gene on a chromosome is the **locus**; genes occupying the locus on a pair of chromosomes are called **alleles**. If each member of the allele's pair affects a given trait in the same manner, the two alleles are homozygous; if not, they are heterozygous (Hamilton, 2009). During the formation of germ cells, the chromosome pairs are split, so that each resulting cell nucleus receives only one-half of the full number of chromosomes (Beaumont et al., 2010).

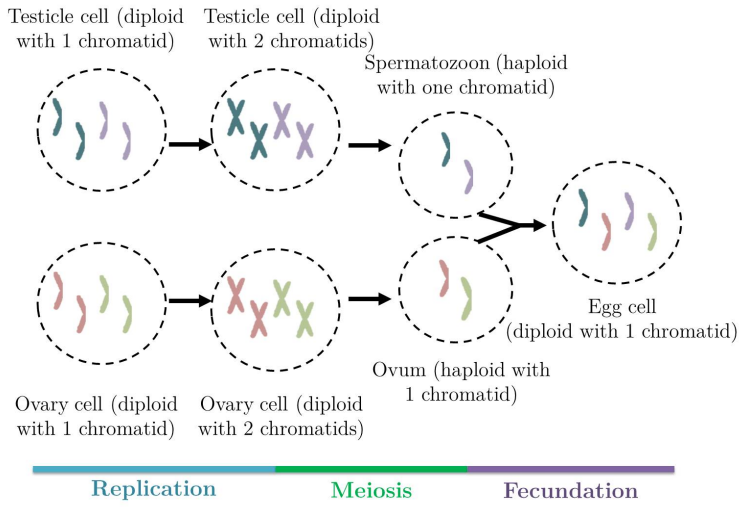


Figure 1: Fertilisation and inheritance of maternal and paternal haplotypes

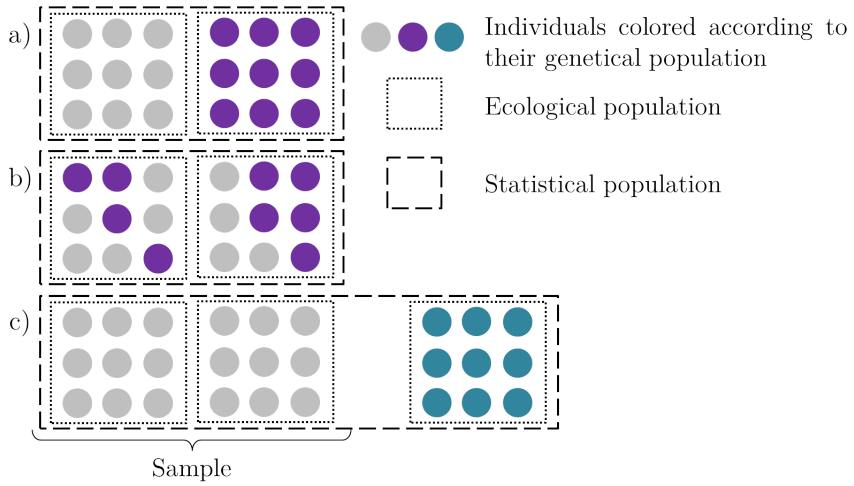
During the separation, several processes, such as recombination, occur which modify the genetic information and create variation. However, there are several others sources of variation of genetic information, like mutation (Smith and Smith, 2001). Genetic variation is a key notion in population genetics which informs on the genetic health of a population and its structure. The genetic variability of a population is a global measure of genetic differences among individuals. These differences are the basis for the numerous observed phenotypes. Genetic variability is the basis of natural selection and so the driving force of evolution. Its estimate is directly linked to a population's capacity to respond to environmental changes (which is also determined by demographic features such as fecundity and natural mortality). For a natural population, its degree of genetic variability is directly associated to its viability. Several indicators are available to characterize the genetic variability and are explained below.

3. Introduction to evolutionary genetics

The marine environment has often been considered as very dispersive and so, marine species have often been supposed as genetically unstructured. However, numerous marine species present strong spatial genetic structure...

BOX 1. It's a population, isn't it ?

The word "population" has been used previously, imprecisely, to designate a group of organisms belonging to the same species. At this point, we need to clarify the concept. At least, three different population definitions can be recognized. In statistics, a population represents all the items under study. In ecology, a population is a group of individuals of the same species within the same habitat at the same time. However, in population genetics, a population designates a group of individuals of the same species which live in a geographic area small enough that each member can reproduce with every other member. So a genetic population encompasses all individuals connected by gene flow (Hartl, 1994; Lowe et al., 2004). The three definitions may coincide, although more frequently they will not:



The figure represents examples for the relationships between statistical (dashed line), ecological (dotted lines) and genetical (colored points) populations. a) the sampled individuals comprise the entire statistical population and the ecological and genetical populations coincide. b) the sampled individuals comprise the entire statistical population but the ecological and genetical populations do not coincide. c) the sampled individuals do not comprise the entire statistical population but ecological and genetical populations coincide. Figure inspired by Lowe et al. (2004).

In this review, we use the genetic definition of a population.

3.1 Some principles of evolutionary biology

Evolutionary biology is a body of statements about the processes of evolution that are believed to have caused the history of evolutionary events (Pigliucci, 2009). Evolution is a change in the gene pool of a population. In order to understand evolution, it is necessary to view populations as a collection of individuals, each harboring a different set of traits. A single organism is never typical of an entire population unless there is no variation within that population. Individual organisms do not evolve, they retain the same genes throughout their life. The differentiation of a species into genetically different populations is a fundamental part of the process of evolution and requires genetic variation (Sober, 1994); several mechanisms exist to increase or create genetic variation as well as to decrease it.

Genetic variation has two components: allelic diversity and non random associations of alleles. In most populations, there are enough loci and enough different alleles that every individual, identical twins excepted, has a unique combination of alleles.

The evolution of the genetic variability of natural populations is driven by four forces (Grant and Waples, 2000):

Mutation

Mutation represents the main source of genetic variability (Hartl, 1994). The cellular machinery that copies DNA sometimes makes mistakes which alter the sequence of a gene. This is called a mutation. There are many kinds of mutations. A point mutation is a mutation in which one "letter" of the genetic code is changed to another. Lengths of DNA can also be deleted or inserted in a gene. Finally, genes or parts of genes can become inverted or duplicated. Mutations are often known for being deleterious (cystic fibrosis, trisomy) but most mutations are thought to be neutral with regards to fitness (but see Kimura, 1990). Mutations provide the raw material for evolution but they are rare in most of the genome: typical rates of mutation are between 10^{-10} and 10^{-12} mutations per base pair of DNA per generation. In contrast, rates of mutation in microsatellite loci are up to 10^{-6}.

Natural selection

Natural selection is the only mechanism of adaptive evolution. It is defined as differential reproductive success of pre-existing classes of genetic variants in the gene pool. Natural selection can maintain or reduce genetic variation depending on how it acts. When selection acts to weed out deleterious alleles, or causes an allele to sweep to **fixation,** it reduces genetic variation. When heterozygotes are more fit than either of the homozygotes, however, selection causes genetic variation to be maintained. Natural selection may not lead a population to have the optimal set of traits. In any population, there would be a certain combination of possible alleles that would pro-

duce the optimal set of traits (the global optimum); but there are other sets of alleles that would yield a population almost as adapted (local optima). Transition from a local optimum to the global optimum may be hindered or forbidden because the population would have to pass through less adaptive states to make the transition. Natural selection only works to bring populations to the nearest optimal point. This idea is Sewall Wright's adaptive landscape (Wright, 1932). This is one of the most influential models that shape how evolutionary biologists view evolution.

Natural selection can be broken down into many components, of which survival is only one. Sexual attractiveness is a very important component of selection, so much that biologists use the term sexual selection when they talk about this subset of natural selection. Sexual selection is natural selection operating on factors that contribute to an organism's mating success.

Gene flow

New organisms may enter a population by **gene flow** from another population. If they mate within the population, they can bring new alleles to the local gene pool. In marine species, gene flow can result from **dispersal** at different life-history stages, including active movement as adults and juveniles and passive transport during pelagic egg and larval stages (Grant and Waples, 2000) but also by migration, *i.e.* round-trip movements of animals between regions or habitats. In genetics, migration is often considered as movement of individuals between sub-populations without any distinction between dispersal and round-trip movements. In this review, we chose to use the restrictive definition of **migration** with a distinction between dispersal, gene flow and migration.

Genetic drift

Allele frequencies can change due to chance alone. This is called genetic drift where the drift is due to stochastic sampling variability of the gene pool. Otherwise, the alleles that form the next generation's gene pool are a sample of the alleles from the current generation. When sampled from a population, the frequency of alleles differs slightly due to chance alone. The intensity of the genetic drift is a function of population size. In very small populations, genetic drift can lead to a high loss of diversity and to a genetic bottleneck. It can also lead to genetic differentiation between two populations separated in space, by a different trajectory of their allele frequencies. The impact of genetic drift is modulated by the sex ratio and the variance in reproductive success. Moreover, genetic drift can modulate the effects of migration and natural selection (Hamilton, 2009).

3.2 Linking genetics and evolutionary biology

Lamarck published a theory of evolution in 1809 (Lamarck, 1809). He thought that species arose continually from nonliving sources. These species were initially very primitive, but increased

in complexity over time due to some inherent tendency. Lamarck proposed that an organism's acclimatization to the environment could be passed on to descendants by inheritance of acquired characters. Fifty years later, Darwin's contributions include hypothesizing the pattern of common descent and proposing a mechanism for evolution – natural selection (Darwin, 1859). In Darwin's theory of natural selection, new variants arise continually within populations. A small percentage of these variants cause their bearers to produce more offspring than others. Darwin's theory did not accord with older theories of genetics. In Darwin's time, biologists subscribed to the theory of blending inheritance – an offspring was an average of its parents. We now know that the idea of blending inheritance is wrong. At the same time, Gregor Mendel, in his experiments on hybrid peas, showed that genes from a mother and father do not blend (Mendel, 1866). An offspring from a short and a tall parent may be medium sized; but it carries genes for shortness and tallness. The genes remain distinct and can be passed separately to descendants.

BOX 2. Mendel's laws

Between 1856 and 1863, the monk Gregor Mendel carried out experiments with pea plants that demonstrated the concept of particulate inheritance. He showed that phenotypes are determined by units that are inherited intact and unchanged through generations. Mendel used pea seed coat color as a phenotype easily tracked across generations. By establishing yellow and green "pure"-breeding lines of peas and by using them as parents, he crossed green-yellow "impure" lines. This work allowed him to reason on dominant traits (yellow seed coat here) and recessive traits (green seed coat here) in impure or heterozygous individuals.

His work is now well known for his two laws:

- Mendel's first law (law of segregation): Two members of a gene pair (*i.e.* the alleles) segregate separately into gametes.
- Mendel's second law (law of independent assortment): During gamete formation, the segregation of alleles of one gene is independent from the segregation of alleles of another gene.

Between 1900 and about 1925, Darwin's theory was regarded as outdated and replaced by Mendelian genetics - the two approaches were regarded as incompatible rivals. In 1908, Hardy (Hardy, 1908) and Weinberg (Weinberg, 1908) worked independently to formulate a relationship to predict allele frequencies given genotype frequencies. This relationship is well known as the Hardy-Weinberg equation: $p^2 + 2pq + q^2 = 1$ where p and q are allele frequencies for a genetic locus with two alleles (Smith and Smith, 2001; Hamilton, 2009; Beaumont et al., 2010). A single

generation of reproduction for which a set of conditions is met will result in a population that meets the **Hardy-Weinberg** expected genotype frequencies (Ryman and Utter, 1987). The list of conditions is long and includes:

- Diploid organism
- Sexual reproduction
- **Non overlapping generations**
- Allele frequencies identical among both sexes
- Random mating
- Random union between gametes
- Infinite population size
- Negligible gene flow
- Negligible mutation
- Absence of natural selection

These conditions make sense when examined. For example, if natural selection acts within a single generation some genotypes will be more frequent than others, breaking the Hardy-Weinberg (HW) equilibrium. However, it is legitimate to wonder whether a model with so many conditions is relevant and if all the assumptions are likely to be met in real life. The Hardy-Weinberg model was not meant to be an exact description of current populations. We must see it as a null model to which compare current populations (Hamilton, 2009). However, the majority of empirical studies on natural populations demonstrate equilibrium conditions even though population sizes are finite (Waples, 2015).

Infinite poulation size is one of the HW conditions. Population size has profound effects on allele frequencies in biological populations and we accept that all biological populations are finite. Therefore, no current population ever exactly meets the population size condition of HW. Genotype and allele frequencies fluctuate from one generation to the next due to **genetic drift**. After some time, a small population will loose alleles at some loci. Over time, allele frequencies spread out progressively as the proportion of fixed genes increases (see **fixation**). The amount of genetic drift increases as the number of the individuals used to produce the next generation decreases (Ryman and Utter, 1987; Smith and Smith, 2001; Hamilton, 2009; Beaumont et al., 2010). A way to restate the population size condition of HW is to say instead that there is very little or no genetic drift occurring. This brings us to the **Wright-Fisher model** which introduced a schematic of the biological life cycle, with an infinite number of gametes in a finite population. The Wright-Fisher model is not biologically realistic but it allows the process of genetic drift to be modeled in a simple fashion (Hamilton, 2009). It makes assumptions identical to the conditions underlying the Hardy-Weinberg equation in addition to assuming that the **Sex ratio** is one and that each generation is founded by sampling $2N$ gametes from an infinite pool of gametes (Smith and Smith, 2001). Here, we have $2N$ gametes which refers to the diploid state of the organism.

Several applications of the HW equilibrium exist. It is used to predict the expected frequency

of a DNA profile. This application can often be seen in newspapers related to crime scene and suspect identification. A second common application is the test of deviation from the HW equilibrium considered as null model. If a population has genotype frequencies which do not fit HW expectation, this is considered evidence that one or more of the evolutionary processes considered absent in the HW model are active (Holsinger, 2001; Hamilton, 2009).

Finally, many of the concepts of population genetics have been combined with ideas from population ecology to make up the field of evolutionary biology or evolutionary ecology.

3.3 Molecular markers for measuring genetic variation

Genetic variation can be measured and quantified at several levels: differences between sequences of DNA fragments, differences between proteins resulting from DNA coding sequence variation, etc... The first markers of genetic variation were phenotypic even though the DNA structure was discovered in 1953. Variations at protein level were discovered during the 1960s and mitochondrial DNA was developed during the 1980s. Finally, the main emergence of the DNA molecular markers was after the 1990s and the introduction of the PCR technique (for Polymerase Chain Reaction) which allowed cheap and rapid amplification of DNA fragments (Hartl and Jones, 1998; Beaumont et al., 2010). Animal cells contain two types of DNA molecules with distinct characteristics: mitochondrial DNA and nuclear DNA. In this review, we only explicit two types of nuclear **genetic markers**: **microsatellites** and **Single Nucleotide Polymorphisms (SNPs)**.

Microsatellites

Microsatellite loci contain repeated motifs of two to five bases and are scattered throughout the genomes of most eukaryotes (Fig. 2). The number of repeated units contained within a particular microsatellite can vary within a population, producing variation in the length of the locus. They are co-dominant, assumed to be neutral and hypervariable with a fast **mutation rate**.

GT pattern repeated 5 times: A A T G C A C G T G T G T G T G T T T C A T

GT pattern repeated 3 times: A A T G C A C G T G T G T T T C A T

Figure 2: Example of microsatellite sequences. The first sequence presents two alleles G and T repeated 5 times and the second one presents the same allelic pattern repeated 3 times.

Microsatellites are powerful markers for which two decades of experience have established the advantages and limits. Indeed, the assumption of neutrality is called into question and they are limited by the need of a mutation model (Chevolot, 2006). The mutation rate of microsatellites is around 10^{-4} per generation (Brumfield et al., 2003).

Single Nucleotide Polymorphisms

SNPs are the most common form of genetic variation and their occurrence throughout the entire genome makes them ideal for studying the inheritance of genomic regions (Baird et al., 2008). A SNP is a variation at a single nucleotide position in the genome between the maternal haplotype and the paternal haplotype (Fig.3), which occurs in at least 1% of the individuals within a species (Vignal et al., 2002). SNPs occur every 1000 bases or so in the human genome but can be much more frequent in many marine species. Their occurrence throughout the genome also makes them ideal for analyses of speciation and historical demography, especially in light of recent theory suggesting that many unlinked nuclear loci (meaning of unlinked loci is explained later) are needed to estimate population genetic parameters with statistical confidence (Brumfield et al., 2003; Beaumont et al., 2010). SNPs have relatively low mutation rates (10^{-8} to 10^{-9}) (Nielsen, 2000; Brumfield et al., 2003).

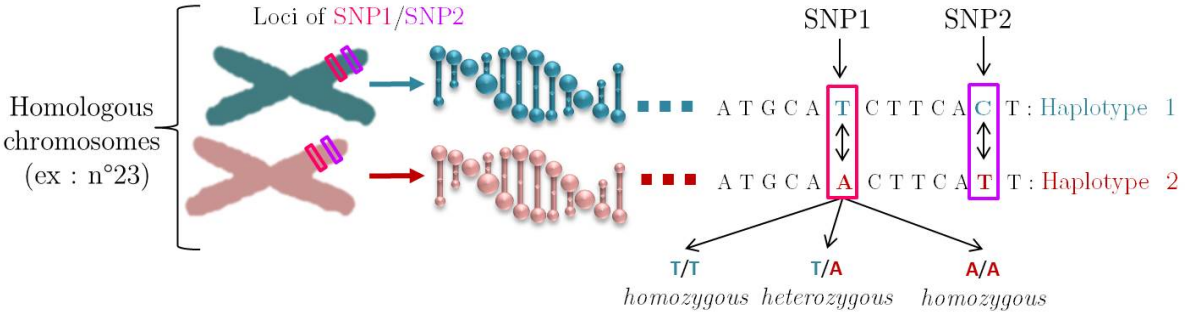


Figure 3: Example of two Single Nucleotide Polymorphisms (SNPs) in a **diploid** organism, heterozygous for the two SNPs presented. For visibility only, chromosomes are represented at mitosis.

Two types of SNP are distinguished: transition, which is a change between two purine or two pyrimidine bases and transversion, which is a change between a purine and a pyrimidine base. In theory, one SNP can show up to the four alleles but in most cases, a SNP is biallelic at each locus (Vignal et al., 2002; Beaumont et al., 2010). Two types of SNP data are available (Morin et al., 2009) phased data, if haplotypes are considered and unphased data if genotypes are considered. For haplotypic data the linkage phase is known, while for genotypic data the linkage phase is unknown (Schaid, 2004). Fig. 4 shows the example of two SNPs: the first one presents the possible allelic states T and A and the second one, the possible allelic states C and A. There are four possible haplotypes (Fig. 4).

The haplotype can be inferred from the genotype in several ways:

- Using biochemistry methods
- Using family information: if the parental genotypes are known, the genotypes of progeny can generally be deduced.

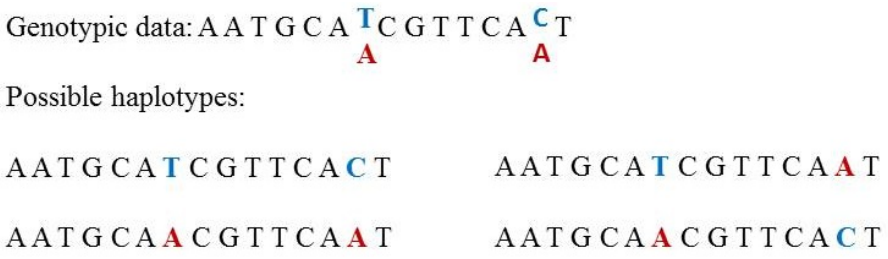


Figure 4: Haplotypes corresponding to genotypic SNP data. Here, we consider two SNPs which possess each two possible allelic states: T or A and C or A.

– Using statistical tools and algorithms of genotypes such as implemented in the program ARLEQUIN (Excoffier et al., 2005).

Several approaches have been developed to infer haplotypes from genotypes with statistical tools. The most known approach is by Clark (1990) which reduces the number of possible haplotypes with a parsimony approaches.

SNPs are simple (bi-allelic) and easy to model which make them powerful contributors to infer population histories. They also present low mutation rates and low scoring error rates. When all population genetic and analytical considerations are weighed, SNPs are superior to microsatellites (Tab. 1).

Molecular markers	**Pros**	**Cons**
SNPs	– Low mutation rates – Bi-allelic – Variation easy to interpret – Deviation tests from neutrality available – Low scoring error rates	– High number of SNPs required – High **Ascertainment bias** – Challenging for computational treatment
Microsatellites	– Few microsatellites needed for population genetics – Display large allelic diversity – Deviation tests from neutrality available	– High mutation rates – Mutation rates variables across loci and across alleles within the same locus – Conducive to **Ascertainment bias**

Table 1: Pros and cons of two types of DNA markers for use in population genetics (Brumfield et al., 2003; Helyar et al., 2011)

Finally, evolutionary genetics (which includes population genetics) is a central discipline

in the study of evolutionary processes. It uses both molecular and classical genetic methods to understand the origin of variation. It describes patterns of genetic variation within and among populations and species, and employs both empirical studies and mathematical theories to discover how this variation is affected by processes such as genetic drift, gene flow, and natural selection.

3.4 Population structure and gene flow

The expectation that genotypes will be present at Hardy-Weinberg equilibrium frequencies depends on the assumption of random mating. Several processes in actual population make this assumption unlikely to hold for many populations (Waples, 2015). For example, within large populations the chances of mating are not uniform but depend on the location of the two mates. This leads to what is called **population structure**. This phenomenon has profound implications for genotype and allele frequencies. Subdivision breaks up a population into smaller units that are genetically independent to some degree. One consequence is that each **subpopulation** has a smaller **effective population size** than the effective size of the entire population if there was random mating. Processes that cause population structure can be considered both as creative or as constraining evolutionary changes (Slatkin, 1987). Genetic isolation, for example, can prevent novel alleles from spreading but can also maintain unique alleles as required for genetic adaptation to local environments. As explained above, a distinction is made between **migration**, **dispersal** and **gene flow**. As dispersal is simply the movement of individuals from one place to another, it may or may not result in gene flow. To confuse matters further, models do not make the distinction and the variable m (for migration rate) is almost universally used to indicate the rate of gene flow.

3.4.1 Studying population structure

In any population study, the ideal first step is to collect samples of the species across its entire range to estimate genetic differentiation within the species as a whole. However, for economic or sampling constraints, most studies focus on limited sampling in specific areas with an economic or conservation interest. Depending on the markers employed, genotypes or haplotypes are scored for the individuals sampled and the data are analyzed in a variety of ways to quantify genetic variation between populations (Beaumont et al., 2010). There is no universal rule for the minimum number of individuals to sample per location. Waples and Do (2008) proposed that 50 individuals appears to be a good trade-off between sampling cost and the bias in estimating population structure.

Several indicators of genetic differentiation exist. One of the most applied indicators is the F-statistics, which was developed by Wright (1949, 1950).

3.4.2 Wright's F-statistics

Before introducing population structure measures, we need to define a key parameter in population genetics: the **fixation index** (Hamilton, 2009), sometimes called the **inbreeding** coefficient (Smith and Smith, 2001). A quantity, symbolized F, is commonly used to compare how much **heterozygosity** is present in an actual population relative to expected levels of heterozygosity under random mating (and other HW equilibrium conditions):

$$F = \frac{H_e - H_0}{H_e} \tag{1}$$

where H_e is the HW expected frequency of heterozygotes based on the population allele frequencies and H_0 is the observed frequency of heterozygotes.

Dividing by the expected heterozygosity puts F on a convenient scale of -1 and $+1$. Negative values indicate an excess of heterozygotes and positive values indicate an excess of homozygotes.

Accounting for divergence of sub-populations, studying the genetic differentiation among populations necessitates several new versions of the fixation index, the so called **F-statistics** (Wright, 1950). The concept of F-statistics was developed by Sewall Wright during the 1920s but the three parameters as we know them now were proposed later (Wright, 1949, 1950). These indicators were designed to describe the population genetic structure of diploid organisms. Basic assumptions are that all populations are of the same size and that there are equal possibilities for any population to exchange individuals with any other population (Beaumont et al., 2010).

Firstly, heterozygosity is calculated for each biallelic loci and then averaged according to the scale considered (total population, subpopulation) (Beaumont et al., 2010). So, a series of hierarchical measures of heterozygosity were defined:

- H_I: mean observed heterozygosity across subpopulations
- H_S: mean expected heterozygosity across subpopulations with random mating within each subpopulation
- H_T: expected heterozygosity with random mating within total population

where subscript T indicates the total population, S the subpopulation and I the individual level (Wright, 1965). Considering a biallelic loci, H_T and H_S have maximum values of 0.5 and H_I can vary between 0 (no observed heterozygotes) and 1 (all observed individuals are heterozygote).

Now, based on H_I, H_S and H_T, three hierarchical F-statistics are defined: F_{IS}, F_{ST} and F_{IT}.

F_{IS}: inbreeding coefficient.

$$F_{IS} = \frac{H_S - H_I}{H_S} \qquad -1 \leq F_{IS} \leq 1 \tag{2}$$

The F_{IS} coefficient represents the difference between the average observed and the HW expected heterozygosity due to non random mating. Thus it is a measure of the extent of genetic inbreeding within subpopulations. At the subpopulation level, it is the correlation between homologous alleles within individuals with reference to the local population (Wright, 1949, 1950, 1965; Beaumont et al., 2010). In other words, F_{IS} is the correlation between homologous alleles within individuals with reference to the local population (*i.e.* the subpopulation under study). A F_{IS} value close to -1 means that all individuals are heterozygous or that there is an excess of heterozygotes compared to the Hardy Weinberg expectation, 0 means that the subpopulations meet the HW assumptions and a value close to $+1$ means that there are no observed heterozygotes (Beaumont et al., 2010).

F_{ST}: fixation index.

$$F_{ST} = \frac{H_T - H_S}{H_T} \qquad 0 \leq F_{ST} \leq 1 \tag{3}$$

The F_{ST} coefficient represents the difference between the average expected heterozygosity of subpopulations and the expected heterozygosity of the total population, so it measures the reduction in heterozygosity due to subpopulation divergence in allele frequency (Lowe et al., 2004; Hamilton, 2009). At a lower level, it is the probability that two alleles sampled at random from a single subpopulation are identical by descent (Smith and Smith, 2001). A F_{ST} value close to 0 means there is no differentiation between subpopulations and a value close to $+1$ that there is complete differentiation between subpopulations.

Although F_{ST} has a theoretical range between 0 and 1, the observed maximum is usually much less than 1. Wright suggested the following qualitative guidelines for the interpretation of F_{ST} (Wright, 1984):

- 0 to 0.05 indicates little genetic differentiation
- 0.05 to 0.15 indicates moderate genetic differentiation
- 0.15 to 0.25 indicates great genetic differentiation
- >0.25 indicates very great genetic differentiation

F_{IT}: overall fixation index.

$$F_{IT} = \frac{H_T - H_I}{H_T} \qquad -1 \leq F_{IT} \leq 1 \tag{4}$$

The F_{IT} coefficient is the correlation between homologous alleles within individuals with reference to the total population (Beaumont et al., 2010). It describes the reduction of heterozygosity within individuals relative to the total population due to non-random mating within subpopulations and population subdivisions (Lowe et al., 2004).

BOX 3. Example of F-statistics calculation

The data shown in the following table are based upon three loci surveyed in three populations (Lowe et al., 2004). Allele frequencies have been calculated assuming HW equilibrium; all means are arithmetic and it is assumed that sample sizes are equal.

For example, for the first locus:
$H_I = (0.4 + 0.4 + 0)/3 = 0.2667$
$H_S = (0.5 + 0.48 + 0)/3 = 0.33$
$H_T = 2 * 0.7 * 0.3$

	Phenotype frequency			Allele frequency		Expected proportion of heterozygotes in randomly mating total population
	a/a	a/b	b/b	a	b	
Locus 1						
Population 1	0.3	0.4	0.3	0.5	0.5	0.5
Population 2	0.4	0.4	0.2	0.6	0.4	0.48
Population 3	1.0	0	0	1.0	0	0
H_I		0.2667				
Mean population allele frequency				0.7	0.3	H_S=0.33
H_T				0.42		

Locus 2						
Population 1	0.3	0.1	0.6	0.35	0.65	0.455
Population 2	0.25	0.5	0.25	0.5	0.5	0.5
Population 3	0.65	0.2	0.15	0.75	0.25	0.375
H_I		0.2667				
Mean population allele frequency				0.53	0.47	H_S=0.44
H_T				0.4982		

Locus 3						
Population 1	1.0	0	0	1.0	0	0
Population 2	1.0	0	0	1.0	0	0
Population 3	1.0	0	0	1.0	0	0
H_I		0				
Mean population allele frequency				1.0	0	H_S=0
H_T				0		

BOX 3. Example of F-statistics calculation - cont.

So the averages are the following:

$H_T = (0.42 + 0.4982 + 0)/3 = 0.31$

$H_S = (0.33 + 0.44 + 0)/3 = 0.26$

$H_I = (0.2667 + 0.2667 + 0)/3 = 0.18$

Wright's F-statistics can be derived as:

$F_{IT} = \frac{H_T - H_I}{H_T} = \frac{0.31 - 0.18}{0.31} = 0.42$

$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{0.31 - 0.26}{0.31} = 0.16$

$F_{IS} = \frac{H_S - H_I}{H_S} = \frac{0.26 - 0.18}{0.26} = 0.31$

Positive F_{IS} and F_{IT} indicate a deficit of heterozygotes with respect to HW expectation. F_{ST} around 0.16 means that 16% of the total genetic variation is between subpopulations, with 84% of the variations within subpopulations. It indicates a great genetic variation between subpopulations.

The three F-statistics are not independent but interrelated according to the formula:

$$F_{ST} = \frac{(F_{IT} - F_{IS})}{(1 - F_{IS})} \tag{5}$$

Finally, Wright's F-statistics provide answers to two different questions:

- for the scored loci, are the genotypes in the proportions predicted by the HW model? (F_{IS} and F_{IT} provide answers)
- for the scored loci, are the allele frequencies different between various populations? (F_{ST} provides answers)

The original formulation of F_{ST} by Wright considered only one biallelic locus. This was extended to accommodate multiple loci termed G_{ST} (Nei, 1973). θ or θ_{ST} (Weir and Cockerham, 1984; Weir, 1996) or ϕ_{ST} (Excoffier et al., 1992) are estimators based on analysis of variance of allele frequencies within and among sub-populations, etc... One of the biggest weaknesses of Wright's F-statistics is the ignorance of the bias due to the number of sampled individuals. The θ estimator by Weir and Cockerham (1984) takes this bias into account; it is often used as an alternative to the F-statistics.

4. Genetic population size

The concept of effective population size appeared for the first time in 1931 proposed by the geneticist Sewall Wright (Wright, 1931). The definition of population size in population genetics

relies on the dynamics of genetic variation in the population. It means that the size of a population is defined by the way genetic variation in the population behaves (Hamilton, 2009; Beaumont et al., 2010; Hare et al., 2011). There are two types of population sizes. One is the real count of individuals in a population (including immatures), called the **census size** N_c. The other one is the genetic size of the population, determined by comparing the genetic drift in a studied real population with the genetic drift in an ideal population, i.e. meeting the Wright–Fisher model which assumes Hardy-Weinberg conditions plus infinite number of gametes and equal sex ratio. The population size of an ideal population that produces the same rate of genetic drift as observed in the current population is the genetic size of the current population or its effective size (Hamilton, 2009). Thus the effective size, denoted N_e, is the size of an ideal population that experiences as much genetic drift as an actual population (regardless of its census size) (Lowe et al., 2004). This comparison with a theoretical ideal population standardizes the measurement of genetic drift and makes N_e comparable across populations with very different life histories (Hare et al., 2011). To summarize and give an ecological interpretation, the census size is the total number of individuals and the effective size is the number of breeding adults that currently contribute to the next generation (Hamilton, 2009). The effective size excludes juveniles, individuals too old to reproduce and those that provide non-contributory gametes. If N_e is very large, the variance in allele frequencies between successive generations is very small (Beaumont et al., 2010). In an ideal population, N_e is equal to the census population size of a generation (Hamilton, 2009).

One of the major difficulties in discussing N_e is that there are two commonly estimated measures of N_e. Variance effective size is the size of an ideal population experiencing genetic drift at the same rate than the actual population and inbreeding effective size is the size of an ideal population losing heterozygosity due to increased relatedness. Typically both are not discussed in detail because for large, stable populations, they are similar. However, the sizes of many wildlife populations are not stable (Leberg, 2005).

4.1 Factors influencing effective population size

Population fluctuations, like bottleneck events have a large impact on effective size but there are several other aspects of biological populations that have the same impact by increasing the sampling error in allele frequency across generations (Lowe et al., 2004) (Tab.2).

Population size fluctuations can lead to changes in alleles represented in later generations which leads to changes in effective size.

An unequal sex ratio has a large influence on genetic drift and so on the effective size of a population. For example, in a polygamous population with a ratio of one breeding male to several females, the offspring will be half or full **siblings**, leading to an increase in genetic drift compared to a case with breeding sex ratio of 1:1 where all offspring are less closely related (Smith and Smith,

2001; Hamilton, 2009).

A third factor is the degree to which adult individuals in the population contribute to the next generation. A population is stable in size over time when each pair of individuals produces on average two progeny. The variance in family size can be used to describe variation among individual reproduction. If the variance in family size increases, the alleles passed to the next generation come increasingly from those parents producing more offspring. When the variance in family size is equal to the average family size ($k = 2$), then the census population size of parents is the effective population size. The Wright-Fisher model assumes that family sizes follow a Poisson distribution (i.e. mean equal to variance). An interesting fact is that in the case where the variance in family size is less than the mean family size, N_e can be larger than N (Hamilton, 2009).

A finite population size can be seen as a form of inbreeding. In small populations, the chance of mating with a relative is large since the number of mates is limited. Genetic drift also occurs due to finite population size. Both phenomenon increase homozygosity and decrease heterozygosity leading to a decrease in effective population size (Lowe et al., 2004; Hamilton, 2009).

Migration and dispersal are also phenomenons acting on the effective population size. If immigrants become part of the breeding population, they introduce a different genetic sample that tends to reduce genetic drift. Information from the field on genetic effects of migration is lacking because of the difficulty to collect data (Smith and Smith, 2001).

Finally, overlapping generations have a great influence on the effective size of a population because age-structured species presents many co-evolutionary processes (Waples et al., 2014). However, the exact effects of overlapping generations remain unclear.

All factors contributing to genetic drift are confounded in the variable N_e, so a small value indicates strong drift but does not identify the causes. Estimating N_e is valuable for harvested population in a stock assessment context but can also provide insights into patterns of connectivity among populations.

4.2 Relationship between N_e and N_c

The first thing to understand is that, for most real populations, effective size and census size are not the same (Leberg, 2005). Commonly, due to genetic bottlenecks for example, the census size is larger than the effective size (Lowe et al., 2004). Many factors can cause $N_e \neq N_c$ including unequal sex ratios, high variation in reproductive success, non random mating, mating system, **Philopatry**, gene flow, overlapping generations and temporal fluctuations in population size (Leberg, 2005).

Event	Consequences for genetic diversity
Population size fluctuations	– Changes alleles present in later generations – Changes effective size
Unequal sex ratio	– Increases genetic drift – Reduces N_e (max if breeding sex ratio 1:1)
Variance in reproductive success	– Stable population: parents produce 2 offspring ($k = 2$) – If $k = 2$: population size of parents is the effective size
Inbreeding	– Increases homozygosity – Decreases effective population size
Migration	– Increases or decreases genetic drift – Increases or decreases effective population size
Overlapping generations	– Effects remain unclear

Table 2: Summary of the effects of different factors influencing genetic drift and effective population size (Smith and Smith, 2001; Hamilton, 2009)

BOX 4. Commonly N_e differs from N_c: example

Assume a population of 100 individuals at time $t-1$, which was reduced to 10 individuals at time t (by a **bottleneck** for example) and increased again to 100 individuals at time $t+1$ (each couple produced offspring)(Hamilton, 2009). The effective size of a population fluctuating over time is calculated as the reciprocal of the average of the sum of reciprocals of the effective number of each generation (Smith and Smith, 2001). This special sort of average is called the harmonic mean of N_e and is given by the following expression:

$$\frac{1}{N_e} = \frac{1}{t}\left[\frac{1}{N_{e(t=1)}} + \frac{1}{N_{e(t=2)}} + ... + \frac{1}{N_{e(t)}})\right]$$

In our example, N_e is equal to 25 and the mean census size is 70. Only those alleles that pass through the genetic bottleneck of 10 individuals are represented in later generations. Using the census size to predict the behavior of allele frequencies will underestimate the genetic drift. We expect allele frequencies in the current population to behave like the allele frequencies in an ideal Wright-Fisher population with a constant size of 25 individuals over three generations (Hamilton, 2009).

The relationship between N_e and N_c is a key to understanding the effects of harvesting: in marine species, N_e/N_c ratios are often as low as 10^{-5}. This very low ratio seems to indicate that enormous marine population may be more sensitive to genetic drift and inbreeding from intensive harvest than census size alone suggests (Hare et al., 2011). However these low ratios can also be a consequence of insufficient sample sizes (Waples, 2016).

Effective population statistics inform us about the genetic health of a population. To be able to handle the loss of genetic variability, inbreeding, a population must reach a critical size. The number of individuals that will ensure the persistence of a population in a viable state for a given interval of time is the **Minimum Viable Population (MVP)** (Soulé and Wilcox, 1980; Shaffer, 1981; Smith and Smith, 2001; Rai, 2006). It must be large enough to cope with environmental changes, genetic drift or variation in individual birth and death. Different studies have been conducted to estimate threshold values for effective size. In 1980, M. Soulé addressed the question of MVP size in order to prevent extinction and concluded a minimum effective size of 50 individuals was needed (with a maximum of inbreeding at 1%). The same year, Franklin suggested that in the long term, genetic variability will be maintained only if population sizes are an order of magnitude higher than 50. This suggestion is based on the assumption that continued evolutionary change is necessary for populations and species survival, and that response to natural selection is limited by small population sizes. Due to this, Franklin proposed that for long term viability the effective size should be 500 in order to account for the expected rapid loss of genetic variance (Franklin, 1980). The two studies gave rise to the 50/500 rule which has been strongly criticized (Caughley, 1994; Lande, 1995).

Given loss of genetic variability can been expressed as the loss in heterozygosity, Meffe and Carroll (1997) argued that an effective size of 50 individuals was not enough for long term conservation (Fig. 5).

Experiments with drosophila showed that a median extinction time of 47.5 generations was necessary to extinct a population of 50 effective individuals (Reed and Bryant, 2000). Such experiments have never been realized for raies and skates and it is hard to predict a minimal effective size for a given minimal time leading to the extinction of a population. Thus, MVP is only a general guideline for the genetic management of endangered species. Past studies showed that applying the 50/500 rule on wild populations can be a dangerous game for species survival. For example, a species with highly density-dependent reproduction, living in a more or less constant environment may persist for a long time despite a decline in genetic diversity. There is no universal rule for MVP (Smith and Smith, 2001).

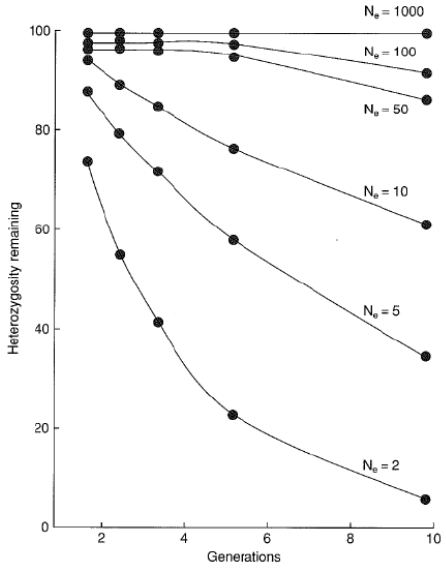


Figure 5: Simulated loss of genetic variability after 10 generations according to the effective size N_e. Populations with large N_e maintain their genetic variability while small ones loose it (Meffe and Carroll, 1997).

4.3 Estimators of effective population size

Genetic methods for estimating N_e have become increasingly practicable by recent advances in genotyping, software and computer processing speed (Hare et al., 2011). Just as there are several models based on different assumptions to describe genetic variation over time, there are several ways to estimate the effective population size. Two main definitions of the effective population size exist (Hamilton, 2009):

- inbreeding effective population size which represents the size of an ideal population that would show the same probability of allele copies being identical by descent as an actual population.
- variance effective population size which represents the size of an ideal population that would show the same sampling variance in allele frequency as an actual population.

For the conservation and the management of wildlife populations, estimates of contemporary N_e are of interest. When estimates are based on single-cohort samples, they reflect the effective number of breeders in one reproductive cycle (N_b) (Dudgeon and Ovenden, 2015). For mixed-age sample, N_e estimates reflect the harmonic mean of the number of breeders over several generations (Waples et al., 2014).

Several estimates of N_e exist which can be summarized into two broad categories (Dudgeon et al., 2012):

- Estimates of long-term or historical effective population size.

– Estimates of contemporary or short term effective population size.

These two estimates of effective population sizes can be derived from samples collected at a single point in time or from temporally spaced samples. In recent years, several estimators based on single-sample method were developed (Waples et al., 2014).

4.3.1 Long-term N_e estimation methods

Estimates of long-term N_e are used if the goal is to understand the long-term N_e of a species prior to contemporary influences. The main method used is the coalescent method which requires only a single random sample of individuals (Hare et al., 2011). This method traces the history of genes in a population back to a common ancestor in order to describe the process of coalescence of allelic copies in the population back in time (Nikolic, 2009). Such models investigate patterns among individuals from the past to the present and aim at reconstructing versions of events such as inbreeding, gene flow or natural selection in the past that could have lead to the current observed population. A coalescent event is defined as the merging back in time of two lineages into a single ancestral lineage. The method aims at connecting current lineages observed in the sample back to a single ancestor in the past, the most recent common ancestor (MRCA) (Hamilton, 2009). For most management applications, this method targets the period just prior to human intervention (Alter et al., 2007).

Historical effective population size has been widely estimated for terrestrial species but also for elasmobranchs such as *Raja typus* (Castro et al., 2007), using the *ARLEQUIN* software (Excoffier et al., 2005), Lemon shark *Negaprion brevistoris* (Schultz et al., 2008), Scalloped hammerhead *Sphyrna lewini* (Nance et al., 2011), using the method developed by Beaumont (1999) and implemented in the software *MSVAR*, and sleeper sharks *Isurus oxyrinchus* (Murray et al., 2008).

4.3.2 Contemporary N_e estimation methods

Contemporary N_e can be estimated using demographic methods or genetic method, here only the genetic methods are detailed. Short-term N_e estimates are usually based on allelic changes between two samples such as loss of heterozygosity, loss of alleles or the increase of the inbreeding coefficient. Commonly, four methods are considered suitable for estimating N_e indirectly from genetic data. All four methods have limited power to estimate N_e when the true value is large, but since conservation is focused on small populations, these methods are increasingly used in a conservation context.

Multiple sample estimation

The temporal method requires population samples separated by at least two generation times

(Dudgeon et al., 2012). The effective population size is estimated by relating the observed amount of change in allele frequencies to that expected under pure genetic drift. This method estimates the variance effective population size using the expected variance in gene frequencies (Grant and Waples, 2000). Several factors may introduce bias into these estimates: sampling error, migration, natural selection, etc... Some developments made during the last decade address these aspects to minimize bias. For example, several formulas are used for calculating N_e depending upon of the sampling scheme: random sampling of the entire population, only reproductive adults, single-cohort of newborns... (Waples and Yokota, 2006). Moreover, Jorde and Ryman (1995) developed a correction factor based on life history traits for accounting for overlapping generations.

Precision and accuracy of N_e estimates can be improved by increasing the number of loci, the number of alleles, the sample size or the amount of time between temporal samples. However, the last option is the most difficult as it requires conducting long-term studies or that old samples are available. As a consequence, it is rarely applied (Leberg, 2005).

Single-sample estimation

Estimations of N_e using a single-time point estimate approach requires only a single sampling of the population. In the last decade, various estimation methods have been developed or improved (Dudgeon et al., 2012).

- **Linkage-Disequilibrium** (LD)
- **Heterozygosity Excess** (HE)
- Molecular Co-ancestry (MC)

4.4 Estimation of effective population based on Linkage-Disequilibrium

4.4.1 Definition and estimation of Linkage-Disequilibrium

In population genetics, Linkage-Disequilibrium (LD) is the non random association of alleles at different loci *i.e.* the presence of statistical associations between alleles at different loci that are different from what would be expected if alleles were independent, based on their individual allele frequencies. When the genotype present at one locus is not independent of the genotype present at another locus there is a Linkage-Equilibrium.

The assumptions for using LD are the same than the Wright Fisher model, but also assume that loci are unlinked. Genes located close to each other on the same chromosome are said to be linked (Fig. 6, a). If two alleles from different genes on the same chromosome tend to be associated in different individuals at a greater frequency than expected due to random association, there is

linkage disequilibrium between these genes. Two genes located on different chromosomes or at large distance on the same chromosome are unlinked (there is Linkage-Equilibrium).

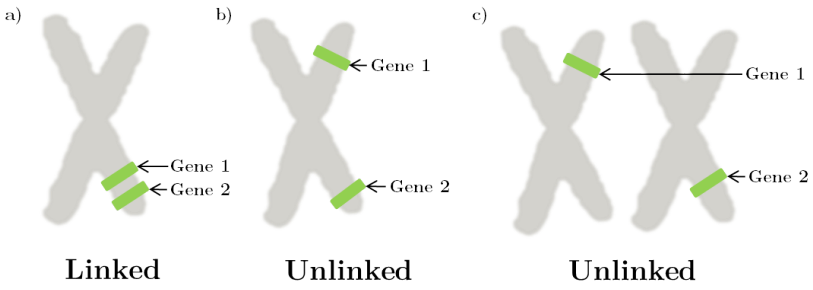


Figure 6: Chromosomes illustrating linkage between loci. Two genes are shown in each situation. Genes 1 and 2 are linked in the first situation and unlinked in the others.

When numerous loci are used, it is usually unavoidable that some will be linked. However, linkage becomes an issue only if genes are very close together or if the recombination rate is very low. The recombination rate can be seen as the proportion of gametes produced by an individual with haplotype differing from the two haplotypes of the individual. For example, an individual A1B1/A2B2 will produce 2 types of gametes: parental type (A1B1 and A2B2) and recombinant type (A1B2 and A2B1). r is the proportion of recombinant type and is bounded between 0.5 (genes are unlinked or very far from each other on the chromosome, Fig. 6 b and c) and 0 (genes are linked Fig. 6, a).

Linkage-Disequilibrium is measured by two statistics, D and r, which can be interpreted as the co-variance and the correlation between loci (see below). The effective population size estimated by the LD method is an inbreeding effective population size.

4.4.2 Origins of Linkage-Disequilibrium

Several factors can create Linkage-Disequilibrium.

- **Genetic drift**: genetic drift conducts to a loss of variability over generations (by random disappearance of alleles or haplotype). Genetic drift is stronger when population size is small.
- **Mutation**: mutation can occur in a haplotype and create a Linkage-Disequilibrium between the mutated locus and this haplotype. This disequilibrium normally reduces over generations but can increase if genetic drift or selection occurs.
- **Gene flow**: LD can also be created by the mixing of populations or by migration. At the beginning, LD is proportional to the allelic differences between populations and is independent of the distance between markers. Over generations, LD decreases but the rate of decrease varies with the link between loci. For independents loci, LD tends to disappear and for linked loci, it persists much longer.

– **Selection**: as genetic drift, selection conducts to a decrease in the number of breeders and so a decrease in the number of haplotypes in the population. Alongside, this decrease leads to an increase of consanguinity.

– **Recombination**: the recombination rate is not constant across the genome. LD is strong in regions with a low recombination rate and high in regions with low recombination rate.

4.4.3 Assessing the effective population size with the Linkage-Disequilibrium method

LD can be illustrated by considering two SNP markers A and B on the same chromosome: A possess the alleles A_1,A_2 and B the alleles B_1,B_2. Four haplotypes are possible: A_1B_1, A_1B_2, A_2B_1 and A_2B_2 (Fig. 7).

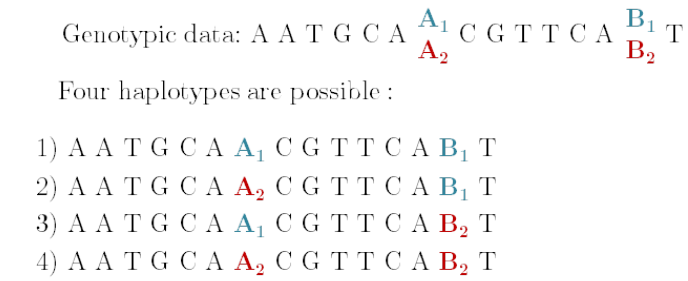


Figure 7: Illustration of the four haplotypes possible by considering a heterozygous individual at two SNPs markers A and B.

If each allele occurs at a frequency of 0.5 and Hardy Weinberg conditions are met, each haplotype should have a frequency of 0.25. Any deviation to this haplotypic frequency reflects a Linkage-Disequilibrium. Several measures of LD are available between alleles with two loci and all depend on the following equation where D is the coefficient of LD, p_{A_1} represents the frequency of the allele A_1 and $p_{A_1B_1}$ the frequency of the haplotype A_1B_1:

$$D = D_{A_1B_1} = p_{A_1B_1} - p_{A_1}p_{B_1} \tag{6}$$

Linkage-Disequilibrium between A_1B_2, A_2B_1 and A_2B_2 can also be calculated. When the genetic markers used are bi-allelic (like SNPs), results between all calculations are exchangeable ($D_{A_1B_1} = -D_{A_1B_2} = -D_{A_2B_1} = D_{A_2B_2}$). This does not apply to multi-allelic markers, such as microsatellites.

D is a non normalized equation. Other statistics have been derived and are more commonly used. Hill and Robertson (1968) defined the correlation coefficient (r^2) between alleles for bi-allelic loci.

$$r^2 = \frac{D^2}{p_{A_1}p_{A_2}p_{B_1}p_{B_2}}, \qquad 0 \leq r^2 \leq 1 \tag{7}$$

Under the null hypothesis ($H_0 : D_{ij} = 0$) this statistics follows a χ_2 distribution. Moreover, the correlation coefficient is less sensitive to the population size than other statistics. Lewontin (1964) proposed the D' (widely used in medicine):

$$D' = \frac{|D|}{D_{max}}, \qquad D_{max} = \begin{cases} min[p_{A_1}p_{B_2}, p_{A_2}p_{B_1}] & if D > 0 \\ max[p_{A_1}p_{B_1}, p_{A_2}, p_{B_2}] & if D < 0 \end{cases} \tag{8}$$

The r^2 and D' as presented here are statistics of LD at two loci but it is more useful to consider LD on a chromosomal region. One of the classical approaches is to calculate a local average of LD for pairs of loci.

BOX 5. Linkage-Disequilibrium with unphased data

In diploid species, it is much easier to determine genotypes (unphased data) than haplotype (phased data). Consequently, which nucleotides, corresponding to SNP alleles in fig. 7, occur together on an individual DNA molecule (chromosome) is usually unknown. This makes it difficult to estimate statistical associations such as LD among loci, as shown below. Let us consider two SNPs on the same pair of **homologous chromosomes** at two loci. A possess the alleles A_1, A_2 and B, B_1 and B_2. All the possible genotypes and the exemplified number of sampled individuals are summarized in the following table:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	0	1	0
A_1A_2	0	1	0
A_2A_2	0	0	0

Genotypes of sampled individuals are:

Individual 1	A_1A_2	B_1B_2
Individual 2	A_1A_1	B_1B_2

Therefore, there are four possible haplotypes for individual 1: A_1B_1, A_1B_2, A_2B_1 and A_2B_2 and:

$$D_{A_1B_1} = f_{A_1B_1} - f_{A_1}f_{B_1}$$

$$f_{A_1} = 0.75 \quad f_{B_1} = 0.5$$
$$f_{A_2} = 0.25 \quad f_{B_2} = 0.5$$

But $f_{A_1B_1}$ is not directly available.

As explained in Box 5, estimating Linkage-Disequilibrium in case of genotypic data is not a direct calculation. Methods currently used to estimate LD bypass this issue by inferring haplotype frequencies using a maximum likelihood estimator (Hill, 1974; Rogers and Huff, 2009) or by using the composite Burrows method (Weir, 1979, 1996). The first inference method developed by Hill (1974) requires that random union of gametes occurs in the population, in which case population genotypic frequencies can be replaced by products of gametic frequencies. Such cases are discussed and Weir (1979) recommended the use of the Burrows method which does not require the random mating assumption and provides a straightforward calculation. Dr Peter Burrows (unpublished work but see Cockerham and Weir, 1977) considered a composite measure of LD: Δ. It performs better than an alternative maximum likelihood estimator (Weir, 1979; Schaid, 2004) and is estimated directly from genotype counts as follow (Russell and Fewster, 2009):

$$\hat{\Delta}_{A_1B_1} = \frac{n_{A_1B_1}}{n} - 2\hat{p}_{A_1}\hat{p}_{B_1} \tag{9}$$

where $n_{A_1B_1}$ are the sample counts of possible genotypes (see Box 5) and n the total number of counts (sampled individuals). $\hat{p}_{A_1}$ and $\hat{p}_{B_1}$ are proportions of alleles A_1 and B_1 in the sample of n individuals genotyped at both loci. A small-sample correction factor of $n/(n-1)$ should be applied to $\hat{\Delta}_{A_1B_1}$ (Weir, 1979, 1996).

The corresponding correlation coefficient $\hat{r}_{A_1B_1}$ can be estimated as:

$$\hat{r}_{A_1B_1} = \frac{\hat{\Delta}_{A_1B_1}}{\sqrt{(\hat{p}_{A_1}(1-\hat{p}_{A_1}) + (\hat{h}_{A_1A_1} - \hat{p_{A_1}}^2))(\hat{p}_{B_1}(1-\hat{p}_{B_1}) + (\hat{h}_{B_1B_1} - \hat{p_{B_1}}^2))}} \tag{10}$$

where $\hat{h}_{A_1A_1}$ and $\hat{h}_{B_1B_1}$ are the observed proportions of A_1A_1 and B_1B_1 homozygotes in the sample of size n .

The correlation coefficient r has $E(r) = 0$ for unlinked loci; in finite populations however, the correlation is likely to take non-zero values with small populations giving the largest values so the expectation of its square is non-zero and is a function of the effective population size, N_e (Russell and Fewster, 2009). The expression for $E(r^2)$ depends upon the mating structure and recombination fraction c in a population, and is also affected by sample size S (Waples, 2006; Russell and Fewster, 2009) The distribution of r^2 is not known but Weir and Hill (1980) showed that the expectation of squared disequilibrium coefficients is strongly affected by the mating system and recombination fraction (c) between the studied loci. For dioecious random mating:

$$E(\hat{r}^2) = \frac{(1-c)^2 + c^2}{2N_e c(2-1)} + \frac{1}{S} \tag{11}$$

If the loci are unlinked (c=0.5), this equation simplifies to:

$$E(\hat{r}^2) = \frac{1}{3N_e} + \frac{1}{S} \tag{12}$$

$E(\hat{r}^2)$ can be expressed as the sum of a term due to finite population size, $E(\hat{r}^2_{drift})$, and a term due to sampling a finite number of individuals, $E(\hat{r}^2_{sample})$, (Waples, 2006; Russell and Fewster, 2009).

$$E(\hat{r}^2) = E(\hat{r}^2_{drift}) + E(\hat{r}^2_{sample}) \tag{13}$$

Waples (2006) suggested that if S varies among loci, the harmonic mean should be used. Replacing $E(\hat{r}^2)$ with its estimate, $\hat{r}^2$, and rearranging leads to an estimator for N_e (for dioecious random mating):

$$\hat{N}_e = \frac{1}{3(\hat{r}^2 - 1/S)} \tag{14}$$

Waples (2006) investigated the fact that equation 14 does not provide an unbiased estimate of effective population size for the range of S/N_e ratios likely to occur in the study of natural populations. $\hat{N}_e$ is downwardly biased if S is less than about $2N_e$, and the bias is substantial for $S < N_e$, particularly for samples of small census population size. By examining data from samples taken only in generation 0 (which mimics samples from a population of infinite size) and so with a $E(\hat{r}^2_{drift}) = 0$, Waples found that $1/S + 3.19/S^2$ can be a good approximation to $E(\hat{r}^2_{sample})$ if S>30 and $0.0018 + 0.907/S + 4.44/S^2$ otherwise. As $E(\hat{r}^2_{sample})$ and r^2 can both be calculated directly from data, $\hat{r}^2_{drift}$ can be deduced and the effective population size can be estimated as follows (here for large samples $S > 30$):

$$\hat{N}_e = \frac{1/3 + \sqrt{1/9 - 2.76\hat{r}^2_{drift}}}{2\hat{r}^2_{drift}} \tag{15}$$

For calculating confidence intervals, the distribution of r^2 is approximated by a χ^2 distribution with $M = L(L-1)/2$ degrees of freedom and L the number of loci (Hill, 1981; Waples, 1991). Confidence limits for r^2 are estimated with:

$$(1-\alpha)CI = (\hat{r}^2 M/\chi^2_{(\alpha/2),M}, \hat{r}^2 M/\chi^2_{(1-\alpha/2),M}) \tag{16}$$

As explained above, the method assumes that loci are neutral (non-selected) and physically unlinked ($c = 0.5$). Microsatellite loci are highly suitable for the Linkage-Disequilibrium method (Schwartz et al. 1998), because they are highly polymorphic and nearly selectively neutral, although this may be compromised by **genetic hitchhiking**. SNPs can also be used for

Linkage-Disequilibrium. Do et al. (2014a) showed that the precision of the LD method is better with 200 SNPs compared with 20 microsatellites. The relationship between the estimated r^2 and N_e obtained with LD method has the form of a hyperbolic curve. When $\hat{r}^2$ is less than $1/n$, negative estimates of N_e are obtained. In these cases, which are most likely to arise when the sample size is small, the contribution of genetic drift to Linkage-Disequilibrium is swamped by the contribution from statistical sampling. Because it is not possible for N_e to be negative, the conventional way of interpreting a negative N_e is to replace it with an estimate of infinity (Waples, 1991). In other words, negative estimates occur when the genetic results can be explained entirely by sampling error without invoking any genetic drift (Waples and Do, 2010).

4.4.4 Application of the Linkage-Disequilibrium method for calculating effective population size

The LD method is based on a number of simplifying assumptions such as selective neutrality, closed populations, and discrete generations that may not apply to many natural populations. The LD method also assumes that of the four evolutionary forces (mutation, selection, migration and genetic drift), only genetic drift is responsible for the signal in the data (Waples and Do, 2010). The consequences of violating these assumptions have been widely studied during the last five years. Effects of overlapping generations (Waples and Yokota, 2006; Robinson and Moyer, 2013; Waples et al., 2014), migration (Waples and England, 2011), Hardy-Weinberg assumptions (Waples, 2015), physical linkage between loci (Waples et al., 2016), sample size (Waples, 2006; Waples and Do, 2010) and a comparison of the genetic markers used are synthesized in Tab. 3 and Fig. 8 (Waples and Do, 2010; Do et al., 2014a). Nevertheless, the neutrality assumption needs to be evaluated more precisely when a large number of SNP loci is used. Tenesa et al. (2007) studied the estimation of human effective population size with 1,000,000 of SNPs. In this study, for each chromosome, pairwise r^2 was calculated, only for SNP pairs between 5 kb and 100 kb apart from each other to avoid the influence of **gene conversion** on observed LD at SNPs that are closer and to minimize the effect of a very recent expansion of the human effective population size on LD.

Tested effects	Protocole	Estimators	Reference
Sample size, number of loci, concurrent effects of both parameters	Random subsamples of the total data set	NeEstimator	Dudgeon and Ovenden (2015)
Migration	Populations simulated with and without migration. Comparison of estimates and simulated N_e values using RMSE (Root Mean Square Error).	Colony; CoNe; MLNe; NeEstimator; ONeSAMP; TMVP	Gilbert and Whitlock (2015)
Overlapping generations	Simulations of age-structured genetic data which tracked demographic and genetic processes. Study of associated biases.	LDNe (old version of NeEstimator)	Waples et al. (2014)
Overlapping generations and age-structure	Simulation of several life-history scenarios. Comparison of the CV of estimates and simulated N_e values.	LDNe	Robinson and Moyer (2013)
Migration	Populations simulated with and without migration. Comparison of estimates and simulated N_e values.	LDNe	Waples and England (2011)
Comparison of SNPs and microsatellites	Simulations of the two types of genetic data. Study of CV of N_e' estimates and of MSE (Mean Square Error).	Unavailable program	Waples and Do (2010)
Sample size, number of loci, number of alleles per locus.	Simulations of genetic data. Study of CV of N_e' estimates and of MSE (Mean Square Error).	Unavailable program	Waples and Do (2010)
Allele exclusion criteria	Simulations of genetic data. Study of CV of N_e' estimates and of MSE (Mean Square Error).	Unavailable program	Waples and Do (2010)
Sample size, allele exclusion criteria	Simulations of genetic data. Comparison of estimates and true values.	LDNe	Macbeth et al. (2013)

Table 3: Summary of the main biases and sensitivity analyses conducted for the LD method. $N_eEstimator$: Do et al. (2014a), $ONeSAMP$: Tallmon et al. (2008), MLN_e: Wang and Whitlock (2003), $COLONY$: Jones and Wang (2010), CoN_e: Anderson (2005), $TMVP$: Anderson (2005), $LDNe$: Waples and Do (2008).

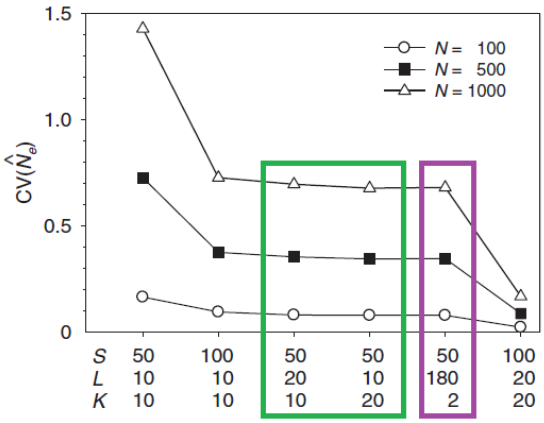


Figure 8: Effects of independently doubling sample size of individuals (S), number of loci (L), and number of alleles per locus (K) on the coefficient of variation (CV) of $\hat{N}_e$. Results are shown for three different ideal population sizes of N individuals (here N corresponds to N_e) (redrawn from Waples and Do (2010)). Violet box can be interpreted as 180 SNPs and green box as 10 and 20 microsatellites with 20 and 10 alleles respectively.

Software

The most widely used software for estimating the effective population size is *NeEstimator* (Do et al., 2014a) which calculates contemporary unbiased estimates of effective population size based on LD. Moreover, the software provides estimates of N_e based on others methods: Heterozygosity Excess, Molecular Co-ancestry, temporal method. The LD method with *NeEstimator* has been most widely used with microsatellite data but can be used with SNP data as well. Accurate estimates of N_e can be obtained with non overlapping generations by using 10-25 microsatellites, which corresponds to approximately 180 SNPs and samples of at least 25-50 individuals. Precision can be improved by sampling more individuals or by sampling more genetic markers. However, precision can be improved more by doubling the number of individuals, especially for SNPs (Waples and Do, 2010).

In 2015, a new tool was developed to infer historical N_e trends based on SNP data only: SNeP (Barbato et al., 2015). This software is a complementary tool for investigating demography through Linkage-Disequilibrium.

Applications

The Linkage-Disequilibrium method has been widely used for estimating effective population size and together with the temporal method, the Linkage-Disequilibrium (LD) method is the most widely used method. As early as 1991, Robin Waples studied the feasibility of using the LD method for cetacean populations and concluded that the LD method may provide meaningful information if a high number of polymorphic gene loci can be resolved (Waples, 1991). Beyond its application to marine mammals, the LD method has been used for several species including humans (Tenesa

et al., 2007), ruminants (Cervantes et al., 2011; Do et al., 2014b; Barbato et al., 2015) and cougar (Juarez et al., 2015) populations but also insects (Francuski and Milankov, 2015), reptiles (Bishop et al., 2009; Monzón-Argüello et al., 2015) and of course, fishes (Macbeth et al., 2013; Van Doornik et al., 2013; Wilson et al., 2014; Dudgeon and Ovenden, 2015; Pilger et al., 2015; Perrier et al., 2015).

4.5 Estimation of effective population with Heterozygosity Excess

4.5.1 Definition and measures of the Heterozygosity Excess method

This method is based on the following principle: when the effective number of breeders (N_{eb}) in a population is small, the allele frequencies will by chance be different in males and females, which causes an excess of heterozygotes in the progeny with respect to Hardy-Weinberg equilibrium expectations (Pudovkin et al., 1996; Luikart and Cornuet, 1999; Leberg, 2005). In other words, when the number of breeders is small, the allele frequencies in males and females will be different due to binomial sampling variability, which generates an excess of heterozygotes in the progeny relative to the HW expected proportions (Robertson, 1965; Rasmussen, 1979).

The proportion of heterozygotes expected in an offspring produced by a small and equal number of males and females can be calculated as follows:

$$H' = 2pq + pq/n = 2pq(1 + 1/(2n)) \tag{17}$$

where n is the number of haploid genomes in the mothers or fathers and p and q are the frequencies of alleles at a locus.

Pudovkin et al. (1996) used the notation H_{obs} instead of H' and noted H_{exp} the expected proportion of heterozygotes in the base population under Hardy-Weinberg proportions (H_{exp}= $2pq$). So:

$$\hat{N}_{eb} = \frac{H_{exp}}{2(H_{obs} - H_{exp})} \tag{18}$$

If the sample size is finite, H_{exp} must be estimated using the unbiased estimator $\frac{2N(2pq)}{2N-1}$ where N is the number of progeny sampled (Luikart and Cornuet, 1999).

Pudovkin et al. (1996) proposed another estimator of N_{eb} using the Selander (1970) index D, which is the excess of deficiency of heterozygotes, corresponding to the reciprocal of the ratio $H_{exp}/(H_{obs} - H_{exp})$. So the following more exact equation was derived:

$$\hat{N}_{eb} = \frac{1}{2D} + \frac{1}{2(D+1)} \;\; with \;\; D = \frac{H_{obs}}{H_{exp}} \tag{19}$$

4.5.2 Application of the Heterozygosity Excess method for calculating effective population size

Luikart and Cornuet (1999) conducted a simulation study on Heterozygosity Excess (HE). Their results showed that when the method is applied to natural populations, it often gives estimates of N_{eb} equal to infinity with poor precision. According to the authors, these results were not surprising as only a few polymorphic loci were analyzed for a small number of progeny. They concluded that additional empirical evaluations are needed, but it is extremely difficult to find large data sets containing individuals produced from a known number of parents. Pudovkin et al. (2010) studied the sampling properties of the method. The larger the number of loci surveyed or the larger the sample size, the narrower is the confidence interval. For the vast majority of simulations with large N_eb, the upper limits of the 95% confidence intervals reaches infinity. The authors concluded that the Heterozygosity Excess estimator is quite effective for very small numbers of parents, up to 30 individuals, if the sample size is larger than 200 and the cumulative number of alleles is more than 80. Larger numbers of alleles or loci can compensate for smaller sample size. If N_eb is large (50–100 individuals), samples size and the number of independent alleles should be much larger (500 – 1000 individuals, 450 – 900 independent alleles). To summarize, the Heterozygosity Excess (HE) method is a little biased estimator with very low precision. The estimator is useful only for very small random mating populations when many markers are genotyped from a large sample (Wang, 2005; Leberg, 2005). The assumptions of no mutation and no selection in the HE method are valid in general, because only one generation is concerned and markers are "neutral". The assumption of a single isolated population without immigration is violated in some natural populations. The strongest assumption made by the HE method seems to be random mating. When mating is not at random, the HE generated by genetic drift can easily be overwhelmed by that generated by non random mating (Wang, 1996, 2005). However, the method has some interesting properties: it requires only one single sample in time and is straightforward to compute (Balloux, 2004).

Software

As for the LD method, the most widely used software is *NeEstimator*, mostly used with microsatellite data (Do et al., 2014a). The other software available is Nb_HetEx (Zhdanova and Pudovkin, 2008).

Applications

The Heterozygosity Excess method for estimating effective population size has been less used than the two previous methods (Launey et al., 2001; Nomura, 2009; Zhivotovsky et al., 2015).

4.6 Estimation of effective population with Molecular Co-ancestry

The molecular co-ancestry between two individuals is the probability that two randomly sampled alleles from the same locus in two individuals are identical. It possesses a straightforward relationship with genealogical co-ancestry and can be used to assess genetic diversity within and between populations (Nomura, 2008).

Let f_t be the coancestry among two randomly sampled individuals in generation t, and P be the probability that two randomly sampled alleles each from different individuals in generation t come from the same individual in generation $t-1$.

$$f_t = P\left(\frac{1+F_{t-1}}{2}\right) + (1-P)f_{t-1} \tag{20}$$

where F_{t-1} is the inbreeding coefficient of individuals in generation $t-1$.

The effective number of breeders, N_{eb} can be described as

$$N_{eb} = \frac{1}{P} \tag{21}$$

Like the Heterozygosity Excess method, the Molecular Co-ancestry method can estimate the effective population size of breeders from a simple formula if the actual N_{eb} is very small. It also suffers from a poor precision of estimates (Luikart et al., 2010; Gilbert and Whitlock, 2015).

As for the previous methods, the most widely used software is *NeEstimator*, mostly used with microsatellite data (Do et al., 2014a). Another software is available *Molkin* (Gutierrez et al., 2005).

4.7 Summary

Method	Principle	Limits	Software & references
Multiple samples			
Temporal	Changes in allele frequencies between samples separated in time from a population, which reflect genetic drift, are used to calculate the effective population size.	Requires population samples separated by at least two generations.	$N_e estimator$ (Do et al., 2014a)
Single sample			
Linkage-Disequilibrium	Measures the non random association of alleles at different loci. LD measure is related to a signal from the genetic drift inside the population and to a signal related to the finite sample. Those signals are also related to the effective population size which allow its inference.	Potentially strongly biased by substructures, overlapping generations, migration and small samples.	$N_e estimator$ (Do et al., 2014a), $OneSAMP$ (Tallmon et al., 2008)
Heterozygosity Excess	The proportion of heterozygotes in the population is measured and compared to the expected proportion of heterozygotes As the proportion of heterozygotes is linked to the effective population size, it can be inferred.	Effective only for a very small number of parents. Very low precision.	$N_b HetEx$ (Zhdanova and Pudovkin, 2008), (Balloux, 2004)
Molecular co-ancestry	The estimator is obtained from a simple parameter (molecular co-ancestry) of allele sharing among sampled individuals. The molecular co-ancestry between two individuals is the probability that two randomly sampled alleles from the same locus in two individuals are identical by state.	Poor precision and applicable to small populations only.	Nomura (2008)

Table 4: Summary of the main methods available for estimating contemporary effective population size in natural populations

5. Synthesis of tools available for genetic studies

The aim of this part is to review some of the available tools for studying population genetics. Literature reviews, open source software developed by scientist and R-packages are described. This review is non-exhaustive.

5.1 Review articles

The following table summarizes the main literature reviews on the use of genetic material in ecology, fisheries or for estimating effective population size.

Year	Title	Reference
2012	A review of the application of molecular genetics for fisheries management and conservation of sharks and rays	Dudgeon et al. (2012)
2012	Molecular markers: progress and prospects for understanding reproductive ecology in elasmobranchs	Portnoy and Heist (2012)
2011	Understanding and estimating effective population size for practical application in marine species management	Hare et al. (2011)
2010	Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches	Luikart et al. (2010)
2005	Genetic approaches for estimating the effective size of populations	Leberg (2005)

Table 5: Summary of main review articles.

5.2 Review of available population genetics software and PYTHON modules

5.2.1 Softwares and PYTHON modules

Numerous softwares are available for studying populations genetics. Almost all of them can be freely downloaded from the internet. The following list is non exhaustive but aims to give a

short review of the software available, their use, the input data type, the type of results provided, the most recent version available (at the time of this review) ...

▷ *ARLEQUIN*

ARLEQUIN (Excoffier and Lischer, 2010) is a powerful genetic package performing a wide variety of tests which can be divided into two categories: intra-population and inter-population methods. Here are some examples:

- Calculation of standard indices
- Estimation of haplotype frequency
- Test for Hardy-Weinberg equilibrium
- AMOVA (Analyses of MOlecular VAriance)
- ...

The current version (3.5.2.2) is available at the following address: http://cmpg.unibe.ch/software/arlequin35/. ARLEQUIN can be launched from R with an R package. The graphical interface is designed to allow users to rapidly select the different analyses they want to perform on their data. It can handle several types of data either in haplotypic or genotypic form: DNA sequences, microsatellite data, allele frequency data, etc...

Several data formats are accepted (GENEPOP, Biosys, Phylip, Fstat, ...) but will be saved in an Arlequin format *.arp*.

▷ *Bottleneck*

Bottleneck (Piry, 1999) is a program for detecting recent effective population size reductions from allele data frequencies. The principle as described on the internet page (http://www1.montpellier.inra.fr/CBGP/software/Bottleneck/pub.html) is the following. It computes for each population sample and for each locus the distribution of the heterozygosity expected from the observed number of alleles, given the sample size under the assumption of mutation-drift equilibrium. This distribution is obtained through simulating the coalescent process of n genes under three possible mutation models. This enables the computation of the average (H_{exp}) which is compared to the observed heterozygosity (H_{obs}, in the sense of Nei's gene diversity) to establish whether there is an Heterozygosity Excess or deficit at this locus. The distribution obtained through simulation enables also the computation of a P-value for the observed heterozygosity.

Bottleneck accepts several data formats, all are text files. The two main formats are the GENEPOP and GENETIX formats. The current version 1.2.02 is available at the following address: http://www1.montpellier.inra.fr/CBGP/software/Bottleneck/.

▷ *BottleSim*

BottleSim (Kuo and Janzen, 2003) is a computer simulation program for simulating the process of population bottlenecks. It can implement an overlapping-generation model and simulate a wide range of scenarios regarding changes in population sizes. An option of generating raw genotypic data output is also available. The raw genotypic data output file contains the genotypic data from the last year of each iteration. The genotypic data output is in GENEPOP format.

The current version 2.6 is available at the following address: http://chkuo.name/software/BottleSim.html.

▷ *CERVUS*

CERVUS (Kalinowski et al., 2007) is a computer program for the assignment of parents to their offspring using genetic markers. It analyses genetic data from **codominant genetic markers** such as microsatellites (STRs) and SNPs. It assumes that the species is diploid and that markers are autosomal, although sex-linked markers can be used for some analyses. It also assumes that markers are inherited independently of each other, in other words that they are in Linkage-Equilibrium. The statistical method used is maximum likelihood.

The last version available is the 3.0.7 from http://www.fieldgenetics.com/pages/login.jsp.

▷ *COLONY*

COLONY (Jones and Wang, 2010) implements a maximum likelihood method to assign sibship and parentage jointly, using individual multiloci genotypes at a number of co-dominant or dominant marker loci. It can be used for estimating full- and half-sib relationships, assigning parentage, inferring mating system (polygamous/monogamous) and reproductive skew in both diploid and haplo-diploid species.

The last update is COLONY2 available at http://www.zsl.org/science/software/colony.

▷ *CoNe*

CoNe (Anderson, 2005) computes the likelihood of N_e given data from two temporally spaced genetic samples. The statistical model used is based on the coalescent method using Markov chain Monte Carlo.

The last version 1.0.1 is available at https://swfsc.noaa.gov/textblock.aspx?Division=FED&ParentMenuId=54&id=3436.

▷ *DnaSP*

DNA sequence polymorphism (Librado and Rozas, 2009) is an interactive computer program for the analysis of DNA polymorphism from nucleotide sequence data. It calculates several

measures of DNA sequence variation within and between populations and also provides some neutrality tests. The program can also conduct computer simulations based on the coalescent process. *DnaSP* can read unphased data and can reconstruct the haplotype phases. *DnaSP* can automatically read several type of formats.

The last version 5.10.1 is available at http://www.ub.edu/dnasp/.

▷ *DIY ABC*

DIYABC (Cornuet et al., 2014) is a software package for analyzing the population history using approximate Bayesian computation for DNA polymorphism data. It allows DNA sequences, microsatellites and SNPs. The program allows considering complex population histories including any combination of population divergence events, admixture events and changes in past population size (with population samples potentially collected at different times).

The last version 2.1.0 is available at http://www1.montpellier.inra.fr/CBGP/diyabc/.

▷ *EASYPOP*

EASYPOP (Balloux, 2001) is a computer program allowing to simulate population genetics datasets. It allows generating genetic data for haploid, diploid, and haplodiploid organisms under a variety of mating systems. It includes various migration and mutation models. Output can be generated for the FSTAT, GENEPOP, and ARLEQUIN genetic analysis packages.

The latest version 2.0.1 is available on the following page http://www.unil.ch/dee/en/home/menuinst/softwares--dataset/softwares/easypop.html.

▷ *FSTAT*

FSTAT (Goudet, 2001) is a computer package for PCs which estimates and tests gene diversity and differentiation statistics from co-dominant genetic markers. It can provide allelic richness per locus and sample, the F-statistics per locus and can perform jackknifing and bootstrapping over loci. The current version 2.9.3.2 is available at http://www2.unil.ch/popgen/softwares/fstat.htm. The input file must be in FSTAT or GENEPOP format. The output format FSTAT.

▷ *GENEPOP*

GENEPOP (Rousset, 2008) is a population genetics software package which computes exact tests from Hardy-Weinberg equilibrium, population differentiation or F-statistics estimates for example. The input format is the GENEPOP format and the program allows the conversion into several widely used formats, e.g. FSTAT.

The current version 4.2 is available at http://genepop.curtin.edu.au/.

▷ *GENETIX*

GENETIX (Belkhir et al., 1996) computes several basic parameters of population genetics such as Nei's D and H, Wright's F-statistics. It also proposes permutation-based statistical inference procedures, jackknifing and bootstrapping. The program handles only multiloci genotypes on diploid organisms. Data can be written directly in the program or can be import from FSTAT, GENEPOP formats.

The current version 4.05 is available at http://www.genetix.univ-montp2.fr/genetix/constr.htm#download.

▷ *LAMARC*

LAMARC (Likelihood Analysis with Metropolis Algorithm using Random Coalescence) (Kuhner, 2006) is a program which estimates population-genetic parameters such as population size, population growth rate, recombination rate, and migration rates. It approximates a summation over all possible genealogies that could explain the observed sample, which may be sequence, SNP, microsatellite, or electrophoretic data. All methods used in the program are derived from the coalescence theory. It requires random samples from each sub-population. A converter integrated to the program can convert PHYLIP files.

The latest version 2.1.0 is available at http://evolution.genetics.washington.edu/lamarc/lamarc_download.html.

▷ *MLNE*

MLNE (Wang and Whitlock, 2003) is a program for calculating maximum likelihood estimates of effective population size (N_e) and migration rate from the observed temporal and spatial differences in marker allele frequencies.

The software is available at the following page http://www.zsl.org/science/software/mlne.

▷ *MolKin*

MolKin (Gutierrez et al., 2005) is a population genetics computer program that conducts several genetic analyses on multilocus information. Primary functions carried out by *MolKin* are the computation of the between individuals (and populations) Molecular Co-ancestry coefficients, the Kinship distance at individual and population levels. Additionally, users can compute with *MolKin* a set of among populations, genetic distances and F-statistics (Wright, 1950) from multilocus information.

The latest version 3.0 is available at https://pendientedemigracion.ucm.es/info/prodanim/html/JP_Web.htm.

▷ *ONeSAMP*

ONeSAMP (Tallmon et al., 2008) uses approximate Bayesian computation to estimate effective population size from a sample of microsatellite genotypes. It requires an input file of sampled individuals' microsatellite genotypes along with information about several sampling and biological parameters. The program provides an estimate of effective population size, along with 95% credible limits.

Unavailable (04/05/2018).

▷ *N_eEstimator*

N_eEstimator (Do et al., 2014a) is a tool for estimating contemporary effective population size (N_e) using multi-locus diploid genotypes from population samples. Data can be microsatellites or SNPs. Four methods are available: three single-sample (point estimation) methods and one two-sample (temporal) method. The single sample methods are the Linkage-Disequilibrium method, the Heterozygosity Excess method and the Molecular co-ancestry method. The user needs to provide genotypic data in one of the accepted formats (FSTAT, GENEPOP).

The last version 2.01 is available at http://www.molecularfisherieslaboratory.com.au/neestimator-software/.

▷ *POPGENE*

POPGENE (Yeh and Boyle, 1997) is a user-friendly computer freeware for the analysis of genetic variation among and within populations using co-dominant and dominant markers. It computes both comprehensive genetic statistics (e.g., allele frequency, gene diversity, genetic distance, G-statistics, F-statistics) and complex genetic statistics (e.g., gene flow, neutrality tests, Linkage-Disequilibrium, multi-locus structure).

The current version 1.32 is available at https://www.ualberta.ca/~fyeh/popgene_download.html.

▷ *simuPOP*

simuPOP (Peng and Kimmel, 2005) is a general-purpose individual-based forward-time population genetics simulation environment. It models individuals with genotypes and simulates the transmission of individual genotype when a population evolves generation by generation. Although the basic evolutionary scenario follows a discrete non-overlapping generation model, aged structured populations can be mimicked using special non-random mating schemes. simuPOP evolves populations forward in time, subject to arbitrary number of genetic and environmental forces such as mutation, recombination, migration and population/subpopulation size changes. Statistics of populations can be calculated and visualized dynamically. simuPOP is provided as a number of Python modules, which provide a large number of Python objects and functions, including population, mating schemes, operators (objects that manipulate populations) and simulators to

coordinate the evolutionary processes. Users have to write a Python script to link these modules and run simulations.

The procedure to install simuPOP is available at the following address: http://simupop.sourceforge.net/Main/Download.

▷ *STRUCTURE*

STRUCTURE (Pritchard et al., 2000) is a software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPS, microsatellites, RFLPs and AFLPs.

The current version 2.3.4 is available at http://pritchardlab.stanford.edu/structure.html.

5.2.2 Use of population genetics software for estimating contemporary N_e

The software used in some published studies for estimating contemporary Ne are listed in table 6. The table compiles a few examples and is in no way an exhaustive review.

Case studies and publication	*N_eEstimator*	*OneSAMP*	*COLONY*	Other software
Cetaceans populations Waples (1991)				PP
Human populations Tenesa et al. (2007)				PP
Crocodile populations Bishop et al. (2009)	X X			TM3.1
Ruminant populations Cervantes et al. (2011)	X			Molkin
Gecko populations Hoehn et al. (2012)	X X	X		
Mackerel populations Macbeth et al. (2013)	X			
Salmon populations Van Doornik et al. (2013)	X			SALMONNb
Sturgeon populations Wilson et al. (2014)	X X X			
Shark populations Dudgeon and Ovenden (2015)	X			
Fly population Francuski and Milankov (2015)	Methods not precised			
Simulation only Gilbert and Whitlock (2015)	X X X X			
Cougar populations Juarez et al. (2015)		X	X	*CoN_e*; *MLN_e*
Snake populations Monzón-Argüello et al. (2015)	X X X			
Salmon populations Perrier et al. (2015)	X			
Endemic fishes populations Pilger et al. (2015)	X		X	

Table 6: Software used for calculating the contemporary effective population size and the associated method: Linkage-Disequilibrium, Heterozygosity Excess, Co-ancestry method, Temporal estimation method. Personnal Programm: PP, *N_eEstimator*: Do et al. (2014a), *ONeSAMP*: Tallmon et al. (2008), *COLONY*: Jones and Wang (2010), *SN_eP*: Barbato et al. (2015), *TM*3.1: Berthier et al. (2002), *SALMONNb*: Waples et al. (2006), *CoN_e*: Anderson (2005), *MLN_e*: Wang and Whitlock (2003), *MolKin*: (Gutierrez et al., 2005).

5.3 R-packages for studying population genetics

As for software, several R package exist for studying population genetics but only a few for estimating effective population size. Here, we only present packages which are still available on CRAN (https://cran.r-project.org/) or which can be downloaded directly from the programmer's website. The following package descriptions are those provided on the CRAN R Project web page of the package or the programmer's own description.

Only two packages are available for calculating the effective population size. The first one uses a two samples method and the second one calculates the effective population size of simulated populations with known demographic data. No R-package which calculates the effective population size of natural population with single sample methods was found when this review was compiled (2015). However, several packages can estimates Linkage-Disequilibrium or Heterozygosity Excess.

▷ *adegenet*: tool-set for the exploration of genetic and genomic data.

Adegenet provides formal classes for storing and handling various genetic data, including genetic markers with varying ploidy and hierarchical population structure, alleles counts by populations, and genome-wide SNP data. It also implements original multivariate methods (DAPC, sPCA), graphics, statistical tests, simulation tools, distance and similarity measures, and several spatial methods. A range of both empirical and simulated data sets is also provided to illustrate various methods (Jombart, 2008; Jombart and Ahmed, 2011).

User manual: https://cran.r-project.org/web/packages/adegenet/adegenet.pdf.

▷ *ape*: Analyses of Phylogenetics and Evolution

This package provides functions for reading, writing, plotting, and manipulating phylogenetic trees, analyses of comparative data in a phylogenetic framework, ancestral character analyses, analyses of diversification and macro-evolution, computing distances from allelic and nucleotide data (Paradis et al., 2004).

User manual: https://cran.r-project.org/web/packages/ape/ape.pdf.

▷ *diveRsity*: A comprehensive, general purpose population genetics analysis package

This package allows the calculation of both genetic diversity partition statistics, genetic differentiation statistics, and locus informativeness for ancestry assignment. It also provides users with various option to calculate bootstrapped 95% confidence intervals both across loci, for pairwise population comparisons, and to plot these results interactively. Parallel computing capabilities and pairwise results without bootstrapping are provided. Weir and Cockerham's 1984 F-statistics are also calculated. Various plotting features are also provided, as well as Chi-square tests of genetic heterogeneity are also provided. Functionality for the calculation of various diversity parameters

is possible for RAD-seq derived SNP data sets containing thousands of marker loci. A shiny application for the development of microsatellite multiplexes is also available.

User manual: https://cran.r-project.org/web/packages/diveRsity/diveRsity.pdf.

▷ *gap*: genetic analysis package

This package is designed as an integrated package for genetic data analysis of both population and family data. Currently, it contains functions for sample size calculations of both population-based and family-based designs, probability of familial disease aggregation, kinship calculation, statistics in linkage analysis, and association analysis involving genetic markers including haplotype analysis with or without environmental covariates (Zhao, 2015).

User manual: https://cran.r-project.org/web/packages/gap/gap.pdf.

▷ *Geneland*: detection of structure from multilocus genetic data

This package provides tools for stochastic simulations and MCMC inferences of structure from genetic data.

User manual: https://cran.r-project.org/web/packages/Geneland/Geneland.pdf.

▷ *genetics*: population genetics

The package *genetics* provides classes and methods for handling genetic data. It includes classes to represent genotypes and haplotypes at single markers up to multiple markers on multiple chromosomes. Functions include allele frequencies, flagging homo/heterozygotes, flagging carriers of certain alleles, estimating and testing for Hardy-Weinberg disequilibrium, estimating and testing for Linkage-Disequilibrium...

User manual: https://cran.r-project.org/web/packages/genetics/genetics.pdf.

▷ *hapassoc*: inference of trait associations with SNP haplotypes and other attributes using the EM algorithm

The package is used for inference of trait associations with haplotypes and other covariates in generalized linear models. The functions are developed primarily for data collected in cohort or cross-sectional studies. They can accommodate uncertain haplotype phase and handle missing genotypes at some SNPs (Burkett et al., 2006).

User manual: https://cran.r-project.org/web/packages/hapassoc/hapassoc.pdf.

▷ *haplo.stat*: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous

A suite of R routines for the analysis of indirectly measured haplotypes. The statistical methods assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown

linkage phase of the genetic markers).

User manual: https://cran.r-project.org/web/packages/haplo.stats/haplo.stats.pdf.

▷ *hwde*: models and tests for departure from Hardy-Weinberg equilibrium and independence between loci

Fits models for genotypic disequilibrium.

User manual: https://cran.r-project.org/web/packages/hwde/hwde.pdf.

▷ *LDcorSV*: Linkage-Disequilibrium corrected by the structure and the relatedness

The package provides a set of functions which aim is to propose four measures of Linkage-Disequilibrium: the usual r^2 measure, the r^2_S measure (r^2 corrected by the structure sample), the r^2_V (r^2 corrected by the relatedness of genotyped individuals), the $r^2_V S$ measure (r^2 corrected by both the relatedness of genotyped individuals and the structure of the sample).

User manual: https://cran.r-project.org/web/packages/LDcorSV/LDcorSV.pdf.

▷ *NB*: Maximum Likelihood method in estimating effective population size from genetic data

The allele frequencies in a closed population change over time, which is known as genetic drift. The magnitude of change is directly related to N_e. This package aims to estimate this quantity from genetic samples collected over multiple time points.

User manual: https://cran.r-project.org/web/packages/NB/NB.pdf.

▷ *NEff*: Calculating effective sizes based on known demographic parameters of a population

This package estimates effective population size with data obtained within less than a generation but considering demographic parameters is possible. This individual based model uses demographic parameters of a population to calculate annual effective sizes and effective population sizes (per generation). A defined number of alleles and loci will be used to simulate the genotypes of the individuals. Step-wise mutation rates can be included. Variations in life history parameters (sex ratio, sex-specific survival, recruitment rate, reproductive skew) are possible. These results will help managers to define whether existing populations as viable or not.

Demographic parameters of a population are used as input values to calculate annual effective sizes and effective population sizes (per generation). The demographic input are the age of sexual maturity, the number of offspring per female per year, the sex ratio, the female and male variances in reproductive success and the female, male and juvenile survival rates. Heterozygosity over time is observed and used to calculate effective sizes. The population can be adapted to every life history parameter combination or genetic variation.

```
> population(max.repeat=3, max.time=50, Na=200, n.recruit.fem=2, surv.ad.fem=0.7,
+ N.loci=190, N.allele=2)
I've done the repeat number 1
I've done the repeat number 2
I've done the repeat number 3


Summary table of an optimal behaviour of your population:
$parameters
                                    parameter    value 95% lower CI 95% upper CI
1       slope of heterozygosity loss per year  -0.0011      -0.0013       -9e-04
2 slope of heterozygosity loss per generation  -0.0033      -0.0036      -0.0029
3                       mean generation length      3.2        2.694        3.706
4                              Ny[simulation] 457.4719     421.0097      493.934
5                                   Ny[calc]  402.224         <NA>         <NA>
6                              Ne[simulation] 153.7963     132.7237     174.8688
7                                   Ne[calc]  125.695         <NA>         <NA>
```

Figure 9: Example of application of the package *NEff*. A population of 200 individuals is simulated over 50 years. The survival rate of the female is 0.7 and the number of offspring per female per year is 2. Genetic data used are simulated as 490 bi-allelic SNPs. The simulated effective population size is estimated at 154 (IC 95%: [133; 175]) and the calculated effective population size is estimated at 125.7.

User manual: https://cran.r-project.org/web/packages/NEff/NEff.pdf.

▷ *NeON*: R-package to estimate human effective population size and divergence time from patterns of Linkage-Disequilibrium between SNPs

The *NeON* R package has been designed to explore population's LD patterns in order to reconstruct two key parameters of human evolution: the effective population size and the divergence time between populations. *NeON* starts with binary or pairwise-LD PLINK files, and allows (a) to assign a genetic map position using HapMap (NCBI release 36 or 37) (b) to calculate the effective population size over time exploiting the relationship between N_e and the average squared correlation coefficient of LD ($r2LD$) within predefined recombination distance categories, and (c) to calculate the confidence interval about N_e based on the observed variation of the estimator across chromosomes. This package also allows to estimating the divergence time between populations given the Ne values calculated from the within-population LD data and a matrix of between-populations F_{ST}. These routines can easily be adapted to any species whenever genetic map positions are available.

Package and user manual available at http://www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software.

▷ *pegas*: population and evolutionary genetics analysis system

This package provides functions for reading, writing, plotting, analysing, and manipulating allelic and haplotypic data, and for the analysis of population nucleotide sequences and microsatellites including coalescence analyses (Paradis, 2010).

User manual: https://cran.r-project.org/web/packages/pegas/pegas.pdf.

▷ *Popgen*: statistical and population Genetics

This is a package that implements a variety of statistical and population genetic methodology, for example, LD measures from genotypes or haplotypes, clustering of SNPs, inferences of population structure, etc...

User manual: https://cran.r-project.org/web/packages/popgen/popgen.pdf.

▷ *poppr*: Population genetic analyses for hierarchical analysis of partially clonal populations built upon the architecture of the 'adegenet' package.

This package provides tools for population genetic analysis, which include genotypic diversity measures, genetic distances with bootstrap support, native organization and handling of population hierarchies, and clone correction (Kamvar et al., 2014).

User manual: https://cran.r-project.org/web/packages/poppr/poppr.pdf.

Bibliography

Alter, S. E., Rynes, E., and Palumbi, S. R. (2007). DNA evidence for historic population size and past ecosystem impacts of gray whales. *Proceedings of the National Academy of Sciences*, 104(38):15162–15167.

Anderson, E. C. (2005). An Efficient Monte Carlo Method for Estimating Ne From Temporally Spaced Samples Using a Coalescent-Based Likelihood. *Genetics*, 170(2):955–967.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10):e3376.

Balloux, F. (2001). EASYPOP (Version 1.7): A Computer Program for Population Genetics Simulations. *Journal of Heredity*, 92(3):301–302.

Balloux, F. (2004). Heterozygote excess in small populations and the heterozygote excess effective population size. *Evolution*, 58(9):1891–1900.

Barbato, M., Orozco-terWengel, P., Tapio, M., and Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, 6.

Beaumont, A. R., Boudry, P., and Hoare, K. (2010). *Biotechnology and genetics in fisheries and aquaculture*. Blackwell, Chichester ; Ames, Iowa, 2nd ed edition.

Beaumont, M. A. (1999). Detecting population expansion and decline using microsatellites. *Genetics*, (153):2013–2029.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme, F. (1996). GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations.

Berthier, P., Beaumont, M. A., Cornuet, J.-M., and Luikart, G. (2002). Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetic Society of America*, 160(2):741–751.

Bishop, J. M., Leslie, A. J., Bourquin, S. L., and O'Ryan, C. (2009). Reduced effective population size in an overexploited population of the Nile crocodile (*Crocodylus niloticus*). *Biological Conservation*, 142(10):2335–2341.

Brumfield, R. T., Beerli, P., Nickerson, D. A., and Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, 18(5):249–256.

Burkett, K., Graham, J., and McNeney, B. (2006). hapassoc: software for likelihood inference of trait associations with SNP haplotypes and other attributes. *J Stat Soft*, 16(2):1–19.

Castro, A. L. F., Stewart, B. S., Wilson, S. G., Hueter, R. E., Meekan, M. G., Motta, P. J., Bowen, B. W., and Karl, S. A. (2007). Population genetic structure of Earth's largest fish, the whale shark (*Rhincodon typus*). *Molecular Ecology*, 16(24):5183–5192.

Caughley, G. (1994). Directions in conservation biology. *The Journal of Animal Ecology*, 63(2):215.

Cervantes, I., Pastor, J., Gutiérrez, J., Goyache, F., and Molina, A. (2011). Computing effective population size from molecular data: The case of three rare Spanish ruminant populations. *Livestock Science*, 138(1-3):202–206.

Chevolot, M. (2006). *Assessing genetic structure of thornback ray, Raja clavata : A thorny situation?* PhD thesis, University of Groningen.

Cockerham, C. C. and Weir, B. S. (1977). Digenic descent measures for finite populations. *Genetical Research*, 30(02):121.

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., and Estoup, A. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30(8):1187–1189.

Darwin, C. (1859). *The origin of species.* Collins classic. Harper Press, 2011 edition.

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., and Ovenden, J. R. (2014a). NeEstimator V2: re-implementation of software for the estimation of contemporary effective population size Ne from genetic data. *Molecular Ecology Resources*, 14(1):209–214.

Do, K.-T., Lee, J.-H., Lee, H.-K., Kim, J., and Park, K.-D. (2014b). Estimation of effective population size using single-nucleotide polymorphism (SNP) data in Jeju horse. *Journal of Animal Science and Technology*, 56(1):28.

Dudgeon, C. L., Blower, D. C., Broderick, D., Giles, J. L., Holmes, B. J., Kashiwagi, T., Krück, N. C., Morgan, J. A. T., Tillett, B. J., and Ovenden, J. R. (2012). A review of the application of molecular genetics for fisheries management and conservation of sharks and rays. *Journal of Fish Biology*, 80(5):1789–1843.

Dudgeon, C. L. and Ovenden, J. R. (2015). The relationship between abundance and genetic effective population size in elasmobranchs: an example from the globally threatened zebra shark *Stegostoma fasciatum* within its protected range. *Conservation Genetics*, 16(6):1443–1454.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, (1):47–50.

Excoffier, L. and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3):564–567.

Excoffier, L., Smouse, P., and Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2):479–491.

Francuski, L. and Milankov, V. (2015). Assessing spatial population structure and heterogeneity in the dronefly: spatial population structure in the dronefly. *Journal of Zoology*, pages n/a–n/a.

Franklin, I. (1980). Evolutionary change in small populations. In *Conservation biology: an evolutionary-ecological perspective*, pages 135–149. Sinauer Associates, Sunderland, Mass.

Gilbert, K. J. and Whitlock, M. C. (2015). Evaluating methods for estimating local effective population size with and without migration: estimating N_e in the presence of migration. *Evolution*, 69(8):2154–2166.

Goudet, J. (2001). FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3).

Grant, W. and Waples, R. S. (2000). Scales of temporal and spatial genetic variability in marine fishes: implications for fisheries oceanography. In *Fisheries Oceanography: Fish Biology and Aquatic Resources*, pages 61–93. Blackwell science edition.

Gutierrez, J. P., Royo, L. J., Alvarez, I., and Goyache, F. (2005). MolKin v2.0: A Computer Program for Genetic Analysis of Populations Using Molecular Coancestry Information. *Journal of Heredity*, 96(6):718–721.

Hamilton, M. B. (2009). *Population genetics*. Wiley-Blackwell, Chichester, UK ; Hoboken, NJ.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science*, (28):49–50.

Hare, M. P., Nunney, L., Schwartz, M. K., Ruzzante, D. E., Burford, M., Waples, R. S., Ruegg, K., and Palstra, F. (2011). Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology*, 25(3):438–449.

Hartl, D. L. (1994). *Génétique des populations*. Sciences et histoire. Médecine Sciences Flammarion.

Hartl, D. L. and Jones, E. W. (1998). *Genetics: principles and analysis*. Jones and Bartlett Publishers, Sudbury, Mass, 4th ed edition.

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., and Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges: analytical approaches. *Molecular Ecology Resources*, 11:123–136.

Hill, W. (1974). Estimation of Linkage-Disequilibrium in randomly mating populations. *Heredity*, 33(2):229–239.

Hill, W. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics research*, 38(03):209–216.

Hill, W. and Robertson, A. (1968). Linkage Disequilibrium in finite populations. *Theoretical and applied genetics*, 38(226-231).

Hoehn, M., Gruber, B., Sarre, S. D., Lange, R., and Henle, K. (2012). Can Genetic Estimators Provide Robust Estimates of the Effective Number of Breeders in Small Populations? *PLoS ONE*, 7(11):e48464.

Holsinger, K. (2001). Hardy–Weinberg Law. In *Encyclopedia of Genetics*, pages 912–914. Elsevier.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–1405.

Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21):3070–3071.

Jones, O. R. and Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10(3):551–555.

Jorde, P. and Ryman, N. (1995). Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics*, 139(2):1077–1090.

Juarez, R. L., Schwartz, M. K., Pilgrim, K. L., Thompson, D. J., Tucker, S. A., Smith, J. B., and Jenks, J. A. (2015). Assessing temporal genetic variation in a cougar population: influence of harvest and neighboring populations. *Conservation Genetics*.

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment: CERVUS likelihood model. *Molecular Ecology*, 16(5):1099–1106.

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). *Poppr* : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2:e281.

Kimura, M. (1990). *The neutral theory of molecular evolution*. Cambridge Univ. Press, Cambridge.

Kuhner, M. K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, 22(6):768–770.

Kuo, C.-H. and Janzen, F. J. (2003). BottleSim: a bottleneck simulation program for long-lived species with overlapping generations. *Molecular Ecology Notes*, 3(4):669–673.

Lamarck, J. (1809). *Philosophie géologique*. Dentu, Paris.

Lande, R. (1995). Mutation and conservation. *Conservation Biology*, 9(4):782–791.

Launey, S., Barre, M., Gerard, A., and Naciri-Graven, Y. (2001). Population bottleneck and effective size in Bonamia ostreae-resistant populations of Ostrea edulis as inferred by microsatellite markers. *Genetics Research*, 78(03):259.

Leberg, P. (2005). Genetic approaches for estimating the effective size of populations. *Journal of Wildlife Management*, 69(4):1385–1399.

Lewontin, R. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49:49–67.

Librado, P. and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11):1451–1452.

Lowe, A., Harris, S., and Ashton, P. (2004). *Ecological genetics: design, analysis, and application*. Blackwell Pub, Malden, MA, USA.

Luikart, G. and Cornuet, J.-M. (1999). Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, 151(3):1211–1216.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K., and Allendorf, F. W. (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics*, 11(2):355–373.

Macbeth, G. M., Broderick, D., Buckworth, R. C., and Ovenden, J. R. (2013). Linkage Disequilibrium Estimation of Effective Population Size with Immigrants from Divergent Populations: A Case Study on Spanish Mackerel (Scomberomorus commerson). *G3: Genes/Genomes/Genetics*, 3(4):709–717.

Meffe, G. K. and Carroll, C. R. (1997). *Principles of conservation biology*. Sinauer, Sunderland, Ma, 2nd ed edition.

Mendel, G. (1866). Versuche über Plflanzen-hybriden. (Bateson translation). *Verhandlungen des naturforschenden Ver-eines in Brünn*, (4):3–47.

Monzón-Argüello, C., Patiño-Martínez, C., Christiansen, F., Gallo-Barneto, R., Cabrera-Pérez, M. n., Peña-Estévez, M. n., López-Jurado, L. F., and Lee, P. L. M. (2015). Snakes on an island: independent introductions have different potentials for invasion. *Conservation Genetics*, 16(5):1225–1241.

Morin, P. A., Martien, K. K., and Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, 9(1):66–73.

Murray, B. W., Wang, J. Y., Yang, S.-C., Stevens, J. D., Fisk, A., and Svavarsson, J. (2008). Mitochondrial cytochrome b variation in sleeper sharks (Squaliformes: Somniosidae). *Marine Biology*, 153(6):1015–1022.

Nance, H. A., Klimley, P., Galván-Magaña, F., Martínez-Ortíz, J., and Marko, P. B. (2011). Demographic processes underlying subtle patterns of population structure in the scalloped hammerhead shark, *Sphyrna lewini*. *PLoS ONE*, 6(7):e21459.

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proc. Nat. Acad. Sci. USA*, 70(12):3321–3323.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942.

Nikolic, N. (2009). *Diversité génétique et taille efficace chez les populations de poissons sauvages : le cas du Saumon atlantique un poisson migrateur amphihalin menacé.* PhD thesis, Université Toulouse III – Paul Sabatier, Toulouse.

Nomura, T. (2008). Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications*, 1(3):462–474.

Nomura, T. (2009). Interval Estimation of the Effective Population Size from Heterozygote-Excess in SNP Markers. *Biometrical Journal*, 51(6):996–1016.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26(3):419–420.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290.

Peng, B. and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687.

Perrier, C., April, J., Cote, G., Bernatchez, L., and Dionne, M. (2015). Effective number of breeders in relation to census size as management tools for Atlantic salmon conservation in a context of stocked populations. *Conservation Genetics*.

Pigliucci, M. (2009). An Extended Synthesis for Evolutionary Biology. *Annals of the New York Academy of Sciences*, 1168(1):218–228.

Pilger, T. J., Gido, K. B., Propst, D. L., Whitney, J. E., and Turner, T. F. (2015). Comparative conservation genetics of protected endemic fishes in an arid-land riverscape. *Conservation Genetics*, 16(4):875–888.

Piry, S. (1999). Computer note. BOTTLENECK: a computer program for detecting recent reductions in the effective size using allele frequency data. *Journal of Heredity*, 90(4):502–503.

Portnoy, D. S. and Heist, E. J. (2012). Molecular markers: progress and prospects for understanding reproductive ecology in elasmobranchs. *Journal of Fish Biology*, 80(5):1120–1140.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pudovkin, A., Zaykin, D., and Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, 144(1):383–387.

Pudovkin, A. I., Zhdanova, O. L., and Hedgecock, D. (2010). Sampling properties of the heterozygote-excess estimator of the effective number of breeders. *Conservation Genetics*, 11(3):759–771.

Rai, U. K. (2006). Minimum sizes for viable population and conservation biology. *Our Nature*, 1(1).

Rasmussen, D. I. (1979). Sibling Clusters and Genotypic Frequencies. *The American Naturalist*, 113(6):948.

Reed, D. H. and Bryant, E. H. (2000). Experimental tests of minimum viable population size. *Animal Conservation*, 3(1):7–14.

Robertson, A. (1965). The interpretation of genotypic ratios in domestic animal populations. *Animal Production*, 7(03):319–324.

Robinson, J. D. and Moyer, G. R. (2013). Linkage disequilibrium and effective population size when generations overlap. *Evolutionary Applications*, 6(2):290–302.

Rogers, A. R. and Huff, C. (2009). Linkage Disequilibrium Between Loci With Unknown Phase. *Genetics*, 182(3):839–844.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8(1):103–106.

Russell, J. C. and Fewster, R. M. (2009). Evaluation of the Linkage Disequilibrium method for estimating effective population size. In Thomson, D. L., Cooch, E. G., and Conroy, M. J., editors, *Modeling Demographic Processes In Marked Populations*, pages 291–320. Springer US, Boston, MA.

Ryman, N. and Utter, F., editors (1987). *Population genetics & fishery management*. Washington Sea Grant Program : Distributed by University of Washington Press, Seattle.

Schaid, D. J. (2004). Linkage disequilibrium testing when linkage phase is unknown. *Genetics*, 166:505–512.

Schleif, R. F. (1993). *Genetics and molecular biology*. Johns Hopkins University Press, Baltimore, 2nd ed edition.

Schultz, J. K., Feldheim, K. A., Gruber, S. H., Ashley, M. V., Mcgovern, T. M., and Bowen, B. W. (2008). Global phylogeography and seascape genetics of the lemon sharks (genus *Negaprion*). *Molecular Ecology*, 17(24):5336–5348.

Schwartz, M., Luikart, G., and Waples, R. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution*, 22(1):25–33.

Selander, R. (1970). Behaviour and genetic variation in natural populations. *American Zoologist*, 10:53–66.

Shaffer, M. L. (1981). Minimum population sizes for species conservation. *BioScience*, 31(2):131–134.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236(4803):787–792.

Smith, R. L. and Smith, T. M. (2001). *Ecology & field biology*. Benjamin Cummings, San Francisco, 6th ed edition.

Sober, E., editor (1994). *Conceptual issues in evolutionary biology*. MIT Press, Cambridge, Mass, 2nd ed edition.

Soulé, M. E. and Wilcox, B. A., editors (1980). *Conservation biology: an evolutionary-ecological perspective*. Sinauer Associates, Sunderland, Mass.

Tallmon, D. A., Koyuk, A., Luikart, G., and Beaumont, M. A. (2008). Computer programs: onesamp: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, 8(2):299–301.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526.

Van Doornik, D. M., Eddy, D. L., Waples, R. S., Boe, S. J., Hoffnagle, T. L., Berntson, E. A., and Moran, P. (2013). Genetic Monitoring of Threatened Chinook Salmon Populations: Estimating Introgression of Nonnative Hatchery Stocks and Temporal Genetic Changes. *North American Journal of Fisheries Management*, 33(4):693–706.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34(3):275–305.

Wang, J. (1996). Deviation from Hardy–Weinberg proportions in finite populations. *Genetical Research*, 68(03):249.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1395–1409.

Wang, J. and Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1):429–446.

Waples, R., K., Larson, W., A., and Waples, R. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, 117(4):233–240.

Waples, R. S. (1991). Genetic method for estimating the effective size of Cetacean populations. *Rep. int. Whal. Commn*, 13:279–300.

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conservation Genetics*, 7(2):167–184.

Waples, R. S. (2015). Testing for Hardy-Weinberg Proportions: Have We Lost the Plot? *Journal of Heredity*, 106(1):1–19.

Waples, R. S. (2016). Tiny estimates of the N_e/N ratio in marine fishes: Are they real? *Journal of Fish Biology*, 89(6):2479–2504.

Waples, R. S., Antao, T., and Luikart, G. (2014). Effects of overlapping generations on Linkage Disequilibrium estimates of effective population size. *Genetics*, 197(2):769–780.

Waples, R. S. and Do, C. (2008). ldne : a program for estimating effective population size from data on linkage disequilibrium: COMPUTER PROGRAMS. *Molecular Ecology Resources*, 8(4):753–756.

Waples, R. S. and Do, C. (2010). Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, 3(3):244–262.

Waples, R. S. and England, P. R. (2011). Estimating contemporary effective population size on the basis of Linkage Disequilibrium in the face of migration. *Genetics*, 189(2):633–644.

Waples, R. S., Masuda, M., and Pella, J. (2006). salmonnb: a program for computing cohort-specific effective population sizes (Nb) in Pacific salmon and other semelparous species using the temporal method: PROGRAM NOTE. *Molecular Ecology Notes*, 7(1):21–24.

Waples, R. S. and Yokota, M. (2006). Temporal estimates of effective population size in species with overlapping generations. *Genetics*, 175(1):219–233.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen (English translations in Boyer 1963 and Jameson 1977. *Jahresh. Ver. Vaterl. Naturkd. Württemb.*, (64):369–382.

Weir, B. and Cockerham, C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(06):1358–1370.

Weir, B. S. (1979). Inferences about Linkage-Disequilibrium. *Biometrics*, 35:235–254.

Weir, B. S. (1996). *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates, Sunderland, Mass.

Wilson, C. C., McDermid, J. L., Wozney, K. M., Kjartanson, S., and Haxton, T. J. (2014). Genetic estimation of evolutionary and contemporary effective population size in lake sturgeon (*Acipenser fulvescens* Rafinesque, 1817) populations. *Journal of Applied Ichthyology*, 30(6):1290–1299.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, (16):97–159.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, pages 355–366.

Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

Wright, S. (1950). Genetical Structure of Populations. *Nature*, 166(4215):247–249.

Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, 19(3):395.

Wright, S. (1984). *Variability within and among natural populations*. Number Sewall Wright ; Vol. 4 in Evolution and the genetics of populations. Univ. of Chicago Press, Chicago, Ill., paperback ed edition.

Yeh, F. and Boyle, T. (1997). Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany*, 129(157).

Zhao, J. (2015). gap: Genetic Analysis Package. R package version 1.1-16.

Zhdanova, O. L. and Pudovkin, A. I. (2008). Nb_hetex: A Program to Estimate the Effective Number of Breeders. *Journal of Heredity*, 99(6):694–695.

Zhivotovsky, L. A., Yurchenko, A. A., Nikitin, V. D., Safronov, S. N., Shitova, M. V., Zolotukhin, S. F., Makeev, S. S., Weiss, S., Rand, P. S., and Semenchenko, A. Y. (2015). Eco-geographic units, population hierarchy, and a two-level conservation strategy with reference to a critically endangered salmonid, Sakhalin taimen Parahucho perryi. *Conservation Genetics*, 16(2):431–441.

Glossary

Allele Variant or alternative form of the DNA sequence at given locus. 5

Allele exclusion criteria Exclusion criteria of rare allele. Commonly fixed at 0.05, 0.02 or 0.01. 33

Ascertainment bias A possible form of bias that occurs when genetic loci are assumed to represent population (or species) wide genetic variation but are actually a feature of a small subset of the population (or species). 14

Bottleneck Sudden reduction in population size including a loss in genetic variation. It increases allele frequencies sampling error and has a disproportionate impact on the effective population size in later generations even if census sizes increases. 22

Census size The total number of individuals in a population including immatures. 19

Codominant genetic marker A genetic marker in which both alleles are expressed, thus heterozygous individuals can be distinguished from either homozygous state. 41

Dioecious Species having the sexual organs (male and female) upon distinct individuals. 30

Diploid A nucleus or individual having two copies of each chromosome. 12

Dispersal Leaving an area of birth or activity for another area. 8, 14

Effective population size The number of individuals in an ideal Wright-Fisher population reproducting and contributing to the alleles present in the next generation. This population experiences as much genetic drift as an actual population regardless of census size. 4, 14

evolution Change in gene frequency through time resulting from natural selection and producing cumulative changes in characteristics of a population. 3

F-statistics A set of statistics used to estimate deviations from the Hardy-Weinberg model in populations and to estimate the degree to which a group of populations is genetically subdivided. 15

Fixation The loss of alleles from a polymorphic population until only one remains, i.e., becomes monomorphic. 7, 10

Fixation index The proportion by which heterozygosity is reduced or increased relative to the heterozygosity in a randomly mating population with the same allele frequencies. 15

Gene Specific nucleotide sequence of DNA that codes for a particular protein, tRNA or rNA, usually means an exon or series of exons. 5

Gene conversion Nonreciprocal transfer of genetic information between homologous chromosome. 32

Gene flow Exchange of genetic material between populations. 8, 14, 27

Genetic drift Change in allele frequencies within a population over time due to the sampling effect of small population size. 3, 10, 27

Genetic hitchhiking Occurs when an allele changes frequency not because it itself is under natural selection, but because it is near another gene on the same chromosome that is undergoing a selective sweep. 31

Genetic marker A sequence of DNA or protein than can be screened to reveal key attributes of its state or composition and thus used to reveal genetic variation. Example : SNP, microsatellites,. . . . 11

genotype State for a particular genetic locus of an organism. 4

Haplotype Genetic data from a single chromosome. 5

Hardy-Weinberg equilibrium The proposition that genotypic ratios resulting from random mating remained unchanged from one generation to another, provided natural selection, genetic drift and mutation are absent. 10

Heterozygosity Can be seen as the probability of a gene to be heterozygous. 15

Heterozygosity Excess excess of heterozygotes in the progeny relative to the HW expected proportions due to differences in allele frequencies in males and females caused by binomial sampling error. 26

Homologous chromosomes Corresponding chromosomes from male and female parents that pair during meiosis. 29

Inbreeding Reproduction between closely related individuals; include autogamy. 15

Linkage-Disequilibrium Two alleles from different loci on the same chromosome co-occuring at a significantly greater frequency than expected by a random association. 26

Locus A specific region or position on the genome or chromosome. 5

Mating system Behavioral mechanism involved in the acquisition of a mate, including the number of mates acquired, the manner in which they are acquired, the nature of the pair bond and provision of parental care. 3

Microsatellite Short tandem repeats of a short sequence of (typically two to four) nucleotides randomly distributed throughout the genome. 11

Migration Intentional, directional, usually seasonal movements of animals between two regions or habitats; involving departure and return of the same individual; *i.e.* a round trip movement. 3, 8, 14

Minimum Viable Population The minimum effective population required to persist despite genetic drift, demographic and environmental stochasticity. 22

Mutation Transmissible change in structure of a gene or chromosome. 3, 27

Mutation rate The frequency at which a particular mutation occurs in a genome. 11

Non overlapping generations Population where individuals of each generation die before the birth of individuals from the next generation. 10

Philopatry Tendency of an organism to stay in, or return to, its home area for breeding. 22

Population (Genetics) All the individuals connected by gene flow, i.e., the gene pool. 3

Population genetic Studies of populations' reproduction and not individuals' reproduction, *i.e.* the studies of the distribution and the evolution of the alleles and genotypes frequencies in populations. 4

Population structure Heterogeneity in allele frequencies across a population caused by limited gene flow. 3, 14

Recombination Exchange of genetic material between two homologous chromosomes to break up linkage groups and yield allelic combinations not occurring in parental generations. 3, 27

Selection The influence of the environment in determining which individuals will breed and pass their genes on to the next generation and which will not breed. 3, 27

Sex ratio The relative number of males and females in a population. 10

Siblings Individuals having the same mother and the same father. 20

Single Nucleotide Polymorphism (SNP) Single base pair substitution distributed throughout the nuclear genome. The majority of SNPs are biallelic with low chances of homoplasy. 11

Subpopulation Spatially distinct unit of a population. 14

Wright-Fisher model A simplified version of the biological life cycle where all sampling to found the next generation occurs from an infinite pool of gametes built from equal contributions of all individuals. This approximation is commonly employed to model genetic drift. 10