UPPSALA
UNIVERSITET

GST and other standardized differentiation measures for a finite linear population

Mattias Siljestam

Degree project in biology, Bachelor of science, 2013
Examensarbete i biologi 15 hp till kandidatexamen, 2013
Biology Education Centre, Uppsala University
Supervisors: Martin Lascoux and Mattias Jakobsson

Index

1. Abstract

Measuring the differentiation between populations has been an active research area since Sewall Wright's publication of Wright's fixation index F_{ST} in 1943. Many researchers have focused on deriving F_{ST} estimators for the migration pattern of the island model, including Sewall Wright himself. As a complement to this commonly used model, this thesis will derive G_{ST} (the common estimator of F_{ST}) for a model having a linear population, namely subpopulations lying in a row with migration between the neighbour subpopulations. This could give a better match to reality in many cases.

As G_{ST} has been criticized as a measure of differentiation two other standardized measure are also derived: a standardized G_{ST} using the principles presented by Hedrick (2005), and the differentiation measure D_{ST} presented by Jost (2008). I extend the derivations of G_{ST}, the standardized G_{ST} and D_{ST} to cover the general case of n subpopulations lying in a row with approximate equations having good accuracy for the whole parameter space of mutation rate, migration rate and population size.

2. Introduction

2.1 Population genetics – a mathematical approach

Population genetics is the set of principles to describe and investigate speciation, population history and selection by using genetic sequences. The way of investigation is commonly comparing empirical data with mathematical models. The key founders of the area were Sewall Wright, JBS Haldane and R. A. Fisher, followed three decades later by Motoo Kimura. Kimura is well known for introducing the neutral theory of molecular evolution in 1968. The theory states that most of genetic diversity observed at the molecular level can be explained by a balance between neutral mutations and genetic drift (Kimura 1979). Today, however, most scientists agree that selection is an important factor too in many species, but that cases also exist where the evolution is nearly neutral (not caused by natural selection).

The mathematical models

In population genetics, if selection is ignored, single locus models are mainly made up of three important parameters: mutation rate, migration rate and coalescence rate. The mutation rate is the probability of a mutation taking place per allele and generation (μ). The migration rate is the probability of migration per allele and generation (m). And finally the coalescence rate is the probability that two alleles picked from the same population share the same ancestor allele in the previous generation. The coalescence rate is scaled by the effective population size (N) and is often seen in the form of $\frac{1}{2N}$. We have 2N in the denominator of the coalescence rate because of having 2N alleles in the population, assuming diploid organisms. The coalescence events are always referring to a backward perspective, and correspond to the rate of genetic drift in a forward perspective (Gillespie 2004).

In the mathematical models, populations can be divided into three types: unstructured populations, metapopulations and subpopulations. An unstructured population is assumed to have random mating among individuals within and no migration to nearby populations. A metapopulation is assumed to be structured up to several subpopulations, where each of the subpopulations having migration in some way between them and random mating within. All three types of populations are always assumed to be in equilibrium between coalescence, mutation and migration.

The mean number of generations back until coalescence occurs between two alleles is called the coalescence time (t), and is expected to be 2N in an unstructured population. The mean number of segregating sites between two alleles (θ) is estimated to be 2μt, or equivalently 4Nμ, in a random mating population. The parameter θ could also be explained as the mean number of mutation that will had occur in the two picked alleles during the t generations back to their common ancestor (Gillespie 2004).

Homozygosity (G) in a population is usually defined as the probability that two randomly picked alleles from the population are identical by state, where ”identity by state” is defined as having identical gene sequences. Heterozygosity (H) on the other hand is defined as the probability that two randomly picked alleles from the population are different by state, and equals 1 – G. The homozygosity for an unstructured population (G_U) is expected to be the following (Gillespie 2004):

Heterozygosity in unstructured populations

$$\mathrm{G_U} = \frac{1}{1+\theta} = \frac{1}{1+4\mathrm{N}\mu} \quad (1)$$

2.2 Wright's F_{ST} statistic

Sewall Wright's fixation index, F_{ST}, is frequently used to measure genetic differentiation among populations or population structure (Novembre and Di Rienzo 2009, Meirmans and Hedrick 2011). Values of F_{ST} close to zero correspond to low genetic differentiation and low population structure and high F_{ST} values to fixation of different alleles in the subpopulations, (Holsinger and Weir 2009). The interpretation and what F_{ST} exactly measure will be discussed later on.

Sewall Wright's F_{ST} statistic is actually based on his inbreeding statistic F_I. The index "ST" comes from that the F_{ST} statistic compares the genetic diversity within the **s**ubpopulations with the diversity within the **t**otal population (equivalent to the metapopulation). Whereas the index "I" comes from that the F_I statistic compares the genetic diversity for within the Individual with the diversity within their population (Kaj and Lascoux 1999, Gillespie 2004).

Estimating F_{ST}

To estimate F_{ST}, genetic data is gathered from a number of sampled subpopulations. There is then numerous ways to estimate F_{ST}. One of the most common estimators is G_{ST} (Nei 1973). This statistics is found under slightly different definitions. One is comparing the within subpopulations heterozygosity (H_S) with the heterozygosity in the total population (H_T). I will refer to this statistic with a circumflex above the G ($\hat{G}_{ST}$) to avoid confusion with the other versions (note the circumflex is not referring $\hat{G}_{ST}$ as an estimator of G_{ST}):

$\hat{G}_{ST}$, using the total population diversity

$$\widehat{\mathrm{G}}_{\mathrm{ST}} = \frac{\mathrm{H_T} - \mathrm{H_S}}{\mathrm{H_T}} \quad (2)$$

H_S is defined as the probability that two alleles randomly picked from the same subpopulation are different by state. H_T on the other hand is defined as the probability that two alleles randomly picked from all of the subpopulation are different by state (Slatkin 1991).

But when the numbers of sampled subpopulations (k) is low, $\hat{G}_{ST}$ gives an underestimation of the F_{ST} value (Meirmans and Hedrick 2011). As this article will focus on pairwise comparisons between subpopulations, the number of samples subpopulation will always be low (k = 2). I will therefore use one of the other common estimators, which I will refer to as G_{ST} (without circumflex).

G_{ST}, using the between subpopulation diversity

$$G_{ST} = \frac{H_D - H_S}{H_D} \quad (3)$$

G_{ST} compares the within subpopulation heterozygosity (H_S) with the between subpopulation heterozygosity (H_D). H_D is defined as the probability that two alleles randomly picked from different subpopulation are different by state (Weir and Cockerham 1987). This means that H_D is independent of H_S and will thereby give a generally higher value then H_T when the number of sampled subpopulations is low. This can be explained by the fact that H_T gets a strong dependency of H_S when k is low, as the probability of picking two alleles from the same subpopulation when defining H_T is equal to 1/k. In this way G_{ST} (in contrast to $\hat{G}_{ST}$) compensates for some of the underestimation for low k values.

2.3 Usage and interpretation of F_{ST}

When Sewall Wright introduced F_{ST} in his article "Isolation by distance" 1943, he presented the famous formula for F_{ST}.

F_{ST} under a Wright island model

$$F_{ST} \approx \frac{1}{1 + 4Nm} \quad (4)$$

This equation showed the existence of a direct relationship between the migration rate and F_{ST}, establishing F_{ST} as a good measurement of population structure. If this were the general case, there would be no problems using F_{ST} as a measurement of population structure as F_{ST} would have a direct relationship with the number of migrating individuals per generation (Nm).

But to obtain the formula in equation 4, among other assumptions, a low mutation rate ($\mu \ll m$) was assumed, as by the time, low-diversity systems were in focus. This is not the general case today as highly variable loci as satellite regions are frequently used. The formula in equation 4 above is actually an approximation of a more general formula that includes the mutation rate (Wright 1943, Meirmans and Hedrick 2011):

F_{ST} during island model, full formula

$$F_{ST} = \frac{1}{1 + 4Nm + 4N\mu} \quad (5)$$

When looking at the full formula in equation 5, we can see that a higher mutation rate will give a lower F_{ST} value, which can be wrongly interpreted as lower differentiation, if the high mutation rate is not accounted for. The rather unintuitive relationship that higher mutation would lower the differentiation according to F_{ST} could be explained by how F_{ST} actually defines differentiation:

F_{ST} is measuring the amount of genetic diversity that is accounted for by genetic differences among populations. This means that a higher genetic variation within a subpopulation will limit the amount of genetic differentiation that could be caused between subpopulation (Holsinger and Weir 2009, Meirmans and Hedrick 2011). This explains this peculiar relationship as the genetic variation within the populations is increased by the number of mutation per generation (Nu), and genetic differentiation between subpopulation is decreased by the number of migrants per generation (Nm).

The question if F_{ST} actually is a good measurement, considering its properties explained above, has been discussed recently. It has been criticized as a measure of differentiation, as it not always is measuring what we expect it to do. Using F_{ST} as a measurement for genetic differentiation or population structure should therefore be done with caution (Meirmans and Hedrick 2011, Jost 2008).

The relationship between F_{ST} and the mutation rate, described above, limits F_{ST} for highly variable genes, making it a biased estimate for differentiation. The common estimator $\hat{G}_{ST}$ (and G_{ST}) cannot for example reach over $1 - H_S$, equaling the within subpopulation homozygosity (Hedrick 2005). This relationship is clearly stated in figure 1 where values of F_{ST} and H_S found in *Molecular Ecology* during 2008 to 2011 are plotted. Note that F_{ST} always is less than 1 – Hs represented by the solid line. This means for example that F_{ST} could never reach over 0.1 when Hs equal 0.9, even if the two populations were totally differentiated (Meirmans and Hedrick 2011, Jost 2008)!

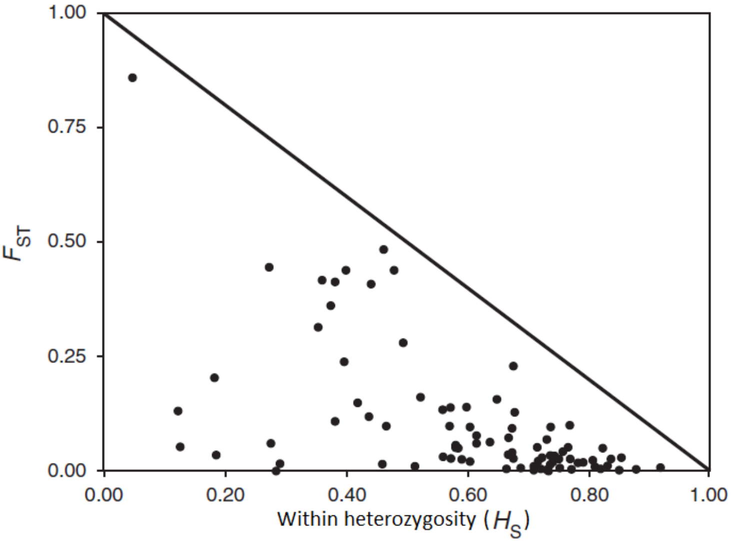


Figure 1 – The circles represents values from 84 species published in Molecular Ecology during 2008-2011. The solid line represents the maximum possible value of FST as the function 1-H_S, showing FST dependency of the within heterozygosity (Meirmans and Hedrick 2011).

2.4 Standardized F_{ST}-related statistics

Hedrick's standardized $\hat{G}_{ST}$

As F_{ST} can't reach high values for highly variable loci even if the populations are totally differentiated, as showed in Figure 1, researchers have tried to find alternative measures. One suggestion made by Hedrick (2005) is to just standardize the F_{ST} estimator by the maximum value it can reach for the within subpopulation heterozygosity, making it able to always reach between 0 and 1. By dividing $\hat{G}_{ST}$ with its max value, $\hat{G}_{ST\,(max)}$, Hedrick got the standardized statistic $\hat{G}'_{ST}$, (where the apostrophe indicates that the statistic is standardized):

Hedrick's standardized $\hat{G}_{ST}$

$$\hat{G}'_{ST} = \frac{\hat{G}_{ST}}{\hat{G}_{ST(max)}} \quad (6)$$

(Hedrick 2005 equation 4b)

Where $\hat{G}_{ST(max)}$ equals:

$$\widehat{G}'_{ST(max)} = \frac{(k-1)(1-H_S)}{k-1+H_S} \quad (7)$$

(Hedrick 2005 equation 4a)

When k is large, $\hat{G}_{ST(max)}$ equals $1 - H_S$, this would give the following equation:

Hedrick's standardized $\hat{G}_{ST}$ when k is large during Island model

$$\widehat{G}'_{ST} = \frac{H_T - H_S}{H_T(1-H_S)} \quad (8)$$

Jost's D_{ST}

Another approach presented by Jost (2008) is to just replace the denominator in the $\hat{G}_{ST}$ with $1 - H_S$. As the numerators max value regarding the within diversity is $1 - H_S$, this will also give a standardized measurement (equation 11 in Jost 2008):

Jost's D_{ST} under the Island model

$$D_{ST} = \left(\frac{k}{k-1}\right)\frac{H_T - H_S}{1 - H_S}$$

By expressing the D_{ST} in terms of H_D instead of H_T, we get rid of the dependency on k, the number of sampled subpopulations (equation 14 in Jost 2008):

Jost's D_{ST}

$$D_{ST} = \frac{H_D - H_S}{1 - H_S} \quad (9)$$

Jost D_{ST} has significantly different properties than $\hat{G}_{ST}$ as it for example increases with higher mutation rate and is unaffected by population size. But it does of course decrease with higher migration just as $\hat{G}_{ST}$ (Jost 2008, Meirmans and Hedrick 2011). This could be seen as more intuitive properties, but Jost's D_{ST} has been criticized due to some odd behaviors. One example is that simulations show that D_{ST} takes much longer time to reach equilibrium than $\hat{G}_{ST}$ and $\hat{G}'_{ST}$ (Meirmans and Hedrick 2011), while Wang criticizes D_{ST} for having tendency to give unintuitive results. In contrast to Jost, some are stating that D_{ST} is not appropriate as a primary statistic instead of $\hat{G}_{ST}$ (Meirmans and Hedrick 2011), while some states it should simply not be used (Wang 2012).

Measurement of the genetic differentiation and the population structure

By analyzing the statistics and comparing them with simulations D_{ST} has been showed to be more suitable for measuring genetic differentiation while $\hat{G}_{ST}$ and $\hat{G}'_{ST}$ being better for measuring population structure (Meirmans and Hedrick 2011).

2.5 Aims

The aim with this study is to derive a model for G_{ST}, standardized G_{ST} and D_{ST} for a finite linear population with migration and mutation, which is accurate for the whole parameter space of mutation rate (μ), migration rate (m) and coalescence rate $\left(\frac{1}{2N}\right)$.
That is to say to make a model, for all the three statistics mentioned, which is accurate for both small and large populations sizes, low or high mutation rates and low or high migration rates for a given number of subpopulations lying in a row.

Part one

This thesis will be divided into two parts, Part one and Part two, as the conclusions made in the discussion of Part one leads to the methods made in Part two.

3. Methods – Part one

3.1 The finite linear population model

The finite linear population model is the model this thesis will focus on and derive the statistics for. In the finite linear population model you have a metapopulation consisting of n equally sized subpopulations lying in a row (where $n \geq 2$). Each subpopulation has random mating within and a population size of N diploid individuals. This gives each subpopulation an allele pool of 2N alleles.

The subpopulations on the edges of the metapopulation will be referred to as edge subpopulations while the others will be called middle subpopulations. The edge subpopulations will only have one neighbour while the middle subpopulations will have two. For each neighbour subpopulation the alleles will have a probability of m to migrate to that neighbour subpopulation each generation. This means that the edge population will undergo less migration flow as they only have one neighbour, while the middle subpopulation having two subpopulations will undergo more migration. At each generation, each allele has a probability of μ to undergo a mutation. The genetic sequence for each allele is assumed to be long, making each mutation to generate unique alleles.

Note that this model has many similarities with the one dimensional case of the stepping stone model (Kimura and Wiess 1964).

Relative positions

The term relative positions will be used frequently, to describe the positions of the two picked alleles when estimating the heterozygosity. With relative position, it is meant that we do not differentiate between cases referring to identical scenarios. As the metapopulation is symmetrical by being linear we do not differentiate between mirror cases. We would for instance not differentiate the case where the two picked alleles are in the same top edge subpopulation or in the same bottom edge subpopulation; we would just say that the alleles are in the same edge subpopulation.

We, in the same way, do not differ between the two picked alleles in question. We would for instance not differentiate between the case where the first picked allele is in the top and the second in bottom edge subpopulation from the case where the first picked allele is in the bottom, and the second picked allele is in the top subpopulation. We would just state that the alleles are in different edge subpopulations.

This means that each relative position corresponds to a unique scenario. For example, in the two subpopulations scenario, when n equals 2, there are two edge subpopulations and no middle subpopulation. This will give only two different relative positions of the two alleles in question. The alleles can be in the same edge subpopulation (SE), or be in different edge subpopulations (DE). Where a migration of one of the alleles will make them get from SE position to DE position or vice versa.

When adding on more subpopulations the number of different relative populations the two alleles could be in will increase rapidly (in a second order polynomial fashion). How to find out the different relative positions for metapopulation with more than two subpopulations will be described later on when investigating the three subpopulations scenario.

3.2 The statistics G_{ST}, G'_{ST} and D_{ST} for the finite linear population model

In this study, three statistics will computed, estimating the differentiation for the finite linear population model: G_{ST}, standardized G_{ST} and D_{ST} for estimating the differentiation between the two edge subpopulation.

Using Hedrick's standardization method for G_{ST}

As G_{ST} will be used rather than $\hat{G}_{ST}$ in this study (see equation 2 and 3), I will apply Hedrick's standardizing method seen in equation 6, for G_{ST}. The standardized G_{ST} will be noted as G'_{ST} where the apostrophe indicated that the statistic is standardized.

G'_{ST} gets a simpler expression for the general case than $\hat{G}'_{ST}$, where G'_{ST} is independent of k (the number of subpopulations), in contrast to $\hat{G}_{ST}$ (see equation 6 and 7). This follow from the fact that $G_{ST(max),}$ the maximum value G_{ST} can reach is the within heterozygosity, always equals 1 – H_S and is independent of k. This is explained by G_{ST} using H_D instead of H_T, as H_D always can assume the value of one, independent of the number of sampled subpopulations (k) and H_S. This is in contrast to H_T (used by $\hat{G}_{ST}$) which is heavily dependent on H_S when k is low. When estimating H_T one picks two alleles from the same subpopulation (giving H_S) in $\frac{1}{k}$ of the cases and pick the two alleles from different population in the remaining $\frac{k-1}{k}$ of the cases (giving H_D). This leads to the following relationship where H_T is a product of the two independent terms H_D and H_S.

$$H_T = \frac{(k-1)H_D + H_S}{k} \quad (10)$$

During the no migration scenario H_D will by logical reasons equal 1 as two alleles picked from two unconnected populations always will be different by state as we assume equilibrium (note that equilibrium could take long time to reach for alleles with low mutation rate). The no migration scenario for H_T derived from equation 10 thereby gives:

No migration

$$H_T = \frac{k-1-H_S}{k}, \quad \text{when } m = 0 \rightarrow H_D = 1$$

Anyway as H_D has this simple behaviour compared with H_T, $G_{ST(max)}$ will always equals 1 - H_S, as H_D = 1 gives G_{ST} = 1 – H_S. The standardized statistic G'_{ST} will thereby lead to a simple expression for the general case (similar to the specific case for $\hat{G}'_{ST}$ when k is large):

Standardized G_{ST}

$$G'_{ST} = \frac{G_{ST}}{G_{ST(max)}} = \frac{H_D - H_S}{H_D(1-H_S)} \quad (11)$$

Relationship between G_{ST}, G'_{ST} and D_{ST}

We can see an interesting relationship between the three statistics above:

$$\mathrm{G'_{ST} = G_{ST} + D_{ST} - G_{ST} \times D_{ST}} \quad (12)$$

As:

$$\mathrm{G_{ST} + D_{ST} - G_{ST} \times D_{ST} = \frac{H_D - H_S}{H_D} + \frac{H_D - H_S}{(1 - H_S)} - \frac{H_D - H_S}{H_D} \times \frac{H_D - H_S}{(1 - H_S)}}$$

$$\mathrm{= \frac{(H_D - H_S)(1 - H_S + H_D - H_D + H_S)}{H_D(1 - H_S)} = \frac{H_D - H_S}{H_D(1 - H_S)} = G'_{ST} \;\; Q.E.D}$$

Equation 12 could be written in the following way, showing that G'$_{ST}$ follows the behaviour of G_{ST} when G_{ST} is large, and it follows D_{ST} when G_{ST} is small.

$$\mathrm{G'_{ST} = G_{ST} + (1 - G_{ST})D_{ST}} \quad (13)$$

The relationship could equivalently be described from the perspective of D_{ST} instead of G_{ST} as the behaviour is symmetrical in that sense:

$$\mathrm{G'_{ST} = D_{ST} + (1 - D_{ST})G_{ST}} \quad (14)$$

Defining the statistics for the finite linear population

All three statistics, G_{ST}, G'$_{ST}$ and D_{ST} are calculated with only the between subpopulation heterozygosity (H_D) and the within subpopulation heterozygosity (H_S). As the differentiation will be calculated between the two edge subpopulations, H_S equals the heterozygosity when picking two alleles within the same edge subpopulation (SE) and H_D the heterozygosity when picking two alleles from the different edge subpopulations (DE). The three statistics for the finite linear population model was defined in the following way, directly derived from theirs equation (see equation 3, 9, 11, 13):

The three statistics for the finite linear population

$$\mathrm{G_{ST(flp)} = \frac{H_{DE} - H_{SE}}{H_{DE}}} \quad (15)$$

$$\mathrm{D_{ST(flp)} = \frac{H_{DE} - H_{SE}}{1 - H_{SE}}} \quad (16)$$

$$\mathrm{G'_{ST(flp)} = \frac{H_{DE} - H_{SE}}{H_{DE}(1 - H_{SE})}} \quad (17)$$

Alternative G'$_{ST}$ formula

$$\mathrm{G'_{ST(flp)} = G_{ST(flp)} + \left(1 - G_{ST(flp)}\right)D_{ST(flp)}} \quad (18)$$

3.3 Model approximations

Before the derivation of the three statistics for the finite linear population model can begin, one last thing must be clarified. That is which approximation will be used to make the calculations of the modelling workable, and what type of bias the approximation will yield.

With the insight that the mutation rate (μ), migration rate (m) and the coalescence rate ($\frac{1}{2\mathrm{N}}$) is numbers much lesser than one, only one consistent approximation rule will be needed for the derivation, which I will call the small value approximation.

The Small Value Approximation

$$\text{If } x \ll 1$$

$$x^n \pm x^{n+1} \approx x^n, n \in \mathbb{R}$$

$$\text{e.g.} \quad 1 + 2m \approx 1$$

$$2\mu + \mu^2 \approx 2\mu$$

Where x could be any of μ, m or $\frac{1}{2\mathrm{N}}$.

The bias from the Small Value Approximation

The strength with only using the small value approximation is that the mutation rate, migration rate and the coalescence rate are not compared with each other. This means that the approximation will not give the model any biased parameter intervals like when you for instance assume mutation rate to be much lower than the migration rate (e.g. equation 4). The small value approximation by itself should not give any noticeable bias as long as the mutation rate, migration rate and coalescence rate by them self are much less than one.

Only used in the first step

I will later on compute the statistics for the finite linear population model. My first step will be to make a probability tree, from where I extract loops which, when computing scenarios with more than two subpopulations, will be expressed as a linear equation system. But during all these steps, the small value approximation will only be needed in the first step while creating the probability tree. This fact results from that once you have neglected all negligible terms in the first step, no more negligible terms will appear in further calculations.

Terms to neglect

There are a few terms and events that will be neglected in the probability tree. The probability that at least one of the two alleles mutate is actually $2\mu - \mu^2$, but will with the small value approximation be:

$$2\mu - \mu^2 \approx 2\mu$$

All type of double events during same generation will in the same way be neglected. For instance, the probability that migration and coalescence occur within the same generation $\left(p = \frac{2m}{2N}\right)$ is ignored relatively to the probability of just migration occurring $(p = 2m)$:

$$2m + \frac{2m}{2N} = 2m\left(1 + \frac{1}{2N}\right) \approx 2m$$

The probability that both alleles picked migrate during the same generation ($p = m^2$) will be neglected in same way.

If you do not neglect the possibility for these double events from the beginning, these extra terms will in the end anyway be neglected with the small value approximation, giving the same answer as if you neglected these double events from the beginning.

Before going further, I just want to emphasize once more that the small value approximation does not in any way compare the three parameters (mutation, migration and coalescence rate). You cannot for example neglect anything in the following expressions:

The Small Value Approximation examples

$$2m + \mu^2 = 2m + \mu^2 \quad (no\ neglection\ available)$$
$$1 + \frac{m}{\mu} = 1 + \frac{m}{\mu} \quad (no\ neglection\ available)$$

3.4 G_{ST}, G'_{ST} & D_{ST} for the finite linear population model with two subpopulations

The three statistics G_{ST}, G'_{ST} and D_{ST} will first be computed for the two-subpopulations scenario (n = 2), the simplest scenario of the finite linear population model. As stated before, two things will be needed in order to calculate all three statistics: the heterozygosity when picking two alleles from the same edge subpopulation (H_{SE}) and the heterozygosity when picking two alleles from different edge subpopulations (H_{DE}).

The two-subpopulations scenario gives two relative positions for the two picked alleles. They could be in: same edge subpopulation (SE) and different edge subpopulation (DE). The insights gained by solving this simple scenario, will be used to be able to solve the more complex scenarios with more than two subpopulations for the linear population model.

The Question loop

The heterozygosity in the model is gained by calculating the probability that two randomly picked alleles are different by state (having different gene sequences). As two randomly picked alleles always will share a common ancestor back in time, the question if they are different by state is answered by if any mutation has occurred in any of their generations back to their common ancestor. The alleles will be different by state and therefore heterozygote if at least one mutation has occurred in the generations leading back to their common ancestor.

So the question giving the heterozygosity can be formulated as:
"When picking two alleles, what is the probability that at least one mutation had occurred in any of the generations back leading to their common ancestor".

To answer that question a Question Loop using a backward perspective was used, that later on was reformulated to a probability tree. The Question Loop (see below) starts in the current generation by picking two alleles and then goes backwards one generation at a time. The loop stops if mutation occur, making the alleles heterozygous or if coalescence occur, making the alleles reach their common ancestor before mutation leading to homozygosity. The third event is migration making the alleles change their relative position. In the two-subpopulations scenario for the finite linear population model, migration will change the relative position of the alleles from being on the same edge subpopulation to being on different edge subpopulation or vice versa. When adding on more subpopulations later on, the migration pattern becomes more complicated, and this will be explained later on.

To get the H_{SE}, the two alleles are picked (starts) in the same edge subpopulation, and to get the H_{DE}, the two alleles starts in the two different edge subpopulations. After choosing the start positions of the picked alleles, The Question Loop starts by asking Question 1:

The Question Loop

Question 1: "Did mutation occur in any of the two alleles, looking one generation back?"
If the answer is yes, we stop and conclude that the alleles are different by state (therefore heterozygous) as at least one mutation occurred before reaching their common ancestor.
But if the answer is no, we ask Question 2:

Question 2: "Did coalescence occur between the two alleles, looking one generation back?"
If the answer is yes, we stop and conclude that the alleles are identical by state as no mutation had occurred during their generations back to their common ancestor. Note that this event of coalescence is only possible if the alleles are in the same subpopulation, else they can't share a parent and the answer will automatically be no.
If the answer is no, we ask Question 3:

Question 3: "Did migration occur for any of the two alleles, looking one generation back?"
If the answer is yes, one of the alleles had migrated, and their relative position will change (as explained above).

After asking Question 3, redo the loop from Question 1, which correspond to going back one generation in time. The loop continues until it stops due to mutation in Question 1 (leading to heterozygosity) or coalescence in Question 2 (leading to homozygosity).

The probability tree for the two-subpopulations model

The Question Loop was directly reformulated into a looping probability tree, corresponding to the two-subpopulations scenario. The small value approximation was used for all probabilities in the tree (Fig 2).

The probability tree gives the H_{SE} by starting in the SE point and the H_{DE} by starting in the DE point, as this corresponds to pick two alleles from the same edge subpopulation (SE) or in different edge subpopulations (DE). Each time the probability tree is looping back to one of the blue points (having the relative position indexes SE or DE), it is going back one generation. It will just like the Question Loop go one generation a time backwards until it stops due to mutation or coalescence leading the alleles to be different or same by state. The heterozygosity is obtained by putting together all the probabilities for the alleles being different by state, and homozygosity is on the other hand obtained by putting together all the probabilities for the alleles being same by state.

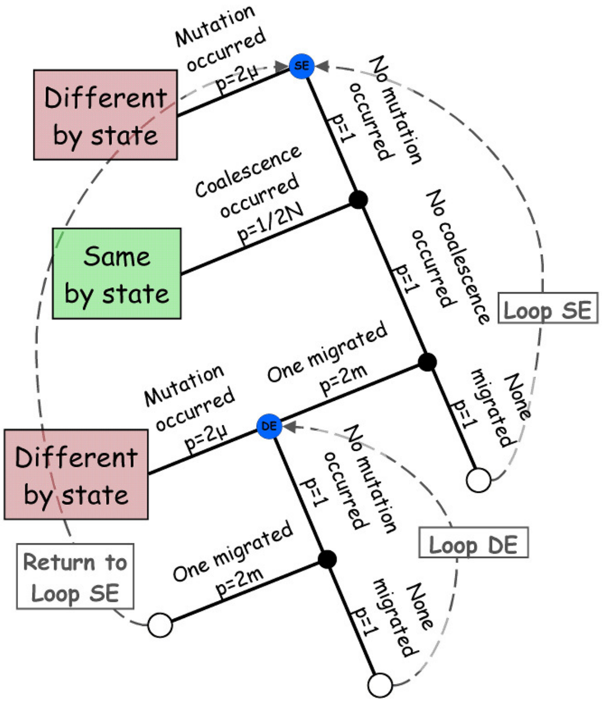


Figure 2 – The probability tree for a two subpopulation model for a finite linear population. It illustrates the probability that two randomly picked alleles will be same by state or different by state depending on if coalescence or mutation occurs first. The probability tree has two loops, Loop SE, when the alleles are located in the same edge subpopulation and Loop DE, when the alleles are located in the two different edge subpopulations. The event of migration switches between the loops.

The problem with summarizing the probabilities for being different by state (or identical by state) is that you would draw the whole probability tree. By drawing out all the loops, the probability tree would get infinite, with infinite number of "different by state boxes" (or "identical by state boxes") to summarize. To solve this problem to get a summarized probability for being different by state, a more abstract mathematical approach will be needed. This leads us to the loop scheme approach.

The Loop Scheme

The object of investigation was once more reformulated, now from a probability tree to a loop scheme, to make it even more mathematically workable. The goal was still to calculate H_{SE} and H_{DE}, by getting the summarized probability for the alleles being different by state if picked within the same edge subpopulation (SE) and if picked from different edge populations (DE).

As you see in the probability tree, two loops are marked out, Loop SE and Loop DE. By reformulating the probability tree we now consider the sequence of events as just those two loops instead of a probability tree. When switching to the loop scheme you are no longer looking one generation at a time backwards like in the probability tree. Instead you only focus on which of the three events (mutation, coalescence and migration) and which to occur first.

The loop scheme has one loop for each relative position, where each loop has a certain probability for the first event to be mutation, coalescence or migration. Where p(D_i) is the probability that mutation being the first event of the three to occur when being in loop i (making the alleles different by state). Then we have p(S_i), the probability that coalescence is the first event to occur when being in loop i (making the alleles same by state). And finally we have p(m$_i$) the probability that migration occurs first, making the alleles just change relative positions which corresponds to change loop. In the two subpopulation scenario we only have two edge subpopulations giving two loop indexes, the SE index, when alleles located in same edge subpopulation and the DE index, when alleles located in different edge subpopulation (see Fig 3).

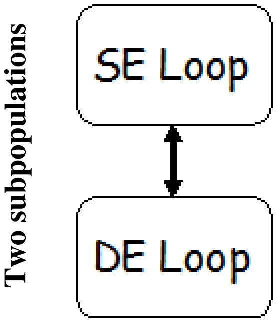


Figure 3 – The loop scheme for the two subpopulation scenario of the finite linear population model. This giving the two picked alleles two different relative positions and thereby two different loops: SE and DE. Each loop has different probabilities for mutation, coalescence and migration to occur. The migration pattern is marked out with arrows.

For example, the probability that mutation being the first event to occur in Loop SE is the probability of mutation occurring divided by the probability of any of the three events occurring during a generation in Loop SE:

$$p(D_{SE}) = \frac{2\mu}{2\mu + 2m + \frac{1}{2N}} = \frac{4N\mu}{4N\mu + 4Nm + 1}$$

In the same way we get all the probability that mutation occurs first p(D$_i$) , coalescence occurs first p(S$_i$) and the probability that migration occurs first p(m$_i$):

$$p(D_{SE}) = \frac{4N\mu}{4N\mu + 4Nm + 1} \qquad p(D_{DE}) = \frac{4N\mu}{4N\mu + 4Nm}$$

$$p(S_{SE}) = \frac{1}{4N\mu + 4Nm + 1} \qquad p(S_{DE}) = \frac{0}{4N\mu + 4Nm} = 0$$

$$p(m_{SE}) = \frac{4Nm}{4N\mu + 4Nm + 1} \qquad p(m_{DE}) = \frac{4Nm}{4N\mu + 4Nm}$$

Achieving the H_{SE} and H_{DE} for the two subpopulation scenario

You can now actually get the H$_{SE}$ and H$_{DE}$ by summarizing the probability that the alleles being different by state through the insights from the loop scheme thinking.

If you pick two alleles at random from the same subpopulation (begin in Loop SE in figure 3 or equivalently beginning at point SE in figure 2) the probability that a mutation occurs before coalescence or migration is $p(D_{SE})$ making them different by state. But if migration occurs first, with probability $p(m_{SE})$ a mutation can still occur in Loop DE. This has a probability of $p(m_{SE})p(D_{DE})$. But once more, if migration occurs again from Loop DE, mutation can occur in Loop SE, this with a probability of $p(m_{SE})p(m_{DE})p(D_{SE})$. If we keep on doing this we will get an expression for the within heterozygosity for the edge population:

$$\begin{aligned} H_{SE} = {} & p(D_{SE}) + p(m_{SE})p(D_{DE}) \\ & +p(m_{SE})p(m_{DE})p(D_{SE}) + p(m_{SE})^2 p(m_{DE})p(D_{DE}) \\ & +p(m_{SE})^2 p(m_{DE})^2 p(D_{SE}) + p(m_{SE})^3 p(m_{DE})^2 p(D_{DE}) \\ & +p(m_{SE})^3 p(m_{DE})^3 p(D_{SE}) + p(m_{SE})^4 p(m_{DE})^3 p(D_{DE}) \\ & + \cdots + p(m_{SE})^n p(m_{DE})^n p(D_{SE}) + p(m_{SE})^{n+1} p(m_{DE})^n p(D_{DE}) \end{aligned}$$

The expression above can be simplified as follow:

$$H_{SE} = p(D_{SE}) \sum_{i=0}^{\infty} p(m_{SE})^i p(m_{DE})^i + p(m_{SE})p(D_{DE}) \sum_{i=0}^{\infty} p(m_{SE})^i p(m_{DE})^i$$

This expression can be simplified once again by introducing a parameter α equalling the migration sums:

$$\alpha = \sum_{i=0}^{\infty} p(m_{SE})^i p(m_{DE})^i \rightarrow$$

$$H_{SE} = \alpha \times p(D_{SE}) + \alpha \times p(m_{SE})p(D_{DE})$$

In the same way as we obtained H_{SE} above, H_{DE} and the corresponding homozygosity G_{SE} and G_{DE} was obtained. This giving the following outcome:

$$H_{DE} = \alpha \times p(D_{DE}) + \alpha \times p(m_{DE})p(D_{SE})$$

$$G_{SE} = \alpha \times p(S_{SE}) + \alpha \times p(m_{SE})p(S_{DE})$$

$$G_{DE} = \alpha \times p(S_{DE}) + \alpha \times p(m_{DE})p(S_{SE})$$

We can now replace the probability $p(D_i)$, $p(S_i)$ and $p(m_i)$ with the three known parameters: mutation rate (μ), coalescence rate ($\frac{1}{2N}$) and migration rate (m):

$$H_{SE} = \alpha \times \frac{4N\mu(4N\mu + 8Nm)}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)}$$

$$H_{DE} = \alpha \times \frac{4N\mu(4N\mu + 8Nm + 1)}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)}$$

$$G_{SE} = \alpha \times \frac{1}{4N\mu + 4Nm + 1}$$

$$G_{DE} = \alpha \times \frac{4Nm}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)}$$

The relationship $H_{DE} = 1 - G_{DE}$, makes us able to solve for α:

$$\begin{cases} H_{DE} = 1 - G_{DE} = 1 - \dfrac{4Nm\alpha}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)} \\ H_{DE} = \dfrac{4N\mu(4N\mu + 8Nm + 1)\alpha}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)} \end{cases} \rightarrow$$

$$\frac{4N\mu(4N\mu + 8Nm + 1)\alpha}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)} = \frac{(4N\mu + 4Nm + 1)(4N\mu + 4Nm) - 4Nm\alpha}{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)} \leftrightarrow$$

$$4N\mu(4N\mu + 8Nm + 1)\alpha = (4N\mu + 4Nm + 1)(4N\mu + 4Nm) - 4Nm\alpha \leftrightarrow$$

$$\alpha = \frac{(4N\mu + 4Nm + 1)(4N\mu + 4Nm)}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} \quad (19)$$

By substituting α in the expression in equation 19, we get H_{SE} and H_{SE} only in terms of the parameters u, m and N:

$$\boxed{\begin{aligned} H_{SE} &= \frac{4N\mu(4N\mu + 8Nm)}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} \\ \\ H_{DE} &= \frac{4N\mu(4N\mu + 8Nm + 1)}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} \end{aligned}}$$

This is the point where the statistics for the two subpopulations scenario can be calculate by using equation 15, 16 and 17:

Precalculations of used terms

$$H_{DE} - H_{SE} = \frac{4N\mu}{4N\mu(4N\mu + 8Nm + 1) + 4Nm}$$

$$1 - H_{SE} = \frac{4N\mu + 4Nm}{4N\mu(4N\mu + 8Nm + 1) + 4Nm}$$

G_{ST} for the two subpopulations scenario

$$G_{ST} = \frac{H_{DE} - H_{SE}}{H_{DE}} =$$

$$= \frac{4N\mu}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} \Bigg/ \frac{4N\mu(4N\mu + 8Nm + 1)}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} = \frac{4N\mu}{4N\mu(4N\mu + 8Nm + 1)} \rightarrow$$

$$\boxed{G_{ST} = \frac{1}{1 + 8Nm + 4N\mu} \quad (20)}$$

D_{ST} for the two subpopulations scenario

$$D_{ST} = \frac{H_{DE} - H_{SE}}{1 - H_{SE}} =$$

$$= \frac{4N\mu}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} \Big/ \frac{4N\mu + 4Nm}{4N\mu(4N\mu + 8Nm + 1) + 4Nm} = \frac{4N\mu}{4N\mu + 4Nm} \rightarrow$$

$$\boxed{D_{ST} = \frac{\mu}{\mu + m} \quad (21)}$$

G'_{ST} for the two subpopulations scenario (from the perspective of G_{ST})

$$G'_{ST} = G_{ST} + (1 - G_{ST})D_{ST} \rightarrow$$

$$\boxed{G'_{ST} = \frac{1}{1 + 8Nm + 4N\mu} + \frac{8Nm + 4N\mu}{1 + 8Nm + 4N\mu} \times \frac{\mu}{\mu + m} \quad (22)}$$

(from the perspective of D_{ST})

$$G'_{ST} = D_{ST} + (1 - D_{ST})G_{ST} \rightarrow$$

$$\boxed{G'_{ST} = \frac{\mu}{\mu + m} + \frac{m}{\mu + m} \times \frac{1}{1 + 8Nm + 4N\mu}}$$

Testing result with the no migration theorem

Some model-unspecific checking can be done of our model-specific result for the three statistics for the finite linear population model with the two subpopulations.

We can use the knowledge that the case of no migration between the subpopulations will always lead to the between heterozygosity equalling one ($m = 0 \rightarrow H_D = 1$) in any migration model. This follows from the fact that the picked allele cannot coalesce with each other as none of them can migrate making them be in the same subpopulation.

If we calculate the three statistics (in there general, model unspecific form) for the no migration case will give the following:

No migration scenario ($m = 0 \rightarrow H_D = 1$)

$$G_{ST} = \frac{H_D - H_S}{H_D} = \frac{1 - H_S}{1} = 1 - H_S = G_S$$

$$D_{ST} = \frac{H_D - H_S}{1 - H_S} = \frac{1 - H_S}{1 - H_S} = 1$$

$$G'_{ST} = \frac{H_D - H_S}{H_D(1 - H_S)} = \frac{1 - H_S}{1(1 - H_S)} = 1$$

By adding the fact that a population structured into subpopulations with no migration between them will make the subpopulations act just like independent unstructured populations, as they do not interact. This leads to that the within homozygosity (G_S) for a subpopulation will in the no migration case equal the homozygosity in the case of an unstructured population (G_U), seen in equation 1.

$$\mathrm{m} = 0 \rightarrow \mathrm{G_S} = \mathrm{G_U} = \frac{1}{4\mathrm{N}\mu + 1}$$

All this can be summarized as the no migration theorem, which will be used several times in this report:

The no migration theorem (model unspecific)

$$\mathrm{G_{ST}} = \frac{1}{4\mathrm{N}\mu + 1}, \quad \text{iff } \mathrm{m} = 0 \quad (23)$$

$$\mathrm{D_{ST}} = 1, \quad \text{iff } \mathrm{m} = 0 \quad (24)$$

$$\mathrm{G'_{ST}} = 1, \quad \text{iff } \mathrm{m} = 0 \quad (25)$$

Testing result

The no migration theorem could be used as a test of our equations: G_{ST}, D_{ST} and G'_{ST} for the two-subpopulations scenario (equation 20, 21, 22). So if we substitute m with zero in these three equations, we should get the same output as in the no migration theorem (equation 23, 24, 25):

$$\mathrm{m} = 0 \rightarrow \mathrm{G_{ST}} = \frac{1}{1 + 4\mathrm{N}\mu}, \qquad \text{correct!}$$

$$\mathrm{m} = 0 \rightarrow \mathrm{D_{ST}} = \frac{\mu}{\mu} = 1, \qquad \text{correct!}$$

$$\mathrm{m} = 0 \rightarrow \mathrm{G'_{ST}} = \frac{1}{1 + 4\mathrm{N}\mu} + \frac{4\mathrm{N}\mu}{1 + 4\mathrm{N}\mu} \times \frac{\mu}{\mu} = 1, \quad \text{correct!}$$

This analytical demonstration shows that the three statistics for the finite linear population model with two subpopulations are without bias for the no migration case. The equation was also more thoroughly checked by comparing it with simulations from the MS-program seen in the “Results – Part one” section below.

3.5 G_{ST}, G'_{ST} & D_{ST} for the finite linear population model with three subpopulations

The statistics was derived for the three subpopulations using the insights gained from deriving the statistics for the two-subpopulations scenario. As the model was extended to the three-subpopulations scenario, a probability tree would have been very big and clunky, because of the increased complexity of the migration pattern. That problem was solved by going directly on the loop scheme approach, as it can describe this more complicated migration pattern in a more compact and simplified way.

Loop scheme for three subpopulations

The thinking behind the loops was the same as in the two subpopulation scenario. The three subpopulations were divided into two edge subpopulations and one middle subpopulation. This giving four different relative positions: same edge subpopulation (SE), different edge subpopulation (DE), edge and middle subpopulations (EM), only middle subpopulation (M), and forms the loop scheme seen in figure 4.

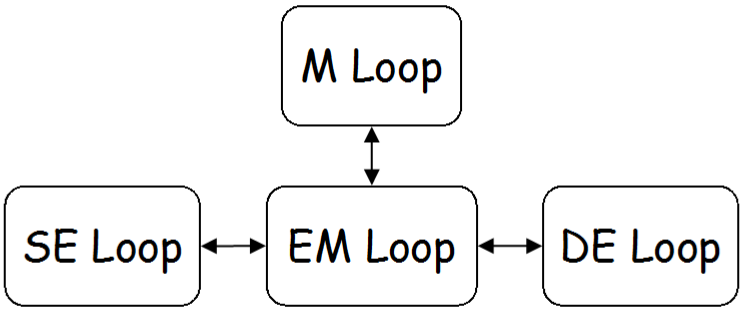


Figure 4 – The loop scheme for the three subpopulation scenario of the finite linier population model. This giving the two picked alleles four different relative positions and thereby four different loops: SE, DE, EM and M. Each loop having different probabilities for mutation, coalescence and migration to occur. The migration pattern is marked out with arrows.

Both H_{DE} and H_{SE} could actually be obtained from here using the same methods as in the two subpopulation scenario. But as this gives much more complicated calculations due to the more complex migration a final method to derive H_{SE} and H_{DE} will be introduced and used. This is the equation system approach, where the Loop chart will be expressed as an equation system having one equation and one unknown variable for each relative position.

But before moving on to the topic of the equation system approach, I will introduce a more rationalized naming system of the relative positions, as the current will be insufficient for scenarios with a larger number of subpopulations. At the same time, the migration pattern between the relative positions for a scenario with an arbitrary number of subpopulations will be described for the finite linear population model. This will be needed for the equation system approach.

Rationalize position indexing

As the number of relative subpopulations will be very large for large number of subpopulations (large n), it won't be sufficient with just giving them all a unique name arbitrated to a few letters (like SE and DE). A consistent numeric naming system will be introduced which also will work as a definition of the relative positions for general case of n subpopulation.

First the subpopulations are given numerical indices, where the indices for the n subpopulations will be 1, 2, ... , n. Then we give the positions of the alleles a numeric position index, where the index i,j indicates that the first picked allele is in subpopulation "i" and the second in subpopulation "j". For example, when n = 3, position index 1,3 would correspond to the DE position index.

But to get just the relative position, we must group up all the identical scenarios. We want to do this to get a more compact equation system in the end. As for example when n = 3, the indexes 1,3 and 3,1 would both describe an identical scenario where the alleles are located in the different edge subpopulations (DE).

The three migration pattern charts in table 1 shows how we extract the relative positions form all the possible position, and in the same way get the migration pattern between them. The migration pattern charts will correspond to the scenario of five subpopulations (n = 5), but will act like an example for the general pattern. The first migration pattern chart is the simplest (table 1A), just giving the all the thinkable positions the alleles could be in, not taking to account that some of the positions are giving identical scenarios. These are thereby not the relative positions.

As there is no difference between the first and the second picked allele, the position index i,j and position index j,i will give identical scenarios. When for example estimating heterozygosity (when n = 5), picking them in the positions 1,5 or in 5,1 would of course give the same result. The same thing would apply when picking the alleles in position 2,3 and 3,2. The general statement is that we do not differentiate between index i,j and j,i, so I will always choose the index i,j, where $i \leq j$.

This make us able to simplify the migration pattern chart by "flipping it" over the diagonal as seen in table 1B. For example, position index 1,1 (the lower left corner) has two migration arrows to position 1,2 instead of having one migration arrow to position 1,2 and one to 2,1.

This step from chart A to chart B in table 1, we remove a lot of identical scenarios, but not all. In the final step, from the chart B to chart C in table 1, we finally get the relative positions, by grouping up all the identical scenarios, making each position in the chart unique. To get there we use the insight that the metapopulation is symmetrical over its center. We thereby do not differentiate between mirror cases as the scenarios when both alleles are in the top subpopulation and when both alleles are in the bottom subpopulation. For example, when n equals 5, we do not differentiate between index 11 and 55 or between 23 and 34.

In other words, we do not differentiate between position index i_1,j_1 and position index i_2,j_2, where $i_2 = n + 1 - i_1$ and $j_2 = n + 1 - j_1$. I will always choose the index i,j, when $i \leq j$ and when $i = i_1 \leq i_2$. This make us able to simplify the migration pattern chart one last time by "flipping it" over the other diagonal giving the chart in table 1C.

The position indices SE and DE are the ones of interest to calculate the statistics, where SE always equal 1,1 and DE always equal 1,n for the finite linear population model.

Table 1 – The table is showing the migration pattern for the finite linear population model with 5 subpopulations. This example is meant to demonstrate the general pattern together with table 2. The arrows show to which position one migration could take the alleles. Each arrow corresponds to a migration probability of m per generation. The bold arrows indicate that the two alleles are within the same subpopulation, and can thereby coalesce. **A:** The 25 different positions the alleles could be found in. Here we do not take in account that many of the positions are corresponding to identical scenarios. **B:** A transformed version of A where we takes into account that the position i,j and the position j,i refer to identical scenario as we do not differentiate between the first and the second allele. **C:** This is a transformed version of B taking in to account that i_1,j_1 and the position j_2,i_2, where i_2 = n + 1 – 1 and j_2 = n + 1 - $j_{1,}$ refers identical scenario as the finite linear population is symmetrical over its center. This takes to account of the last identical scenarios, making each position unique in relation to the others. These positions are called the relative positions.

(A)

Position of the second allele (j)

Position of first allele (i)

	1	2	3	4	5
5	↓→	←↓→	←↓→	←↓→	**←↓**
4	↑↓→	←↑↓→	←↑↓→	**←↑↓→**	←↑↓
3	↑↓→	←↑↓→	**←↑↓→**	←↑↓→	←↑↓
2	↑↓→	**←↑↓→**	←↑↓→	←↑↓→	←↑↓
1	**↑→**	←↑→	←↑→	←↑→	←↑

(B)

Position of the second allele (j)

Position of first allele (i)

	1	2	3	4	5
5					**↓↓**
4				**↓↓→→**	←↑↓
3			**↓↓→→**	←↑↓→	←↑↓
2		**↓↓→→**	←↑↓→	←↑↓→	←↑↓
1	**→→**	←↑→	←↑→	←↑→	←↑

(C)

Position of the second allele (j)

Position of first allele (i)

	1	2	3	4	5
5					
4					
3			**↓↓↓↓**		
2		**↓↓→→**	←↑↓→	←←↓↓	
1	**→→**	←↑→	←↑→	←↑→	←←

The migration pattern chart for the relative positions will always end with a pyramid shape. This pyramid will be pointy when n is an odd value having the top consisting of one position which has four migration arrows down (table 1C). But when the n is an even number, the top will consist of two positions each having two migration arrows down, and two in sideways (see table 2). From table 1C and table 2 together the general pattern or relative positions and the migration pattern between can be extracted, or by just following the description in the text above.

Table 2 – The table is showing the migration pattern for relative positions for the finite linear population model having an event number of subpopulation (n = 6). This gives a complementation to the example above in table 1 where n is odd.

Position of the second allele (j)

Position of first allele (i)

	1	2	3	4	5	6
6						
5						
4						
3			↓↓→→	←←↓↓		
2		↓↓→→	←↑↓→	←↑↓→	←←↓↓	
1	→→	←↑→	←↑→	←↑→	←↑→	←←

Equation system approach

Having the method of getting the relative position and the migration pattern between them, we can move further on to the equation system approach. The scenario of three subpopulations will act as a calculation example. The first step was to get the migration pattern chart for the relative position (table 3). We can see that index 1,1 corresponds to SE, index 1,2 to M, index 1,3 to DE and index 2,2 to EM in the loop chart in figure 4.

Table 3 – The migration pattern for relative positions for the finite linear population model with three subpopulations. The arrows show to which position one migration could take the alleles. Each arrow corresponds to a migration probability of m per generation. The bold arrows indicate that the two alleles can coalesce as they are within the same subpopulation.

Position of the second allele (j)

Position of first allele (i)

	1	2	3
3			
2		↓↓↓↓	
1	→→	←↑→	←←

With the help of the migration pattern chart in table 3, an equation system was constructed, calculating the heterozygosity for the four different relative positions. This equation system has one equation for each loop, describing the probability for the alleles being different by state for that relative position, which equals the heterozygosity. The heterozygosity for the relative position i,j will thereby equal the probability that mutation occurs first in that relative position, plus the probability for the alleles to migrate to other relative position first, each multiplied with that position heterozygosity (it will be clearer when looking at the equations below).

The probability that mutation occurs first in a relative position has 2μ as its numerator. The denominator equals the probability that any event occurs during a generation in that relative position. The denominator thereby equals 2μ plus A×m (where A is the total number of migration arrows in that relative position, see for example table 3) plus $\frac{1}{2N}$ if coalescence is possible (when i = j, these relative positions are marked with bold arrows).

Each migration probability to a certain relative position has a×m as its numerator, where a is the certain number of migration arrows in the migration pattern chart, pointing to the relative position in question. The denominator is the same as described above, where $A = a_1+a_2+....$ All terms was written on the same denominator as they share the same denominator for each equation.

The general appearance of a heterozygosity equation:

$$H_{i,j} = \frac{2\mu + a_1 m H_{i_2,j_2} + a_2 m H_{i_3,j_3} + \cdots}{2\mu + (a_1 + a_2 + \cdots)m + \langle \frac{1}{2N}, \text{if } i = j \rangle}$$

To simplify the calculation both the numerator and the denominator was multiplied with 2N:

$$H_{i,j} = \frac{4N\mu + a_1 2NmH_{i_2,j_2} + a_2 2NmH_{i_3,j_3} + \cdots}{4N\mu + (a_1 + a_2 + \cdots)2Nm + \langle 1, \text{if } i = j \rangle}$$

The expression was now reduces from having three parameters: μ, m and N, to an expression having only two parameters: the average number of mutations per generation and subpopulation: U = 2Nμ, and the average number of migrating alleles per generation and per subpopulation neighbour: M = 2Nm.

$$H_{i,j} = \frac{2U + a_1 MH_{i_2,j_2} + a_2 MH_{i_3,j_3} + \cdots}{2U + (a_1 + a_2 + \cdots)M + \langle 1, \text{if } i = j \rangle}$$

The equation system for the three subpopulations scenario is expressed in equation 26.

The equation system for the finite linear population model with three subpopulations:

$$\begin{cases} H_{11} = \dfrac{2U + 2MH_{12}}{2U + 2M + 1} \\ H_{12} = \dfrac{2U + MH_{11} + MH_{22} + MH_{13}}{2U + 3M} \\ H_{22} = \dfrac{2U + 4MH_{12}}{2U + 4M + 1} \\ H_{13} = \dfrac{2U + 2MH_{12}}{2U + 2M} \end{cases} \qquad (26)$$

Note that the equation system only consists of the parameters M, U and the unknown variables H_{ij}. As both μ, m and $\frac{1}{2N}$ is considered to be small values, we can confirm that no more terms can be neglected with the small value approximation, as both M and U is one “small value” divided by and other (e.g. $U = \mu/\frac{1}{2N}$) .

The heterozygosity in one relative position is dependent on the heterozygosity in all the others when solving the equation system. But the solving gets trivial in spite of this complexity as the numbers of unknown variables equals the number of equations. MATLAB's function `solve(equation_1, equation_2,..,variable_1, variable_2,..)` was used for the solving the equation systems as the H_{ij} tends to get huge expressions when n get larger

By solving the equation system [26] the heterozygosity was obtained for all the relative positions. To calculate the three statistics G_{ST}, G'_{ST} and D_{ST} only H_{SE} and H_{DE} was needed, which in the case of n = 3 is H_{11}, and H_{13}.

$$H_{SE} = H_{11} = \frac{U(36M^3 + 72M^2U + 5M^2 + 44MU^2 + 7MU + 8U^3 + 2U^2)}{U(144M^3 + 288M^2U + 64M^2 + 176MU^2 + 64MU + 5M + 32U^3 + 16U^2 + 2U) + 12M^3 + 2M^2}$$

$$H_{DE} = H_{13} = \frac{U(144M^3 + 288M^2U + 56M^2 + 176MU^2 + 64MU + 5M + 32U^3 + 16U^2 + 2U)}{U(144M^3 + 288M^2U + 64M^2 + 176MU^2 + 64MU + 5M + 32U^3 + 16U^2 + 2U) + 12M^3 + 2M^2}$$

The expressions for H_{SE} and H_{DE} are much more complex than for the two subpopulations scenario, which is intuitive, because the equations take the complex migration pattern into account. The alleles can migrate around in all the four different relative positions before they coalesce or mutate, which gives this complex analytical expression for the heterozygosity.

The statistics G_{ST}, G'_{ST} and D_{ST} were now calculated by using MATLAB's solve for equation 15, 16 and17 to get the statistics.

The three statistics for the finite linear model with three subpopulations

$$G_{ST} = \frac{1}{1 + 2M + 2U} = \frac{1}{1 + 4Nm + 4N\mu} \tag{27}$$

$$D_{ST} = \frac{U(18M^2 + 18MU + 5M + 4U^2 + 2U)}{6M^3 + 22M^2U + 2M^2 + 18MU^2 + 5MU + 4U^3 + 2U^2} \tag{28}$$

$$G'_{ST} = \frac{(M + 2U + 6MU + 4U^2)(6M^2 + 8MU + 2M + 2U^2 + U)}{(1 + 2M + 2U)(6M^3 + 22M^2U + 2M^2 + 18MU^2 + 5MU + 4U^3 + 2U^2)} \tag{29}$$

We can see that G_{ST} (in equation 29) has the same compact expression, but that the complexity for both D_{ST} and G'_{ST} has increased.

3.6 MS-simulation

The three statistics derived for both two and three subpopulation for the finite linear population model were compared with simulation results. Richard R. Hudson's program MS (a program for generation samples under neutral models) was used to simulate a two and three subpopulation scenario for two picked alleles in the same edge and different edge subpopulations giving H_{SE} and H_{DE}.

The code in MS was the following, where each {parameter} was replaced by the numeric values:

Data to calculate H_{SE} for n = 2

```
./msdir/ms 2 {number of simulations} -t {2U} -I 2 2 0 {2M} | ./msdir/sample_stats |
cut -f 4 >datafile_SE
```

Data to calculate H_{DE} for n = 2

```
./msdir/ms 2 {number of simulations} -t {U} -I 2 1 1 {2M} | ./msdir/sample_stats |
cut -f 4 >datafile_DE
```

Data to calculate H_{SE} for n = 3

```
./msdir/ms 2 {number of simulations} -t {U} -I 3 2 0 0 0 -m 1 2 {2M} -m 2 1 {2M} -m
2 3 {2M} -m 3 2 {2M} | ./msdir/sample_stats | cut -f 4 >datafile_SE
```

Data to calculate H_{DE} for n = 3

```
./msdir/ms 2 {number of simulations} -t {U} -I 3 1 0 1 0 -m 1 2 {2M} -m 2 1 {2M} -m
2 3 {2M} -m 3 2 {2M} | ./msdir/sample_stats | cut -f 4 >datafile_DE
```

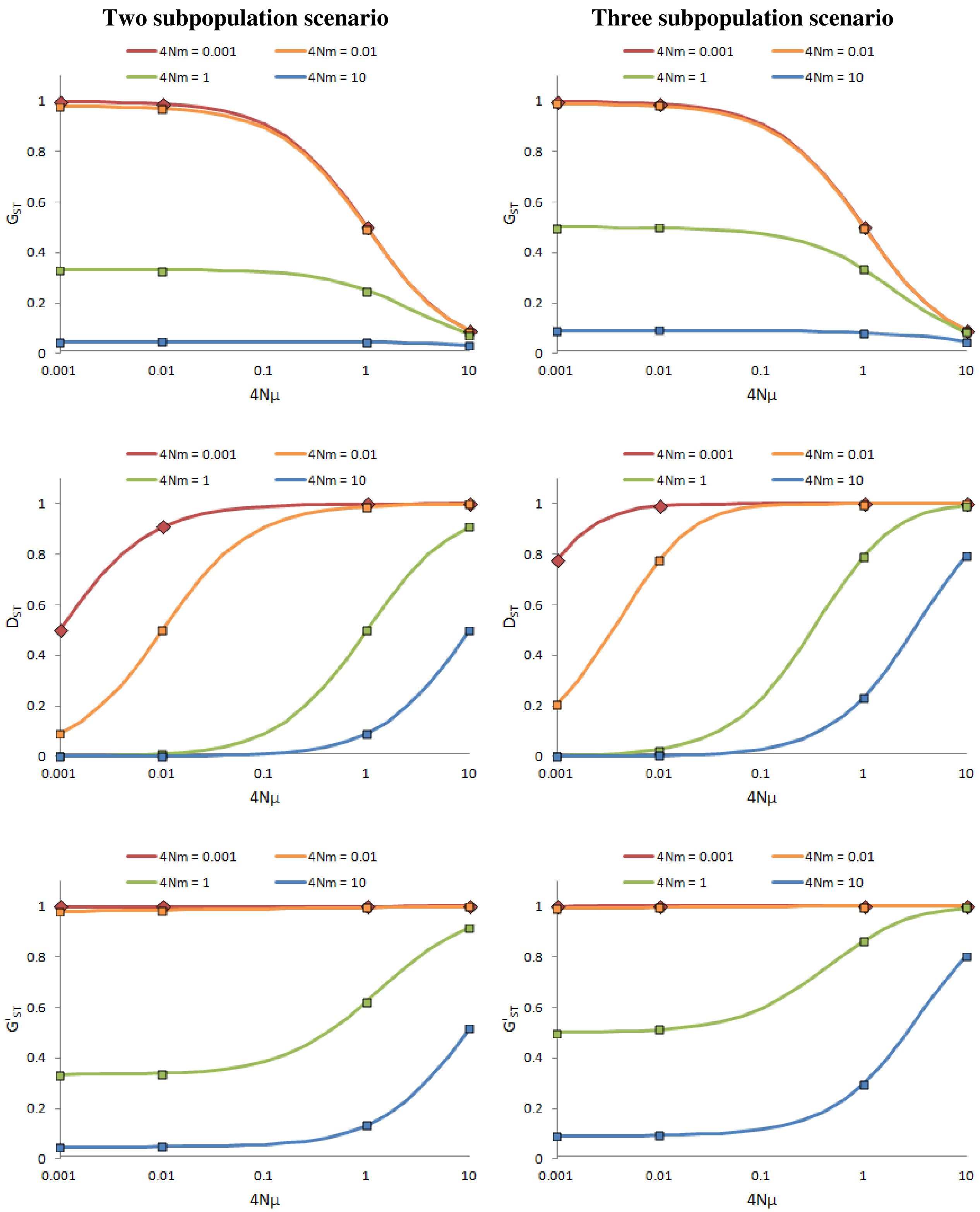
4. Results - Part one

4.1 Comparing the derived equations with MS-simulation

The derived equations for the three statistics, G_{ST},G'_{ST} and D_{ST}, were compared with the output from the MS-simulation.

The H_{SE} and H_{DE} were extracted from the simulation that was run for four different values of U and M. The simulation where run for 4Nμ (2U): 0.001, 0.01, 1 and 10. For each value of 4Nμ the simulation was run for four different values of 4Nm (2M): 0.001, 0.01, 1 and 10. This gives simulations for 16 different parameters sets. The number of simulated picked allele pairs where 10 million when 4Nμ equalled 0.001 and 1 million when 4Nμ equalled 0.01, 1 and 10. A higher number of replicates was used for lower numbers of 4Nμ due to higher variance.

The three statistics were then calculated from the heterozygosity obtained from simulation with equation 15, 16 and 17. The simulated results were then compared with the analytical derived expectations for both the two-subpopulations scenario (equation 20, 21 and 22) and the three-subpopulations scenario (equation 27, 28 and 29) seen in Figure 5.


Two subpopulation scenario

Three subpopulation scenario

Figure 5 – Plots of the three statistics G_{ST}, D_{ST} and G'_{ST} for both the two and three subpopulation scenario of the finite linear population model. The solid line represents the derived equation, based on the parameters Nm and Nμ. The squares represent the statistics calculated from heterozygosity (H_{SE} and H_{DE}) gained from simulations. The replication is 10 million simulated allele pairs for 4Nμ = 0.001 and 0.01, and 1 million simulated allele pairs for 4Nμ = 1 and 10. No error bars are included as the error was almost zero due to the high replication.

5. Discussion – Part one

5.1 Finite linear population model versus MS-simulation

We can see that the derived equations for all the three statistics G_{ST}, D_{ST} and G'_{ST} for both the two and three subpopulation scenario of the finite linear population model were giving the exact same result as the simulation. A perfect match thereby not only given for the simple equation but also for the complex equations gained for D_{ST} and G'_{ST} for the three subpopulation scenario. This is the expected result as the small value approximation was the only approximation made the derivation of the statistics.

The derived equations, giving the expected mean that would be gained from simulation, will thereby be used as the "real answer" acting more like the simulation results in Part two. This comes handy as high precision from simulations could take very long time, especially in certain parameter interval and when simulating high number of subpopulations.

5.2 The different behaviour of the three statistics

The different behaviour of the three statistics seen in figure 5 will be discussed in the Discussion – Part two.

Part two

6. Methods – Part two

6.1 G_{ST}, D_{ST} & G'_{ST} for the general finite linear population model

Deriving G_{ST}, D_{ST} & G'_{ST} for large number of subpopulations

The three statistics G_{ST}, D_{ST} & G'_{ST} could be calculated in the same way as for the three subpopulations scenario for a scenario with any number of subpopulation. This becomes just a trivial extension, making a bigger migration chart and resulting in a larger equation system to solve. The three statistics were calculated for 2 to up 12 subpopulations in this way, with the help of MATLAB.

For the scenario of two subpopulations, all three statistics had quite simple derived equations (equation 20, 21, 22). When adding up to the three subpopulation scenario, both D_{ST} and G'_{ST} got a more complex expression (equation 28, 29) while G_{ST} hade the same small expression, with just halved the value on the constant before the Nm parameter (equation 27).

But all the statistics got larger and larger expressions adding on more subpopulation, even G_{ST} (the example in table 4). For instance, the derived equation for G_{ST} is compact, having a length of 13 symbols (as expressed in MATLAB) for both the two and three subpopulation scenario. But gets an equation length of 127 symbols for the scenario of 4 subpopulations and around 15 thousand symbols for the scenario for the 12 subpopulations scenario (the equation length is taken after MATLAB has simplified the equations repeated times using the function `simplify()`). The expressions for D_{ST} and G'_{ST} grows in similar way, and do for example also get expressions with over ten thousand symbols for the 12 subpopulation scenario.

Table 4 – Example demonstrating the full equation for GST for the six subpopulation scenario

```
GST = (U^9*(4864*M + 256) + U^8*(4352*M + 40448*M^2 + 64) + U^7*(960*M +
32000*M^2 + 193664*M^3) + U*(25512*M^7 + 170744*M^8 + 270720*M^9) +
U^6*(6144*M^2 + 133504*M^3 + 590576*M^4) + U^2*(54316*M^6 + 440700*M^7 +
837328*M^8) + U^5*(21920*M^3 + 348144*M^4 + 1196976*M^5) + U^3*(65036*M^5
+ 644264*M^6 + 1470096*M^7) + U^4*(47724*M^4 + 588352*M^5 + 1630272*M^6) +
5183*M^8 + 28480*M^9 + 37440*M^10 + 256*U^10)/(U^10*(10240*M + 768) +
U^9*(14080*M + 90112*M^2 + 384) + U^8*(6400*M + 112640*M^2 + 459520*M^3 +
64) + U*(25512*M^7 + 201538*M^8 + 472704*M^9 + 321408*M^10) + U^7*(960*M +
46080*M^2 + 516864*M^3 + 1504480*M^4) + U^2*(54316*M^6 + 545344*M^7 +
1619936*M^8 + 1406912*M^9) + U^6*(6144*M^2 + 187904*M^3 + 1503120*M^4 +
3308608*M^5) + U^3*(65036*M^5 + 829848*M^6 + 3098880*M^7 + 3383968*M^8) +
U^5*(21920*M^3 + 477672*M^4 + 2886800*M^5 + 4963840*M^6) + U^4*(47724*M^4
+ 784048*M^5 + 3695136*M^6 + 5049088*M^7) + 5183*M^8 + 31812*M^9 +
57728*M^10 + 29952*M^11 + 512*U^11)
```

The good thing with the derived equations it that they basically give the exact expected result as seen in the Results – Part one. And the expressions will always be in the simple form of a rational function (ratio of two polynomial functions), even for higher number of subpopulations giving enormous expressions (consider the example in table 4 above)

But even if the derived equation gives the exact result, they become difficult to interpret when they get large. The dynamics behind the statistics becomes hard to understand, as they act more like the output gained from simulation where the mechanism behind gets far too complex to be grasped by the human mind.

Testing the no migration theorem

The no migration theorem was also used here to test the acquired equations for the three statistics for the 2 to 12 subpopulation scenarios. The no migration theorem where checked by substituting M with zero (as m = 0 → M = 0) with MATLAB for all the equations.

The expectation were met for all three statistics for all of the subpopulation scenarios:

$$\begin{aligned} &\mathrm{m} = 0 \rightarrow \mathrm{G_{ST}} = \frac{1}{1 + 4\mathrm{N}\mu}, && \text{for n} = 2 \text{ to } 12 \\ &\mathrm{m} = 0 \rightarrow \mathrm{D_{ST}} = 1, && \text{for n} = 2 \text{ to } 12 \\ &\mathrm{m} = 0 \rightarrow \mathrm{G'_{ST}} = 1, && \text{for n} = 2 \text{ to } 12 \end{aligned}$$

So even the huge equation derived for the 12 subpopulation scenario for G_{ST} having around 15 thousand symbols, were reduced to this simple form when the parameter m (and thereby M) were set to zero.

Hypothesis of simplified G_{ST} equations for the finite linear population model

The shape for G_{ST}'s derived equation in the two and three subpopulations scenario had the same form for both cases:

$$\mathrm{G_{ST}} = \frac{1}{b_1 + b_2\mathrm{Nm} + b_3\mathrm{N}\mu} \quad (30)$$

Where the constant b_1 and b_3 equals 1 and 4 for both cases, and b_2 equals 8 for n=3 and equals 4 for n=3 (equation 20, 27). G_{ST} actually also has this simple shape for 2 to 12 subpopulations for the no migration scenario as seen above (b_2 just equals zero). G_{ST}-related equations also tend to get that particular shape for other migration models, for example as it was derived by Wrights (see equation 5).

A hypothesis was made from here, that G_{ST} would approximately follow the same shape seen above in equation 30 for any number of subpopulations in the finite linear population model. When testing this hypothesis, we know according to the no migration theorem that G_{ST} always will equal $\frac{1}{1+4\mathrm{N}\mu}$, when m = 0. We would thereby expect $b_1 = 1$ and $b_3 = 4$ (in equation 30), being independent of n. This would make b_2 be the only constant to change with increasing number of subpopulation (increasing n).

If a relationship between b_2 and n were found, the general equation for G_{ST} for the finite linear population model would follow. This equation would have four parameters: μ, m, N and n (where μ, m and N could be substituted with U and M, giving three parameters instead: U, M and n).

Testing hypothesis of a simplified G_{ST} equation

The hypothesis of simplified G_{ST} equation following the shape seen in equation 30 were tested for n = 2 to 12. It was tested by doing a generalized linear regression with a Gaussian distribution using an inverse link function. The regression was made with two predictor variables X_1 and X_2 and one predictor variable Y. The predictor variable X_1 equalled the parameter U and X_2 the parameter M. The response variable Y equalled the output from the derived full model of G_{ST} with the two parameters U and M specified by the predictor variables.

The predictor variables were given 70 different values each, evenly spread between 0.001 and 100 in a logarithmic fashion. The test was fully crossed between the two predictor variables giving 70×70 = 4900 different values on the response variable.

The output taken from the regressions were three coefficients, one constant term c_1, and two constants c_2 and c_3 determine the slope for X_1 and X_2. The model, displayed in equation 31 below, get the exact same shape as the equation 30, due to the inverse link function.

$$\mathrm{Y} = \frac{1}{c_1 + c_2\mathrm{X}_1 + c_3\mathrm{X}_2} \qquad \begin{matrix} \text{where } \mathrm{X}_1 = \mathrm{U} = 2\mathrm{N}\mu \\ \text{and } \mathrm{X}_2 = \mathrm{M} = 2\mathrm{Nm} \end{matrix} \qquad (31)$$

The estimated coefficients c_1, c_2 and c_3 was in this way gained for n = 2 to 12. The residuals were analysed telling how good the simplified G_{ST} equation matches the output from the derived full G_{ST} equation.

Testing hypothesis of a simplified D_{ST}

The same hypothesis was made for D_{ST}, that the full derived model (being huge for large n) would approximately follow the simple equation shape seen in the scenario of two subpopulations.

D_{ST} can be written in the form of an inverted linear relationship:

$$\mathrm{D}_{\mathrm{ST}} = \frac{\mu}{\mu + \mathrm{m}} = \frac{1}{1 + \frac{\mathrm{m}}{\mu}}$$

Where the equation for the general shape would be:

$$\mathrm{D}_{\mathrm{ST}} = \frac{1}{b_1 + b_2\frac{\mathrm{m}}{\mu}} \qquad (32)$$

As we know that D_{ST} equal 1 when m = 0 according to the no migration theorem, we would expect b_1 to be equal 1, independent of n. This makes b_3 the only constant to change with increasing n.

The hypothesis was tested in the same way as for $G_{ST,}$ using a generalized linear regression with a Gaussian distribution having an inverse link function. The regression was made with one response variable Y equals to the output from the full derived model of D_{ST} using the same parameter set of U and M as in regression for G_{ST}, having 4900 different combinations. But this regression was made with only one predictor variable X equaling M/U (which equals m/μ) making the regression match the shape seen in equation 32.

The output taken from the regressions was the two coefficients, one constant term c_1, and one constant c_2 determine the slope for X, where the model, displayed in equation 33 below, get the exact same shape as the equation 32, just as wanted:

$$\mathrm{Y} = \frac{1}{c_1 + c_2\mathrm{X}}, \quad \text{where } \mathrm{X} = \frac{\mathrm{M}}{\mathrm{U}} = \frac{\mathrm{m}}{\mu} \qquad (33)$$

The output of interest from the regression was the estimated coefficients c_1 and c_2 giving the estimated simplified equation of D_{ST} for n = 2 to 12. The residuals were analysed telling how good the simplified D_{ST} equation matches the output from the full derived D_{ST} equation.

Simplified G_{ST} for the finite linear population model

G'_{ST} couldn't be written in the form of an inverted linear relationship, even for its simple expression gained in the two subpopulation scenario. But this was not a problem as G'_{ST} could just be referred as a product of G_{ST} and D_{ST} as seen in equation 13 (or 12 and 14 expressing the same). The simplified G'_{ST} was defined as:

$$\mathrm{G'}_{\mathrm{ST\_Reg}} = \mathrm{G}_{\mathrm{ST\_Reg}} + \left(1 - \mathrm{G}_{\mathrm{ST\_Reg}}\right)\mathrm{D}_{\mathrm{ST\_Reg}}$$

Where G_{ST_Reg} is the simplified equation estimated by its regression analysis:

$$\mathrm{G}_{\mathrm{ST\_Reg}} = \frac{1}{b_1 + b_2\mathrm{Nm} + b_3\mathrm{N\mu}}$$

Where the constants b_1, b_2 and b_3 equals the coefficients c_1, $2c_2$ and $2c_3$. The multiplication of 2 before c_2 and c_3 comes from that U and M used in the regression equals 2 times Nm and Nμ.

And D_{ST_Reg} is the simplified equation estimated by its regression analysis:

$$\mathrm{D}_{\mathrm{ST\_Reg}} = \frac{1}{b_1 + b_2\frac{\mathrm{m}}{\mathrm{\mu}}}$$

Where the constants b_1 and b_2 equals the coefficients c_1 and c_2. No multiplication needed as M/U equals m/μ.

No regression analysis where thereby made to obtain G'_{ST_Reg}, as it was defined from the regression results giving G_{ST_Reg} and D_{ST_Reg}.

7. Result – Part two

7.1 The simplified equation estimated by regression for the three statistics

The constants for G_{ST_Reg} and D_{ST_Reg} obtained by regression was analyzed and plotted in figure 6. Their behaviors are described in table 5 below.

Table 5 – Analysis of the constants for G_{ST_Reg} and D_{ST_Reg} gained by the regression.

Constant	Expected value	Behavior
b_1 for G_{ST_Reg}	1	b_1 follows the expected value with very high precision. It never deviates more than $1.5*10^{-3}$ from 1.
b_2 for G_{ST_Reg}	Change as n increases. Suggestion: $b_2 = \frac{8}{n-1}$	b_2 start at 8 when n = 2, then equals 4 when n = 3, then almost exact equal 8/3 when n = 4, then almost exact equal 8/4 when n = 5 etc, etc. This could be expressed as the suggestion to the left: $b_2 = \frac{8}{n-1}$. The constant is having a high precision relative to this suggestion, having a maximum deviation of 0.018 for n equal 2 to 12.
b_3 for G_{ST_Reg}	4	b_3 is showing a little bit of strange behavior compared b_1 that had an almost perfect match to the expectation. b_3 starts with the expected value of 4, but then begins to increase when reaching n higher than 3. The increase is 0.090 for n = 3 to n = 4 giving b_3 a value of 4.09. The increasing is continuous but slows down (in what it seems to be an even pace) to having an increase of 0.068 from 11 to 12 subpopulation giving b_3 a value of 4.61.
b_1 for D_{ST_Reg}	1	b_1 follows the expected value with high precision. It is deviating as most when n = 4 having a deviation of 0.016 from the expectation of 1 and then slowly converge against 1 as n increases having a deviation of 0.012 for n = 12.
b_2 for D_{ST_Reg}	Change as n increases. Suggestion: $b_2 = \frac{1}{(n-1)^{1.75}}$	Having a similar pattern as b_2 for G_{ST_Reg} but decreases with a more rapid pace. The constant is having a high precision relative to this suggestion of $b_2 = \frac{1}{(n-1)^{1.75}}$. It is deviating as most when n = 3 having a deviation of 0.016 from the suggestion, but then rapidly converge against the suggested value as n increases having a deviation of $9.7*10^{-4}$ for n = 12.

Suggested formulas

From the analysis presented in table 5 above, we got suggested values for the last constants b_2 for G_{ST_Reg} and b_2 for D_{ST_Reg}. This gives us a full suggested formula for G_{ST} and D_{ST} and thereby G'_{ST}, as we already have expected values for b_1 and b_3 for G_{ST_Reg} and b_1 for D_{ST_Reg}. These suggested formulas will be called G_{ST_Sug} and D_{ST_Sug} and G'_{ST_Sug} (see equation 34, 35 and 36 below).

$$G_{ST} = \frac{H_D - H_S}{H_D} \approx G_{ST\_Reg} = \frac{1}{b_1 + b_2 Nm + b_3 N\mu} \approx G_{ST\_Sug} = \frac{1}{1 + \frac{8}{n-1} Nm + 4N\mu} \qquad (34)$$

$$D_{ST} = \frac{H_D - H_S}{1 - H_S} \approx D_{ST\_Reg} = \frac{1}{b_1 + b_2 \frac{m}{\mu}} \approx D_{ST\_Sug} = \frac{1}{1 + \frac{1}{(n-1)^{1.75}} \frac{m}{\mu}} \qquad (35)$$

$$G'_{ST} = G_{ST} + (1 - G_{ST})D_{ST} \approx G'_{ST\_Reg} = G_{ST\_Reg} + \left(1 - G_{ST\_Reg}\right)D_{ST\_Reg}$$

$$\approx G'_{ST\_Sug} = G_{ST\_Sug} + \left(1 - G_{ST\_Sug}\right)D_{ST\_Sug} \leftrightarrow$$

$$\leftrightarrow G'_{ST\_Sug} = \frac{1 + \frac{8}{n-1} Nm + 4N\mu + \frac{1}{(n-1)^{1.75}} \frac{m}{\mu}}{\left(1 + \frac{8}{n-1} Nm + 4N\mu\right)\left(1 + \frac{1}{(n-1)^{1.75}} \frac{m}{\mu}\right)} \qquad (36)$$

Where equation 35 could be written in the following form, looking like the equation for two subpopulation scenario:

$$D_{ST\_Sug} = \frac{\mu}{\mu + \frac{1}{(n-1)^{1.75}} m}$$

Plotting the suggested constants and the constants estimated by regression

The constants estimated by regression for the simplified equations for G_{ST} and D_{ST} is plotted in figures 6 together with the suggested values of the constants as seen in equation 34 and 35 marked as out as dotted lines.

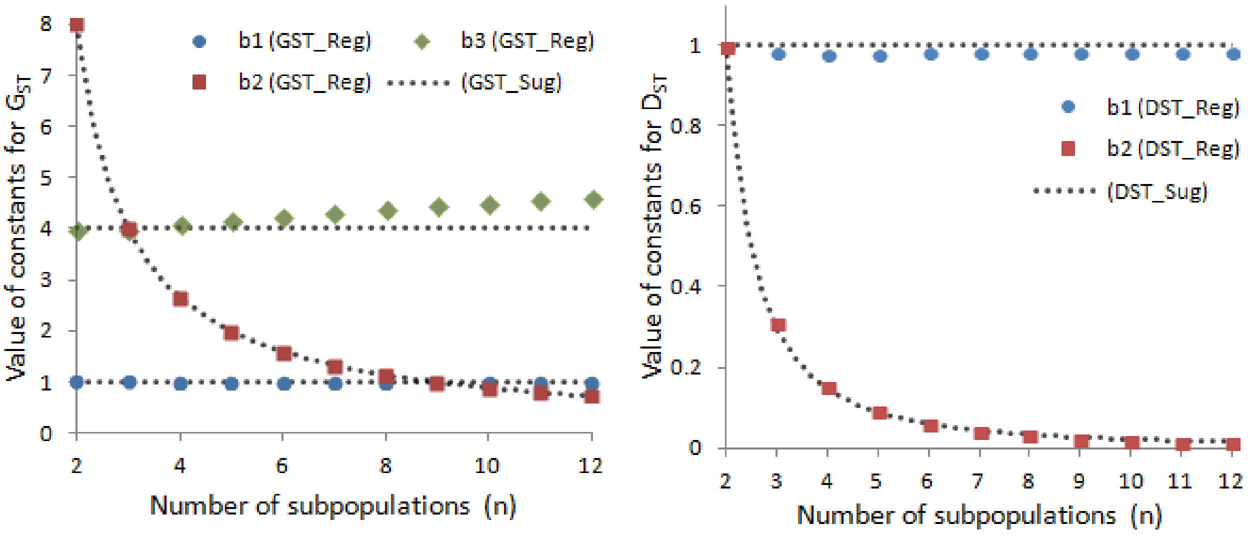


Figure 6 – The coloured markers represents the constants for G_{ST_Reg} and D_{ST_Reg} estimated by regression for the different numbers of subpopulations (n). The dotted line represents the suggested values of the constants (see equation 34 and 35).

The precision of the approximated equations

The deviation is plotted figure in 7 for the three statistics' approximated equations (equation 34, 35, 36) compared with their full derived formulas (see example in table 4). The deviation is presentenced by plotting the mean |residual| and the 95% quantile of the |residuals| for both the formulas estimated with regression (having the extra index of _Reg) and the suggested formula (having the extra index of _Sug). The suggested formulas are using the same response variable as the ones estimated with regression to determine the residuals.

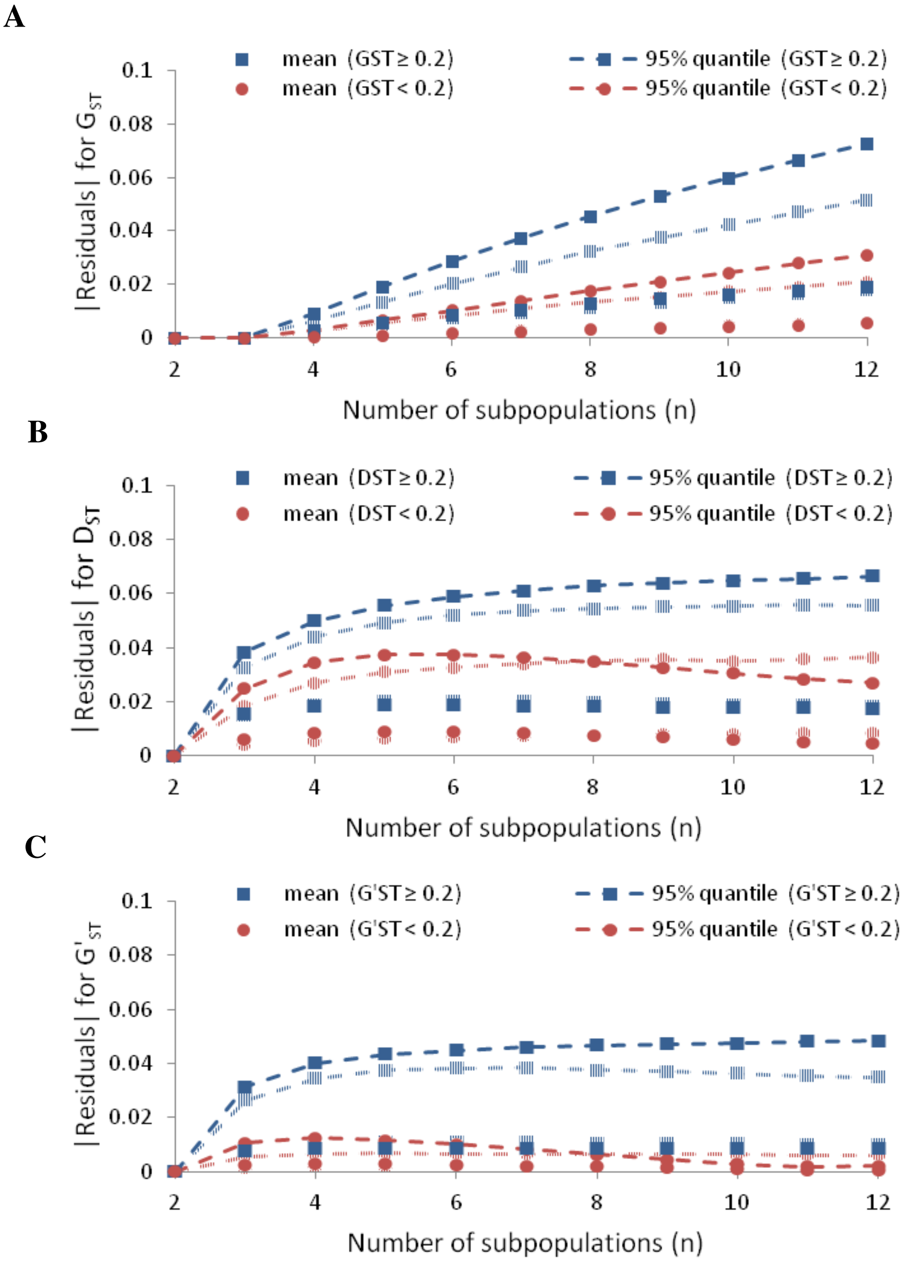


Figure 7 – The deviation in form of mean |residual| and 95% quantile of the |residual| for the approximated formulas compared with their derived full formulas. The deviation is divided into two cases, for high values (full formula statistic ≥ 0.2) and low values (full formula statistic < 0.2). **A.** The deviation for the approximated equations G_{ST_Sug} and G_{ST_Reg} compared with the full derived formula for G_{ST}. G_{ST_Sug} is represented by the solid colored markers and G_{ST_Sug} by the bleach colored markers. **B.** Same as A, but for D_{ST}. D_{ST_Sug} is represented by the solid colored markers and D_{ST_Sug} by the bleach colored markers. **C.** Same as A and B, but for G'_{ST}. G'_{ST_Sug} is represented by the solid colored markers and G'_{ST_Sug} by the bleach colored markers.

8. Discussion – Part two

8.1 Approximated models versus full derived models

We can see in figure 7 that the simplified approximated models (equation 34, 35, and 36) explain the large derived models (see example in table 4) with very high precision. We can see that the mean |residual| for high values always is lower than 0.02 and that the mean deviation for low values is even lower. The approximated models for $G'_{ST,}$ being derived by both G_{ST} and D_{ST}, are interestingly showing a better fit than both G_{ST} and D_{ST}. Especially for low values of the statistic where the deviation tends is to zero!

Using these approximated models when estimating differentiation for metapopulations matching the population structure defined in the finite linear population model could be a good alternative. Especially as the migration pattern of a linear population may give a better match than other migration models, like the island models.

8.2 The formulas estimated by regression versus the suggested formulas

We have two different types of simplified equations for the finite linear population model. One is being the formula estimated by regression, using the regression analysis to defining the constants of the equation. The other is the suggested formula, using logical thinking combined with analysis of the result from the regression formula suggesting a formula that will meet certain criterions.

The formulas are (as seen in figure 7) almost exact the same for all constants except b_3 for G_{ST}. It has a constant value of 4 in the suggested formula while the formula estimated starts giving it a value of 4, but then in a increases the value for increasing number of subpopulations.

We can see in figure 6 that the formulas estimated by regression has a slightly better fit then the suggested formula in terms of lower residual deviation. That is expected as the regression is defined in such a way that it will pick the constants giving the lowest residual deviation. But this doesn't necessarily mean that the estimated by regression is better. We for example know that the constant b_3 for G_{ST}, being the only constant showing important difference, is correct for G_{ST_Sug} when migration is low as the constant b_3 is known to equal 4 for scenarios with no migration according to the no migration theorem.

My conclusion is that I would recommend the suggested formulas (G_{ST_Sug}, D_{ST_Sug} and G'_{ST_Sug}) before the formula estimated by regression (G_{ST_Reg}, $D_{ST\_ Reg}$ and $G'_{ST\_ Reg}$) as they are having almost the same precision seen in figure 7. One other big strength of the suggested formulas is that they give the general formulas for any number of subpopulation, whereas the formulas estimated by regression must be calculated for each number of subpopulation with regression.

8.3 Interpretation of G_{ST}, D_{ST} & G'_{ST} and their different behaviour

As clearly seen in figure 5, the three statistics G_{ST}, D_{ST} and G'_{ST} have very different behaviour, where G_{ST} and D_{ST} in many cases give more or less opposite results.

Population structure

The question is what type of differentiation we want to measure. The definition of differentiation behind G_{ST} gives properties that can be misinterpreted if the properties of G_{ST}

are not fully understood, and these properties may even be unwanted (Meirmans & Hedrick 2011).

One property that one may want the statistic to measure is the degree of population structure and a statistic having a direct relationship with the number of migrants per generation (Statistic ~ Nm) could thereby be appropriate, just as Wright presented F_{ST} in the approximated equation 4. We can see in figure 5 that none of the three statistics have this property. If a statistic in figure 5 would have this property they would have lines without slope, unaffected by increasing mutation rate in the x-axis. But a combination of the statistics G_{ST} and G'_{ST} would give a close answer to the question. This as both statistics are unaffected by mutation when the mutation is low giving a direct relation between the statistic and migration rate. In this unbiased parameter interval G_{ST} will equal G'_{ST}. In the finite linear population model we will have the following relationship under low mutation rate (using the G_{ST_Sug} for approximation):

$$\mathrm{G_{ST}} \approx \mathrm{G'_{ST}} \approx \frac{1}{1+\frac{8}{n-1}\mathrm{Nm}}\,, \text{when } \mu \ll \mathrm{m} \quad (37)$$

When mutation gets higher G_{ST} will drop in value but and G'_{ST} will increase where. This mean that when $G_{ST} < G'_{ST}$ they are biased by the mutation, but a value in the interval between them will satisfy the relationship in equation 37. This general relationship is described in equation 38.

$$\mathrm{G_{ST}} \leq \frac{1}{1+\frac{8}{n-1}\mathrm{Nm}} \leq \mathrm{G'_{ST}} \quad (38)$$

This means that if $G_{ST} \approx G'_{ST}$ we know that equation 37 is valid, and the statistic being directly related with the number of migrants (Nm) and thereby being a good measure of population structure, where Nm can be derived from G_{ST}. But on the other hand, if $G_{ST} << G'_{ST}$ we know that no of them will be a direct measure of population structure, but that a value between G_{ST} and G'_{ST} will.

Genetic differentiation

If only the differentiation of the genetic sequences is the thing that you want to measure, D_{ST} would be preferable. This is easily seen if you consider the case of two subpopulations, making each migration fusing the two populations in question genetically, while each mutation makes them more genetically different:

$$\mathrm{D_{ST}} = \frac{\mu}{\mu + \mathrm{m}}$$

We can see that D_{ST} just becomes a ratio between the two events that make the two subpopulations more genetically alike and more genetically different. This is making D_{ST} to measure something more like a phylogenetic distance between the populations.

But I think it is always good to include the G_{ST} measure as it estimates F_{ST} which is a statistic with a lot of history, even if you just want to measure the genetic differentiation rather than the population structure. This makes both the researcher and the readers able to compare the results with result obtained all the way back to the 1940s.

5. Acknowledgements

I would like to express my gratitude to my supervisors Prof. Martin Lascoux and Mattias Jakobsson for making me able to do this bachelor thesis with their valuable assistance and great knowledge.

6. Reference list

Gillespie JH 2004. Population Genetics. A Concise Guide. 2nd ed. The Johns Hopkins University Press, Maryland.

Hedrick PW. 2005. A standardized genetic differentiation measure. Evolution 59: 1633–1638.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews 10: 639-650.

Jost L. 2008. G_{ST} and its relatives doo not measure differentiation. Molecular Ecology 17: 4015-4026.

Kaj I, Lascoux M. 1999. Probability of Identity by Descent in Metapopulations. Genetics 152: 1217-1228.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217: 624-626.

Kimura M. 1979. The neutral theory of molecular evolution. Scientific American 241: 98-100, 102, 108.

Kimura M and Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49: 561-576.

Meirmans P, Hedrick P. 2011. Assessing population structure: F_{ST} and related measures. Molecular Ecology Resources 11: 5-18.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70: 3321–3323.

Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nature Reviews 10: 745-755.

Slatkin M. 1991. Inbreeding coefficient and coalescence times. Genetical Research 58: 167-175.

Wang J. 2012. On the measurement of genetic differentiation among populations. Genetic Research, Cambridge 94: 275-289.

Weir BS, Cockerham CC. 1987. Correlations, descent measures: Drift with migration and mutation. Proceedings of the National Academy of Sciences 84: 8512-8514.

Wright S. 1943. Isolation by distance. Genetics 28: 114-138.